From Canopy to Crown: High-Fidelity Tree Façade Synthesis from Nadir LiDAR data

Raghav Sharma[1], Frank Zhang[2], Jane Liu [3], Baoxin Hu [4]

[1,2] Department of Computer Science, University of Fraser Valley, Abbotsford, BC V2S 7M7, Canada,
[3] Department of Geography and Planning, University of Toronto, Toronto, ON M5S 3G3, Canada, janejj.liu@utoronto.ca
[4] Department of Earth and Space Science and Engineering, York University, 4700 Keele St., Toronto, ON M3J 1P3, Canada, baoxin@yorku.ca

Keywords: Diffusion models, tree crown synthesis, LiDAR, remote sensing, novel view synthesis, forest inventory

Abstract

Synthesizing realistic façade views of individual trees from nadir-view remote sensing data would transform large-scale forest analysis, yet remains unsolved due to data scarcity and task ambiguity. We present the first conditional diffusion model to generate structurally plausible façade views of individual tree crowns from single nadir-view LiDAR rasters, leveraging the FOR-species20K benchmark dataset. Our approach integrates nadir projections with tree species and height within a U-Net-based denoising diffusion framework. Experiments demonstrate that nadir imagery alone is insufficient, but conditioning on species and height enables synthesis of visually realistic, species-specific façade views. The fully conditioned model achieves substantial gains in perceptual (LPIPS: 0.184) and structural (SSIM: 0.576) similarity, outperforming nadir-only baselines by more than twofold. Our results establish that ancillary attributes critically constrain the solution space, enabling diffusion models to infer plausible structures from ambiguous nadir input. This work demonstrates a scalable path to enriching nadir-based forest inventories with synthesized structural detail, reducing the need for resource-intensive ground surveys.

1. Introduction

Deep generative models have shown exceptional capability in producing high-quality synthetic data across diverse modalities (Goodfellow et al., 2014; Ho et al., 2020; Rombach et al., 2022). Among them, diffusion models have achieved state-of-the-art performance in visual content generation due to their stability, fidelity, and ability to model complex data distributions (Song et al., 2021), recently extending to novel view synthesis from limited input imagery.

In remote sensing of individual trees, the most accessible perspective is the nadir view, acquired via UAVs, airborne LiDAR, or satellites. In contrast, façade views require oblique or ground-based acquisitions, which are logistically demanding. Consequently, large-scale individual tree crown (ITC) datasets are dominated by nadir views, constraining applications in forest inventory, urban forestry, and ecological modeling.

No prior study has applied diffusion models to synthesize façade views of ITCs from nadir data, partly due to lack of large, paired nadir–façade datasets. The recent FOR-species20K dataset (Puliti et al., 2025) addresses this gap by providing approximately 20,000 high-resolution, proximally sensed laser-scanned ITCs from which both nadir and façade rasterizations can be derived.

This study investigates whether nadir-view ITC imagery, augmented with ancillary attributes such as species identity and tree height, contains sufficient information to synthesize visually realistic and structurally plausible façade views. Specifically, we aim to: (1) assess the feasibility of façade-view generation from single nadir-view ITC LiDAR rasterizations, and (2) quantify the contribution of ancillary attributes to perceptual and structural quality.

2. Methods

2.1 Study Area and Data Acquisition

Data were sourced from the FOR-species20K benchmark dataset, a large, open-source collection of individual tree point clouds developed for machine learning applications (Puliti et al., 2025). The dataset contains 20,158 high-quality, manually segmented individual tree point clouds compiled from 25 separate data collections.

The geographic scope is extensive, with 90% of trees scanned in temperate (61%), boreal (25%), and Mediterranean (7%) biogeographic regions across Europe, with additional data from Canada, Australia, and New Zealand to broaden ecological and structural diversity.

Data acquisition was performed using three proximal laser scanning modalities: terrestrial laser scanning (TLS) (70% of trees, using 12 different sensors including Leica P20, RIEGL VZ-400, and Faro Focus 3D), uncrewed aerial vehicle laser scanning (ULS) (22%), and mobile laser scanning (MLS) (8%). This multi-platform approach results in significant heterogeneity in point cloud characteristics, including variations in point density, completeness, and occlusion patterns.

