SVDTree: Semantic Voxel Diffusion for Single Image Tree Reconstruction (Supplementary Materials)

Yuan Li[1†] Zhihao Liu[2†] Bedrich Benes[3] Xiaopeng Zhang[1,4] Jianwei Guo[1,4*]
[1]MAIS, Institute of Automation, Chinese Academy of Sciences [2]The University of Tokyo
[3]Computer Science, Purdue University [4]School of Artificial Intelligence, University of Chinese Academy of Sciences

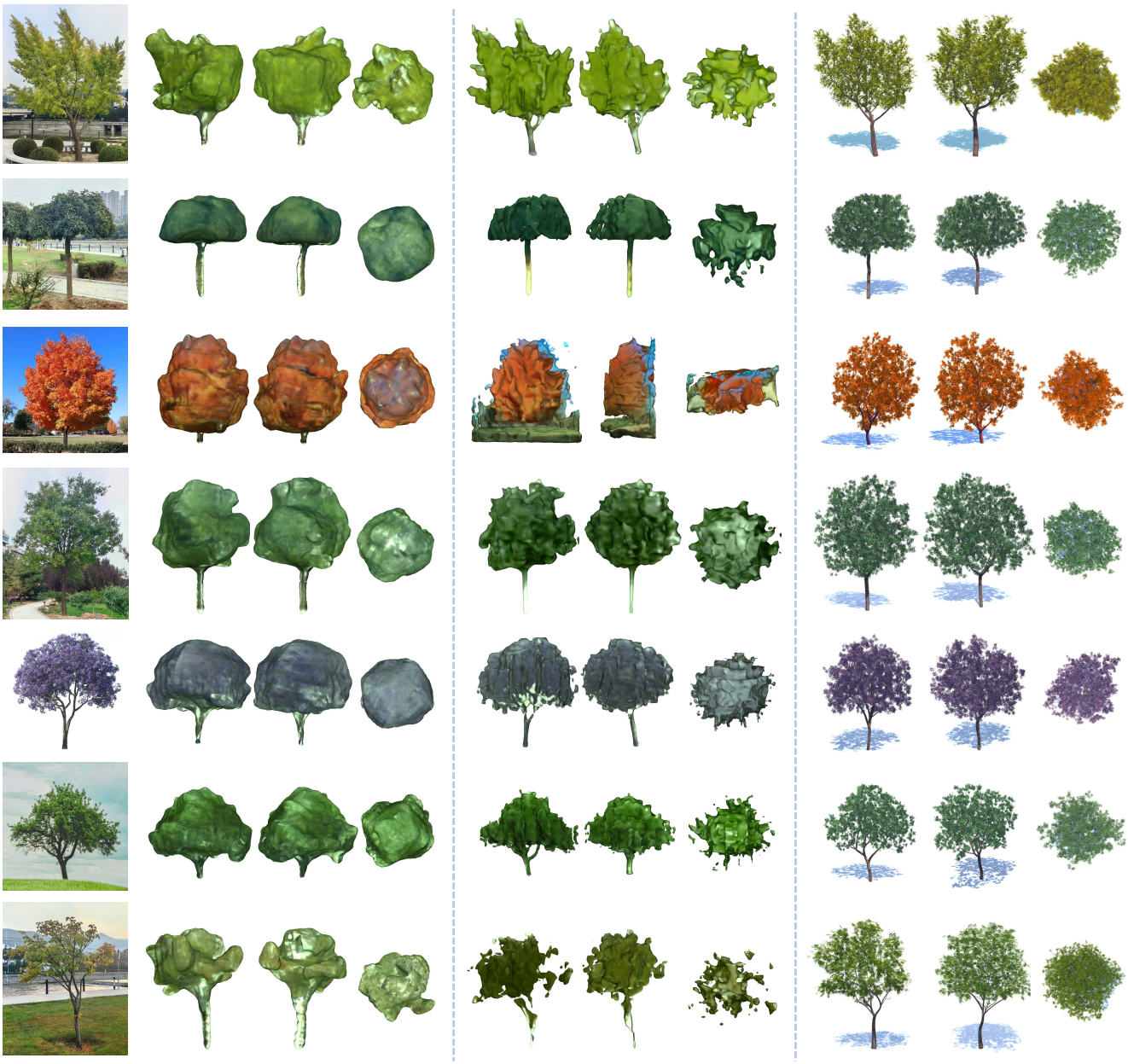

Figure 1. We compare *SVDTree* with two state-of-the-art diffusion-based single-view reconstruction methods, including One-2-3-45 [1] and Wonder3D [3]. For each method, we show the front, side, and top views of the reconstruction results from left to right.

1. Comparison to Diffusion-based Single-view Reconstruction

Recent advances in large 2D diffusion models, which are pre-trained on Internet-scale image datasets, have significantly improved the performance of single-image 3D reconstruction to create high-quality textured meshes. For example, One-2-3-45 [1] first utilizes a view-conditioned 2D diffusion model (Zero123 [2]) to generate multi-view images from the input single image, then produces 3D geometry using an SDF-based generalizable neural surface reconstruction. A recent state-of-the-art method, Wonder3D [3], also follows this pipeline: they first generate multi-view 2D images [4], and then generate multi-view consistent normal maps and their corresponding color images with a cross-domain diffusion model. By modeling the joint distribution of two different domains (*i.e.*, normals and colors), they reconstruct high-fidelity textured meshes from a single-view image.

