EUROGRAPHICS 2026 / B. Masia and J. Thies
(Guest Editors)

COMPUTER GRAPHICS *forum*
Volume 45 (2026), Number 2

TreeON: Reconstructing 3D Tree Point Clouds from Orthophotos and Heightmaps

Angeliki Grammatikaki[1], Johannes Escher[1], Pedro Hermosilla[1], Oscar Argudo[2], Manuela Waldner[1]

[1]TU Wien, Institute of Visual Computing and Human-Centered Technology, Vienna, Austria
[2] Universitat Politècnica de Catalunya, Department of Computer Science, Barcelona, Spain

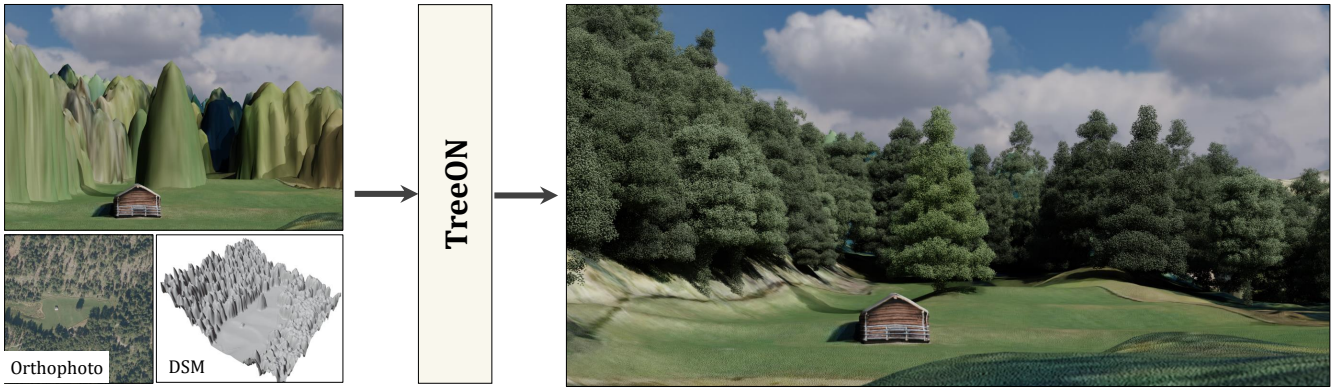


Figure 1: *Our neural-based framework, TreeON, can be used to reconstruct coherent point clouds for the trees in a real dataset given an orthophoto and a Digital Surface Model (DSM).*

Abstract

We present TreeON, a novel neural-based framework for reconstructing detailed 3D tree point clouds from sparse top-down geodata, using only a single orthophoto and its corresponding Digital Surface Model (DSM). Our method introduces a new training supervision strategy that combines both geometric supervision and a differentiable shadow and silhouette loss to learn point cloud representations of trees without requiring species labels, procedural rules, detailed terrestrial reconstruction data, or ground laser scan data. To address the lack of ground truth data, we generate a synthetic dataset of point clouds from procedurally modeled trees and train our network on it. Quantitative and qualitative experiments demonstrate better reconstruction quality and coverage compared to existing methods, as well as strong generalization to real-world data, leading to visually appealing and structurally plausible tree point cloud representations that can be integrated into interactive digital 3D maps. The codebase, synthetic dataset, and pretrained model are publicly available at `https://angelikigram.github.io/treeON/`.

CCS Concepts

• ***Computing methodologies*** → *Neural networks; Point-based models; Reconstruction;*

1. Introduction

Digital 3D maps are widely used to communicate spatial and environmental information, offering intuitive and immersive representations of real landscapes [She05]. Unlike urban settings, where terrestrial scans [BFGS06] or multi-view photogrammetry [SSS06] are often used to generate detailed 3D models, high-resolution datasets remain limited or region-specific in natural environments. In rural areas, digital 3D maps are therefore mostly generated from aerial imagery, including orthophotos combined with Digital Surface Models (DSMs; raster heightmaps encoding terrain elevation and above-ground objects such as vegetation and buildings).. As a result, rural 3D maps are, in fact, often limited to a 2.5D surface representation, which lacks detail in close-up oblique views. This limitation is most evident for vegetation, where fine struc-

tural detail is lost (see Figure 1 left) , which is perceived as non-appealing [GEL*25].

We hypothesize that much of the missing 3D information for vegetation, especially for isolated trees, can be inferred from the currently available 2.5D inputs. This would enable lightweight yet visually appealing reconstructions without requiring additional data collection. Such an approach could address a key design dilemma: how can we create tree models that are both visually compelling and computationally lightweight, and hence scalable enough for large outdoor environments, and effective in data-sparse conditions?

This hypothesis is motivated by findings from studies in virtual environments, which show that trees with distinctive shapes not only improve perceived realism but also facilitate orientation and wayfinding, particularly when they serve as recognizable landmarks [JNP21]. Just as buildings and monuments serve as key reference points in urban navigation, trees can provide essential visual anchors in natural terrain. Although simplified vegetation can still serve as landmarks, prior work shows that users strongly prefer the visual appeal of more detailed models and perceive them as more reliable for navigation [GEL*25].

Prior methods for vegetation reconstruction have relied on data-rich techniques involving dense LiDAR point clouds [XGC07; DLL*19], multi-view terrestrial or aerial imagery [SRDT01; INJN19; YLX*21; HLZ*17], photogrammetry and remote sensing approaches [CCD*19], or biologically-inspired procedural modeling based on species-specific rules [XM15]. While these approaches can produce structurally detailed results, they tend to be computationally expensive, difficult to scale, and reliant on specialized data collection. Consequently, they are ill-suited for wide-area deployment in natural landscapes, where acquiring such detailed datasets is often infeasible. This motivates the need for lightweight, scalable solutions that can generate realistic vegetation models from standard geodata sources alone.

An ideal method for large-scale rural maps should satisfy several requirements: (1) high visual appeal in close-up oblique views, (2) computational efficiency for large-scale deployment (particularly for rendering), (3) structural plausibility for different tree types, and (4) automation without the need for additional input data, like species information, photographs, or user sketches. These requirements reveal why existing methods fall short: some achieve realism but are inefficient or data-hungry, while others are lightweight yet lack plausibility.

To fill this gap, we propose a novel approach that reconstructs visually plausible 3D tree point clouds from sparse geodata by exploiting multiple complementary cues present in DSMs and orthophotos. These cues include canopy shape, overall height, and height-to-width ratio from the DSM, as well as canopy texture and shadows from the orthophoto. Our pipeline constructs a 3D point cloud representation of each tree from a single orthophoto and its corresponding DSM. We adopt a point cloud output as a pragmatic, generator-agnostic representation for large-scale 3D mapping, avoiding inference of parameters tied to a specific procedural model, while still allowing higher-level structure to be recovered from the point cloud when needed (TreeStructor [ZLB*25]).

Our method integrates geometric and visual cues via a PointNet-style terrain encoder and a CNN-based orthophoto encoder, with a neural decoder predicting occupancy at arbitrary 3D query points. The training process is supervised through five complementary loss terms: a binary cross-entropy loss to enforce structural plausibility, a color consistency loss to match orthophoto appearance, and three differentiable perceptual losses [ZIE*18] that supervise alignment with observed shadows, and top and side view silhouettes. This multi-modal supervision enables the model to infer plausible and expressive tree shapes, yielding realistic reconstructions in environments where more detailed 3D data is unavailable. In summary, our work makes three main contributions:

1. A synthetic dataset of procedurally generated trees with paired orthophotos and DSMs,
2. A network architecture for reconstructing detailed 3D tree point clouds from sparse top-down geospatial data, and
3. A training supervision strategy that combines occupancy prediction with shadow- and silhouette-based losses.

In addition, we demonstrate through quantitative and qualitative evaluation that, with these contributions, our approach fulfills the requirements of realistic and visually appealing tree reconstruction better than alternative architectures and supervision strategies.

2. Related Work

Modeling realistic trees has been central to both computer graphics and remote sensing. Traditional methods often depend on dense 3D terrestrial laser scans to extract explicit branch skeletons and generate foliage-rich models [XGC07]. However, such data are typically unavailable in remote or rural areas, where only orthophotos and elevation models exist. These methods satisfy structural plausibility and visual appeal, but fail scalability and automation due to costly data acquisition.

To overcome the need for rich input data, researchers have explored multi-view image reconstructions, including volumetric and point cloud assemblies [SRDT01; BNB13], and sketch-based techniques requiring user guidance [TZW*07]. Argudo et al. [ACA16] developed a single-image method to infer plausible structure, but it struggles under significant occlusion. All these approaches either rely on multi-view imagery or human intervention, limiting scalability to large, sparsely mapped landscapes. Thus, while they improve realism, they fall short in efficiency and automation.

Procedural generation synthesizes tree forms from parametric or stochastic growth models [YWH22], which can be inverted from visual cues [SPK*14] or placed into landscapes using vegetation classification from aerial images [ACC*18; GSF*24]. However, such classification is often inaccurate or unreliable from aerial data alone [QLH*23; HOM*24]. As a result, these methods may generate highly detailed but biologically implausible tree models, which risks creating unfounded trust in the reconstructions despite low realism. While valuable for producing training data and large-scale vegetation, their realism often depends on species-specific templates or predefined rules, limiting flexibility compared to data-driven reconstruction. In terms of requirements, procedural models are efficient and automated, but often lack structural plausibility and visual appeal without species priors.

