EG UK Computer Graphics & Visual Computing (2026)
Y. Sheng and M. Elshehaly (Editors)

TreesFormer: Multimodal Grammar-Based 3D Tree Reconstruction from Sparse Geodata

Angeliki Grammatikaki[1,2], Milena Vuckovic[2], Manuela Waldner[1]

[1]TU Wien, Institute of Visual Computing and Human-Centered Technology, Vienna, Austria
[2]VRVis GmbH, Vienna, Austria

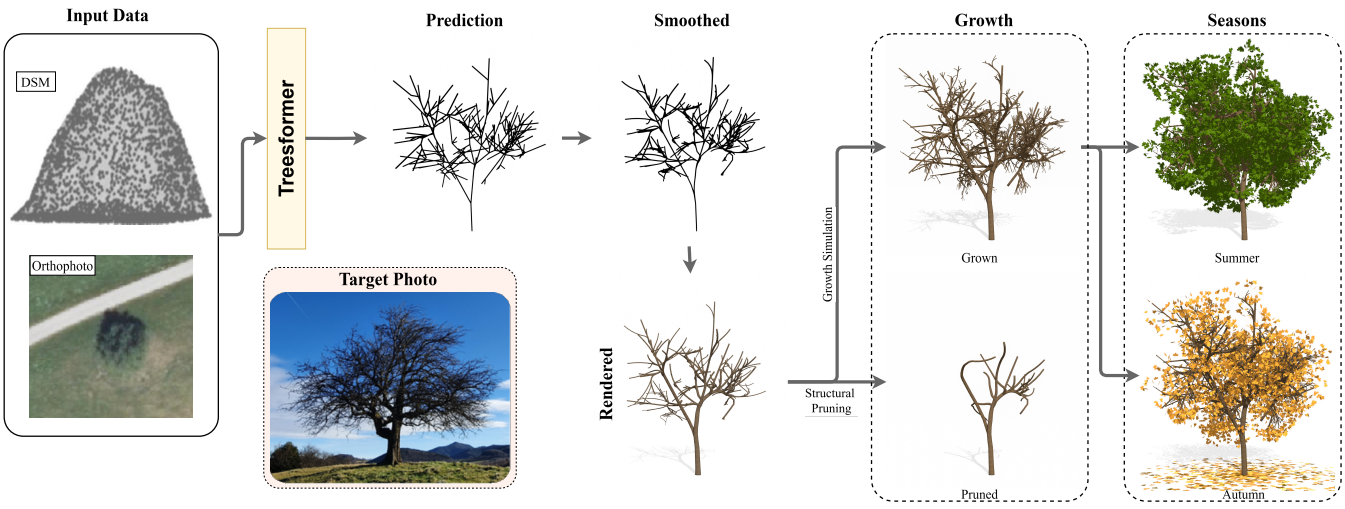


Figure 1: *TreesFormer predicts parametric L-system grammars from DSM and orthophoto inputs, enabling structural smoothing, mesh rendering, growth simulation, and seasonal foliage.*

Abstract

We present TreesFormer, the first grammar-based framework for reconstructing hierarchical 3D tree structures directly from sparse top-down geodata using only a single orthophoto and its corresponding Digital Surface Model (DSM). It employs a multimodal autoregressive transformer that predicts compact parametric L-system grammars from DSM point clouds and orthophoto features, jointly predicting symbolic structure and geometric parameters while enforcing grammar constraints during decoding. To enable supervision in the absence of real-world grammar annotations, we introduce a synthetic multimodal dataset of procedurally generated trees with aligned aerial inputs and ground-truth L-system labels. Experiments show that DSMs drive overall geometric accuracy and crown shape, while orthophoto conditioning improves structural regularity and branching depth; their combination consistently outperforms either modality alone. The model generalizes to real-world Austrian and French aerial data, producing interpretable branching structures suitable for large-scale rural 3D mapping. The codebase, synthetic dataset, and pretrained model are publicly available at https://angelikigram.github.io/treesformer/.

CCS Concepts

• ***Computing methodologies*** → *Neural networks; Mesh models; Reconstruction;*

1. Introduction

Digital 3D maps have become a fundamental medium for representing spatial environments in visualization, simulation, environmental monitoring, and navigation [She05]. While urban scenes are increasingly reconstructed using dense terrestrial LiDAR, structured light scanning, or large-scale multi-view image pipelines [BFGS06; SSS06], natural landscapes are typically mod-

© 2026 The Author(s).
Proceedings published by Eurographics - The European Association for Computer Graphics.
This is an open access article under the terms of the Creative Commons Attribution License, which permits use, distribution and reproduction in any medium, provided the original work is properly cited.

DOI: 10.2312/aeta,0.04/..6

delivered by EUROGRAPHICS DIGITAL LIBRARY www.eg.org diglib.eg.org

eled from aerial observations alone. In practice, publicly available rural datasets primarily consist of orthophotos and Digital Surface Models (DSMs), capturing appearance and elevation from a top-down perspective. Although sufficient for terrain reconstruction, these modalities do not explicitly encode the hierarchical and topological structure of vegetation. Consequently, trees in large-scale rural 3D maps are often approximated as height extrusions or billboard assets, resulting in limited realism in oblique views and reduced structural plausibility.

Vegetation reconstruction has traditionally relied on data-intensive approaches. High-resolution LiDAR acquisition [XGC07; DLL*19], multi-view terrestrial or aerial imagery [SRDT01; YLX*21], and detailed photogrammetric workflows [CCD*19] can recover fine geometric detail and branch-level structure, but require specialized sensors, extensive acquisition campaigns, or controlled imaging conditions. Procedural modeling techniques based on botanical growth rules generate structurally plausible trees, yet typically depend on manual parameter tuning, species priors, or handcrafted rule sets [XM15; PL90]. As a result, a gap persists between scalable aerial data sources and interpretable, structurally expressive representations of vegetation.

We argue that DSMs and orthophotos together provide sufficient geometric and photometric cues to infer plausible branching topology. DSMs encode canopy height and crown footprint, while orthophotos provide appearance and shadow cues that constrain 3D structure; together they yield the strongest reconstructions. In this work, we adopt a grammar-based formulation for tree reconstruction. Tree morphology exhibits recursive branching patterns that can be compactly represented using parametric L-systems [PL90] – symbolic representations that are interpretable, generative, editable, and memory-efficient, making them particularly suitable for large-scale digital mapping. However, learning such sequences from aerial observations requires (1) preserving syntactic validity, (2) maintaining balanced branching structures, and (3) ensuring geometric consistency with observed elevation and appearance cues.

To address these challenges, we propose a multimodal autoregressive transformer that predicts parametric L-system grammars conditioned on DSM point clouds and orthophotos. The architecture integrates cross-attention mechanisms to align symbolic decoding with visual features, and employs a factorized decoder that decouples token-type prediction from geometric parameter estimation. Grammar-constrained decoding enforces syntactic validity throughout generation, while an auxiliary geometric supervision head grounds predicted sequences in explicit three-dimensional spatial context. Since large-scale datasets pairing aerial imagery with ground-truth grammar annotations do not exist, we construct a synthetic benchmark of procedurally generated trees with aligned aerial modalities and ground-truth L-system labels. The main contributions of this work are:

1. The first method to infer parametric L-system grammars from sparse top-down aerial data using a multimodal autoregressive transformer with grammar-constrained decoding and auxiliary geometric supervision.
2. A synthetic multimodal dataset of procedurally generated trees with aligned DSMs, orthophotos, and ground-truth L-system annotations — the first to enable supervised grammar inference directly from aerial observations.
3. A multi-objective supervision strategy for symbolic tree grammar learning that jointly optimizes token classification, discretized geometric parameters, and geometric state consistency for end-to-end training from aerial inputs.

