SVDTree: Semantic Voxel Diffusion for Single Image Tree Reconstruction

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

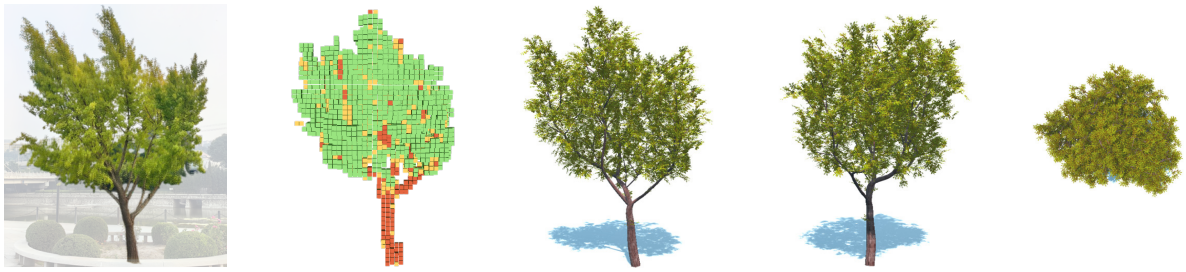


Figure 1. *SVDTree* for single image tree reconstruction. Given a masked image, we use a diffusion model to automatically infer a semantic voxel structure of the tree, which guides a hybrid geometry reconstruction algorithm to produce a 3D tree with high visual fidelity.

Abstract

Efficiently representing and reconstructing the 3D geometry of biological trees remains a challenging problem in computer vision and graphics. We propose a novel approach for generating realistic tree models from single-view photographs. We cast the 3D information inference problem to a semantic voxel diffusion process, which converts an input image of a tree to a novel Semantic Voxel Structure *(SVS) in 3D space. The SVS encodes the geometric appearance and semantic structural information (*e.g.*, classifying trunks, branches, and leaves), which retains the intricate internal tree features. Tailored to the SVS, we present SVDTree a new hybrid tree modeling approach by combining structure-oriented branch reconstruction and self-organization-based foliage reconstruction. We validate SVDTree by using images from both synthetic and real trees. The comparison results show that our approach can better preserve tree details and achieve more realistic and accurate reconstruction results than previous methods.*

1. Introduction

Vegetation is an indispensable part of natural and urban scenes. However, capturing the vegetation is a complex task that is dominated by procedural models [46, 48, 52, 66]. Recently, plant reconstruction methods have seen a significant improvement and have found applications in areas such as plant geometry and topology for vision-assisted plant phenotyping [19, 37], forestry [24] or counting [44].

Research on 3D tree model acquisition has received considerable attention for decades, but accurately reconstructing trees is still challenging because of the tree's topological and geometric complexity. Previous works often separate the branch and foliage reconstruction, by first to reconstructing high-level branching structures from sensor data and then modeling the foliage by synthesizing twigs and leaves using a procedural model. For example, reconstruction from LiDAR point clouds achieves faithful skeletal branches using graph-based methods [32, 63], while the foliage can be approximated as 3D envelopes for populating geometry details with predefined rules [33]. The 2D skeletons can be efficiently estimated and fused into a 3D branching structure from multi-view photographs [22, 34, 57]. However, scans often generate incomplete point clouds due to occlusion [3]. Furthermore, image matching and registration from multiple images do not perform robustly on trees with complexity, translucency, and self-occlusions, leading to an unsatisfying point cloud with large reconstruction errors. Moreover, multiple views are not always available, making the single-view reconstruction often a more flexible and cost-effective solution to generate 3D tree models, especially on a large scale [2].

We introduce the *SVDTree* framework that reconstructs high-fidelity trees from single-image photographs. Our approach is motivated by the record-breaking performance of diffusion models in many digital content generation tasks [64]. We formulate the estimation problem as a semantic voxel diffusion process to recover 3D information

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

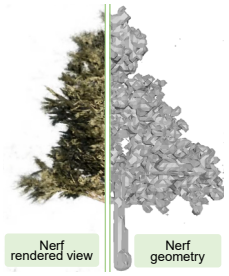
from a single image. Specifically, we utilize the diffusion model to convert an input image of a tree to a *Semantic Voxel Structure* (SVS), in which each voxel encodes the geometric and semantic information of the tree by detailing the tree trunk, branches, and leaves. The SVS representation accurately captures the overall geometric tree appearance, and it retains the intricate internal features by encoding the semantic structural information. Moreover, it allows us to infer hidden branches that are not visible in the input images to faithfully reconstruct the complete shapes.

We develop a new hybrid tree geometry construction approach tailored to the SVS data structure. Guided by structural information within the SVS, we introduce a structure-oriented branch reconstruction approach and then combine it with the self-organization-based foliage reconstruction [46] to reconstruct the final geometric model. We validate our approach on both synthetic images and in-the-wild real images and also show that our diffusion model can be trained efficiently on a library of synthetic 3D tree models. Compared with previous methods that predict depth maps via cGAN [31] or learn fixed-size 3D bounding volumes [26], our approach does not rely on prior knowledge of the tree species and provides more detailed tree structure. In summary, we make the following contributions:

1. We introduce a novel structured voxel representation of trees in 3D space to encode the structural information of trunks, branches, and leaves, thereby expressing detailed features such as the trunk's primary topology and the foliage's spatial morphology.
2. We propose a semantic voxel diffusion model to faithfully generate 3D tree structures from single photographs. We show its benefits in recovering more complete and detailed reconstructions.
3. We develop a novel tree geometry construction algorithm based on a hybrid modeling approach (combining structure-oriented branch reconstruction and self-organization-based foliage reconstruction), which is specially designed for our customized data representation.

2. Related work

3D tree geometry reconstruction. Tree reconstruction generates models from real-world sensor data. Previous methods aimed at the reconstruction of 3D tree skeletons can be classified into three categories: (1) procedural reconstruction, where input data is used to guide the tree modeling (*e.g.*, rule-based [18, 60] or particle-flow modeling [42]) based on a significant set of geometric parameters, (2) geometry-based extraction, which estimates the branching structure from point clouds by extracting its skeleton through minimal spanning graphs [13, 32] or deep learning [25, 30, 70], and (3) image-based modeling, which extracts 2D branching structures from a set of images [4, 22, 37, 53, 57], which are then converted into a 3D skeleton. Most closely related to our work is single image tree reconstruction [1, 26, 31, 58]. These methods have several limitations. For example, Tan *et al*. [58] and Liu *et al*. [31] require user interactions to identify the main branches or the crown shape. Li *et al*. [26] automatically learns tree species and *radial bounding volumes* (RBV) as a lightweight representation of tree models, which is then used to guide tree growth carefully. Our method does not require any user input, and our novel structural voxel representation captures more fine-grained details, including small interior branches and spatial morphology of the foliage. Recent Nerf method [39] can produce realistic rendered views, seeming it can also work for 3D tree reconstruction. However, their nice-looking results are powered by the volumetric rendering, but the geometries are still rough and full of noises (see the wrapped figure). Thus, it cannot create high-quality structured 3D trees for practical applications.