A few works have addressed tree reconstruction using minimal geospatial input data. Earlier approaches include Mayr et al. [MM99], who fitted ellipsoidal primitives to DSM data to model isolated rural and leafless urban trees, and Hirschmugl et al. [HORS07], who combined DSM-based height peaks with spectral cues from orthophotos for tree detection. More recently, Grammatikaki et al. [GEL*25] proposed a fully automatic pipeline that models landmark trees directly from DSM and orthophotos, producing varying levels of detail. Though insightful, these approaches either approximate trees with coarse shapes or detect trees rather than reconstructing full per-tree structures. They therefore only partially meet structural plausibility and do not provide the visual appeal required for landmark-level detail.

In response to these limitations, general single-view reconstruction and view-synthesis methods, such as NeRFDiff [GTL*23] and EscherNet [KLL*24], as well as recent text- and image-conditioned generative models for object- and scene-level 3D synthesis [WXC*25], including DreamFusion [PJBM22], Magic3D [LGT*23], ProlificDreamer [WLW*23], Zero123++ [SCZ*23], 3DTopia-XL [CTD*25], and TRELLIS [XLX*25], address related 3D inference problems but assume perspective imagery and object- or scene-centric priors, making them unsuitable for geospatial reconstruction from top-down orthophotos and DSMs.

Recent years have seen the rise of neural network-driven vegetation modeling methods that directly target tree and forest reconstruction. DeepTrees [ZLB*23] models tree growth using situated latent representations that capture spatial and structural dependencies during a biologically inspired generative process. Foliager [TB25] procedurally generates forest scenes from natural language and ecological data, combining AI planning with scientific growth rules. In parallel, neural radiance field (NeRF)-based methods have been adapted to vegetation [HTC23]. They aim to reconstruct individual canopies from sparse aerial views, though the results often remain noisy and miss fine branching detail. Diffusion-based pipelines, such as Tree-D Fusion [LLB*24a] and SVDTree [LLB*24b], produce high-fidelity branch-leaf structures from a single image by leveraging learned semantic voxel encodings and generative priors.

In contrast, TreeStructor [ZLB*25] introduces a retrieval-and-ranking approach that assembles forest reconstructions from a library of neural components. While these approaches significantly advance the realism of tree modeling, they often require multi-view imagery, genus-level priors, or dense semantic cues-factors that are unavailable in our single-view, label-free geospatial setting, and, in the case of species classification, unreliable [FLS*16]. While these methods strongly address realism and plausibility, they typically require multi-view inputs, species priors, or dense semantic cues, limiting their scalability and automation from sparse rural geodata.

In contrast, our work addresses this gap by reconstructing spatially explicit 3D tree forms from minimal geospatial input, aiming to satisfy all four requirements simultaneously: visual appeal, efficiency, structural plausibility, and automation, without relying on semantic labels, species priors, or dense image coverage.

3. Overview

Our method integrates synthetic data generation, multi-modal neural networks, tailored supervision, and postprocessing for accurate 3D tree reconstruction. We generate a training dataset with aligned DSMs, orthophotos, and ground-truth colored point clouds from procedural trees. These inputs are encoded by image and point cloud backbones and decoded into occupancy and per-point color. Supervision combines geometric, photometric, and projection-based losses, while postprocessing densifies and renders the final reconstructions.

3.1. Synthetic Training Dataset Generation Pipeline

Whilst our method relies on a multi-modal supervision, such paired data is rarely available in real-world settings. Therefore, we construct a diverse synthetic dataset tailored for our 3D tree reconstruction. Each training instance includes three aligned modalities: a colored point cloud representing a 3D tree (Figure 2), a Digital Surface Model (DSM), and a top-down orthophoto.

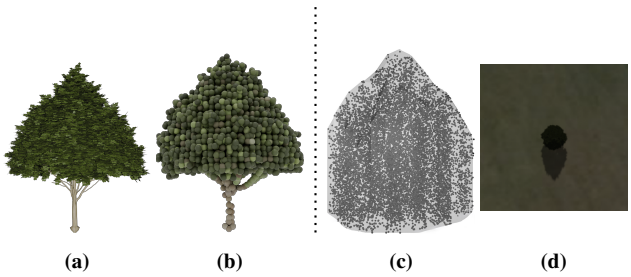


Figure 2: *Tree generation pipeline starting from (a) procedural mesh: sampled colored point cloud (b), artificial DSM point cloud (visualized as surface with interior points) (c) and orthophoto (d).*

Tree Generation Our tree generation pipeline begins with a procedurally generated tree mesh created using Blender's Grove add-on [The25] (Figure 2 (a)), although other modeling tools could be used. Each tree mesh contains inline vertex colors, which encode species-specific bark and crown appearance. For training, we require point clouds that are compact enough for efficient learning yet still representative of detailed tree geometry and color. We therefore apply Poisson-disk sampling [Bri07] to the mesh vertices, which yields uniformly distributed samples across the crown and trunk. This reduces each mesh to approximately $K \approx 6{,}000$ points, which we found to be a good trade-off between geometric fidelity and computational efficiency. Larger values of K increased memory usage without improving reconstruction quality, while smaller values reduced structural coverage. The resulting colored point cloud serves as the *ground truth target* in training (Figure 2 (b)).

Orthophoto and DSM Rendering For each synthetic tree, we aim to replicate the data modalities commonly available from aerial imagery and airborne LiDAR. Therefore, we render orthophotos and DSMs from the detailed tree mesh (Figure 2 (a)). As illustrated in Figure 3, we generate data by rendering the processed 3D mesh in Blender from a controlled top-down view. For realism, trees are placed on terrain patches sampled from high-resolution

DEMs, ensuring diverse contexts such as flatland, hilly, or forested areas. Locations were selected such that terrain features (e.g., valleys, clearings) support plausible vegetation growth, while regions with unsuitable land cover (e.g., water bodies, snow) were avoided. To avoid shadow mismatches during supervision, any visible surrounding vegetation is removed using an inpainting neural network [YFF*23]. In some scenes, we placed additional synthetic trees around the central target tree to create more realistic orthophotos with surrounding vegetation. A top-down orthographic camera is then positioned directly above the target tree, aligned along the Z axis, and centered at the tree's bounding box origin. The orthographic scale is fixed by the camera height, centering the tree crown while including surrounding terrain for context. We sample ten light directions per tree to diversify shadow cues during training. Azimuth angles are drawn from $[0^\circ, 360^\circ)$, while elevation angles are restricted to $[40^\circ, 70^\circ]$, consistent with typical orthophoto acquisitions around mid-day in summer, yet providing sufficient variation for robust shadow supervision.

The **orthophotos** are rendered at a configurable resolution consistent with aerial imagery standards. These images simulate top-down remote sensing captures (e.g., aerial photography or satellite imagery). For each condition, the orthographic top-down camera renders the tree together with its cast shadow on the ground plane. Shadows are generated in Blender's Cycles renderer with a directional light source, where the sun's angular diameter $\alpha \sim U(0.5, 10)$ controls the penumbra size, producing both hard and soft shadows. Shadow masks encode occlusion intensity ($0 =$ fully shadowed, $1 =$ illuminated), with Gaussian noise, channel variations, and clipping applied for realism following Morales et al. [MHT19]. The resulting orthophotos thus capture naturalistic shadow transitions and crown textures under varying illumination. For each rendered orthophoto, the corresponding sun direction vector is also recorded and stored in a metadata file. This allows us to consistently reproduce shadows under the same lighting setup during supervision.

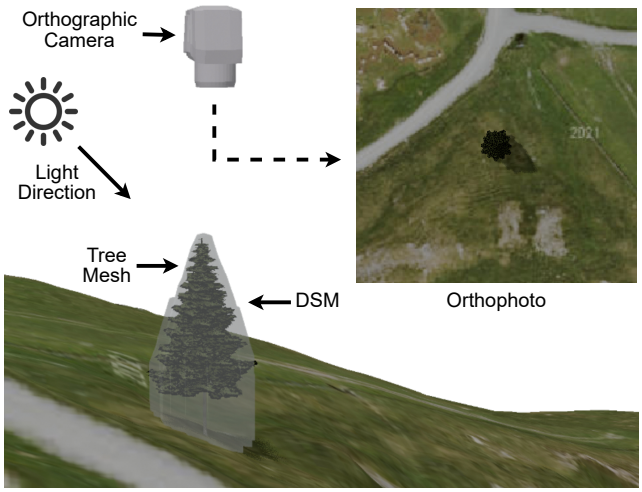


Figure 3: *Orthophoto and DSM generation from a top-down orthographic camera. Orthophotos include trees and shadows under directional lighting, while DSMs DSM (visualized as a surface for illustration) are derived from per-pixel depth. Sun positions vary in azimuth and elevation.*

To produce the **DSM**, we emulate a LiDAR-like surface scanning process. Specifically, we extract the Z-buffer from the same orthographic camera view and convert per-pixel depth values into real-world elevation. This is done by subtracting the depth from the camera's world-space height, resulting in a per-pixel heightmap that captures the vertical profile of all visible geometry. The heightmap is normalized by subtracting the local ground elevation and scaling heights to the unit range [0,1], ensuring consistent DSM encoding while preserving relative crown geometry. Each DSM effectively represents the tree's canopy shape and elevation in image form, matching the modality of publicly available DSMs. The resolution of the DSM rendering is configurable and reported for each dataset in Section 4.