To compare with these diffusion model based methods, we conduct experiments using real-world images. We show the qualitative comparison in Fig. 1 and display the reconstruction results from different views. Although these methods have good capabilities to capture the overall shape of a tree, they only generate closed meshes whose geometries are rough and full of noise. Also, as they do not take into account the knowledge of tree development, they lack the semantic structure information. They can not be used in downstream applications, such as tree analysis, editing, and simulation. In comparison, we do not rely on generating multi-view images; instead, we use the diffusion model to generate a semantic voxel structure, which is an intuitive but more efficient approach. Therefore, our method not only ensures more complete reconstructions but also can better preserve rich tree details of geometric and semantic information. Moreover, it achieves more realistic texturing results by treating the branchings and the crown separately.

2. Qualitative and Quantitative Analysis

To analyze the robustness of our approach, we first show the reconstructions of a non-leafy tree and a bush in Fig. 2. Although our dataset does not contain such data, we still have a certain capability to handle them. We would like to extend our datasets to further improve the reconstruction accuracy of such plants in future work.

To further conduct a quantitative numerical evaluation of the reconstruction error, we captured multiple images of a real tree and then used the Multi-View-Stereo (MVS) approach to obtain the dense point cloud $\mathcal{P}$ as an approximate ground truth. The reconstruction error is computed as the average distance from all the points in $\mathcal{P}$ to the reconstructed 3D model. Fig. 3 shows both the qualitative and quantitative results, as well as the comparison to a SOTA diffusion-based method.

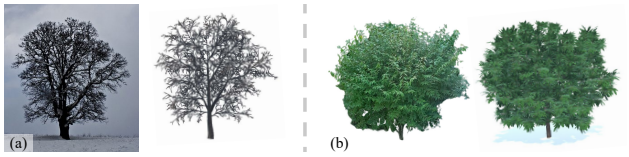

Figure 2. Two special examples: (a) non-leafy tree, (b) bush.

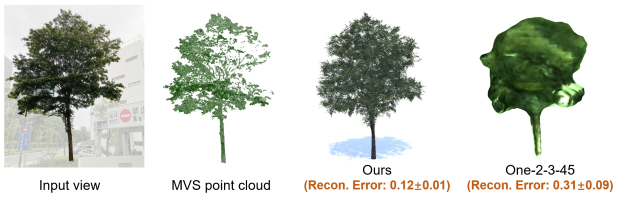

Figure 3. Visual and quantitative evaluation wrt. MVS point cloud. We report reconstruction error and variance. Lower error is better.

3. 3D Illustration of Hybrid Tree Geometry Reconstruction

In the main paper, we have used a 2D illustration to explain our hybrid tree geometry reconstruction algorithm. To better illustrate how visual 3D tree models are constructed from SVS, we now use a concise 3D real example to show the process of 3D tree reconstruction. Fig. 4 shows each step of our approach. Please refer to our main paper for a detailed explanation of each step.

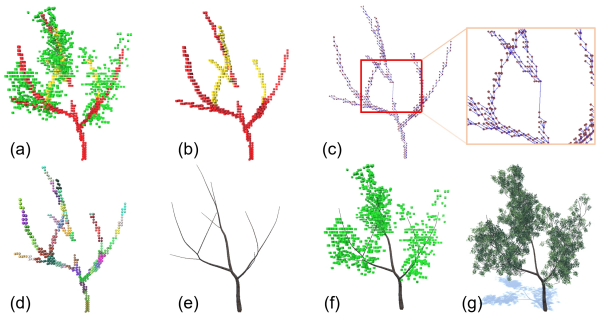

Figure 4. 3D illustration of reconstructing the skeleton (b-e) and leaves (f-g).

References

[1] Minghua Liu, Chao Xu, Haian Jin, Linghao Chen, Zexiang Xu, Hao Su, et al. One-2-3-45: Any single image to 3d mesh

† Joint first authors with equal contributions.
* Corresponding author: jianwei.guo@nlpr.ia.ac.cn

in 45 seconds without per-shape optimization. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2023. 1, 2

[2] Ruoshi Liu, Rundi Wu, Basile Van Hoorick, Pavel Tokmakov, Sergey Zakharov, and Carl Vondrick. Zero-1-to-3: Zero-shot one image to 3d object. In *IEEE International Conference on Computer Vision (ICCV)*, pages 9298–9309, 2023. 2

[3] 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. *arXiv preprint arXiv:2310.15008*, 2023. 1, 2

[4] Yichun Shi, Peng Wang, Jianglong Ye, Mai Long, Kejie Li, and Xiao Yang. Mvdream: Multi-view diffusion for 3d generation. *arXiv preprint arXiv:2308.16512*, 2023. 2