Single-view 3D reconstruction. Reconstructing a 3D shape from only a single image is a difficult problem that requires strong prior knowledge of the real world. In the past several years, single-view 3D shape reconstruction using deep learning has seen rapid growth [14]. Existing methods are based on the 3D representation used as an output, such as voxel grids [10, 51, 62], octree [59], point cloud [8, 15], mesh [28, 47, 61, 65], and implicit fileds [9, 36]. Among them, pixel-to-voxel methods are straightforward and popular, which usually follow the encoder-decoder pattern where an image is encoded into a learned feature vector and then decoded into a voxel representation of the target shape. However, these methods are usually limited to simple or regular (man-made) objects. We show that an efficient volumetric representation encoding rich geometric and semantic information can be learned directly from a diffusion model [20], enabling us to reconstruct complex tree shapes.

3D generative models. Recent advancements in deep generative architectures enable the creation of high-fidelity 3D content [56]. Many 3D generative models have been proposed to synthesize 3D shapes and appearances, such as Variational Autoencoders (VAEs) [40, 67], Generative Adversarial Networks (GANs) [6, 7, 11, 17, 45], and Diffusion Models (DMs) [21, 35, 43, 69]. In particular, diffusion models with various input conditions (*e.g.*, text [29, 38, 50], image [41]) are effective at providing user control over the generation process while retaining the diversity of the generated shapes. In this work, we use a semantic diffusion model to predict the 3D SVS representation of trees.

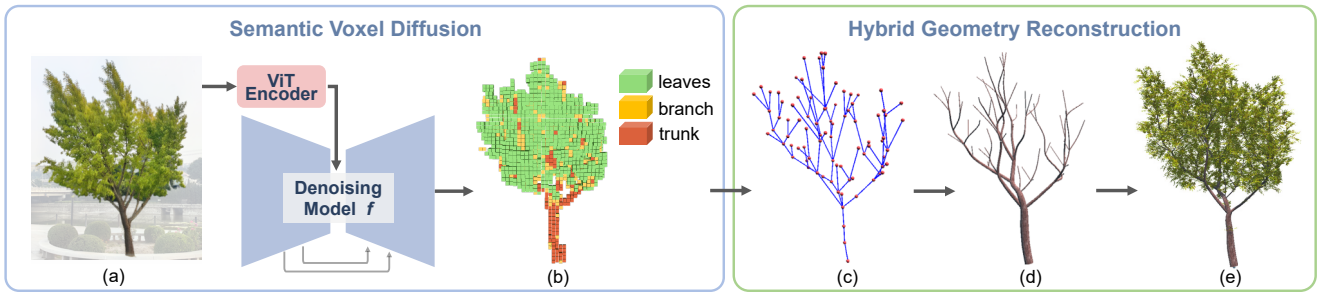


Figure 2. **SVDTree framework pipeline.** Starting from a single photograph, we use an instance segmentation module to detect the tree mask (a). Then, the embedding feature encoded from the masked image is fed to a denoising diffusion model to infer a semantic voxel structure (SVS) (b). Based on the predicted SVS, we propose to use a skeletonization approach and a space colonization approach to reconstruct the tree skeleton (c) and branch (d), respectively. (e) shows the final 3D tree model rendered with leaves and textures.

3. Overview

Given a single image of a biological tree, we aim to generate a corresponding 3D geometric model. This task is challenging due to the inherently ill-posed nature of 2D-to-3D translation and the intricate tree branch geometry. The key to our framework is a specialized data structure named *Semantic Voxel Structure* (SVS), which represents a tree with a set of 3D semantic voxels in a lightweight manner. The SVS abstracts the overall shape of the intricate tree structure but still maintains sufficient semantic and topological features to aid the neural network inference effectively.

As shown in Fig. 2, our solution to the single-image tree reconstruction is three-fold. Given a single tree image, our method begins with extracting the tree instance mask (Fig. 2 (a)) by utilizing an advanced image segmentation foundation model, called Segment Anything Model (SAM) [23], which effectively separates the tree pixels from the background. Next, we integrate the image features encoded from the masked tree instance and feed them into a semantic diffusion model to predict the SVS of the tree (Fig. 2 (b)).

Finally, given the structured voxel representation, we develop a hybrid tree geometry reconstruction algorithm. In this step, the tree skeleton, representing branching structures, is extracted based on the construction of a spanning tree from the trunk and branch voxels (Fig. 2 (c)). The crown is synthesized by a space colonization approach [46] based on the leaf voxels (Fig. 2 (d)). Our framework faithfully captures the key visual features observed in real trees while accentuating the intricate details of tree structures.

4. Method

4.1. Semantic Voxel Structure

We propose the semantic voxel structure (SVS). A volumetric representation for capturing the nuanced structure of tree models in a manner that retains crucial internal tree features. SVS is a 3D voxel grid of uniform size, where each voxel,

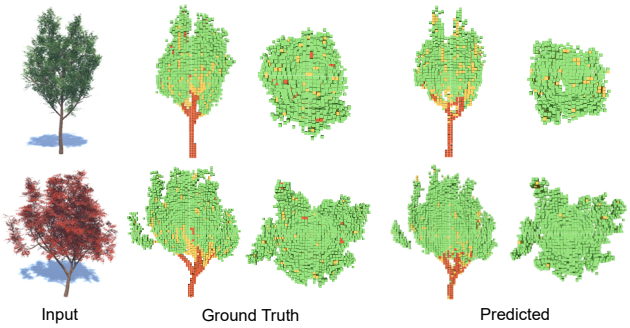


Figure 3. The ground truth and predicted SVS from the side and the top viewpoints. In each SVS, tree trunk, branches, and leaves are encoded by red, orange, and green colors.

represented as $\mathcal{V} = (x, y, z, C)$, records the semantic label C of the tree shape at the spatial position (x, y, z). We classify the 3D space containing a tree into four types of semantic labels: *tree trunk*, *tree branch*, *leaf*, and *empty* (representing the background). We assign different values to the semantic labels: 1:trunk, 0.5:branch, 0:leaf, and -1:empty to facilitate regression and classification.

Fig. 3 shows two examples of SVS representation. SVS not only captures the overall geometric appearance of a tree but also retains the intricate internal features by encoding the structural information. Compared to binary voxels (*i.e.*, 1:tree and 0:background) or neural bounding volumes [26], SVS prioritizes encoding the internal composition within the tree shape, enabling us to infer hidden branches that are occluded by the foliage. Consequently, the absence of SVS in the current neural networks that predict binary voxels limits their ability to learn the details of internal structure, restricting them to merely reconstructing the outer contour shape. To balance the reconstruction accuracy and network training speed, we set the resolution of SVS to 64^3.