For use in the network, we back-project DSM pixels into 3D, yielding a point cloud representation. This preserves sharp height discontinuities and explicit 3D geometry, and aligns naturally with point-based encoding and occupancy supervision at arbitrary 3D query locations. We sample uniformly within the DSM volume until reaching about $K \approx 6{,}000$ points per tree to obtain a fixed-size and stable PointNet input. Fewer points led to unstable gradients, while larger sets increased memory and computation without measurable performance gains.

3.2. Network Architecture

The model processes three inputs: an RGB orthophoto of size $3 \times 224 \times 224$, a DSM-based point cloud of shape $K \times 3$, and a set of N 3D query points sampled within a unit cube $[0,1]^3$ (see Figure 4). We fix $K \approx 6{,}000$ points, following the voxel-based downsampling in our synthetic dataset (Section 3.1). The number of query points is set to $N = 25{,}000$, which we found sufficient to balance coverage of inside/outside regions with computational tractability. For each query point, the model predicts whether it lies inside or outside the tree surface, while also estimating per-point RGB appearance. This implicit formulation avoids using a fixed voxel grid and allows for continuous shape reconstruction. To ensure supervision across both occupied and free-space regions, query points are sampled to balance inside/outside labels rather than concentrating only near the DSM surface.

The orthophoto is encoded using a ResNet-18 [HZRS16] backbone pre-trained on ImageNet. Intermediate features are pooled and passed through a projection head with batch normalization [IS15] and LeakyReLU activation [Maa13], yielding a 1024-dimensional image embedding normalized via $\ell 2$ norm. Simultaneously, the DSM input represented as a sparse 3D point cloud is encoded using a PointNet-style [QSMG17] architecture with 1D convolutions (64, 128, 1024 channels), batch normalization, and LeakyReLU activations. The resulting per-point features are aggregated via max-pooling into another 1024-dimensional latent vector, also $\ell 2$-normalized. This design leverages the strengths of each encoder: ResNet extracts texture and shadow cues from 2D orthophotos, while PointNet captures the unordered 3D structure of DSM point clouds, aligning with the occupancy formulation. Then, the im-

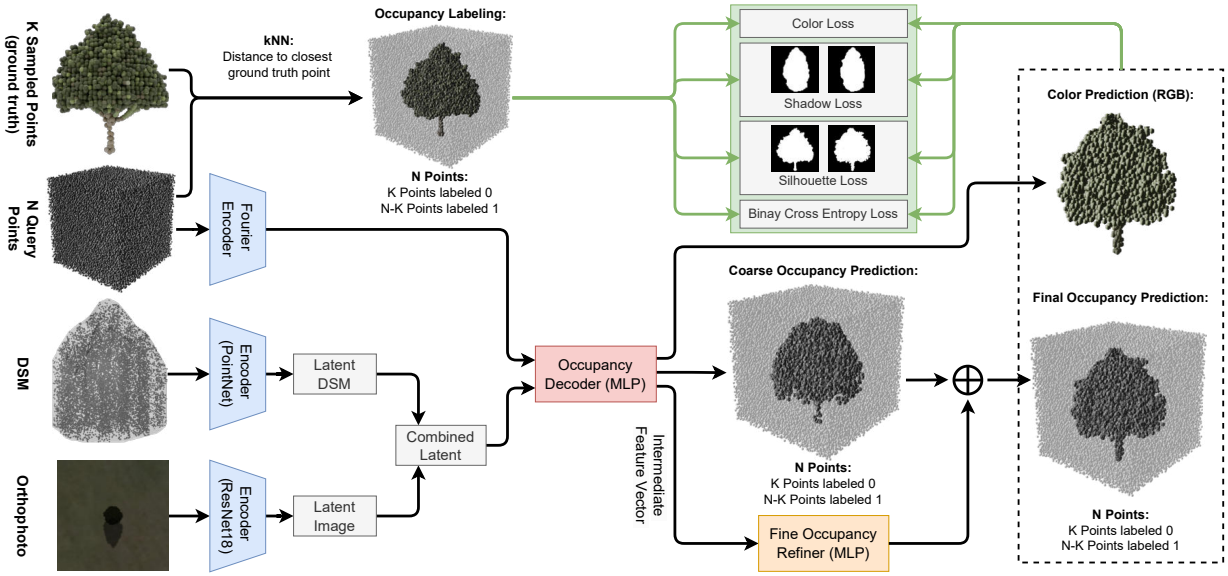


Figure 4: *Overview of the network architecture and training supervision (green).*

age and point cloud embeddings are concatenated into a 2048-dimensional latent representation. The fused image–DSM latent is injected into every decoder layer together with Fourier-encoded query points, conditioning local occupancy predictions on global context.

Each 3D query point is passed through a Fourier positional encoder [TSM*20] with 25 frequency bands, producing a 153-dimensional representation. Both the global latent vector from the image and point cloud encoders and the Fourier-encoded query points are then provided as inputs to the occupancy decoder, where they are combined internally, yielding an effective per-point input size of 2201. The occupancy decoder is a deep multilayer perceptron with six fully connected layers, each incorporating Layer Normalization [BKH16], dropout [SHK*14], LeakyReLU activation, and residual connections [HZRS16]. It branches into two output heads: one predicts a scalar occupancy logit for each query point, and the other predicts a 3-channel RGB color vector by combining the learned features with the original Fourier-encoded input.

During initial experiments, we observed that the coarse MLP decoder tended to produce overly smooth occupancy transitions at the tree surface, failing to capture crisp details such as fine-branch boundaries (see Figure 5). To address this, we integrated a lightweight refinement module: three small residual MLP layers that compute a corrective delta added to the coarse occupancy logits. This module acts as a local corrector: it reduces false positives and negatives near the surface boundary and recovers fine-scale geometric details that may be lost in the coarse prediction, while keeping the main decoder compact. We refine only geometry since the main bottleneck is surface sharpness, while color already benefits from dense per-point supervision and does not suffer from the same over-smoothing problem. Similar architectures have been adopted in related contexts, e.g., attention-based refinement modules improve surface fidelity in human reconstruction tasks [DCZ*24], and progressive refinement was proposed for occupancy prediction in diffusion-style decoders [WWT*24].

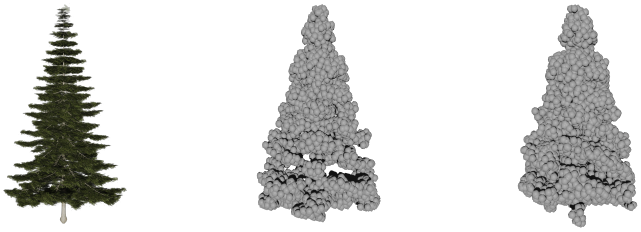


Figure 5: *(a) Ground-truth tree, (b) reconstruction without refinement, and (c) reconstruction with refinement.*

The final outputs of the model are the refined occupancy logits over the query points and per-point RGB color predictions. This architecture supports end-to-end learning of both geometry and appearance, conditioned on visual input and terrain structure. An overview is shown in Figure 4.

3.3. Supervision Strategy

We employ a combination of point-level, color, and projection losses to jointly supervise geometry and appearance. Point-level losses enforce correct occupancy at sampled 3D coordinates, while a color loss supervises the predicted per-point RGB values. Projection-based supervision uses differentiable shadow and silhouette rendering, complemented with a perceptual LPIPS loss [ZIE*18] on the silhouettes to capture structural differences beyond pixel-level matching.

We supervise the model at the point level using a binary cross-entropy (BCE) loss $\mathcal{L}_{\text{occ}}$, a standard choice for implicit occupancy networks [MON*19; PNM*20]. This loss penalizes discrepancies between predicted occupancy logits and ground-truth labels, guiding the network to distinguish occupied (inside) from free-space (outside) points. To obtain ground-truth occupancy labels for this loss, first, we compute the distance to the nearest neighbor in the reference point cloud. Points closer than a threshold to a ground-truth sample are then labeled as occupied (1), while all others are labeled as free space (0). Given the flat structure of leaves and the thinness of branches, tree meshes typically lack significant volume, making this a reasonable approximation. Additionally, we supervise the predicted color of each point with a standard L2 loss.