The dataset features 33 tree species across 19 genera, with representation reflecting realistic abundance distributions. The most common species are Pinus sylvestris (3,296 trees) and Fagus sylvatica (2,482 trees), while rarer species such as Abies alba (119 trees) and Prunus avium (50 trees) are also included. Tree heights range from 0.3 m to over 56 m, encompassing significant intra-specific variation in crown architecture due to differing growth conditions.

2.2 Data Pre-processing and Rasterization

The raw data underwent rigorous pre-processing and quality control to prepare a high-quality, standardized dataset. The

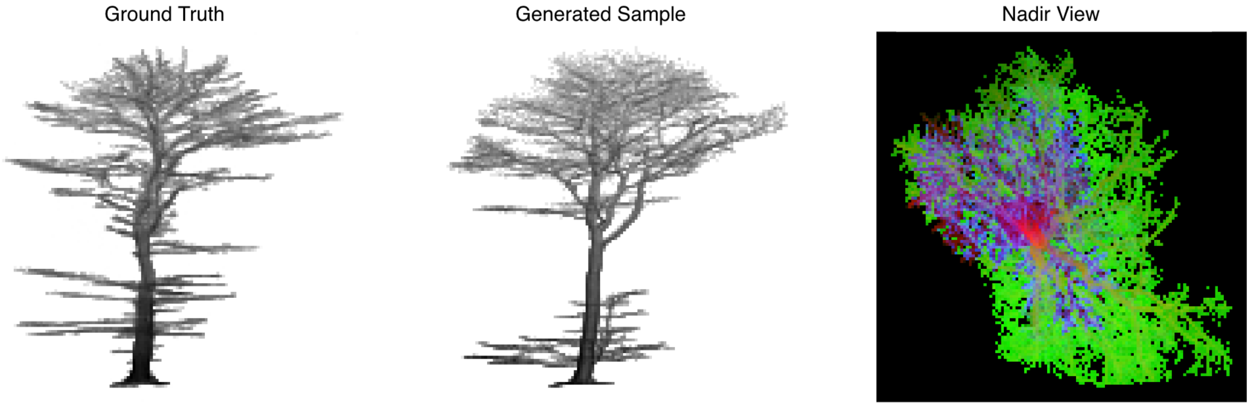

Figure 1. Illustration of the nadir-to-façade synthesis challenge. Right: widely available nadir (top-down) view of an individual tree crown. Left: the corresponding façade (side) view, which is logistically expensive to acquire but contains critical structural information for forest analysis. Center: Our approach synthesizes the façade view from the nadir view using conditional diffusion models.

pipeline isolated clean, information-rich point clouds and removed statistical outliers, producing paired, co-registered nadir and façade raster images for each tree.

Initial Selection: Only TLS-acquired data were retained to ensure high point cloud density and structural detail. Trees with height greater than 3 meters were included, excluding saplings and small understory trees lacking distinct structural features.

Quality Control: An automated audit was performed based on preliminary rasterized façade projections. Key metrics included point count, projection sparsity (proportion of non-zero pixels), mean pixel intensity, and standard deviation. Samples were removed if they met any of the following quantile-based criteria: façade sparsity in the lowest 5th percentile; mean pixel value in the lowest 2nd percentile; sparsity or mean value in the highest 1st percentile; or top 5th percentile for both mean and standard deviation (indicating extreme contrast artifacts).

Paired Rasterization: Each point cloud generated a pair of co-registered 128×128 pixel images: a multi-channel nadir view (input) and a single-channel façade view (target).

The nadir view was created by projecting points onto a top-down (X-Y) plane, encoded as a three-channel RGB image:

- **Red Channel (Point Density):** Spatial distribution with logarithmic contrast scaling to preserve canopy texture
- **Green Channel (Normalized Mean Height):** Average height per pixel, normalized by tree's min-max height
- **Blue Channel (Normalized Height Variation):** Standard deviation of point heights per pixel, highlighting structural complexity

The façade view was generated by projecting onto a side-on (X-Z) plane, encoded as single-channel grayscale representing point density with logarithmic contrast enhancement. Both views share a common coordinate system, ensuring spatial coherence.

Final Dataset: The pipeline produced 11,440 paired images, split 95/5 into training (10,868 pairs) and testing (572 pairs) subsets with fixed seed (42) for reproducibility.

2.3 Conditional Diffusion Model

Our generative model is a conditional U-Net operating within a denoising diffusion probabilistic model (DDPM) framework (Ho et al., 2020), designed to predict the noise component of a corrupted façade image at timestep t. The architecture features a symmetric encoder-decoder with skip connections to preserve high-resolution spatial features (Ronneberger et al., 2015).