The symbolic output further enables pruning, growth simulation, and seasonal rendering without rerunning inference (Figure 1).

2. Related Work

Early reconstruction approaches relied on dense terrestrial laser scanning to extract explicit branch skeletons and generate high-fidelity structural models [XGC07; DLL*19]. While such methods achieve strong geometric accuracy, they depend on costly acquisition campaigns and controlled conditions, making them unsuitable for large-scale rural environments. Multi-view image-based techniques, including volumetric fusion and point cloud aggregation from calibrated imagery [SRDT01; BNB13], reduce sensor requirements but still demand controlled multi-view capture. TreePartNet [LGB*21] learns a neural cylindrical decomposition of scanned point clouds to reconstruct branch geometry, but requires dense input point clouds. Li et al. [LKK*21] reconstruct botanical trees from single perspective photographs using a pipeline of three neural networks that estimate species, semantic segmentation, and a radial bounding volume used to parametrize a procedural growth model; their method, however, requires species identification and frontal perspective images, assumptions that do not hold in nadir aerial settings. Sketch-based and image-guided methods further enable interactive modeling from sparse cues [TZW*07], though they require user intervention and do not generalize to aerial observations. Our setting differs fundamentally: we reconstruct trees from a nadir orthophoto and a DSM without species labels, multi-view input, or dense terrestrial scans.

Procedural plant modeling synthesizes tree structures through parametric or stochastic growth rules, often based on L-systems or botanical simulations [YWH22; PL90]. Inverse procedural modeling recovers procedural parameters from existing geometric representations. Stava et al. [SPK*14] propose a landmark method that estimates parameters of a developmental tree model from polygonal input using Monte Carlo Markov Chains, demonstrating that procedural parameters can be inferred rather than manually designed. However, their approach requires a full 3D geometric tree as input and cannot operate from sparse aerial observations. Guo et al. [GJB*20] learn L-system representations directly from 2D images of branching structures using deep learning, generating compact textual rewriting rules from visual input. Their formulation is conceptually close to ours in that it recovers symbolic L-system descriptions from perceptual data, but it operates on orthographic 2D images of isolated branching patterns rather than aerial 2.5D geospatial modalities. More recently, Lee et al. [LLB23] demonstrate that a transformer can encode hierarchical L-string structures and generate 3D trees by learning distributions over L-system strings from a large dataset of procedural models; this approach is unconditional and does not infer grammars from visual observations. In contrast, our method conditions grammar prediction on DSM point clouds and orthophotos, and must infer symbolic struc-

© 2026 The Author(s).
Proceedings published by Eurographics - The European Association for Computer Graphics.

ture from a single aerial view without any 3D geometric input or explicit L-system supervision derived from visual parsing.

Several works have specifically targeted reconstruction from minimal aerial inputs. Mayr et al. [MM99] approximated isolated trees with ellipsoidal primitives fitted to DSM height data, while Hirschmugl et al. [HORS07] combined elevation peaks with spectral orthophoto cues for tree detection and coarse modeling. Grammatikaki et al. [GEL*25] proposed an automatic pipeline for modeling landmark trees directly from DSMs and orthophotos. TreeON [GEH*26] extended this to neural point cloud reconstruction from the same inputs, recovering dense geometry but without explicit branching topology or the compactness, editability, and hierarchical organization of a symbolic representation. Vegetation placement has been guided by aerial classification methods [ACC*18; GSF*24], though species-level classification from aerial imagery alone remains unreliable [QLH*23; HOM*24], limiting the applicability of species-conditioned procedural pipelines. While these approaches reconstruct plausible tree geometry from aerial inputs, none produce a compact symbolic representation that is interpretable, editable, and directly compatible with grammar-driven growth and disturbance modeling [PGG*24]. Our method addresses this gap by inferring parametric L-system grammars from the same aerial modalities.

Recent neural approaches have advanced vegetation modeling considerably. DeepTrees [ZLB*23] learns situated latent representations that locally predict branch growth, while SVDTree [LLB*24b] and Tree-D Fusion [LLB*24a] leverage diffusion priors to produce high-fidelity branch-leaf structures and simulation-ready trees from single images. Foliager [TB25] focuses specifically on foliage modeling, and TreeStructor [ZLB*25] assembles forest reconstructions via neural part retrieval. Beyond vegetation-specific methods, diffusion-based and NeRF-based general 3D generation approaches [PJBM22; XLX*25; SCZ*23] learn powerful object-level generative priors that produce compelling geometry. However, all of these methods generally assume perspective imagery, multi-view inputs, semantic labels, or genus-level priors, and represent trees as surface meshes, volume fields, or point clouds rather than compact symbolic structures. Structured program-based representations have been explored in ProcGen3D [ZGLD25], which learns neural procedural graphs conditioned on visual observations, and 3D-GPT [SHD*24], which generates procedural descriptions from textual prompts via large language models. While these highlight the potential of neuro-symbolic 3D generation, they operate in object-centric or language-conditioned settings rather than from top-down aerial geodata.

In contrast to all prior work, our method predicts hierarchical parametric L-system grammars from DSM point clouds and orthophotos without species labels, multi-view imagery, or terrestrial acquisition — to our knowledge the first approach to learn inverse L-system grammar prediction from aerial geodata.

3. Overview

We learn a grammar-based 3D tree representation from aerial observations. Given an orthophoto and a corresponding DSM point cloud, our model autoregressively predicts a parametric L-system describing branching structure and growth geometry. Supervision is enabled through a synthetic dataset of procedurally generated trees with aligned aerial renderings and ground-truth L-system labels.

3.1. Synthetic Training Dataset Generation Pipeline

Since paired aerial data with ground-truth grammar annotations is unavailable, we construct a synthetic dataset of procedurally generated trees with aligned DSMs, orthophotos, and L-system labels. Point clouds serve as a unified 3D representation: DSM back-projection naturally yields sparse 3D points preserving metric elevation, while ground-truth geometry is sampled to match this representation.

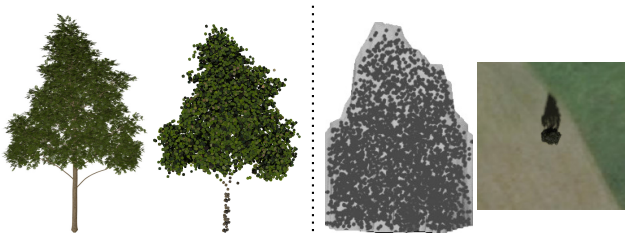


Figure 2: *Tree generation pipeline: (a) procedural mesh, (b) sampled colored point cloud, (c) rendered DSM point cloud (visualized as surface with interior samples), and (d) orthophoto.*

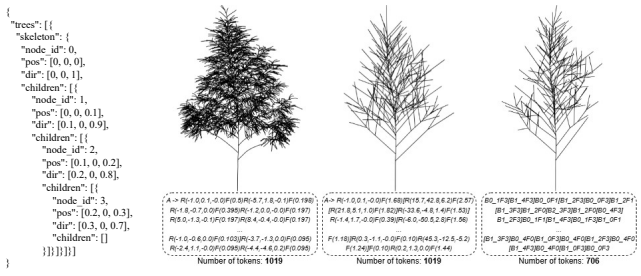
Tree, DSM, and Orthophotos Generation Following the synthetic generation pipeline of Grammatikaki et al. [GEH*26], tree geometries are created using the Grove add-on [The25], producing botanically plausible branching structures with explicit hierarchical topology (Figure 2). Poisson-disk sampling [Bri07] reduces each mesh to $K \approx 2{,}000$ points (Figure 2(b)). Orthophotos and DSMs are rendered from a top-down orthographic camera under ten sampled light directions (azimuth $\sim U(0^\circ, 360^\circ)$, elevation $\sim U(40^\circ, 70^\circ)$) on terrain patches spanning flat, hilly, and forested regions; DSMs are extracted from the Z-buffer, normalized to $[0,1]$, and back-projected into 3D point clouds sampled to $K \approx 2{,}000$ points per tree. The novel contribution of our pipeline is the extraction of ground-truth L-system labels from the resulting branch hierarchies, described below.