While BCE is widely adopted in prior point cloud and implicit surface reconstruction methods, our contribution lies in extending this supervision with novel projection-based constraints. Specifically, we introduce differentiable shadow and silhouette projections to regularize tree structure in ways not captured by point-level occupancy alone. For shadow supervision, the occupancy values of query points are projected into the top-down orthophoto plane through a soft differentiable renderer [LLCL19], yielding a shadow that encodes the 3D shape along the illumination direction. The shadow loss $\mathcal{L}_{\text{shadow}}$ compares using LPIPS the predicted shadow against a ground-truth shadow rendered under the same lighting direction as the orthophoto. This encourages alignment with the real shadow observed in the top-down imagery. For silhouettes, projections are generated from five canonical viewpoints (0°, 45°, 90°, 135°, and top view) to constrain the global crown outline across viewpoints, independent of lighting. The silhouette loss $\mathcal{L}_{\text{silh}}$ constrains the global structure by aligning the predicted and ground-truth projections using the LPIPS loss.

The final training loss $\mathcal{L}_{\text{total}}$ is defined as a weighted sum of these four components:

$$\mathcal{L}_{\text{total}} = \lambda_{\text{occ}}\mathcal{L}_{\text{occ}} + \lambda_{\text{col}}\mathcal{L}_{\text{col}} + \lambda_{\text{sh}}\mathcal{L}_{\text{sh}} + \lambda_{\text{silh}}\mathcal{L}_{\text{silh}}. \quad (1)$$

where $\mathcal{L}_{\text{occ}}$ is the occupancy loss, $\mathcal{L}_{\text{col}}$ is the color loss, and $\mathcal{L}_{\text{silh}}$ is the silhouette loss. The weights $\lambda_{\text{occ}}, \lambda_{\text{col}}, \lambda_{\text{sh}}, \lambda_{\text{silh}}$ control the relative influence of each term. In our experiments, we use $\lambda_{\text{occ}} = 0.14$, $\lambda_{\text{col}} = 0.71$, $\lambda_{\text{sh}} = 0.07$, and $\lambda_{\text{silh}} = 0.08$, since we found them to produce the best trade-off between detail, smoothness, and structural plausibility. These combined losses guide the model to produce reconstructions that are not only geometrically accurate but also visually consistent with natural shadows and silhouettes.

3.4. Postprocessing and Rendering

While during training, we query random points within the volume, at inference time, we only output a set of points inside the tree surface. We achieve this by ranking all query points by their predicted occupancy probability and retaining the top-K points (highest logits after sigmoid). This yields a compact fixed-size reconstruction aligned with the ground-truth size ($K \approx 6{,}000$), while discarding low-confidence points. This step is non-differentiable and used only for visualization and downstream evaluation. Note that while thresholding on occupancy might be an alternative, it produces highly variable point counts across trees. Our approach, on the other hand, ensures consistent reconstructions and fair comparisons.

Additionally, to improve visual density while preserving predicted colors, we upsample the sparse set of occupied points. This is done by discretizing the point cloud into a volumetric representation via a voxel grid and then randomly resampling the volume at a higher density. Both the resampling density and the number of voxels are dynamically derived from the reconstructed tree's volume, computed via a convex-hull approximation of the original point cloud. This ensures that larger trees receive more points, yielding reconstructions that are both scale-aware and visually consistent. During upsampling, the original `Color` attribute is propagated: new points inherit interpolated colors from their nearest neighbors in the original set. This preserves the canopy's spatially coherent texture while adapting the point density to the tree's size.

For rendering, each point is visualized as a shaded sphere with its assigned radius and color, producing a dense and natural appearance that conveys both the reconstructed geometry and its authentic tonal variation. Surface normals used for shading are estimated via local PCA on the upsampled point cloud. Finally, although higher tree density increases ambiguity from top-down inputs, TreeON reconstructs trees individually and composes them into full scenes (Figure 1) using the pipeline of Grammatikaki et al. [GEL*25], which ensures correct scaling, placement, and terrain alignment.

4. Experiments

We evaluate our method on both synthetic and real-world datasets, using quantitative and qualitative analyses to assess reconstruction accuracy, completeness, and generalization.

4.1. Datasets

We generated 3,000 synthetic tree models spanning 17 species (Beech, Oak, Birch, Hornbeam, Alder, Aspen, Poplar, Ash, Linden, Elm, Maple, Field Maple, Plane Tree, Fir, Austrian Pine, Scots Pine, and Stone Pine) with proportions based on alpine forestry data [Fed23]: approximately 60% conifers and 40% deciduous species. An overview of representative species is shown in Figure 6. To simulate diverse environments, each tree is placed on one of 12 sampled terrain patches covering 3×3 km. From each terrain, we crop a localized 30×30 m region centered on the tree, ensuring that the tree is embedded in a realistic but computationally tractable environment.

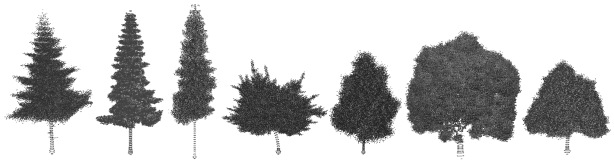


Figure 6: *Examples of tree species (coniferous and deciduous) generated from Grove, shown as 3D point clouds.*

Orthophotos are rendered at a resolution consistent with aerial imagery standards (30 cm/pixel in the Austrian dataset, equivalent to Zoom Level 19 in the basemap.at WMTS service [bas25]), while

DSMs are normalized to 8-bit grayscale images, where black corresponds to ground level and white to the tree apex. To ensure realistic sampling density, DSMs are rendered at 1-meter resolution, consistent with BEV's Open Data Austria height models [Off25]. Each DSM is then back-projected into a point cloud of about $K \approx 6{,}000$ points per tree, matching the ground-truth point cloud size for consistency.

To test generalization, we use a dataset of Austrian landmark trees [Gra24], which contains isolated trees with orthophotos, DSMs, and photographs but no 3D ground truth, restricting evaluation to qualitative comparisons. Additionally, we validate our reconstructions against data from the French National Institute of Geographic and Forest Information (IGN) [IGN25], including orthophotos (20 cm/pixel), dense LiDAR point clouds (≥ 10 pts/m^2), and derived elevation models (DSMs and DTMs). In selected Pyrenees regions, isolated trees were identifiable in both orthophotos and point clouds.

4.2. Implementation Details

We train our model for 700 epochs using the Adam optimizer with an initial learning rate of 10^{-2} and a batch size of 16. A learning rate scheduler `ReduceLROnPlateau` halves the learning rate if the validation loss does not improve for 20 epochs. The training/validation split follows an 85/15 ratio over the 2,500 training trees, ensuring no overlap of tree instances between sets. The remaining 500 trees are held out for testing, including ablation studies and baseline comparisons. All models are trained on a single NVIDIA GeForce RTX 3070 GPU with 8 GB memory, requiring about 24 hours for 700 epochs. Inference on a single tree takes about 0.3 seconds.

4.3. Evaluation Protocol

Our evaluation follows the four requirements defined in Section 1. Structural plausibility is measured quantitatively through Chamfer Distance (CD), Normalized Chamfer Distance (NCD), and F1-Score, which capture geometric accuracy and completeness. Visual appeal is evaluated qualitatively via comparisons to baselines and real imagery, since perceptual realism cannot be fully captured by metrics alone. Automation is demonstrated by the fully automatic pipeline requiring no species priors or multi-view input. Finally, computational efficiency is measured through training cost, inference speed, and scalability across thousands of trees.

Chamfer Distance (CD) is a widely used metric in 3D shape reconstruction that quantifies the geometric error between p redicted and ground-truth point clouds [FSG17; MON*19]. Given the predicted, P, and the ground truth, Q, point clouds, the CD is computed as:

$$\mathrm{CD}(P,Q) = \frac{1}{|P|}\sum_{x\in P}\min_{y\in Q}\|x-y\|_2^2 + \frac{1}{|Q|}\sum_{y\in Q}\min_{x\in P}\|y-x\|_2^2. \quad (2)$$

Since CD averages squared Euclidean distances, its unit is m^2 when coordinates are expressed in meters. CD values are non-negative, with smaller values indicating closer alignment between prediction and ground truth.

To enable fair comparison across trees of different sizes, we normalize CD by the cube root of the ground-truth tree volume v, yielding the **Normalized Chamfer Distance (NCD)** [LBB*23]:

$$\mathrm{NCD}(P,Q) = \frac{\mathrm{CD}(P,Q)}{v^{2/3}}. \quad (3)$$

Here, $v^{1/3}$ provides a characteristic length scale of the tree (with units of meters), and $v^{2/3}$ has units of m^2. Dividing CD (m^2) by $v^{2/3}$ (m^2) produces a unitless NCD. Like CD, NCD takes values ≥ 0, with lower values indicating higher geometric accuracy.

To evaluate both reconstruction accuracy and completeness, we use the **F1-Score**, which combines precision and recall within a fixed distance threshold ε [ALL25; SJS06]. In our evaluation, ε ranges between 0.02 and 0.04 m, depending on the tree size and sampling density. The threshold is tied to the scale of the query point sampling and the dataset's spatial resolution, and is configurable: smaller values enforce stricter matching, while larger values tolerate greater deviation. The metrics are computed as:

$$\mathrm{Precision} = \frac{|\{x\in P : \min_{y\in Q}\|x-y\| < \varepsilon\}|}{|P|}, \quad (4)$$

$$\mathrm{Recall} = \frac{|\{y\in Q : \min_{x\in P}\|y-x\| < \varepsilon\}|}{|Q|}, \quad (5)$$

$$\text{F1-Score} = \frac{2\cdot\mathrm{Precision}\cdot\mathrm{Recall}}{\mathrm{Precision}+\mathrm{Recall}}. \quad (6)$$

Precision, Recall, and F1-Score each range from 0 to 1, with higher values indicating better alignment. A score of 1 indicates that all predicted points match ground-truth points within ε, and vice versa.