The encoder consists of four down-sampling blocks progressively increasing feature depth from 128 to 256, 384, and 512 channels at the bottleneck, with two residual layers per block. The decoder mirrors this structure, up-sampling while reducing channel depth, outputting a single-channel noise prediction.

Conditioning Mechanisms:

1. **Image-based Conditioning:** The three-channel nadir view is concatenated with the single-channel noisy façade view at the initial layer, creating a four-channel tensor ensuring spatial information availability throughout the network.
2. **Attribute Conditioning:** Tree height is processed by a small MLP; species labels are converted to dense vectors via an embedding layer.
3. **Timestep Conditioning:** Noise level timestep t is converted to a sinusoidal embedding.

Three Experimental Models:

- **Model 1 (Fully-Conditioned):** Nadir + Species + Height. Conditioning vector sums timestep, species, and height embeddings, injected into each residual block.
- **Model 2 (Height-Conditioned):** Nadir + Height. Conditioning vector sums only timestep and height embeddings.
- **Model 3 (Nadir-Only):** Nadir view only. Conditioning vector contains only timestep embedding.

This setup enables direct comparison to quantify the contribution of species and height information to façade synthesis quality.

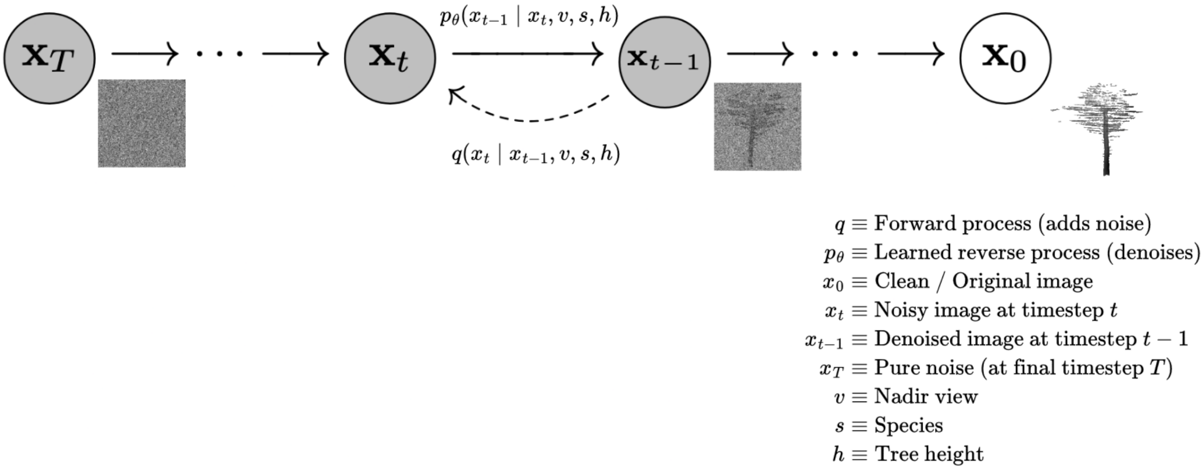
$\mathbf{x}_T$
$\mathbf{x}_t$
$p_\theta(x_{t-1} \mid x_t, v, s, h)$
$q(x_t \mid x_{t-1}, v, s, h)$
$\mathbf{x}_{t-1}$
$\mathbf{x}_0$
$q \equiv$ Forward process (adds noise)
$p_\theta \equiv$ Learned reverse process (denoises)
$x_0 \equiv$ Clean / Original image
$x_t \equiv$ Noisy image at timestep t
$x_{t-1} \equiv$ Denoised image at timestep $t-1$
$x_T \equiv$ Pure noise (at final timestep T)
$v \equiv$ Nadir view
$s \equiv$ Species
$h \equiv$ Tree height

Figure 2. Conditional U-Net architecture showing the three conditioning mechanisms: image-based conditioning, attribute conditioning (species and height embeddings), and timestep conditioning. The conditioning vector is injected into each residual block of the encoder-decoder path.

2.4 Training and Implementation

Models were trained independently to minimize Mean Squared Error (MSE) between predicted and actual noise using Adam optimizer ($\epsilon = 10^{-5}$) (Kingma & Ba, 2014). All models trained for 50 epochs with batch size 64 (∼9,000 steps).