L-System Reconstruction We derive a compact parametric L-system description from the branch hierarchy extracted from each Grove-generated mesh, ensuring the model learns a species-agnostic grammar formalism independent of any specific procedural engine and generalizing to real-world trees. Each tree is represented as a hierarchy of branches, where every branch consists of an ordered sequence of 3D points forming a polyline (Figure 3a). We convert this hierarchy into a parametric L-system of the form:

$$\mathcal{A} \rightarrow R(\alpha, \beta, \gamma)\, \mathcal{S}, \tag{1}$$

where $\mathcal{A}$ is the axiom (root symbol initiating the derivation), $\mathcal{S}$ is the derived command string, $F(l)$ moves the turtle cursor – a virtual pen that traces branch geometry by executing the grammar commands one step at a time – forward by length l producing a branch

© 2026 The Author(s).
Proceedings published by Eurographics - The European Association for Computer Graphics.



(a) (b) (c) (d)

Figure 3: *Transformation from hierarchical tree structure to the discrete symbolic sequence used for training: (a) JSON hierarchy, (b) continuous parametric L-system $R(\alpha,\beta,\gamma)$, $F(l)$, (c) simplified L-system, (d) discretized spherical tokens $B_{b_\theta,b_\phi}F_{b_\ell}$.*

segment, and $\mathcal{R}(\alpha,\beta,\gamma)$ rotates its orientation. The symbols [and] push and pop the cursor state to open and close side branches. The points $\mathbf{p_i}$ are the 3D positions visited by the cursor as it executes $\mathcal{S}$.

For each consecutive segment $(\mathbf{p}_i,\mathbf{p}_{i+1})$, we compute its length $l = \|\mathbf{p}_{i+1} - \mathbf{p}_i\|$ and the minimal rotation aligning the current forward direction with the normalized segment direction. The alignment rotation is derived from the rotation matrix mapping a canonical forward vector to the target direction and converted to Euler angles (XYZ convention). Side branches are encoded using bracket tokens inserted recursively to preserve hierarchy. To improve compactness and numerical stability, we remove short segments, merge nearly collinear segments, eliminate negligible rotations, and prune branches below a minimum length threshold or with fewer than a minimum number of child segments, yielding a syntactically valid parametric L-system that faithfully approximates the branch hierarchy while remaining tractable for neural sequence modeling.

Token Budget Simplification Raw reconstructed sequences may contain tens of thousands of tokens, which is infeasible for autoregressive transformer training due to quadratic memory complexity and long-range dependency accumulation. We therefore enforce a maximum length of 1024 tokens per tree. As shown in Figure 3, an adaptive simplification pipeline merges consecutive forward commands, collapses repetitive rotation–forward chains, removes negligible rotations, prunes short branches, and limits branching depth. If necessary, pruning thresholds are progressively relaxed until the token budget is met. This preserves the trunk and primary branching structure while discarding high-frequency geometric detail.

Absolute Spherical Re-Parameterization The simplified L-system represents growth as incremental rotations $R(\alpha,\beta,\gamma)$ followed by forward commands $F(l)$. Although spherical directions could be computed directly from branch polylines, we re-parameterize after simplification because geometry-aware operations (e.g., collinearity merging, rotation pruning, branch filtering) are more naturally expressed in incremental rotation space, and early discretization would introduce quantization artifacts. Moreover, incremental rotations accumulate numerical noise and induce long-range dependencies during autoregressive modeling; our re-parameterization to absolute spherical coordinates eliminates this accumulation entirely. We therefore simulate the turtle interpreter over the simplified sequence to compute the absolute forward direction $\mathbf{d}_t = \mathbf{O}_t(0,0,1)^\top$, where $\mathbf{O}_t$ denotes the accumulated orientation at timestep t, and convert it to spherical coordinates $\theta_t = \arccos(d_{t,z})$, $\phi_t = \text{atan2}(d_{t,y}, d_{t,x})$, $\phi_t \in [0, 2\pi)$. Each rotation–forward pair is then replaced by a compact token:

$$B_{b_\theta,b_\phi}F_{b_\ell}, \tag{2}$$

where B and F are the direction and forward token symbols respectively, and $b_\theta \in \{0,\dots,N_\theta-1\}$, $b_\phi \in \{0,\dots,N_\phi-1\}$, $b_\ell \in \{0,\dots,N_\ell-1\}$ are discrete bin indices obtained by uniformly quantizing θ, ϕ, and ℓ into N_θ, N_ϕ, and N_ℓ bins respectively. This representation also reduces sequence length by collapsing rotation–forward pairs into single tokens, resulting in an average sequence length of ≤ 700 tokens per tree.

3.2. Network Architecture

We propose a multimodal autoregressive transformer (Figure 4) that generates discretized parametric L-system sequences conditioned on aerial observations. The model predicts symbolic branching structure and geometric growth parameters in a factorized manner while enforcing syntactic validity through grammar-constrained decoding.

The L-system vocabulary consists of five token types $\{F, [,], \text{EOS}, \text{PAD}\}$. The token F represents a growth segment, [and] denote push and pop stack operations for branching, EOS marks sequence termination, and PAD is used for fixed-length window padding. Each F token is associated with three discretized geometric parameters and is represented as $F(\ell,\theta,\phi)$, where ℓ denotes a forward length bin, $\theta \in [0,\pi]$ the polar angle bin, and $\phi \in [0,2\pi]$ the azimuth angle bin. Angles are defined in an absolute spherical coordinate system; consequently, each predicted segment specifies an absolute 3D direction rather than an incremental rotation.

Generation is conditioned on two input modalities: a normalized DSM point cloud and a corresponding orthophoto image. The orthophoto is processed by a ResNet-18 backbone truncated before the final classification layer, producing a spatial feature map that is linearly projected to the transformer model dimension d and reshaped into a sequence of image tokens. The DSM point cloud of N 3D points – represented in metric elevation space rather than as a rasterized heightmap to preserve sharp height discontinuities – is processed independently per point using a PointNet-style [QSMG17] multilayer perceptron. The resulting per-point features are projected to dimension d and subsampled to a fixed bottleneck size to control memory complexity. The image tokens and DSM tokens are concatenated to form a unified visual memory matrix $\mathbf{M} \in \mathbb{R}^{N_v \times d}$, which remains fixed during decoding as the key–value memory for cross-attention.

Sequence generation follows teacher-forced autoregressive training. At timestep t, the decoder receives embeddings of previously generated tokens (shifted right during training). Each input embedding consists of a learned token-type embedding combined with learned embeddings of associated geometric value bins when applicable. The backbone decoder comprises L transformer layers with causal self-attention and feed-forward networks. Self-attention employs key–value caching for efficient autoregressive inference, and

© 2026 The Author(s).
Proceedings published by Eurographics - The European Association for Computer Graphics.