To quantify output diversity in the context of multi-sample reconstruction, we employ the **Coverage Score (COV)** [ADMG18; ZMM*22]. Let $\{P_i\}$ denote a set of predicted point clouds and $\{Q_j\}$ the ground-truth set. COV is computed as the fraction of unique ground-truth shapes that are closest (under CD) to at least one prediction:

$$\mathrm{COV} = \frac{|\{Q_j : \exists i \text{ such that } Q_j = \arg\min_j \mathrm{CD}(P_i,Q_j)\}|}{|\{Q_j\}|} \quad (7)$$

COV values also range from 0 to 1, with higher values indicating that predictions cover a larger fraction of the ground-truth shape space. A value of 1 means that every ground-truth shape is matched by at least one prediction, while lower values indicate mode collapse or limited diversity.

Finally, to additionally assess appearance fidelity, we report the **CIEDE2000 color difference** (ΔE00) [MLCM19], a perceptual metric that quantifies color discrepancies between predicted and ground-truth reconstructions. ΔE00 values correspond to perceptual thresholds: values below 1 are imperceptible to the human eye, values of 2–3 are perceptible upon close observation, while values above 5 are clearly visible differences [LCR01].

5. Results

We report quantitative and qualitative results. First, we analyze the impact of input modalities and supervision in an ablation study. We then assess generalization on Austrian landmark trees and IGN LiDAR data. Finally, we compare our method against state-of-the-art baselines retrained on our dataset.

	DSM					Orthophoto					DSM + Orthophoto				
Losses	CD ↓	NCD ↓	F1 ↑	COV ↑	ΔE00 ↓	CD ↓	NCD ↓	F1 ↑	COV ↑	ΔE00 ↓	CD ↓	NCD ↓	F1 ↑	COV ↑	ΔE00 ↓
Shadow	2.1964	0.4850	0.1335	38.7%	10.18	1.9938	0.4484	0.1835	25.4%	9.20	2.0699	0.4550	0.1885	25.8%	8.08
Silhouettes	2.1046	0.4624	0.1510	35.6%	8.84	1.9328	0.4566	0.1428	32.3%	7.81	2.1061	0.4446	0.1641	35.7%	**7.59**
Shadow + Silh.	2.2125	0.4410	0.2174	30.2%	8.83	1.5358	0.3723	0.2239	35.9%	7.06	1.9371	0.4531	0.2969	30.9%	8.77
BCE	1.3553	0.2789	0.7488	70.5%	8.77	1.3750	0.3427	0.5208	50.6%	8.25	1.3096	0.2765	0.8010	65.7%	8.15
BCE + Shadow	1.2277	0.2618	0.7372	65.4%	8.76	1.4297	0.3108	0.5879	60.2%	8.34	1.1738	0.2688	0.8201	75.8%	7.97
BCE + Silh.	1.1104	0.2446	0.8569	75.2%	**8.05**	1.4585	0.3200	0.5301	50.8%	7.19	1.0754	0.2456	0.8728	78.6%	8.83
Mixed	**0.9994**	**0.2296**	**0.8958**	**85.9%**	8.75	**1.3371**	**0.2913**	**0.6376**	**70.4%**	**6.91**	**0.9699**	**0.2239**	**0.8846**	**90.7%**	8.23

Table 1: *Ablation study across loss functions and input modalities. Each cell reports Chamfer Distance (CD in m), Normalized CD, F1 Score, Coverage Score (COV in %), and Mean ΔE00 color error (ΔE00). Arrows (↓/↑) indicate whether lower or higher values are better.*

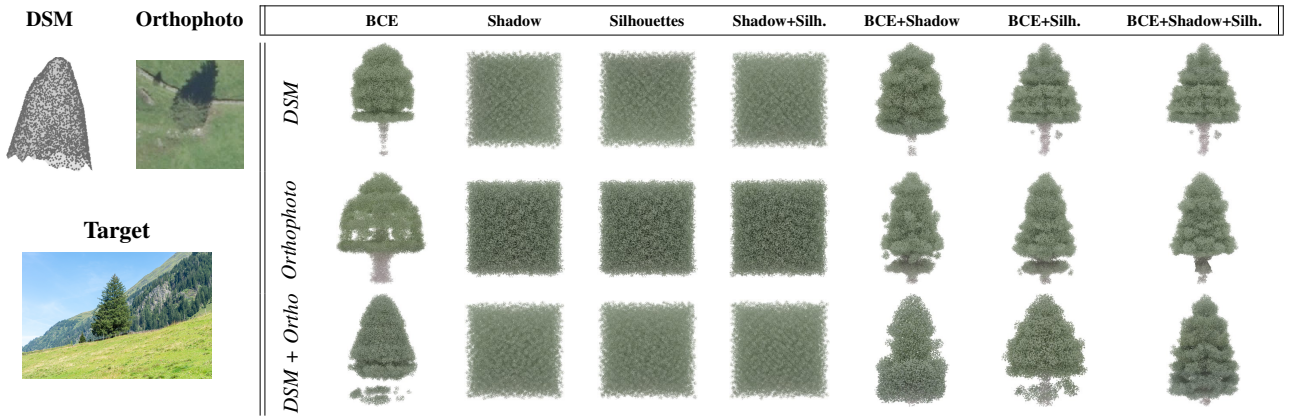


Table 2: *Qualitative ablation study of different loss functions (columns) and input modalities (rows). Left: DSM and orthophoto inputs with target photo. Right: Rows show reconstructions from DSM, orthophoto, or both inputs, while columns compare supervision strategies.*

5.1. Ablation Study

Table 1 presents an ablation study evaluating performance across four key metrics: Chamfer Distance (CD), Normalized Chamfer Distance (NCD), F1-Score (F1), and Coverage (COV) over different input modalities and loss combinations. Specifically, we evaluate models trained on (1) DSM only, (2) orthophoto only, and (3) both modalities, each under five supervision regimes: (i) BCE only, (ii) BCE + shadow, (iii) BCE + silhouette, (iv) shadow + silhouette, and (v) BCE + shadow + silhouette.

Quantitative The results presented in Table 1 reveal several important insights into the impact of loss functions and input modalities on tree shape reconstruction performance. Among all configurations, the mixed supervision with DSM and Orthophoto input stands out as the most effective approach, achieving the lowest Chamfer Distance (CD = 0.9699), the highest F1 Score (F1 = 0.8846), and the most complete reconstructions in terms of Coverage (COV = 90.7%).

An interesting trend emerges when comparing supervision types. Loss functions based solely on shadows or silhouettes perform the worst across all input configurations. For instance, using only the shadow loss with DSM input yields a CD of 2.1964 and a meager F1 of 0.1335, indicating a high degree of geometric error and almost no structural correctness. This shows that shadows or silhouettes in isolation lack sufficient signal to guide accurate reconstruction, particularly when they are not paired with 3D-aware or pixel-wise losses, as projection-based cues alone under-constrain depth and allow multiple degenerate 3D solutions.

In contrast, introducing BCE supervision — despite its simplicity — results in a large performance jump. On DSM input alone, BCE improves CD by nearly 1 meter (from 2.2125 to 1.3553) and increases F1 from 0.2174 to 0.7488. This effect is consistent across all modalities, underscoring BCE's importance in guiding point-level precision. Adding shadow or silhouette supervision on top of BCE further improves results modestly in most cases, especially in terms of Coverage, suggesting that these additional losses help regularize and fill out the reconstructed shape.

Between the two types of projection losses, silhouettes generally prove more beneficial than shadows: they align directly with canopy contours and provide a stable global constraint, whereas shadows vary with lighting direction and length, making them less reliable. For DSM or mixed inputs, silhouettes consistently lead to more complete reconstructions and higher F1 scores. The only exception is with orthophoto input, where BCE and shadow supervision slightly outperforms BCE and silhouette, suggesting that in image-based settings, shadows can provide complementary cues that silhouettes alone may not capture.

Overall, the mixed supervision setting with DSM and Orthophoto input demonstrates the value of combining all available cues: DSM provides geometric structure, orthophotos supply texture and appearance, BCE ensures point-level accuracy, and shadow/silhouette losses add shape variety. This combination not

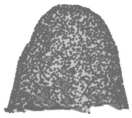
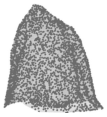
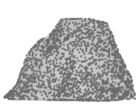
	Typical Reconstructions				Difficult Cases			
	DSM	Orthophoto	Target	Output	DSM	Orthophoto	Target	Output
1					1			
2					2				
3					3				
4					4				
5					5				

Table 3: *Qualitative comparison of reconstructed landmark trees against target photograph of the real tree. Left column: reconstructions. Right column: difficult cases: (1) DSM underestimates tree size, (2) orthophotos are cluttered by neighboring trees, (3) shadows are weak or missing, and (4) dead tree in the orthophoto. In each case, the model compensates by leveraging the complementary input signal. (5) Failure case: DSM shape and weak ortho cues mislead the model into reconstructing a conifer instead of the deciduous target.*

only improves per-sample accuracy but also reduces mode collapse, leading to the highest COV scores (90.7%) by covering a larger fraction of the ground-truth shape space. The synergies between these modalities and losses are critical, and future work could explore how to further enhance such fusion, possibly with learned weighting strategies or attention-based mechanisms that adaptively prioritize each signal.