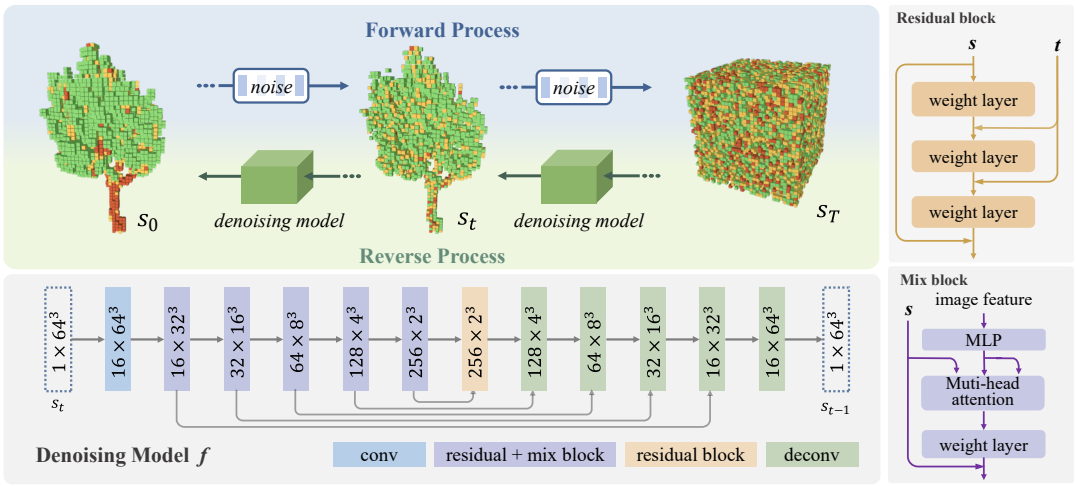


Figure 4. Network architecture of our semantic voxel diffusion model. Forward and reverse processes are specifically trained to gradually predict the SVS from a sampled noisy prior. The input image is encoded into the mix blocks to guide the denoising process.

4.2. Instance Segmentation

We use the Segment Anything Model (SAM) [23], a state-of-the-art network designed for promptable segmentation tasks, to accurately identify masks of tree instances. Unlike the previous method [26] that uses a specific synthetic dataset for training a semantic segmentation network, SAM has been trained on a vast dataset comprising over one billion annotations, primarily focused on natural images. It adopts a Vision Transformer model [5] with thoughtfully considered trade-offs to ensure real-time performance. Additionally, SAM facilitates zero-shot transfer to various tasks through prompt engineering.

We aim to identify valid tree masks to reconstruct trees from photographs faithfully. However, similar texture information is abundant in tree photographs, which complicates the segmentation of trees. We enhance the segmentation accuracy by feeding simple prompts to SAM, *i.e.*, just pointing out the trunk and foliage areas. SAM can also distinguish tree instances with minimal human intervention if the input image contains multiple trees. After segmenting the tree out precisely, we replace the background in the original image with a white color.

4.3. Learning to Reconstruct SVS

We use a diffusion model for the pixel-to-voxel conversion to estimate the SVS representation from the masked image.

4.3.1 Synthetic Training Dataset

Considering the difficulty of collecting and reconstructing real trees, generating synthetic tree datasets for training neural networks of different tasks (*e.g.*, reconstruction [22, 26, 30], segmentation [16], evaluation [49]) is widely adopted. We follow this idea and employ the space colonization algorithm [46] to create 4,000 tree models of diverse tree species automatically. We then generate corresponding single-view images (256^2) and their associated SVS (64^3) to train our network.

Image rendering. There may be a domain gap because the data distributions of real tree photographs and the rendering images may not align. To address this potential issue, we randomly position the camera around the tree, add leaf and branch textures, and adjust the rendering lighting intensity as shown in an example in Fig. 3. We then apply a Gaussian filter with a kernel size of three to smooth the rendered images.

SVS generation. We simplify the multi-layer tree structure representation introduced in [68] to annotate different semantic parts and decompose a tree into the trunk, the main branches, and the upper canopy. We then densely sample points on the surface meshes of 3D tree models and voxelize this semantic point cloud to obtain the ground truth SVS representation (Fig. 3).

4.3.2 Semantic Voxel Diffusion Model

We aim to predict the mapping from a masked tree image to its corresponding SVS. A network designed to address this task can be defined as follows:

$$f_{SVS}(I) : \mathcal{I} \rightarrow \mathcal{S},$$

where $I \in \mathcal{I}$ is the masked image, and $\mathcal{S}$ denotes the SVS.

Our objective is to assign a correct semantic label to each voxel in SVS using a neural network, which is a typical classification problem. However, the distribution of semantic labels in the training dataset is not uniform, *i.e.*, the number of tree voxels is too small compared to the number of

empty voxels, overwhelming training an efficient classification network. To overcome this issue, we cast it as a regression problem, meaning that we predict the classification values, which are then quantized and truncated to determine the semantic label.

Semantic voxel diffusion model f. The neural network trained to estimate SVS is based on a denoising diffusion model [20]. We call it a semantic voxel diffusion model because it outputs the 3D shape in the form of voxels and predicts the classification value for each voxel. Our diffusion model comprises forward and reverse processes (Fig. 4). Given an initial SVS s_0, the forward process yields a sequence of noised data $\{s_t|t \in (0,T)\}$ by adding increasing Gaussian noise to s_0:

$$s_t = \sqrt{\alpha(t)}s_0 + \sqrt{1-\alpha(t)}\epsilon, \tag{1}$$

where $\epsilon \sim \mathcal{N}(0,I)$, $t \sim \mathcal{U}(0,1)$, and $\alpha(t)$ is a monotonically decreasing function from one to zero. $\mathcal{N}$ and $\mathcal{U}$ denote Gaussian distribution and uniform distribution, respectively. We set$\alpha(t) = cos^{-2}(\frac{\pi}{2}\frac{t+s}{s+1}) - 1, s = 0.008$.

The reverse chain iteratively denoises the corrupted SVS, *i.e.*, recovering s_{t-1} from s_t by predicting the added random noise ϵ. In this process, a denoising model $f(s_t, t)$ models the prediction from s_t to s_0. The training of this network relies on the denoising loss, formulated as follows:

$$\mathcal{L}_{s_0} = \mathbb{E}_{\epsilon\sim\mathcal{N}(0,I),t\sim\mathcal{U}(0,1)}\|f(s_t,t) - s_0\|_2^2, \tag{2}$$

where s_t is sampled using Eq. 1.

Our semantic voxel diffusion model is based on the U-Net architecture [54]. The U-Net comprises six levels with voxel resolutions of 64^3, 32^3, 16^3, 8^3, 4^3, and 2^3, and corresponding feature dimensions are 16, 16, 32, 64, 128, and 256, respectively. Each level comprises a residual block that incorporates the noise level information. An additional residual block is added in the bottleneck of the U-Net. Furthermore, we add five mix blocks to integrate the image features obtained from the ViT encoder [12] into the training process. Finally, five deconvolution layers with kernel size four for each upsampling layer are attached at the end of the network to map the image features at the finest level to the SVS of size 64^3.

SVS inference. We initialize s_t with Gaussian noise, which is fed into the above denoising model to regress the classification values of SVS by using the DDPM sampling strategy [20]. The semantic label $C_{\mathcal{V}_i}$ for each voxel $\mathcal{V}_i$ is then determined by the regressed classification value $S_{\mathcal{V}_i}$: $C_{\mathcal{V}_i}$ is trunk if $S_{\mathcal{V}_i} \in [0.8, 1]$, $C_{\mathcal{V}_i}$ is branch if $S_{\mathcal{V}_i} \in [0.4, 0.8]$, $C_{\mathcal{V}_i}$ is leaf if $S_{\mathcal{V}_i} \in [0, 0.4]$.

4.4. Hybrid Tree Geometry Reconstruction

Once the SVS is predicted, we synthesize the fine-grained 3D tree geometry, and Fig. 5 illustrates how the 3D branching structure is progressively constructed from SVS. Our algorithm consists of two sub-steps: (a-e) constructing the initial main skeleton from non-leaf voxels and (f) propagating tiny twigs and leaves to detail the tree crown.