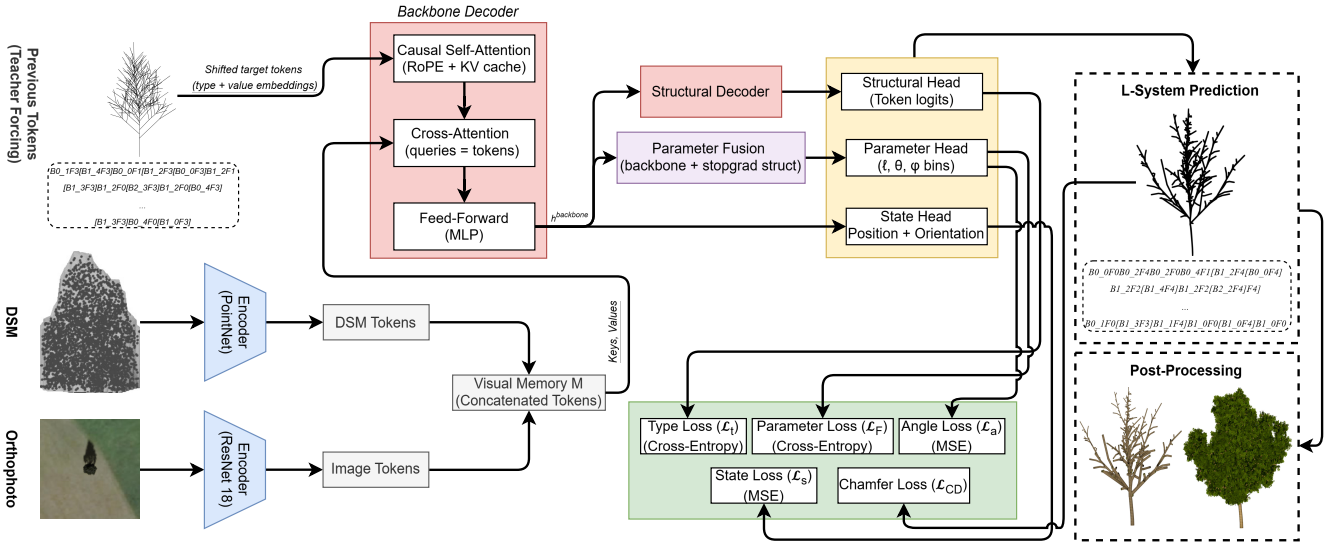


Figure 4: *Overall architecture of the multimodal L-system generator. A multimodal visual encoder conditions an autoregressive backbone decoder factorized into a structural branch for token prediction and a parameter branch for geometric prediction, with supervision (green).*

positional information is encoded using Rotary Position Embeddings (RoPE), applied directly to query and key vectors within the attention mechanism. Cross-attention layers are interleaved with self-attention layers and applied in every second decoder layer. In these layers, decoder hidden states attend to the fixed visual memory $\mathbf{M}$, enabling multimodal conditioning while controlling computational cost.

Let $\mathbf{h}_t^{\text{backbone}} \in \mathbb{R}^d$ denote the backbone hidden representation at timestep t. To decouple symbolic grammar prediction from geometric parameter prediction, this representation is factorized into two branches. The structural branch first projects $\mathbf{h}_t^{\text{backbone}}$ to a reduced dimensionality and processes it through a lightweight decoder composed solely of causal self-attention layers. This branch produces structural features $\mathbf{h}_t^{\text{struct}}$ used for token-type prediction.

Geometric parameter prediction is conditioned on both multimodal grounding and structural context. Specifically, the parameter branch receives the concatenation of the backbone representation and a detached version of the structural features, $\tilde{\mathbf{h}}_t = \text{Proj}\big([\mathbf{h}_t^{\text{backbone}}, \text{stopgrad}(\mathbf{h}_t^{\text{struct}})]\big)$, where the stop-gradient operation prevents parameter supervision from influencing structural learning. The projected representation is normalized and fed into three independent linear classifiers that predict length bin logits, polar angle bin logits, and azimuth angle bin logits.

In addition to symbolic and geometric predictions, the model produces a 9D geometric state vector directly from $\mathbf{h}_t^{\text{backbone}}$, comprising the current 3D position and the two basis vectors of the cursor's local coordinate frame (branch growth direction and its perpendicular axis), from which the full orientation is recoverable. This auxiliary prediction grounds the decoder in explicit 3D spatial context by supervising the position component directly.

During decoding, token logits are masked using a fixed grammar transition matrix $\mathcal{M} \in \{0,1\}^{V \times V}$, where invalid transitions are assigned $-\infty$ prior to softmax. Additional runtime constraints enforce balanced brackets, restrict maximum branch depth, prevent branch closure before emitting a growth token, and allow EOS only at root depth. These constraints are applied during both training-time sampling and inference to guarantee syntactically valid L-system sequences.

3.3. Supervision Strategy

Training uses teacher forcing over fixed-length windows of 700 tokens. Let Ω denote valid (non-PAD) timesteps and $\Omega_F \subset \Omega$ the subset where the ground-truth token is F.

Token classification At each $t \in \Omega$, the structural decoder predicts a distribution over the token vocabulary. This loss drives learning of symbolic tree structure – branching, growth, and termination – with grammar masking applied before softmax to ensure invalid transitions carry no probability mass: $\mathcal{L}_t = \frac{1}{|\Omega|} \sum_{t \in \Omega} \text{CE}(\hat{y}_t, y_t)$.

Parameter classification At each $t \in \Omega_F$, three independent classifiers predict discretized length ℓ, polar angle θ, and azimuth ϕ. This enforces geometric correctness independently of token-type learning, preventing parameter errors from corrupting structural gradients: $\mathcal{L}_F = \frac{1}{|\Omega_F|} \sum_{t \in \Omega_F} \big[\text{CE}(\hat{\ell}_t, \ell_t) + \text{CE}(\hat{\theta}_t, \theta_t) + \text{CE}(\hat{\phi}_t, \phi_t)\big]$.

Soft angle supervision At each $t \in \Omega_F$, an L1 loss on soft-expected bin values $\bar{\theta}_t = \sum_i i \cdot \text{softmax}(\hat{\theta}_t)_i$ penalizes angular error proportionally to its magnitude, complementing the discrete cross-entropy in $\mathcal{L}_F$: $\mathcal{L}_a = \frac{1}{|\Omega_F|} \sum_{t \in \Omega_F} \big[|\bar{\theta}_t - \theta_t| + d_\phi(\bar{\phi}_t, \phi_t)\big]$, where d_ϕ denotes circular distance over azimuth bins.

Geometric state supervision The model predicts a 9D turtle state comprising 3D position, 3D growth direction, and a 3D perpendicular axis at each timestep; the position component is supervised directly via MSE while the orientation axes provide auxiliary structural context to the decoder. This counteracts the recursive

© 2026 The Author(s).
Proceedings published by Eurographics - The European Association for Computer Graphics.

drift that accumulates in geometry-blind autoregressive decoders. Ground-truth positions are normalized per tree for scale-invariant supervision: $\mathcal{L}_s = \frac{1}{|\Omega|}\sum_{t\in\Omega}\|\hat{\mathbf{p}}_t - \mathbf{p}_t\|_2^2$.

Geometric consistency We execute a differentiable turtle simulation – with segment directions and lengths as soft-weighted sums over bins to allow gradient flow – to obtain $\hat{P}$, supervised against the ground-truth point cloud P^* (Figure 2(b)) via Chamfer Distance: $\mathcal{L}_{CD} = \frac{1}{|\hat{P}|}\sum_{x\in\hat{P}}\min_{y\in P^*}\|x-y\|_2^2 + \frac{1}{|P^*|}\sum_{y\in P^*}\min_{x\in\hat{P}}\|y-x\|_2^2$, providing a holistic shape-level signal that complements the token-level supervision losses.