A further dimension of the ablation is provided by the color error (ΔE00). As expected, orthophoto input improves color fidelity compared to DSM-only training, with the lowest error observed when using orthophotos with mixed supervision (6.91). In contrast, DSM-only runs consistently exhibit higher color errors ($\geq$ 8.75), reflecting the lack of spectral information in elevation data. Overall, ΔE00 values range between 6.91 and 10.18 across experiments, indicating that while reconstructions capture broad color trends, visible deviations remain due to projection mismatches between orthophoto colors and 3D geometry, as well as smoothing effects in the learned color representation.

Qualitative To better understand how different inputs and supervision signals affect reconstruction quality, we conducted a qualitative ablation study shown in Table 2. The previously discussed trends, including the impact of individual inputs and the benefits of combining supervision signals, are visually evident in the table.

5.2. Results on Austrian Landmark Trees

Individual tree instances (orthophoto and DSM patches) are extracted from large scenes using the automatic segmentation and localization pipeline of Grammatikaki et al. [Gra24], ensuring consistent centering and scale across inputs. To understand how the model generalizes to real world data we validated our approach using the dataset of landmark trees in Austria [Gra24].

Qualitative evaluation In Table 3 each row corresponds to a single tree instance. From left to right, we display the DSM and orthophoto inputs used for reconstruction, a real-world photograph of the corresponding tree, and a rendering of the predicted 3D point cloud.

Despite the sparse input data, our model is able to reconstruct 3D tree shapes that match key morphological traits observed in the target photos-such as tree height, crown size, and overall silhouette. In addition, we observe substantial variation in the reconstructed shapes, which reflects natural diversity across species (e.g., coniferous vs. deciduous forms). This suggests that the model generalizes well, even in the absence of species labels or multi-view supervision.

An important strength of our approach is its ability to remain reliable when one of the input signals is degraded. Table 3 illustrates three such cases (1, 2, and 3) in the right column. In case 1, the DSM underestimates tree size, but the orthophoto compensates. In

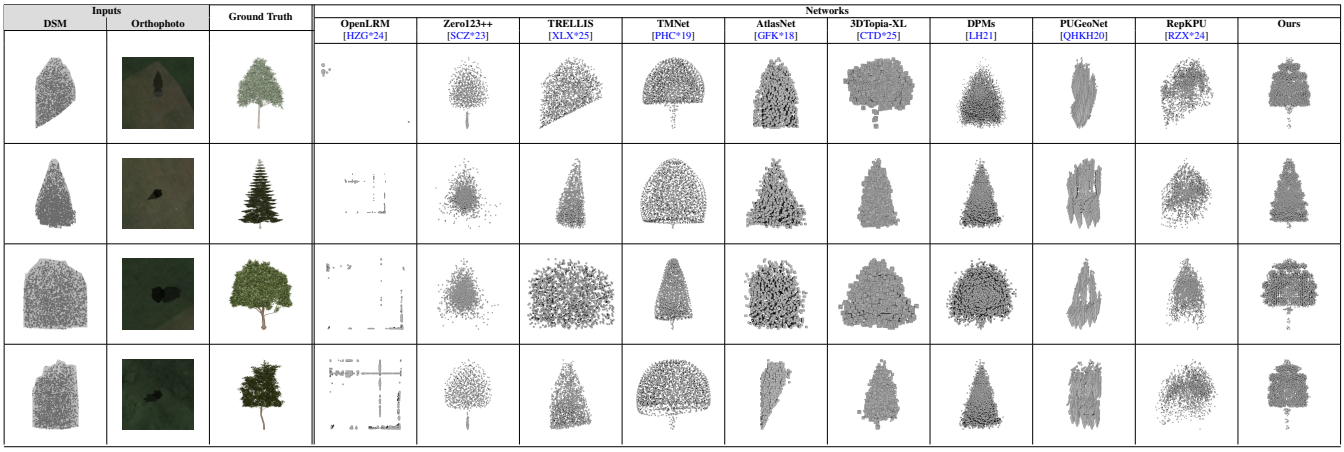


Table 4: *Visual comparison across baseline models with our approach.*

case 2, cluttered orthophotos are resolved by relying on the DSM, and in case 3, weak shadows in the orthophoto are balanced by the DSM structure, yielding coherent reconstructions in all cases. These examples confirm that the dual-input design allows the network to adaptively rely on the most reliable signal, ensuring robust results even when individual inputs are imperfect.

Quantitative evaluation Table 5 reports image-based quantitative metrics for the landmark trees in Table 3. Silhouette IoU measures crown geometry agreement, while normalized CIELAB color error [Kim97] and LPIPS assess appearance consistency. Typical cases achieve higher IoU and lower appearance error, while difficult cases degrade when input cues are misleading (cases 4–5). Case 2 reflects near-ideal DSM and orthophoto cues, while case 4 corresponds to a dead tree with ambiguous geometry and weak appearance signals, explaining the best and worst performance.

	Typical Reconstructions				Difficult Cases		
	IoU ↑	Color ↓	LPIPS ↓		IoU ↑	Color ↓	LPIPS ↓
1	0.755	0.177	0.302	1	0.753	0.149	0.252
2	**0.841**	**0.132**	**0.212**	2	0.735	0.160	0.323
3	0.804	0.239	0.273	3	0.808	0.171	0.267
4	0.759	0.152	0.298	4	**0.470**	**0.317**	**0.475**
5	0.809	0.179	0.266	5	0.568	0.272	0.381
Mean	0.794	0.176	0.270	Mean	0.667	0.214	0.340

Table 5: *Quantitative evaluation for the landmark trees shown in Table 3. Row indices correspond to Table 3. Each cell reports IoU, normalized Color Error (CIELAB), and LPIPS. Arrows (↓/↑) indicate whether lower or higher values are better.*

5.3. Comparison to State-of-the-Art Models

Table 4 and table 6 compare our method with representative baselines from four categories: single-view reconstruction networks (AtlasNet [GFK*18], TMNet [PHC*19], Zero123++ [SCZ*23]), point-based upsampling networks (PUGeoNet [QHKH20], RepKPU [RZX*24]), generative 3D models, including diffusion-based approaches (Diffusion PC (DPMs) [LH21], 3DTopia-XL [CTD*25]) and rectified-flow-based models (TRELLIS [XLX*25]), and large reconstruction models (OpenLRM [HZG*24]). We selected point-based upsampling methods (PUGeoNet, RepKPU) because they represent a natural baseline for refining sparse 3D inputs such as DSMs; single-view reconstruction networks (AtlasNet, TMNet, Zero123++), as they aim to infer full geometry from a single projected input, analogous to our orthophoto setting; generative 3D models, including diffusion-based models (DPMs, 3DTopia-XL) and rectified-flow-based models (TRELLIS), as they have recently emerged as a strong paradigm for unconditional or weakly conditioned 3D generation; and large reconstruction models (OpenLRM) that assume multi-view calibrated imagery but represent the current frontier in scaling 3D reconstruction. Together, these categories span the main strategies that could, in principle, be adapted to reconstruct tree geometry from our DSM and orthophoto inputs.

Most of these baselines were not originally developed for DSM or orthophotography data. Instead, they were originally designed for object-centric benchmarks, such as ShapeNet, or synthetic point cloud datasets, where the goal is to reconstruct rigid CAD-like objects from multi-view RGB images with known camera intrinsics. AtlasNet and TMNet are optimized for mesh generation and deformation from single-view images, while Zero123++ performs single-view view synthesis.; PUGeoNet and RepKPU target point cloud upsampling on synthetic shapes; diffusion-based models focus on unconditional or multi-view 3D generation, while rectified-flow-based models use a different flow-based generative paradigm; and large models such as OpenLRM assume multi-view imagery and camera calibration. None of these assumptions holds in our setting, where input comes from top-down orthophotos and DSMs of irregular tree crowns.

To ensure fairness, we retrain all baselines from scratch on our dataset using the same DSM and orthophoto inputs. Nonetheless, their architectural priors remain poorly aligned with our domain: tree crowns are highly irregular, lack rigid symmetries, and cannot be observed from side or front views. As a result, reconstructions often oversimplify canopy geometry, introduce noise, or miss structural details such as trunks. Qualitatively (Table 4), single-modality

baselines (DSM or orthophoto only) produce biased reconstructions, either oversimplifying or missing crown details. Single-view image-based reconstruction and view-synthesis models (AtlasNet, TMNet, Zero123++) reduce Chamfer Distance but oversmooth the canopy, while point-based upsampling methods (PUGeoNet, RepKPU) generate noisy or fragmented shapes, reflected in negligible F1 scores. Generative models (Diffusion PC, 3DTopia-XL, TRELLIS) achieve low Chamfer Distance (1.148, 1.351, and 1.412) but suffer from unstable structure and poor coverage (32.8%, 24.4%, and 33.3%). OpenLRM achieves very high coverage (75.8%) but almost zero F1, showing that while outputs are diverse, they lack geometric consistency; this is expected, as OpenLRM assumes calibrated multi-view side and front imagery, which is incompatible with our top-down DSM-orthophoto inputs. In contrast, our approach achieves the best balance (Table 6): low CD (0.969), the lowest NCD (0.224), and the highest F1 (0.884), while maintaining strong coverage (90.7%). This demonstrates that our model captures sharper crowns and more coherent geometry, aligning closely with ground truth.