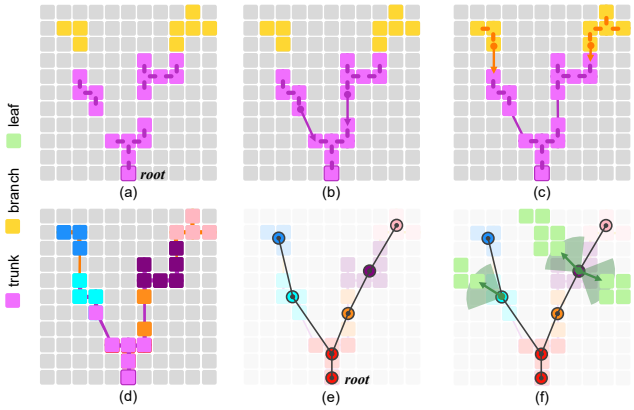


Figure 5. **2D illustration of hybrid tree geometry reconstruction.** (a) Construct subgraphs for trunk voxels. (b) Sequentially connect each subgraph to the lowest one. (c) Extend to the branch voxels to form the complete connected graph. (d) Cluster the voxels into a set of groups based on the graph adjacency. (e) Form the main skeleton by computing centroids for each voxel group. (f) Synthesize the crown inside the area of leaf voxels.

Main Skeleton Construction. Our strategy to construct the main skeleton from non-leaf voxels is inspired by a heuristic modeling algorithm [63]. As shown in Fig. 5 (a), we take the lowest trunk voxel in SVS as the tree root. In this step, each trunk voxel is connected to all other voxels within its direct local neighborhood (threshold=1.0) to construct a set of subgraphs. Then, starting from the lowest subgraph, we traverse all the graphs and connect them sequentially to the main graph (Fig. 5 (b)). The same operation is applied to the branch voxels (Fig. 5 (c)). We construct the connected graph separately for trunk and branch voxels because these two types typically exhibit different prediction noises. Thus, if the trunk voxels, which are often predicted more accurately, are considered first, the resulting main graph can have a more structurally correct connection accordingly.

We then use a minimum spanning tree algorithm to the constructed graph and cluster all non-leaf voxels into a series of groups (marked in different colors in Fig. 5 (d)) based on the graph adjacency. Finally, the centroid of each group of voxels is calculated (Fig. 5 (e)). The centroids serve as the tree nodes. Finally, the main skeleton is constructed by connecting the centroid points of the adjacent voxel groups.

Crown Synthesis. After constructing the main skeleton, we apply a developmental growth model to complete the tree crown (Fig. 5 (f)). This step is inspired by a space colonization algorithm [55], which simulates the competition of branches for space to grow branches inside a given 3D space. We use all the leaf voxels as an intersection volume. Each newly generated branch outside the volume of

leaf voxels is removed directly. The branch growth process stops until all the leaf voxels are touched by branches with a threshold of 0.5. As a result, the entire 3D tree skeleton is strictly inside the SVS. Finally, the leaves are randomly distributed along the branches within the area of leaf voxels.

5. Results and Evaluation

Implementation. The algorithm for generating the synthetic dataset and our hybrid tree reconstruction has been implemented in C++ with OpenGL. By employing various random parameters in procedural modeling, we generate 4,000 data pairs containing single images and corresponding SVS to train our diffusion model. Tree species that are difficult to synthesize by [46], such as palms, are excluded from the dataset. For SAM instance segmentation in Sect. 4.2, we utilized the publicly available PyTorch implementation[1]. Our semantic voxel diffusion model was developed in PyTorch and run on an Intel Xeon Gold 6226R at 2.90GHz with 8× Nvidia GeForce RTX 2080 GPUs.

Diffusion model training. The network is trained using a Mean Absolute Error (MAE) loss for regression, employing the Adam optimizer with a learning rate 10^{-4} and a batch size 2. We dynamically adjusted the learning rate using a cosine annealing schedule. The training process takes approximately three days.

5.1. Results

Evaluation on synthetic data. We additionally generate 100 testing images using procedural modeling to evaluate the accuracy of our SVS predictions, where the tree model parameters differ from those employed in the training dataset. These testing examples serve as the SVS ground truth, facilitating qualitative and quantitative evaluations. On average, the instance segmentation using SAM takes 12.6 seconds to process one image, SVS prediction using the diffusion model takes 4.4 seconds, and procedural geometry reconstruction requires about 13.7 seconds. Tab. 1 shows the statistics and the distribution of different classes in the predicted SVS, from which we can observe the class imbalance between empty voxels and tree voxels.

Fig. 3 shows two examples of the ground truth SVS and predicted SVS. The predicted SVS faithfully maintains its original relative structure, aligning closely with the ground truth. This indicates that the semantic voxel diffusion model effectively extracts tree information, even the missing branch details, from single images, providing an approximate predicted geometric appearance and structural information using SVS representation.

Fig. 6 shows branching geometry reconstruction using our hybrid reconstruction algorithm described in Sect. 4.4. The first column shows SVS containing semantic information about the tree structure, encompassing the tree trunk and main branches. The second and third columns represent the ground truth tree model and the tree model reconstructed with our algorithm based on the SVS. The figure not only demonstrates the capability of SVS to represent intricate tree structures required for tree models but also validates the effectiveness of our hybrid tree modeling algorithm in extracting semantic structural information from SVS for accurate tree reconstruction.

[1]https://github.com/facebookresearch/segment-anything

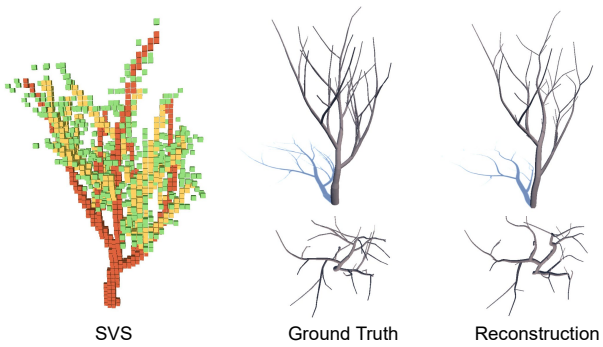


Figure 6. Our proposed tree geometry reconstruction algorithm can generate high-fidelity branching structures from SVS.

Table 1. Voxel distribution of different classes in the predicted SVSs of all synthetic testing images.

Methods	#Trunk	#Branch	#Leaf	#Empty	Total
Voxel Num.	715	2,357	9,946	24,9126	26,2144 (*i.e.*, 64^3)
Ratio	0.27%	0.94%	3.79%	95%	100%

Evaluation on real-world data. We conduct experiments on a collection of real-world images from public sources or captured using a smartphone to demonstrate the robustness and generalization ability of our approach.

Fig. 7 presents several leafy trees by showing the step-by-step results of our reconstruction process. The SVSs are obtained by employing the same diffusion model trained on the synthetic dataset, and the final tree models are generated by our hybrid reconstruction algorithm. Each predicted SVS (second column) captures the overall geometric shape of the input tree, while branching structures (third column) reveal that tree skeletons can be faithfully extracted benefiting from the semantic structure information provided by SVS. Finally, the rendering images from the front, side and top views demonstrate that all resulting trees closely resemble the target images. Our approach reproduces realistic 3D tree models in preserving intricate tree details, yielding accurate reconstructions.