The total objective is a weighted sum, with weights chosen empirically to balance symbolic and geometric supervision:

$$\mathcal{L}_{total} = \lambda_t\mathcal{L}_t + \lambda_F\mathcal{L}_F + \lambda_a\mathcal{L}_a + \lambda_s\mathcal{L}_s + \lambda_{\text{CD}}\mathcal{L}_{\text{CD}}, \quad (3)$$

with $\lambda_t = 0.23$, $\lambda_F = 0.16$, $\lambda_a = 0.24$, $\lambda_s = 0.14$, and $\lambda_{\text{CD}} = 0.23$.

3.4. Post-Processing

The predicted L-system is not a static output – it can be pruned, grown, and seasonally rendered as direct grammar operations without rerunning inference (Figure 1). The raw predicted grammar is first smoothed by interpolating branch directions across sub-steps using a power-curve weighted average, reducing angular discontinuities and producing natural branch curvature. The smoothed grammar is then rendered in Blender as a full 3D mesh: branches are built using radius-decayed cylindrical sweeps with parallel-transport frames to avoid twisting, with radius tapering applied per segment. From the rendered mesh, two grammar-level operations are available without rerunning inference: *structural pruning*, which removes branches below a target hierarchy depth to simulate younger or sparser trees; and *growth simulation*, which injects stochastic sub-branches via a level-dependent spawning rule to densify the canopy. Finally, leaves are instanced procedurally at terminal segments with size and density controls, and a hue-saturation shift applied to leaf textures derived from the corresponding orthophoto patch via histogram matching produce summer and autumn seasonal variants.

4. Experiments

We evaluate TreesFormer on synthetic and real-world datasets through quantitative and qualitative evaluations, measuring reconstruction accuracy, structural completeness, and generalization.

Datasets We use a synthetic dataset of 2,620 trees spanning 17 species with an alpine-proportioned conifer/deciduous split [Fed23], rendered at 30 cm/pixel orthophoto and 1 m DSM resolution following the protocol of Grammatikaki et al. [GEH*26]. For real-world generalization, we evaluate on Austrian landmark trees [Gra24] (orthophotos, DSMs, and photographs; qualitative only) and IGN aerial data from the French Pyrenees [IGN25] (orthophotos at 20 cm/pixel, dense LiDAR ≥ 10 pts/m^2, and DSMs).

Implementation Details We train for 700 epochs using the AdamW optimizer ($lr = 10^{-4}$, batch size 16) with `ReduceLROnPlateau` halving the learning rate after 20 epochs without validation improvement. An 85/15 train/validation split is applied over 2,120 trees, with 500 held out for testing. Training operates on windowed subsequences of length $T = 700$. Shorter sequences are padded with `PAD` tokens excluded from loss computation. All models are trained on a single NVIDIA RTX 3070 (8 GB); training takes ∼24 hours. Inference on a single tree takes ∼0.2 s.

5. Results

We evaluate TreesFormer through ablation studies, qualitative comparison on real-world data, and quantitative comparison against state-of-the-art methods adapted to our aerial input setting.

5.1. Ablation Study

Metrics We report three primary metrics: *Chamfer distance* (↓), the mean bidirectional nearest-neighbor distance between predicted and reference point clouds, measuring geometric fidelity; *Coverage* (↑), the fraction of reference points within a fixed radius of at least one predicted point, reflecting reconstruction completeness; and *Validity* (↑), the proportion of generated L-system sequences that are syntactically well-formed, assessing grammatical correctness – a criterion unique to symbolic reconstruction.

Condition	Ch.↓	Cov.↑	Val.↑
DSM	0.238	67.3%	77.0%
Ortho	0.302	72.9%	68.4%
DSM+Ortho	**0.187**	**82.2%**	**87.0%**

(a) Modality ablation.

Temp.	Ch.↓	Cov.↑	Val.↑
T=0.5	0.314	73.1%	69.4%
T=1.0	**0.187**	**82.2%**	**87.0%**
T=1.5	0.259	71.8%	67.3%

(b) Temperature ablation.

Table 1: Ablation studies.

Modality Ablation Table 1a compares DSM-only, orthophoto-only, and full multimodal conditioning. The full model achieves the best performance across all three metrics (Chamfer 0.187, Coverage 82.2%, Validity 87.0%), confirming that the two modalities carry complementary information. The DSM-only model provides stronger geometric accuracy than orthophoto-only (Chamfer 0.238 vs. 0.302), consistent with elevation being the primary spatial cue, while the orthophoto-only model produces deeper branching hierarchies, suggesting appearance cues contribute more to structural regularity than to spatial precision. Figure 5 confirms this qualitatively: the DSM-only model captures trunk alignment and canopy height but produces shallower branching, while the orthophoto-only model generates richer sub-branching but misestimates absolute geometry; the multimodal model best balances both.

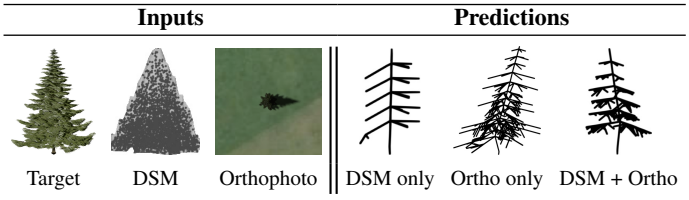


Figure 5: *Qualitative modality ablation on a representative tree.*

Temperature Ablation Table 1b evaluates sampling temperature for the full multimodal model at T=0.5, 1.0, and 1.5. T=1.0 performs best (Chamfer 0.187, Coverage 82.2%, Validity 87.0%);

© 2026 The Author(s).
Proceedings published by Eurographics - The European Association for Computer Graphics.

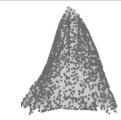
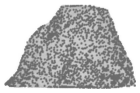
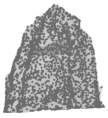
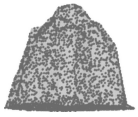
	Typical Reconstructions				Difficult Cases			
	DSM	Orthophoto	Target	Output	DSM	Orthophoto	Target	Output
1					1				
2					2				
3					3				
4					4				

Table 2: *Reconstructed landmark trees (left) and difficult cases (right). Difficult cases: (1) DSM underestimates tree size, (2) cluttered orthophoto, (3) dead tree, (4) failure – simultaneous modality ambiguity.*

higher temperature degrades all metrics (0.259, 71.8%, 67.3%), while lower temperature reduces coverage and validity without improving Chamfer (0.314, 73.1%, 69.4%), suggesting reconstruction quality is primarily governed by multimodal conditioning rather than sampling stochasticity. We adopt $T=1.0$ as the default.

5.2. Results on Austrian Landmark Trees

Individual tree instances are extracted from large aerial scenes using the automatic segmentation and localization pipeline of Grammatikaki et al. [Gra24], ensuring consistent centering and scale across inputs. Table 2 shows reconstructions of real Austrian landmark trees alongside difficult cases, each displaying DSM and orthophoto inputs, a ground-truth photograph, and the predicted 3D tree. Despite training exclusively on synthetic data, the model recovers L-system grammars that reproduce key morphological traits – crown silhouette, height, and branching character – without any species supervision. In case 1, a flat DSM underestimates tree size but the orthophoto compensates; in case 2, a cluttered orthophoto is resolved by DSM elevation structure; in case 3, a dead tree with bare branches is guided by the narrow DSM crown toward a sparse L-system output. Case 4 illustrates a failure: simultaneous ambiguity in both modalities misleads the model into predicting a conifer grammar for a deciduous target. Together, these cases confirm that the dual-input design allows the model to leverage the more reliable modality when one is degraded.