Method	CD ↓	NCD ↓	F1 ↑	COV ↑
OpenLRM	10.604	0.735	0.001	75.8%
Zero123++	3.474	0.354	0.006	17.3%
TMNet	1.436	0.338	0.016	47.6%
AtlasNet	2.031	0.310	0.204	85.2%
3DTopia-XL	1.351	0.301	0.212	24.4%
Diffusion PC	1.148	0.277	0.353	32.8%
TRELLIS	1.412	0.327	0.124	33.3%
PUGeoNet	5.759	0.330	0.002	35.2%
RepKPU	1.828	0.968	0.038	31.9%
Ours	**0.969**	**0.224**	**0.884**	**90.7%**

Table 6: *Comparison of baseline models on our dataset.*

5.4. Comparison with high-resolution LiDAR data

We further validate our reconstructions against IGN LiDAR scans from the Pyrenees. Table 7 shows that our outputs align well with the LiDAR shapes, capturing overall height and crown spread. While fine details (e.g., branching) differ, the consistency across modalities confirms that our method recovers realistic tree geometry from sparse geodata.

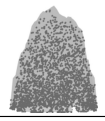
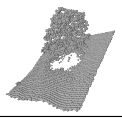
	DSM	Orthophoto	LiDAR	Ours
1				
2				

Table 7: *Qualitative comparison of our reconstructions with IGN LiDAR and orthophoto data in the French Pyrenees. Point cloud density from aerial LiDAR is 37 and 25 pts/m^2, respectively.*

5.5. Computational Efficiency

In addition to accuracy and realism, our method achieves high computational efficiency. A full pipeline run, including postprocessing, takes on average 0.3 s per tree on a single GPU. Each reconstructed tree requires only ∼0.2 MB of memory on disk, compared to tens of MBs for high-detailed meshes. This is an order of magnitude faster and lighter than the high-detail trees from Grammatikaki et al. [GEL*25], the only prior non-network method using the same DSM and orthophoto inputs to reconstruct trees, reported render times of up to 15 s per tree for 30M-triangle meshes and storage requirements in the tens of MB range. In contrast, TreeON performs end-to-end reconstruction and directly predicts a compact 6k-point colored point cloud at mapping-relevant resolution, rather than relying on a staged classification–procedural pipeline. As both approaches rely on the same Grove-generated source meshes, subsampling effects are already reflected in Table 1, enabling a fair comparison at mapping-relevant resolution.

6. Limitations

A key limitation of our approach, as with any supervised ML method, is its dependence on the quality and representativeness of the training data. In our case, synthetic meshes define which geometric structures the model can learn to reconstruct. If certain species, sizes, or crown shapes are underrepresented, the model may be biased toward more common patterns and struggle with rare or unseen structures. An instance of this limitation is the seasonal bias in our dataset: Most samples are drawn from summer orthophotos, which expose the model mainly to trees with full foliage under consistent lighting conditions. This reduces robustness to other seasonal appearances, such as leafless winter trees or dead trees (a reconstruction example is shown in Table 3, case 4 on the right), snow cover, or different vegetation stages, that are rarely included in the training data. While our current dataset does not account for this, seasonal variability could in principle be simulated directly in our training pipeline and learned by the model. Exploring such synthetic augmentation would be an interesting direction for future work, enabling the model to handle leafless or snow-covered trees.

Another challenge is the difficulty of quantitatively assessing generalizability to real-world data. While we provide qualitative examples and landmark tree reconstructions, a systematic benchmark on real orthophotos and DSMs is challenging due to the lack of high-quality paired ground truth. This underscores the motivation for using synthetic training data in the first place, but it also indicates that evaluating domain transfer remains only partially addressed.

7. Conclusion

We introduced TreeON, a neural framework for reconstructing detailed 3D tree point clouds from sparse geospatial inputs. DSMs provide strong geometric cues, while orthophotos complement them with appearance and shadow information. Combining both modalities with occupancy, shadow, and silhouette supervision achieves higher reconstruction quality than single-image or upsampling methods.

Where previous work typically fulfills only some requirements for rural 3D mapping: visual appeal, structural plausibility, efficiency, and automation — our framework addresses all four simultaneously. This is enabled by three key ingredients: (1) the strong geometric foundation from DSMs enriched by orthophoto-based appearance and shadow cues, (2) a multi-loss training strategy combining occupancy, shadow, and silhouette supervision, and (3) a compact implicit representation that allows fast inference and lightweight outputs.

The method provides structurally plausible reconstructions on a large scale, with light inference times (0.3s per tree) and lightweight outputs (0.2MB). This combination of quality and efficiency makes it particularly well-suited for integration into digital 3D maps. It effectively bridges the gap between oversimplified 2.5D surfaces and more data-rich, computationally intensive reconstructions, prioritizing visual plausibility over biological fidelity.

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. This research has been funded by WWTF project ICT22-055 - Instant Visualization and Interaction for Large Point Clouds, and Grant PID2021-122136OB-C21 funded by MICIU/AEI/10.13039/501100011033 and ERDF/EU.

References

[ACA16] ARGUDO, OSCAR, CHICA, ANTONIO, and ANDÚJAR, CARLOS. "Single-picture reconstruction and rendering of trees for plausible vegetation synthesis". *Computers & Graphics* 57 (2016), 55–67. ISSN: 0097-8493. DOI: https://doi.org/10.1016/j.cag.2016.03.005 2.

[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 2.

[ADMG18] ACHLIOPTAS, PANOS, DIAMANTI, OLGA, MITLIAGKAS, IOANNIS, and GUIBAS, LEONIDAS. *Learning Representations and Generative Models for 3D Point Clouds*. 2018 7.

[ALL25] ALONSO, PEDRO, LI, TIANRUI, and LI, CHONGSHOU. *Less is More: Efficient Point Cloud Reconstruction via Multi-Head Decoders*. 2025 7.

[bas25] BASEMAP.AT. *basemap.at*. https://basemap.at/. Accessed: 2025-11-03. 2025 6.

[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.

[BKH16] BA, JIMMY LEI, KIROS, JAMIE RYAN, and HINTON, GEOFFREY E. *Layer Normalization*. 2016 5.

[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 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.

[CTD*25] CHEN, ZHAOXI, TANG, JIAXIANG, DONG, YUHAO, et al. *3DTopia-XL: Scaling High-quality 3D Asset Generation via Primitive Diffusion*. 2025 3, 10.

[DCZ*24] DU, BANG, CHEN, KUNYAO, ZHANG, HAOCHEN, et al. "Modeling Detailed Human Geometry with Adaptive Local Refinement". en. *2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*. Seattle, WA, USA: IEEE, June 2024, 5620–5630. ISBN: 9798350365474. DOI: 10.1109/CVPRW63382.2024.00571 5.

[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 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.

[FLS*16] FASSNACHT, FABIAN EWALD, LATIFI, HOOMAN, STEREŃCZAK, KRZYSZTOF, et al. "Review of studies on tree species classification from remotely sensed data". *Remote Sensing of Environment* 186 (2016), 64–87. ISSN: 0034-4257. DOI: https://doi.org/10.1016/j.rse.2016.08.013 3.

[FSG17] FAN, HAOQIANG, SU, HAO, and GUIBAS, LEONIDAS J. "A point set generation network for 3D object reconstruction from a single image". *CVPR*. 2017 7.

[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 2, 3, 6, 11.

[GFK*18] GROUEIX, THIBAULT, FISHER, MATTHEW, KIM, VLADIMIR G., et al. *AtlasNet: A Papier-Mâché Approach to Learning 3D Surface Generation*. 2018 10.

[Gra24] GRAMMATIKAKI, ANGELIKI. *Landmark trees in Austria*. Version 1.0.0. TU Wien, Mar. 2024. DOI: 10.48436/nsj20-6ka24 7, 9.

[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 2.

[GTL*23] GU, JIATAO, TREVITHICK, ALEX, LIN, KAI-EN, et al. *NerfDiff: Single-image View Synthesis with NeRF-guided Distillation from 3D-aware Diffusion*. 2023 3.

[HLZ*17] HU, SHAOJUN, LI, ZHENGRONG, ZHIYI, ZHANG, et al. "Efficient Tree Modeling from Airborne LiDAR Point Clouds". *Computers & Graphics* 67 (May 2017). DOI: 10.1016/j.cag.2017.04.004 2.

[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 2.

[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.

[HTC23] HUANG, HONGYU, TIAN, GUOJI, and CHEN, CHONGCHENG. *Evaluating the point cloud of individual trees generated from images based on Neural Radiance fields (NeRF) method*. 2023 3.