Next, we conduct an experiment focusing on multi-tree reconstruction (Fig. 8). Leveraging the image segmentation model of SAM, we can extract accurate segmented mask information to generate three cropped images (b). These cropped images are then individually fed into our seman-

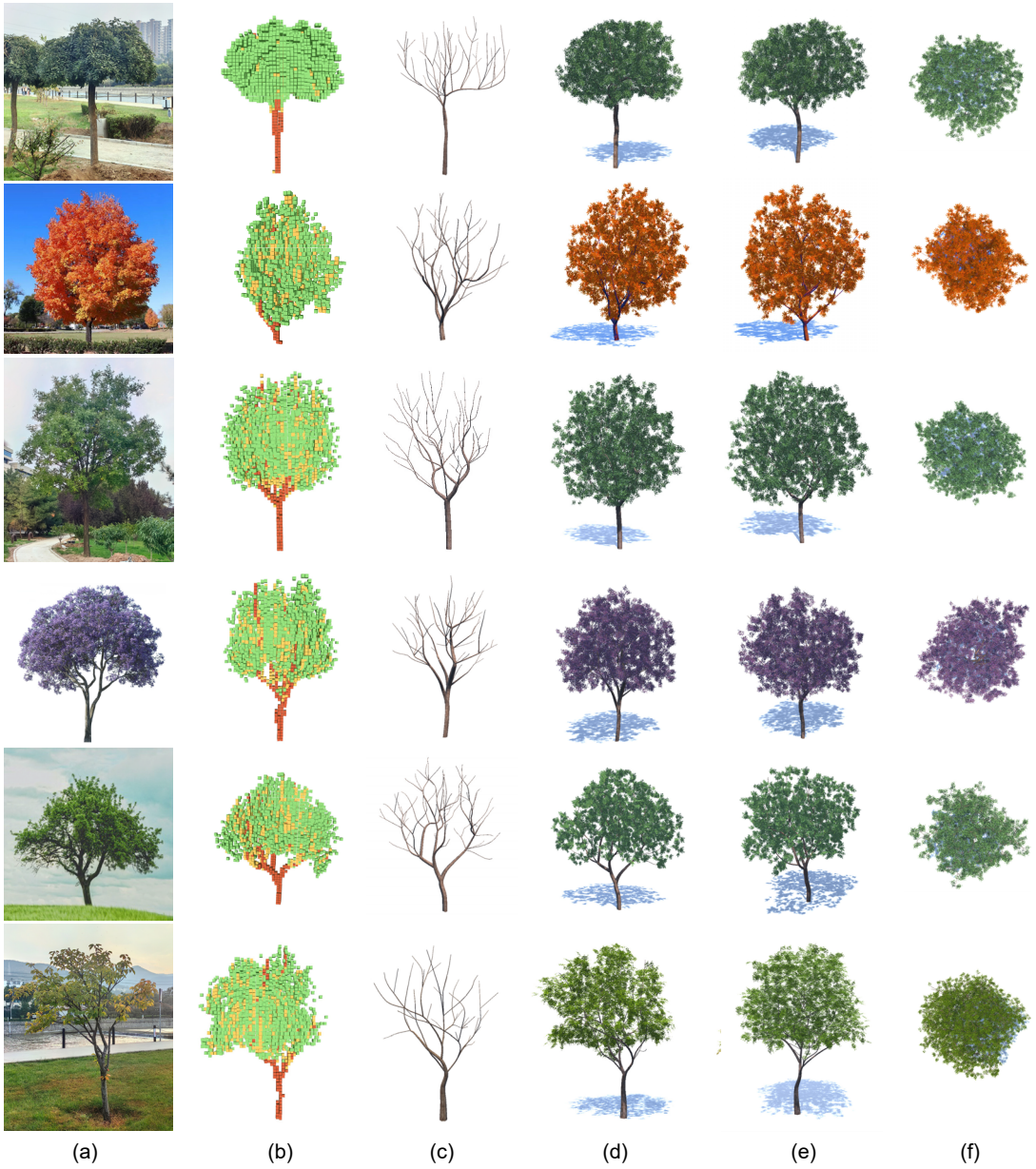


Figure 7. Reconstruction of real trees. For each example, we show input photograph (a), predicted SVS (b), reconstructed branch structures (c), and the final full models rendered from front, side, and top views (d-f).

tic voxel diffusion model to predict SVSs, whose positions correspond to the locations of multiple trees in the original image (c). Final tree models are obtained by applying our hybrid reconstruction algorithm to the SVSs (d). Note that we do not aim to learn the relative positioning of trees, therefore the 3D trees are arranged manually.

5.2. Comparisons

Reconstructing trees from point clouds [30, 32] or multiple images [18, 22, 57] is popular and usually generates more accurate results than our approach. However, collecting such input data is costly and time-consuming, and multiple views are not often available. For single-image tree reconstruction, the semi-automatic methods proposed by Tan et al. [58] and Liu et al. [31] rely on user annotations to identify the main branches or the crown shape. Li et al. [26] propose the first automatic tree reconstruction from single images, which is most closely related to ours.

Fig. 9 compares the reconstruction results between [26] and our method. Given single photographs, [26] auto-

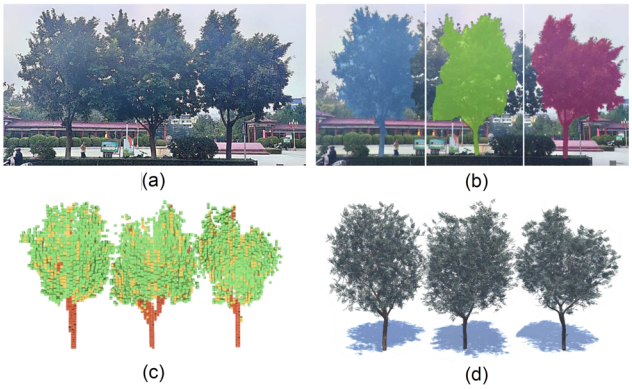


Figure 8. Reconstruction of multiple trees from a single image.

Table 2. Ablation study of the diffusion model and mix block with attention. We report precision for the class of trunk (AP_T), branch AP_B, leaf AP_L, empty AP_E, and the mean average precision mAP over all classes. AE_{mean} and AE_{sd} represent the average and the standard deviation of absolute errors, respectively.

Methods	Precision					Absolute Error	
	AP_T	AP_B	AP_L	AP_E	mAP	AE_{mean}	AE_{sd}
cGAN	0.130	0.033	0.137	0.886	0.297	0.0468	0.244
Ours (W/o Att.)	0.515	0.061	0.313	0.952	0.460	0.0169	0.208
Ours (Full)	**0.538**	**0.176**	**0.461**	**0.991**	**0.546**	**0.0134**	**0.198**

matically learns radial bounding volumes (RBV) that are then utilized to guide tree growth, and a species identification network provides parameter values for a developmental tree model. Our approach leverages SVS representation, which captures finer details, including the intricate interior branches and spatial morphology of foliage. Moreover, our reconstruction approach does not rely on manually provided parameters for tree model development, surpassing the reconstruction efficacy of the RBV method.