5.3. Comparison with State-of-the-Art Methods

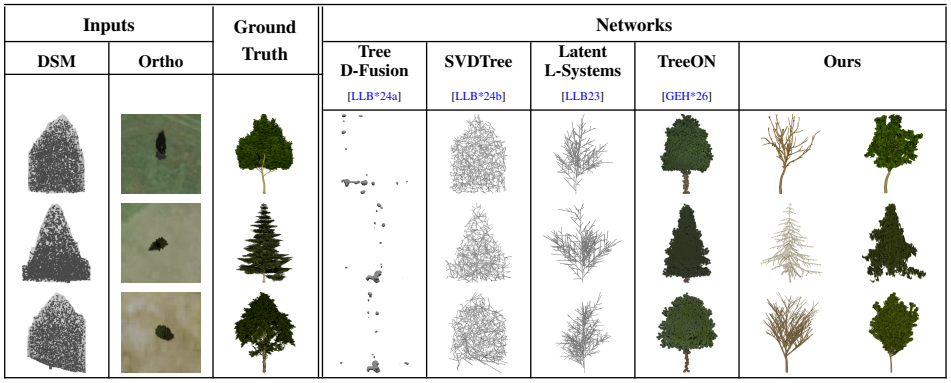
We compare against Tree-D Fusion [LLB*24a] and SVDTree [LLB*24b] (perspective-image mesh reconstruction), Latent L-Systems [LLB23] (unconditional L-system generation), and TreeON [GEH*26] (DSM + orthophoto point cloud reconstruction), retrained or adapted to our aerial inputs: Tree-D Fusion with orthophotos replacing side-view images; SVDTree's skeleton extraction on DSM point clouds instead of dense scans; Latent L-Systems via 100-candidate sampling with nearest-DSM-match selection under Chamfer Distance. All methods are evaluated against the tree point cloud (Figure 2(b); for TreesFormer the foliage-rendered L-system is used as prediction) under a shared protocol (GT-bbox normalisation, 2,048-point subsampling) using NCD, F1, and COV [LBB*23; ALL25; ADMG18].

TreesFormer achieves the best NCD (0.21) and F1 (0.81) overall, with grammar-constrained multimodal conditioning yielding both precise and diverse reconstructions (COV 75%). SVDTree and Tree-D Fusion degrade under aerial inputs – the former because sparse DSM clouds lack the density required for reliable skeleton extraction, the latter because orthophotos cannot replace the structural cues of side-view imagery. Latent L-Systems is unconditional and not directly comparable: its metrics reflect nearest-sample retrieval from a species distribution, as visible in row 2 of Table 3a, where the conifer's DSM footprint leads to a broad deciduous match. TreeON shares our inputs and closely matches canopy geometry (NCD 0.24), and its higher COV (86% vs. 75%) is expected given that a dense point cloud covers more of the GT shape space than a sparse symbolic output; however, its unstructured representation encodes no branching topology, precluding the downstream editing, growth simulation, and seasonal rendering that follow directly from our symbolic output (Figure 1).

5.4. Comparison with High-Resolution LiDAR

We validate our reconstructions against IGN LiDAR scans from the French Pyrenees (Table 4, point cloud density 37 and 25 pts/m^2). Our predicted L-system grammars recover overall crown shape and height consistent with the LiDAR reference, despite using only DSM and orthophoto as input. Fine-grained branch-level detail naturally differs, as LiDAR captures full 3D structure while our

© 2026 The Author(s).
Proceedings published by Eurographics - The European Association for Computer Graphics.



(a) Visual outputs for test trees.

Method	NCD↓	F1↑	COV↑
Tree D-Fusion	0.66	0.01	16%
SVDTree	0.31	0.58	72%
Latent L-Systems	0.33	0.57	44%
TreeON	0.24	0.76	**86%**
Ours	**0.21**	**0.81**	75%

(b) Quantitative results.

Table 3: *Comparison with state-of-the-art methods. (a) Visual outputs; DSM, orthophoto, and ground truth shown for reference; our two columns show branches and foliage. (b) Quantitative results on 500 trees; NCD: normalized Chamfer distance, F1 and COV at 5% threshold.*

method infers a compact symbolic representation from sparse aerial observations. Nevertheless, the structural agreement across both examples confirms that the predicted grammars encode geometrically plausible trees at mapping-relevant scale.

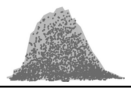
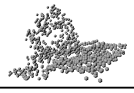
	DSM	Orthophoto	LiDAR	Ours
1				
2				

Table 4: *Qualitative comparison against IGN LiDAR point clouds (37 and 25 pts/m^2) in the French Pyrenees.*

6. Discussion

The grammar-based representation offers substantial advantages in both efficiency and flexibility. Each predicted L-system requires only ∼1 KB of storage and inference takes ∼0.2 s per tree on a single GPU – far smaller and faster than both the 30M-triangle meshes of Grammatikaki et al. [GEL*25] (tens of MB, up to 15 s render time) and the dense point clouds of TreeON [GEH*26] (∼0.2 MB per tree). Beyond efficiency, the symbolic formulation enables post-hoc manipulation unavailable in point cloud or mesh outputs: predicted grammars can be pruned or expanded via stochastic growth rules [PL90], all without rerunning inference (Figure 1).

The reliance on procedurally generated training data introduces *distribution bias*: most samples reflect full-foliage summer appearances under mid-day illumination, which may reduce robustness when the input orthophoto shows atypical conditions such as snow-covered canopies or bare winter crowns, and creates a *sim-to-real domain gap* that cannot currently be quantified as no real-world grammar annotations exist for benchmarking. Reconstruction from nadir inputs is inherently *ill-posed*: branching structure is only partially observable from above, leaving ambiguity where multiple grammars may explain the same input. Autoregressive generation over long sequences entails *long sequence dependencies*, making optimization memory-intensive and sensitive to exposure bias; the recursive nature of L-systems further leads to *error amplification*, where token-level inaccuracies induce disproportionate geometric deviations after expansion. Finally, *limited grammar expressiveness* constrains the range of representable structures, reducing the ability to capture highly irregular or atypical growth patterns; the current formulation models only *woody structure*, and cannot represent highly irregular or atypical crown geometries such as wind-deformed or multi-stemmed trees.

7. Conclusions and Future Work

TreesFormer reconstructs hierarchical parametric L-system grammars directly from DSM point clouds and orthophotos, recovering compact and interpretable symbolic representations suitable for large-scale rural 3D mapping. Unlike prior approaches that produce purely geometric outputs, the symbolic formulation enables downstream editing, growth simulation, and efficient storage without rerunning inference. Future work could extend the input setting to incorporate oblique or street-level imagery for richer structural cues, and extend the grammar formalism to encode foliage density and phenological growth stages for full botanical plausibility. The symbolic output is also directly compatible with real-time VR and simulation environments, enabling interactive rural scene exploration at mapping scale. Potential applications include forestry inventory and disturbance scenario modeling, for which the editable symbolic output provides a natural foundation [PGG*24].

Acknowledgments

The authors acknowledge TU Wien Bibliothek for financial support through its Open Access Funding Programme. The VRVis GmbH is funded by BMIMI, BMWET, Tyrol, Vorarlberg and Vienna Business Agency in the scope of COMET - Competence Centers for Excellent Technologies (904918, 911654) which is managed by FFG.

© 2026 The Author(s).
Proceedings published by Eurographics - The European Association for Computer Graphics.

References

[ACC*18] ARGUDO, OSCAR, COMINO, MARC, CHICA, ANTONIO, et al. "Segmentation of aerial images for plausible detail synthesis". *Computers & Graphics* 71 (2018), 23–34. ISSN: 0097-8493. DOI: https://doi.org/10.1016/j.cag.2017.11.004 3.