Learning rates were determined via learning rate finder (Smith, 2018) for each model: Model 1: 2×10^{-3}; Model 2: 1×10^{-4}; Model 3: 5×10^{-3}. The One-Cycle learning rate schedule was applied, gradually increasing to maximum then decreasing over training.

Implementation used PyTorch (Paszke et al., 2019) with the miniai framework. Training on a single NVIDIA A100 GPU employed automatic mixed-precision (AMP) for efficiency, with each model fully trained in approximately one hour.

2.5 Evaluation Metrics

Models were evaluated on the 572-sample test set. For each model, façade views were generated using a DDPM scheduler over 1000 timesteps. Both generated and ground-truth images were normalized before computing four metrics:

1. **Mean Absolute Error (MAE):** Average absolute pixel-wise difference (lower is better)
2. **Peak Signal-to-Noise Ratio (PSNR):** Logarithmic quality measure comparing maximum signal to noise (higher is better)
3. **Structural Similarity Index (SSIM):** Perceptual similarity based on luminance, contrast, and structure (range: -1 to 1, higher is better) (Wang et al., 2004)
4. **Learned Perceptual Image Patch Similarity (LPIPS):** Deep feature-based perceptual similarity using pre-trained AlexNet (lower is better) (Zhang et al., 2018)

3. Results

To quantitatively assess the impact of ancillary conditioning information on façade-view synthesis, the three models—Fully-Conditioned, Height-Conditioned, and Nadir-Only—were evaluated on the hold-out test set using four distinct metrics. The mean performance across all 572 test samples is summarized in Table 1.

Model	**MAE↓**	**PSNR↑**	**SSIM↑**	**LPIPS↓**
Nadir+Sp.+Ht.	0.077	15.48	0.576	0.184
Nadir+Height	0.087	14.57	0.538	0.259
Nadir Only	0.123	12.95	0.269	0.424

Table 1. Model performance on test set. Arrows indicate better direction.

The quantitative results reveal a clear and consistent performance hierarchy across all evaluation metrics, demonstrating a direct correlation between the amount of conditioning information provided and the fidelity of the generated façade view. The Fully-Conditioned model, which leveraged nadir, species, and height information, decisively outperformed the other configurations. It achieved the lowest (best) scores for both pixel-wise and perceptual error, with a Mean Absolute Error (MAE) of 0.077 and a Learned Perceptual Image Patch Similarity (LPIPS) score of 0.184. Concurrently, it obtained the highest (best) scores for image quality, with a Peak Signal-to-Noise Ratio (PSNR) of 15.48 and a Structural Similarity Index (SSIM) of 0.576.

Performance systematically degraded as conditioning information was removed. The ablation of species information (Height-Conditioned model) resulted in a marked increase in error across all metrics. The Nadir-Only model, which lacked any ancillary data, yielded the poorest results, with a particularly significant drop in structural similarity (SSIM: 0.269) and a more than two-fold increase in perceptual error (LPIPS: 0.424) compared to the fully-conditioned model.