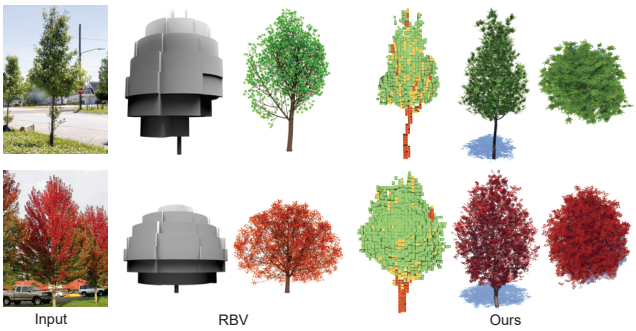


Figure 9. Comparison to RBV [27]: input, RBVs, reconstruction using [27], and our reconstructed SVSs and tree models.

5.3. Ablations

We report the standard metrics, including the precision of predicting each semantic label, and the absolute error between ground truth SVSs and the predicted SVSs, for quantitative evaluations of different network configurations.

Effect of the diffusion model. We first examine the effect of our diffusion model on recovering 3D shape and semantic information. We replace the diffusion model with a conditional Generative Adversarial Network (cGAN) architecture adopted in [31], with the encoder also utilizing ResNet-50 and the decoder employing deconvolution layers to upsample the output voxels to a resolution of 64^3. Tab. 2 summarizes the evaluation results. It reveals that the proposed method achieves the best performance across all metrics. Our precision values of classifying trunks and leaves are 40% and 33% higher than those of using cGAN. In terms of mean average precision and mean absolute error, we still attain significantly higher scores.

Effect of mix block with attention. We propose the mix block with multi-head attention to better integrate image features for SVS prediction. We conducted an experiment using a mix-block without the attention mechanism to assess its efficiency. In this setup, we directly concatenate the image feature obtained by the encoder to the SVS feature for denoising. As shown in Tab. 2, by using the multi-head attention mechanism, the diffusion model improves the regression and classification performance.

6. Conclusion and Future Work

We introduce a novel structured voxel representation of trees in 3D space to encode trunks, branches, and leaves, thereby expressing detailed features such as the trunk's primary topology and the foliage's spatial morphology. We propose a semantic voxel diffusion model to generate such 3D tree structures from single photographs faithfully. Moreover, we develop a new tree geometry construction algorithm based on a hybrid modeling approach (combining structure-oriented branch reconstruction and self-organization-based foliage reconstruction) specially designed for our customized data representation.

In future work, we aim to explore richer information from images, such as learning tree branch growth directions and crown color information. We will also explore the additional reconstruction gained by incorporating 3D prior (e.g., depth). Moreover, we plan to simulate a more diverse range of tree species, expand our training dataset, and consider the reconstruction of multiple trees simultaneously.

Acknowledgments. We thank the anonymous reviewers for their valuable suggestions. This work is partially funded by the National Natural Science Foundation of China (62172416, U22B2034, U21A20515, 32271983, 62262043), and Guangdong Science and Technology Program (2023B1515120026). This work was supported by NRCS grant #NR233A750004G044 to Benes. The findings and conclusions should not be construed to represent any agency determination or policy.

References

[1] Oscar Argudo, Antonio Chica, and Carlos Andujar. Single-picture reconstruction and rendering of trees for plausible vegetation synthesis. *Computers & Graphics*, 57:55–67, 2016. 2

[2] Sara Beery, Guanhang Wu, Trevor Edwards, Filip Pavetic, Bo Majewski, Shreyasee Mukherjee, Stanley Chan, John Morgan, Vivek Rathod, and Jonathan Huang. The auto arborist dataset: A large-scale benchmark for multiview urban forest monitoring under domain shift. In *IEEE Computer Vision and Pattern Recognition (CVPR)*, pages 21294–21307, 2022. 1

[3] Peter B Boucher, Ian Paynter, David A Orwig, Ilan Valencius, and Crystal Schaaf. Sampling forests with terrestrial laser scanning. *Annals of Botany*, 128(6):689–708, 2021. 1

[4] Derek Bradley, Derek Nowrouzezahrai, and Paul Beardsley. Image-based reconstruction and synthesis of dense foliage. *ACM Trans. Graph.*, 32(4):74, 2013. 2

[5] Nicolas Carion, Francisco Massa, Gabriel Synnaeve, Nicolas Usunier, Alexander Kirillov, and Sergey Zagoruyko. End-to-end object detection with transformers. In *European Conference on Computer Vision (ECCV)*, pages 213–229. Springer, 2020. 4

[6] Eric R Chan, Marco Monteiro, Petr Kellnhofer, Jiajun Wu, and Gordon Wetzstein. pi-gan: Periodic implicit generative adversarial networks for 3d-aware image synthesis. In *IEEE Computer Vision and Pattern Recognition (CVPR)*, pages 5799–5809, 2021. 2

[7] Eric R Chan, Connor Z Lin, Matthew A Chan, Koki Nagano, Boxiao Pan, Shalini De Mello, Orazio Gallo, Leonidas J Guibas, Jonathan Tremblay, Sameh Khamis, et al. Efficient geometry-aware 3d generative adversarial networks. In *IEEE Computer Vision and Pattern Recognition (CVPR)*, pages 16123–16133, 2022. 2

[8] Chao Chen, Zhizhong Han, Yu-Shen Liu, and Matthias Zwicker. Unsupervised learning of fine structure generation for 3d point clouds by 2d projections matching. In *IEEE International Conference on Computer Vision (ICCV)*, pages 12466–12477, 2021. 2

[9] Zhiqin Chen and Hao Zhang. Learning implicit fields for generative shape modeling. In *IEEE Computer Vision and Pattern Recognition (CVPR)*, pages 5939–5948, 2019. 2

[10] Christopher B Choy, Danfei Xu, JunYoung Gwak, Kevin Chen, and Silvio Savarese. 3d-r2n2: A unified approach for single and multi-view 3d object reconstruction. In *European Conference on Computer Vision (ECCV)*, pages 628–644, 2016. 2

[11] Yu Deng, Jiaolong Yang, Jianfeng Xiang, and Xin Tong. Gram: Generative radiance manifolds for 3d-aware image generation. In *IEEE Computer Vision and Pattern Recognition (CVPR)*, pages 10673–10683, 2022. 2

[12] Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, et al. An image is worth 16x16 words: Transformers for image recognition at scale. *arXiv preprint arXiv:2010.11929*, 2020. 5

[13] Shenglan Du, Roderik Lindenbergh, Hugo Ledoux, Jantien Stoter, and Liangliang Nan. Adtree: Accurate, detailed, and automatic modelling of laser-scanned trees. *Remote Sensing*, 11(18):2074, 2019. 2

[14] George Fahim, Khalid Amin, and Sameh Zarif. Single-view 3d reconstruction: A survey of deep learning methods. *Computers & Graphics*, 94:164–190, 2021. 2

[15] Haoqiang Fan, Hao Su, and Leonidas J Guibas. A point set generation network for 3d object reconstruction from a single image. In *IEEE Computer Vision and Pattern Recognition (CVPR)*, pages 605–613, 2017. 2