[ADMG18] ACHLIOPTAS, PANOS, DIAMANTI, OLGA, MITLIAGKAS, IOANNIS, and GUIBAS, LEONIDAS. *Learning Representations and Generative Models for 3D Point Clouds*. 2018. arXiv: 1707.02392 [cs.CV]. URL: https://arxiv.org/abs/1707.02392 7.

[ALL25] ALONSO, PEDRO, LI, TIANRUI, and LI, CHONGSHOU. *Less is More: Efficient Point Cloud Reconstruction via Multi-Head Decoders*. 2025. arXiv: 2505.19057 [cs.CV]. URL: https://arxiv.org/abs/2505.19057 7.

[BFGS06] BOSTRÖM, G., FIOCCO, M., GONÇALVES, JOAO, and SEQUEIRA, VÍTOR. "Urban 3D modelling using terrestrial laser scanners". *International Archives of Photogrammetry, Remote Sensing and Spatial Information Sciences* 36.Part5 (Jan. 2006), 279–284 1.

[BNB13] BRADLEY, DEREK, NOWROUZEZAHRAI, DEREK, and BEARDSLEY, PAUL. "Image-based reconstruction and synthesis of dense foliage". *ACM Trans. Graph.* 32.4 (2013), 1–10. ISSN: 0730-0301, 1557-7368. DOI: 10.1145/2461912.2461952 2.

[Bri07] BRIDSON, ROBERT. "Fast Poisson disk sampling in arbitrary dimensions". *ACM SIGGRAPH 2007 Sketches*. SIGGRAPH '07. San Diego, California: Association for Computing Machinery, 2007, 22–es. ISBN: 9781450347266. DOI: 10.1145/1278780.1278807. URL: https://doi.org/10.1145/1278780.1278807 3.

[CCD*19] CHOUDHURY, MD, COSTANZINI, SOFIA, DESPINI, FRANCESCA, et al. "Photogrammetry and Remote Sensing for the identification and characterization of trees in urban areas." *Journal of Physics: Conference Series* 1249 (May 2019), 012008. DOI: 10.1088/1742-6596/1249/1/012008 2.

[DLL*19] DU, SHENGLAN, LINDENBERGH, RODERIK, LEDOUX, HUGO, et al. "AdTree: Accurate, Detailed, and Automatic Modelling of Laser-Scanned Trees". en. *Remote Sensing* 11.18 (Jan. 2019). Publisher: Multidisciplinary Digital Publishing Institute, 2074. ISSN: 2072-4292. DOI: 10.3390/rs11182074. URL: https://www.mdpi.com/2072-4292/11/18/2074 (visited on 08/13/2025) 2.

[Fed23] FEDERAL MINISTRY OF AGRICULTURE, FORESTRY, REGIONS AND WATER MANAGEMENT. *Austrian Forest Report 2023: We Take Care of the Forest*. Tech. rep. Vienna, Austria: Federal Ministry of Agriculture, Forestry, Regions and Water Management, Aug. 2023 6.

[GEH*26] GRAMMATIKAKI, ANGELIKI, ESCHNER, JOHANNES, HERMOSILLA, PEDRO, et al. *TreeON: Reconstructing 3D Tree Point Clouds from Orthophotos and Heightmaps*. 2026. arXiv: 2603.10996 [cs.GR]. URL: https://arxiv.org/abs/2603.10996 3, 6–8.

[GEL*25] GRAMMATIKAKI, ANGELIKI, ESCHNER, JOHANNES, LEDERMANN, FLORIAN, et al. "How to represent landmark trees in digital 3D maps? An automated workflow and user study". *Cartography and Geographic Information Science* 0.0 (2025), 1–18. DOI: 10.1080/15230406.2025.2489543. URL: https://doi.org/10.1080/15230406.2025.2489543 3, 8.

[GJB*20] GUO, JIANWEI, JIANG, HAIYONG, BENES, BEDRICH, et al. "Inverse Procedural Modeling of Branching Structures by Inferring L-Systems". *ACM Trans. Graph.* 39.5 (June 2020). ISSN: 0730-0301. DOI: 10.1145/3394105. URL: https://doi.org/10.1145/3394105 2.

[Gra24] GRAMMATIKAKI, ANGELIKI. *Landmark trees in Austria*. Version 1.0.0. TU Wien, Mar. 2024. DOI: 10.48436/nsj20-6ka24. URL: https://doi.org/10.48436/nsj20-6ka24 6, 7.

[GSF*24] GONG, HAOYU, SUN, QIAN, FANG, CHENRONG, et al. "TreeDetector: Using Deep Learning for the Localization and Reconstruction of Urban Trees from High-Resolution Remote Sensing Images". *Remote Sensing* 16.3 (2024). ISSN: 2072-4292. DOI: 10.3390/rs16030524 3.

[HOM*24] HUANG, YUNMEI, OU, BOTONG, MENG, KEXIN, et al. "Tree Species Classification from UAV Canopy Images with Deep Learning Models". *Remote Sensing* 16.20 (2024). ISSN: 2072-4292. DOI: 10.3390/rs16203836. URL: https://www.mdpi.com/2072-4292/16/20/3836 3.

[HORS07] HIRSCHMUGL, MANUELA, OFNER, MARTIN, RAGGAM, JOHANN, and SCHARDT, MATHIAS. "Single tree detection in very high resolution remote sensing data". *Remote Sensing of Environment* 110.4 (2007). ForestSAT Special Issue, 533–544. ISSN: 0034-4257. DOI: 10.1016/j.rse.2007.02.029 3.

[IGN25] IGN. *LiDAR HD Géoservices*. https://geoservices.ign.fr/lidarhd. Institut national de l'information géographique et forestière (IGN), France. 2025 6.

[LBB*23] LLULL, CRISTIÁN, BALOIAN, NELSON, BUSTOS, BENJAMIN, et al. "Evaluation of 3D Reconstruction for Cultural Heritage Applications". *2023 IEEE/CVF International Conference on Computer Vision Workshops (ICCVW)*. 2023, 1634–1643. DOI: 10.1109/ICCVW60793.2023.00179 7.

[LGB*21] LIU, YANCHAO, GUO, JIANWEI, BENES, BEDRICH, et al. "TreePartNet: neural decomposition of point clouds for 3D tree reconstruction". *ACM Trans. Graph.* 40.6 (Dec. 2021). ISSN: 0730-0301. DOI: 10.1145/3478513.3480486. URL: https://doi.org/10.1145/3478513.3480486 2.

[LKK*21] LI, BOSHENG, KAŁUŻNY, JACEK, KLEIN, JONATHAN, et al. "Learning to reconstruct botanical trees from single images". *ACM Trans. Graph.* 40.6 (Dec. 2021). ISSN: 0730-0301. DOI: 10.1145/3478513.3480525. URL: https://doi.org/10.1145/3478513.3480525 2.

[LLB*24a] LEE, JAE JOONG, LI, BOSHENG, BEERY, SARA, et al. *Tree-D Fusion: Simulation-Ready Tree Dataset from Single Images with Diffusion Priors*. 2024. arXiv: 2407.10330 [cs.CV]. URL: https://arxiv.org/abs/2407.10330 3, 7, 8.

[LLB*24b] LI, YUAN, LIU, ZHIHAO, BENES, BEDRICH, et al. "SVDTree: Semantic Voxel Diffusion for Single Image Tree Reconstruction". *2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. 2024, 4692–4702. DOI: 10.1109/CVPR52733.2024.00449 3, 7, 8.