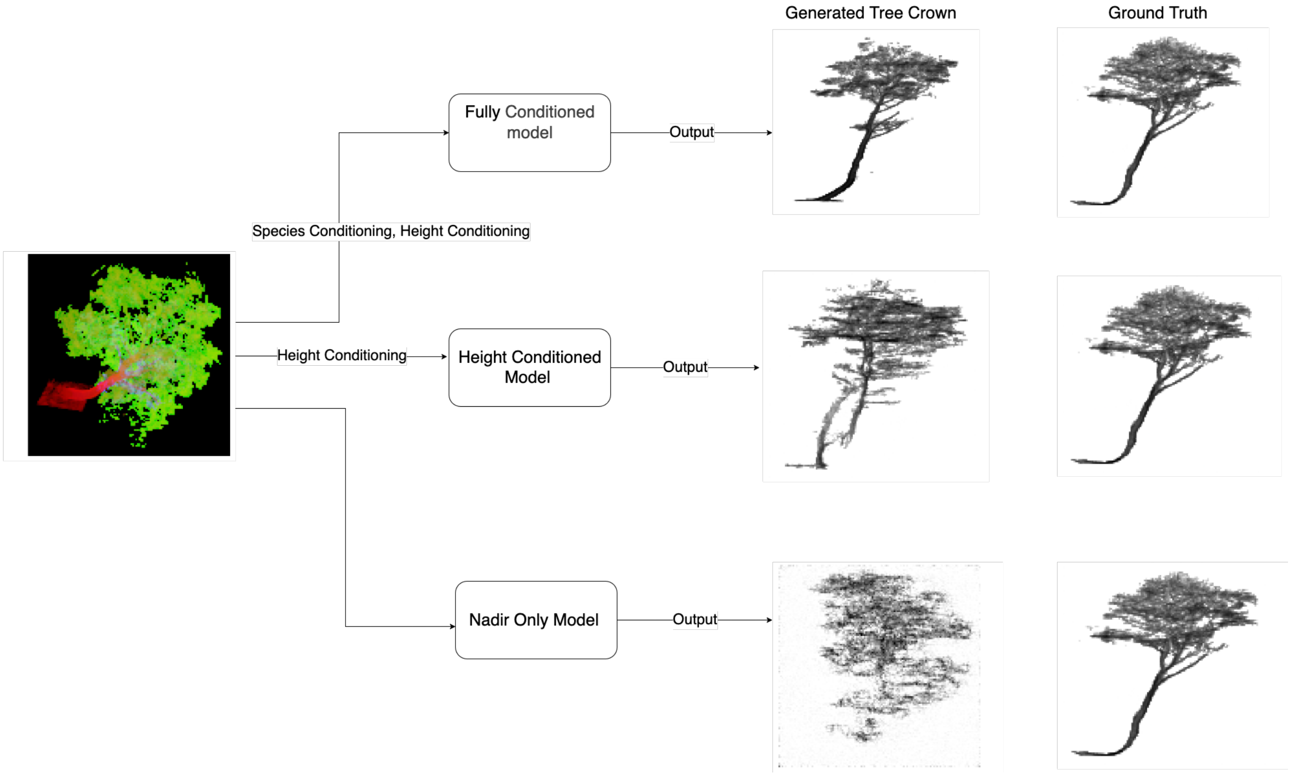

Figure 3. Example outputs from the three model configurations. Top row: Fully-Conditioned model output. Middle row: Height-Conditioned model output. Bottom row: Nadir-Only model output. The progression demonstrates the critical role of ancillary conditioning information in generating structurally plausible façade views.

Qualitative examples, presented in Figure 3, provide a visual counterpart to the quantitative metrics and illustrate the contribution of each piece of ancillary data. The output from the Nadir-Only model typically results in a generic, often amorphous structure, struggling to capture more than the basic presence of an object. The introduction of tree height as a condition (Height-Conditioned model) yields a substantial improvement in overall scale and proportion, but the output often remains generic in form. Finally, the inclusion of the species label in the Fully-Conditioned model enables the synthesis of highly plausible and recognizable tree structures. The model successfully generates characteristic crown shapes and branching patterns consistent with the ground-truth species, demonstrating that all three conditioning inputs are necessary to achieve high-fidelity façade views.

Figure 4 presents the distribution of LPIPS scores for each model across the entire test set, offering insight into their relative performance and consistency. The plot visually confirms the trend observed in the mean metric scores, where performance degrades as conditioning information is removed. The Fully-Conditioned model not only achieves the lowest median LPIPS score but also exhibits the most consistent performance, as indicated by its tight distribution centered around low error values. In contrast, the distributions for the Height-Conditioned and Nadir-Only models are progressively wider and shifted towards higher (worse) LPIPS scores. The Nadir-Only model, in particular, shows significant variance in its outputs, with a broad spread of scores indicating highly inconsistent perceptual quality.

4. Discussion

This study investigated whether the nadir view of an Individual Tree Crown (ITC), supplemented with ancillary data, contains sufficient information to generate a realistic and structurally plausible façade view. Furthermore, we sought to determine the quantity of this ancillary information necessary to achieve high-fidelity synthesis. Our quantitative and qualitative results demonstrate that conditional diffusion models can successfully learn this complex relationship. The findings clearly indicate that while a basic façade can be generated from a nadir view alone, the inclusion of tree height and species labels as conditioning variables dramatically and progressively improves the quality, realism, and structural coherence of the output.