[16] Adnan Firoze, Cameron Wingren, Raymond A Yeh, Bedrich Benes, and Daniel Aliaga. Tree instance segmentation with temporal contour graph. In *IEEE Computer Vision and Pattern Recognition (CVPR)*, pages 2193–2202, 2023. 4

[17] Jun Gao, Tianchang Shen, Zian Wang, Wenzheng Chen, Kangxue Yin, Daiqing Li, Or Litany, Zan Gojcic, and Sanja Fidler. Get3d: A generative model of high quality 3d textured shapes learned from images. In *Advances in Neural Information Processing Systems*, pages 31841–31854, 2022. 2

[18] Jianwei Guo, Shibiao Xu, Dong-Ming Yan, Zhanglin Cheng, Marc Jaeger, and Xiaopeng Zhang. Realistic procedural plant modeling from multiple view images. *IEEE Trans. on Vis. and Comput. Graph.*, 26(2):1372–1384, 2020. 2, 7

[19] Long He and James Schupp. Sensing and automation in pruning of apple trees: A review. *Agronomy*, 8(10):211, 2018. 1

[20] Jonathan Ho, Ajay Jain, and Pieter Abbeel. Denoising diffusion probabilistic models. *Advances in Neural Information Processing Systems*, 33:6840–6851, 2020. 2, 5

[21] Ka-Hei Hui, Ruihui Li, Jingyu Hu, and Chi-Wing Fu. Neural wavelet-domain diffusion for 3d shape generation. In *SIGGRAPH Asia 2022 Conference Papers*, pages 1–9, 2022. 2

[22] Takahiro Isokane, Fumio Okura, Ayaka Ide, Yasuyuki Matsushita, and Yasushi Yagi. Probabilistic plant modeling via multi-view image-to-image translation. In *IEEE Computer Vision and Pattern Recognition (CVPR)*, pages 2906–2915, 2018. 1, 2, 4, 7

[23] 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. *arXiv preprint arXiv:2304.02643*, 2023. 3, 4

[24] Štefan Kohek, Borut Žalik, Damjan Strnad, Simon Kolmanič, and Niko Lukač. Simulation-driven 3d forest growth forecasting based on airborne topographic lidar data and shading. *International Journal of Applied Earth Observation and Geoinformation*, 111:102844, 2022. 1

[25] Jae Joong Lee, Bosheng Li, and Bedrich Benes. Latent l-systems: Transformer-based tree generator. *ACM Trans. Graph.*, 43(1), 2024. 2

[26] Bosheng Li, Jacek Kałużny, Jonathan Klein, Dominik Michels, Wojtek Pałubicki, Bedrich Benes, and Sören Pirk. Learning to reconstruct botanical trees from single images. *ACM Trans. Graph.*, 40(6), 2021. 2, 3, 4, 7

[27] Bosheng Li, Jonathan Klein, Dominik L. Michels, Soeren Pirk, Bedrich Benes, and Wojtek Palubicki. Rhizomorph:

The coordinated function of shoots and roots. *ACM Transaction on Graphics*, 42(4), 2023. 8

[28] Xueting Li, Sifei Liu, Kihwan Kim, Shalini De Mello, Varun Jampani, Ming-Hsuan Yang, and Jan Kautz. Self-supervised single-view 3d reconstruction via semantic consistency. In *European Conference on Computer Vision (ECCV)*, pages 677–693. Springer, 2020. 2

[29] Chen-Hsuan Lin, Jun Gao, Luming Tang, Towaki Takikawa, Xiaohui Zeng, Xun Huang, Karsten Kreis, Sanja Fidler, Ming-Yu Liu, and Tsung-Yi Lin. Magic3d: High-resolution text-to-3d content creation. In *IEEE Computer Vision and Pattern Recognition (CVPR)*, pages 300–309, 2023. 2

[30] Yanchao Liu, Jianwei Guo, Bedrich Benes, Oliver Deussen, Xiaopeng Zhang, and Hui Huang. Treepartnet: neural decomposition of point clouds for 3d tree reconstruction. *ACM Trans. Graph.*, 40(6), 2021. 2, 4, 7

[31] Zhihao Liu, Kai Wu, Jianwei Guo, Yunhai Wang, Oliver Deussen, and Zhanglin Cheng. Single image tree reconstruction via adversarial network. *Graphical Models*, 117: 101–115, 2021. 2, 7, 8

[32] Yotam Livny, Feilong Yan, Matt Olson, Baoquan Chen, Hao Zhang, and Jihad El-Sana. Automatic reconstruction of tree skeletal structures from point clouds. *ACM Trans. Graph.*, 29(6):151, 2010. 1, 2, 7

[33] Yotam Livny, Sören Pirk, Zhanglin Cheng, Feilong Yan, Oliver Deussen, Daniel Cohen-Or, and Baoquan Chen. Texture-lobes for tree modeling. *ACM Trans. Graph.*, 30(4): 1–10, 2011. 1

[34] Luis D Lopez, Yuanyuan Ding, and Jingyi Yu. Modeling complex unfoliaged trees from a sparse set of images. *Comput. Graph. Forum*, 29(7):2075–2082, 2010. 1

[35] Shitong Luo and Wei Hu. Diffusion probabilistic models for 3d point cloud generation. In *IEEE Computer Vision and Pattern Recognition (CVPR)*, pages 2837–2845, 2021. 2

[36] Lars Mescheder, Michael Oechsle, Michael Niemeyer, Sebastian Nowozin, and Andreas Geiger. Occupancy networks: Learning 3d reconstruction in function space. In *IEEE Computer Vision and Pattern Recognition (CVPR)*, 2019. 2

[37] Lukas Meyer, Andreas Gilson, Oliver Scholz, and Marc Stamminger. Cherrypicker: Semantic skeletonization and topological reconstruction of cherry trees. In *IEEE Computer Vision and Pattern Recognition (CVPR)*, pages 6243–6252, 2023. 1, 2

[38] Oscar Michel, Roi Bar-On, Richard Liu, Sagie Benaim, and Rana Hanocka. Text2mesh: Text-driven neural stylization for meshes. In *IEEE Computer Vision and Pattern Recognition (CVPR)*, pages 13492–13502, 2022. 2

[39] Ben Mildenhall, Pratul P Srinivasan, Matthew Tancik, Jonathan T Barron, Ravi Ramamoorthi, and Ren Ng. Nerf: Representing scenes as neural radiance fields for view synthesis. *Communications of the ACM*, 65(1):99–106, 2021. 2

[40] Paritosh Mittal, Yen-Chi Cheng, Maneesh Singh, and Shubham Tulsiani. Autosdf: Shape priors for 3d completion, reconstruction and generation. In *IEEE Computer Vision and Pattern Recognition (CVPR)*, pages 306–315, 2022. 2

[41] Norman Müller, Yawar Siddiqui, Lorenzo Porzi, Samuel Rota Bulo, Peter Kontschieder, and Matthias Nießner. Diffrf: Rendering-guided 3d radiance field diffusion. In *IEEE Computer Vision and Pattern Recognition (CVPR)*, pages 4328–4338, 2023. 2

[42] Boris Neubert, Thomas Franken, and Oliver Deussen. Approximate image-based tree-modeling using particle flows. In *ACM SIGGRAPH 2007 papers*, pages 88–es. 2007. 2