[LLB23] LEE, JAE JOONG, LI, BOSHENG, and BENES, BEDRICH. "Latent L-systems: Transformer-based Tree Generator". *ACM Trans. Graph.* 43.1 (Nov. 2023). ISSN: 0730-0301. DOI: 10.1145/3627101. URL: https://doi.org/10.1145/3627101 2, 7, 8.

[MM99] MAYER, HELMUT and MAYR, WILHELM. "Automatic Extraction of Deciduous Trees from High Resolution Aerial Imagery". *Mustererkennung 1999*. Ed. by FÖRSTNER, WOLFGANG, BUHMANN, JOACHIM M., FABER, ANNETT, and FABER, PETKO. Berlin, Heidelberg: Springer Berlin Heidelberg, 1999, 102–110. ISBN: 978-3-642-60243-6 3.

[PGG*24] PEYTAVIE, ADRIEN, GAIN, JAMES, GUÉRIN, ERIC, et al. "DeadWood: Including Disturbance and Decay in the Depiction of Digital Nature". *ACM Trans. Graph.* 43.2 (Feb. 2024). ISSN: 0730-0301. DOI: 10.1145/3641816. URL: https://doi.org/10.1145/3641816 3, 8.

[PJBM22] POOLE, BEN, JAIN, AJAY, BARRON, JONATHAN T., and MILDENHALL, BEN. *DreamFusion: Text-to-3D using 2D Diffusion*. 2022. URL: https://arxiv.org/abs/2209.14988 3.

[PL90] PRUSINKIEWICZ, PRZEMYSLAW and LINDENMAYER, ARISTID. *The Algorithmic Beauty of Plants*. Springer, 1990 2, 8.

© 2026 The Author(s).
Proceedings published by Eurographics - The European Association for Computer Graphics.

[QLH*23] QUAN, YING, LI, MINGZE, HAO, YUANSHUO, et al. "Tree species classification in a typical natural secondary forest using UAV-borne LiDAR and hyperspectral data". *GIScience & Remote Sensing* 60.1 (2023), 2171706. DOI: 10.1080/15481603.2023.2171706. URL: https://doi.org/10.1080/15481603.2023.2171706 3.

[QSMG17] QI, CHARLES R., SU, HAO, MO, KAICHUN, and GUIBAS, LEONIDAS J. *PointNet: Deep Learning on Point Sets for 3D Classification and Segmentation*. 2017. URL: https://arxiv.org/abs/1612.00593 4.

[SCZ*23] SHI, RUOXI, CHEN, HANSHENG, ZHANG, ZHUOYANG, et al. *Zero123++: a Single Image to Consistent Multi-view Diffusion Base Model*. 2023 3.

[SHD*24] SUN, CHUNYI, HAN, JUNLIN, DENG, WEIJIAN, et al. *3D-GPT: Procedural 3D Modeling with Large Language Models*. 2024. arXiv: 2310.12945 [cs.CV]. URL: https://arxiv.org/abs/2310.12945 3.

[She05] SHEPPARD, STEPHEN R.J. "Landscape visualisation and climate change: the potential for influencing perceptions and behaviour". *Environmental Science & Policy* 8.6 (2005). Mitigation and Adaptation Strategies for Climate Change, 637–654. ISSN: 1462-9011. DOI: https://doi.org/10.1016/j.envsci.2005.08.002. URL: https://www.sciencedirect.com/science/article/pii/S1462901105001188 1.

[SPK*14] STAVA, O., PIRK, S., KRATT, J., et al. "Inverse Procedural Modelling of Trees". *Comput. Graph. Forum* 33.6 (Sept. 2014), 118–131. ISSN: 0167-7055. DOI: 10.1111/cgf.12282. URL: https://doi.org/10.1111/cgf.12282 2.

[SRDT01] SHLYAKHTER, I., ROZENOER, M., DORSEY, J., and TELLER, S. "Reconstructing 3D tree models from instrumented photographs". 21.1 (2001), 53–61. ISSN: 02721716. DOI: 10.1109/38.920627 2.

[SSS06] SNAVELY, NOAH, SEITZ, STEVEN M., and SZELISKI, RICHARD. "Photo tourism: exploring photo collections in 3D". *ACM Trans. Graph.* 25.3 (July 2006), 835–846. ISSN: 0730-0301. DOI: 10.1145/1141911.1141964. URL: https://dl.acm.org/doi/10.1145/1141911.1141964 (visited on 08/13/2025) 1.

[TB25] TODD, GRACE and BAILEY, MIKE. "Foliager: Procedural Forest Generation from Natural Language Using Scientific Data and AI". *Proceedings of the Special Interest Group on Computer Graphics and Interactive Techniques Conference Posters*. SIGGRAPH Posters '25. Association for Computing Machinery, 2025. ISBN: 9798400715495. DOI: 10.1145/3721250.3743024. URL: https://doi.org/10.1145/3721250.3743024 3.

[The25] THE GROVE 3D. *The Grove 3D*. https://www.thegrove3d.com/. Accessed: 2025-03-13. 2025 3.

[TZW*07] TAN, PING, ZENG, GANG, WANG, JINGDONG, et al. "Image-based tree modeling". *ACM Trans. Graph.* 26 (2007), 87 2.

[XGC07] XU, HUI, GOSSETT, NATHAN, and CHEN, BAOQUAN. "Knowledge and heuristic-based modeling of laser-scanned trees". *ACM TOG* 26.4 (2007), 19. ISSN: 0730-0301, 1557-7368. DOI: 10.1145/1289603.1289610 2.

[XLX*25] XIANG, JIANFENG, LV, ZELONG, XU, SICHENG, et al. *Structured 3D Latents for Scalable and Versatile 3D Generation*. 2025. URL: https://arxiv.org/abs/2412.01506 3.

[XM15] XU, LING and MOULD, DAVID. "Procedural Tree Modeling with Guiding Vectors". *Computer Graphics Forum* 34.7 (2015), 47–56. DOI: https://doi.org/10.1111/cgf.12744. eprint: https://onlinelibrary.wiley.com/doi/pdf/10.1111/cgf.12744. URL: https://onlinelibrary.wiley.com/doi/abs/10.1111/cgf.12744 2.

[YLX*21] YOU, HANGKAI, LI, SHIHUA, XU, YIFAN, et al. "Tree Extraction from Airborne Laser Scanning Data in Urban Areas". *Remote Sensing* 13 (Aug. 2021), 3428. DOI: 10.3390/rs13173428 2.

[YWH22] YANG, YINHUI, WANG, RUI, and HUO, YUCHI. *Rule-based Procedural Tree Modeling Approach*. 2022. arXiv: 2204.03237 [cs.GR]. URL: https://arxiv.org/abs/2204.03237 2.

[ZGLD25] ZHANG, XINYI, GAO, DAOYI, LI, NAIQI, and DAI, ANGELA. *ProcGen3D: Learning Neural Procedural Graph Representations for Image-to-3D Reconstruction*. 2025. arXiv: 2511.07142 [cs.CV]. URL: https://arxiv.org/abs/2511.07142 3.

[ZLB*23] ZHOU, XIAOCHEN, LI, BOSHENG, BENES, BEDRICH, et al. *DeepTree: Modeling Trees with Situated Latents*. 2023. arXiv: 2305.05153 [cs.LG]. URL: https://arxiv.org/abs/2305.05153 3.

[ZLB*25] ZHOU, XIAOCHEN, LI, BOSHENG, BENES, BEDRICH, et al. "TreeStructor: Forest Reconstruction With Neural Ranking". *IEEE Transactions on Geoscience and Remote Sensing* 63 (2025), 1–19. DOI: 10.1109/TGRS.2025.3558312 3.

© 2026 The Author(s).
Proceedings published by Eurographics - The European Association for Computer Graphics.