[HZG*24] HONG, YICONG, ZHANG, KAI, GU, JIUXIANG, et al. *LRM: Large Reconstruction Model for Single Image to 3D*. 2024 10.

[HZRS16] HE, KAIMING, ZHANG, XIANGYU, REN, SHAOQING, and SUN, JIAN. "Deep Residual Learning for Image Recognition". *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. 2016, 770–778. DOI: 10.1109/CVPR.2016.90 4, 5.

[IGN25] IGN. *LiDAR HD Géoservices*. https://geoservices.ign.fr/lidarhd. Institut national de l'information géographique et forestière (IGN), France. 2025 7.

[INJN19] INDIRABAI, INDU, NAIR, M. V. HARINDRANATHAN, JAISHANKER, R. NAIR, and NIDAMANURI, RAMA RAO. "Terrestrial laser scanner based 3D reconstruction of trees and retrieval of leaf area index in a forest environment". *Ecological Informatics* 53 (2019), 100986. ISSN: 1574-9541. DOI: https://doi.org/10.1016/j.ecoinf.2019.100986 2.

[IS15] IOFFE, SERGEY and SZEGEDY, CHRISTIAN. *Batch Normalization: Accelerating Deep Network Training by Reducing Internal Covariate Shift*. 2015 4.

[JNP21] JÆGER, PETER, NILSSON, NIELS, and PALAMAS, GEORGE. "Can't see the Forest for the Trees: Perceiving Realism of Procedural Generated Trees in First-Person Games". *EAI Endorsed Transactions on Creative Technologies* (Mar. 2021). DOI: 10.4108/eai.16-3-2021.169029 2.

[Kim97] KIM, DONG-HO. "New Weighting Functions for the Modified CIELAB Colour-Difference Formulae". *Textile Coloration and Finishing* 9 (Jan. 1997) 10.

[KLL*24] KONG, XIN, LIU, SHIKUN, LYU, XIAOYANG, et al. *EscherNet: A Generative Model for Scalable View Synthesis*. 2024 3.

[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.

[LCR01] LUO, MING, CUI, GUIHUA, and RIGG, B. "The development of the CIE 2000 colour-difference formula: CIEDE2000". *Color Research & Application* 26 (Oct. 2001), 340–350. DOI: 10.1002/col.1049 7.

[LGT*23] LIN, CHEN-HSUAN, GAO, JUN, TANG, LUMING, et al. *Magic3D: High-Resolution Text-to-3D Content Creation*. 2023 3.

[LH21] LUO, SHITONG and HU, WEI. *Diffusion Probabilistic Models for 3D Point Cloud Generation*. 2021 10.

[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 3.

[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.

[LLCL19] LIU, SHICHEN, LI, TIANYE, CHEN, WEIKAI, and LI, HAO. "Soft Rasterizer: A Differentiable Renderer for Image-based 3D Reasoning". *Proceedings of the IEEE/CVF International Conference on Computer Vision*. 2019, 7708–7717 6.

[Maa13] MAAS, ANDREW L. "Rectifier Nonlinearities Improve Neural Network Acoustic Models". 2013 4.

[MHT19] MORALES, GIORGIO, HUAMÁN, SAMUEL G., and TELLES, JOEL. "Shadow Removal in High-Resolution Satellite Images Using Conditional Generative Adversarial Networks". *Information Management and Big Data*. Ed. by LOSSIO-VENTURA, JUAN ANTONIO, MUÑANTE, DENISSE, and ALATRISTA-SALAS, HUGO. Cham: Springer International Publishing, 2019, 328–340. ISBN: 978-3-030-11680-4 4.

[MLCM19] MIRJALILI, FERESHTEH, LUO, MING RONNIER, CUI, GUIHUA, and MOROVIC, JAN. "Color-difference formula for evaluating color pairs with no separation: ΔENS". *Journal of the Optical Society of America A* 36.5 (Apr. 2019), 789. ISSN: 1520-8532. DOI: 10.1364/josaa.36.000789 7.

[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.

[MON*19] MESCHEDER, LARS, OECHSLE, MICHAEL, NIEMEYER, MICHAEL, et al. "Occupancy networks: Learning 3D reconstruction in function space". *CVPR*. 2019 6, 7.

[Off25] OFFENE DATEN ÖSTERREICHS. *Offene Daten Österreichs*. https://www.data.gv.at. Accessed: 2025-11-03. 2025 7.

[PHC*19] PAN, JUNYI, HAN, XIAOGUANG, CHEN, WEIKAI, et al. "Deep Mesh Reconstruction from Single RGB Images via Topology Modification Networks". *Proceedings of the IEEE International Conference on Computer Vision*. 2019, 9964–9973 10.

[PJBM22] POOLE, BEN, JAIN, AJAY, BARRON, JONATHAN T., and MILDENHALL, BEN. *DreamFusion: Text-to-3D using 2D Diffusion*. 2022 3.

[PNM*20] PENG, SONGYOU, NIEMEYER, MICHAEL, MESCHEDER, LARS, et al. *Convolutional Occupancy Networks*. 2020 6.

[QHKH20] QIAN, YUE, HOU, JUNHUI, KWONG, SAM, and HE, YING. *PUGeo-Net: A Geometry-centric Network for 3D Point Cloud Upsampling*. 2020 10.

[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 2.

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

[RZX*24] RONG, YI, ZHOU, HAORAN, XIA, KANG, et al. "RepKPU: Point Cloud Upsampling with Kernel Point Representation and Deformation". *2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. 2024, 21050–21060. DOI: 10.1109/CVPR52733.2024.01989 10.

[SCZ*23] SHI, RUOXI, CHEN, HANSHENG, ZHANG, ZHUOYANG, et al. *Zero123++: a Single Image to Consistent Multi-view Diffusion Base Model*. 2023 3, 10.

[She05] SHEPPARD, STEPHEN R.J. "Landscape visualisation and climate change: the potential for influencing perceptions and behaviour". *Environmental Science & Policy* 8.6 (2005), 637–654. ISSN: 1462-9011. DOI: https://doi.org/10.1016/j.envsci.2005.08.002 1.

[SHK*14] SRIVASTAVA, NITISH, HINTON, GEOFFREY, KRIZHEVSKY, ALEX, et al. "Dropout: A Simple Way to Prevent Neural Networks from Overfitting". *Journal of Machine Learning Research* 15.56 (2014), 1929–1958 5.

[SJS06] SOKOLOVA, MARINA, JAPKOWICZ, NATHALIE, and SZPAKOWICZ, STAN. "Beyond Accuracy, F-Score and ROC: A Family of Discriminant Measures for Performance Evaluation". Vol. Vol. 4304. Jan. 2006, 1015–1021. ISBN: 978-3-540-49787-5. DOI: 10.1007/11941439_114 7.

[SPK*14] STAVA, ONDREJ, PIRK, SÖREN, KRATT, JULIAN, et al. "Inverse Procedural Modelling of Trees". *Computer Graphics Forum* 33 (2014). DOI: 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 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 3.

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

[TSM*20] TANCIK, MATTHEW, SRINIVASAN, PRATUL P., MILDENHALL, BEN, et al. *Fourier Features Let Networks Learn High Frequency Functions in Low Dimensional Domains*. 2020 5.

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

[WLW*23] WANG, ZHENGYI, LU, CHENG, WANG, YIKAI, et al. *ProlificDreamer: High-Fidelity and Diverse Text-to-3D Generation with Variational Score Distillation*. 2023 3.

[WWT*24] WANG, GUOQING, WANG, ZHONGDAO, TANG, PIN, et al. *OccGen: Generative Multi-modal 3D Occupancy Prediction for Autonomous Driving*. 2024 5.

[WXC*25] WEN, BEICHEN, XIE, HAOZHE, CHEN, ZHAOXI, et al. *3D Scene Generation: A Survey*. 2025 3.

[XGC07] XU, HUI, GOSSETT, NATHAN, and CHEN, BAOQUAN. "Knowledge and heuristic-based modeling of laser-scanned trees". *ACM Transactions on Graphics* 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 3, 10.

[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 2.

[YFF*23] YU, TAO, FENG, RUNSENG, FENG, RUOYU, et al. *Inpaint Anything: Segment Anything Meets Image Inpainting*. 2023. DOI: 10.48550/arXiv.2304.06790 4.

[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 2.

[ZIE*18] ZHANG, RICHARD, ISOLA, PHILLIP, EFROS, ALEXEI A, et al. "The Unreasonable Effectiveness of Deep Features as a Perceptual Metric". *CVPR*. 2018 2, 5.

[ZLB*23] ZHOU, XIAOCHEN, LI, BOSHENG, BENES, BEDRICH, et al. *DeepTree: Modeling Trees with Situated Latents*. 2023 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 2, 3.

[ZMM*22] ZHANG, YUHANG, MIAO, ZHENWEI, MI, TIEBIN, et al. "Dual adversarial model: Exploring low-dimensional space features for point clouds generating and completing". *Computer Vision and Image Understanding* 223 (2022), 103551. ISSN: 1077-3142. DOI: https://doi.org/10.1016/j.cviu.2022.103551 7.