[43] Alex Nichol, Heewoo Jun, Prafulla Dhariwal, Pamela Mishkin, and Mark Chen. Point-e: A system for generating 3d point clouds from complex prompts. *arXiv preprint arXiv:2212.08751*, 2022. 2

[44] Till Niese, Sören Pirk, Matthias Albrecht, Bedrich Benes, and Oliver Deussen. Procedural urban forestry. *ACM Trans. Graph.*, 41(2), 2022. 1

[45] Roy Or-El, Xuan Luo, Mengyi Shan, Eli Shechtman, Jeong Joon Park, and Ira Kemelmacher-Shlizerman. Stylesdf: High-resolution 3d-consistent image and geometry generation. In *IEEE Computer Vision and Pattern Recognition (CVPR)*, pages 13503–13513, 2022. 2

[46] Wojciech Palubicki, Kipp Horel, Steven Longay, Adam Runions, Radomír Měch, and Przemyslaw Prusinkiewicz. Self-organizing tree models for image synthesis. *ACM Trans. Graph. (Proc. SIGGRAPH)*, 28:58:1–58:10, 2009. 1, 2, 3, 4, 6

[47] Junyi Pan, Xiaoguang Han, Weikai Chen, Jiapeng Tang, and Kui Jia. Deep mesh reconstruction from single rgb images via topology modification networks. In *IEEE International Conference on Computer Vision (ICCV)*, pages 9964–9973, 2019. 2

[48] Sören Pirk, Ondřej Štava, Julian Kratt, Michel Abdul Massih Said, Boris Neubert, Radomír Měch, Bedrich Benes, and Oliver Deussen. Plastic trees: interactive self-adapting botanical tree models. *ACM Trans. Graph.*, 31(4):50:1–50:10, 2012. 1

[49] Tomas Polasek, David Hrusa, Bedrich Benes, and Martin Čadík. Ictree: Automatic perceptual metrics for tree models. *ACM Trans. Graph.*, 40(6):1–15, 2021. 4

[50] Ben Poole, Ajay Jain, Jonathan T Barron, and Ben Mildenhall. Dreamfusion: Text-to-3d using 2d diffusion. In *International Conference on Learning Representations (ICLR)*, 2023. 2

[51] Stefan Popov, Pablo Bauszat, and Vittorio Ferrari. Corenet: Coherent 3d scene reconstruction from a single rgb image. In *European Conference on Computer Vision (ECCV)*, pages 366–383, 2020. 2

[52] P. Prusinkiewicz and A. Lindenmayer. *The Algorithmic Beauty of Plants*. Springer–Verlag, New York, 1990. With J.S.Hanan, F.D. Fracchia, D.R.Fowler, M.J.de Boer, and L.Mercer. 1

[53] Alex Reche-Martinez, Ignacio Martin, and George Drettakis. Volumetric reconstruction and interactive rendering of trees from photographs. *ACM Trans. Graph. (Proc. SIGGRAPH)*, 23(3):720–727, 2004. 2

[54] Olaf Ronneberger, Philipp Fischer, and Thomas Brox. U-net: Convolutional networks for biomedical image segmentation. In *Medical Image Computing and Computer-Assisted Intervention–MICCAI 2015: 18th International Conference, Munich, Germany, October 5-9, 2015, Proceedings, Part III 18*, pages 234–241. Springer, 2015. 5

[55] Adam Runions, Brendan Lane, and Przemyslaw Prusinkiewicz. Modeling trees with a space colonization algorithm. In *Proc. of the Third EG Conf. on Nat. Phenomena*, pages 63–70, 2007. 5

[56] Zifan Shi, Sida Peng, Yinghao Xu, Yiyi Liao, and Yujun Shen. Deep generative models on 3d representations: A survey. *arXiv preprint arXiv:2210.15663*, 2022. 2

[57] Ping Tan, Gang Zeng, Jingdong Wang, Sing Bing Kang, and Long Quan. Image-based tree modeling. *ACM Trans. Graph. (Proc. SIGGRAPH)*, 26(3):87, 2007. 1, 2, 7

[58] Ping Tan, Tian Fang, Jianxiong Xiao, Peng Zhao, and Long Quan. Single image tree modeling. *ACM Trans. Graph. (Proc. SIGGRAPH Asia)*, 27(5):108:1–108:7, 2008. 2, 7

[59] Maxim Tatarchenko, Alexey Dosovitskiy, and Thomas Brox. Octree generating networks: Efficient convolutional architectures for high-resolution 3d outputs. In *IEEE International Conference on Computer Vision (ICCV)*, pages 2088–2096, 2017. 2

[60] Ondřej Šťava, Sören Pirk, Julian Kratt, Baoquan Chen, Radomir Mech, Oliver Deussen, and Bedrich Benes. Inverse procedural modelling of trees. *Comput. Graph. Forum*, 33 (6):118–131, 2014. 2

[61] Nanyang Wang, Yinda Zhang, Zhuwen Li, Yanwei Fu, Wei Liu, and Yu-Gang Jiang. Pixel2mesh: Generating 3d mesh models from single rgb images. In *European Conference on Computer Vision (ECCV)*, pages 52–67, 2018. 2

[62] Haozhe Xie, Hongxun Yao, Shengping Zhang, Shangchen Zhou, and Wenxiu Sun. Pix2vox++: Multi-scale context-aware 3d object reconstruction from single and multiple images. *Int. Journal of Computer Vision*, 128(12):2919–2935, 2020. 2

[63] Hui Xu, Nathan Gossett, and Baoquan Chen. Knowledge and heuristic-based modeling of laser-scanned trees. *ACM Trans. Graph.*, 26(4):19, 2007. 1, 5

[64] Ling Yang, Zhilong Zhang, Yang Song, Shenda Hong, Runsheng Xu, Yue Zhao, Wentao Zhang, Bin Cui, and Ming-Hsuan Yang. Diffusion models: A comprehensive survey of methods and applications. *ACM Computing Surveys*, 2022. 1

[65] Yufei Ye, Shubham Tulsiani, and Abhinav Gupta. Shelf-supervised mesh prediction in the wild. In *IEEE Computer Vision and Pattern Recognition (CVPR)*, pages 8843–8852, 2021. 2

[66] Lei Yi, Hongjun Li, Jianwei Guo, Oliver Deussen, and Xiaopeng Zhang. Tree growth modelling constrained by growth equations. *Comput. Graph. Forum*, 37(1):239–253, 2018. 1

[67] Biao Zhang, Matthias Nießner, and Peter Wonka. 3dilg: Irregular latent grids for 3d generative modeling. *Advances in Neural Information Processing Systems*, 35:21871–21885, 2022. 2

[68] Xiaopeng Zhang, Hongjun Li, Mingrui Dai, Wei Ma, and Long Quan. Data-driven synthetic modeling of trees. *IEEE Trans. on Vis. and Comput. Graph.*, 20(9):1214–1226, 2014. 4

[69] Linqi Zhou, Yilun Du, and Jiajun Wu. 3d shape generation and completion through point-voxel diffusion. In *IEEE International Conference on Computer Vision (ICCV)*, pages 5826–5835, 2021. 2

[70] Xiaochen Zhou, Bosheng Li, Bedrich Benes, Songlin Fei, and Soeren Pirk. Deeptree: Modeling trees with situated latents. *IEEE Transactions on Visualization & Computer Graphics*, (01):1–14, 2023. 2