Interpreting Model Performance: Plausibility vs. Reconstruction. A key finding is the divergence between pixel-wise reconstruction metrics (PSNR, SSIM) and the perceptual metric (LPIPS). The fully-conditioned model, while achieving a respectable SSIM score of 0.576, still shows modest performance on metrics that demand pixel-perfect reconstruction. This is expected, as the task is fundamentally ill-posed: a single 2D nadir view could correspond to a multitude of valid 3D tree structures and, consequently, many possible façade views. The model's objective is not to reconstruct the one, exact ground-truth structure—an impossible task given the input—but to generate a plausible façade that is consistent with the provided information. The strong LPIPS score of 0.184 achieved by the fully-conditioned model confirms its success in this regard, indicating that the model is learning an implicit, species-specific structural prior and successfully inferring likely branching patterns and crown morphologies.

The Critical Role of Conditioning Variables. The progress-

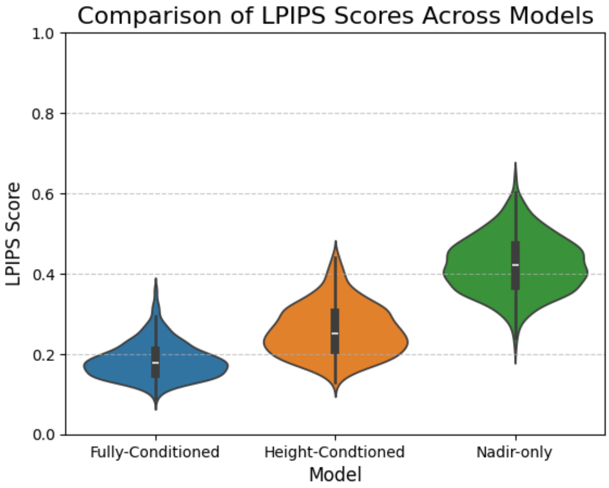

Figure 4. Distribution of LPIPS scores across the test set for each model. The Fully-Conditioned model achieves the lowest median LPIPS score with tight distribution, while the Nadir-Only model shows significant variance indicating inconsistent perceptual quality.

ive improvement across our three models underscores the importance of conditioning. The failure of the Nadir-Only model is particularly instructive. Its outputs, often resembling noisy, amorphous shapes, reveal the profound ambiguity of the task without constraints. Lacking information on scale or morphology, the model struggles to make informed decisions during the denoising process and defaults to a generic, low-structure output. The addition of tree height provides the first critical constraint, allowing the model to infer the correct scale and proportions. Given that tree height is readily derivable from nearly all forms of LiDAR data, including sparse UAV scans, its inclusion represents a highly effective and practical method for enhancing generation quality. The inclusion of the species label provides the final and most powerful constraint, allowing the model to move beyond generating a generic “tree” of a certain height and instead generate a tree with species-specific characteristics. This acts as a strong prior, guiding the model to synthesize the correct crown architecture and branching patterns.

Implications and Significance. This work has significant implications for remote sensing, forestry, and ecology. By demonstrating the ability to generate realistic façade views from nadir data, our method offers a pathway to substantially reduce the cost and logistical challenges associated with ground-based surveys. Scalable nadir data from UAV or aerial platforms could be enriched with synthesized façade views, enabling more comprehensive, 3D-aware forest analysis. Potential applications include enhanced forest inventories with improved estimates of biomass, timber volume, and carbon stocks (Calders et al., 2015); urban forestry assessments of structural health and potential hazards at scale; and ecological modeling with more detailed habitat models for species dependent on specific tree architectures.

Limitations. Despite promising results, this study has several limitations: (1) *Closed-Set Species Recognition*—the model was trained on 30 species and cannot generate species outside its training distribution; (2) *Focus on Isolated Trees*—the model was trained on well-segmented, individual trees and its ability to handle complex, real-world scenes with overlapping canopies has not been evaluated; (3) *Single Façade View*—our method generates a single, fixed façade view and does not yet capture the full 360-degree structure.

Future Directions. Several compelling avenues for future research could significantly extend the model’s capabilities. A primary direction is advancing from structural to photorealistic synthesis by generating full-color RGB façade views, requiring a multi-modal approach leveraging co-registered RGB and LiDAR nadir data. Future iterations could explore more sophisticated architectural improvements such as cross-attention layers, enabling the decoder to dynamically attend to salient spatial features at different generation stages (Rombach et al., 2022). A natural extension is synthesizing complete 3D point clouds from single nadir views, representing a paradigm shift in data acquisition. Finally, addressing the open-set problem by developing methods to recognize and flag unseen species would be crucial for creating a truly robust and deployable tool for large-scale forest analysis.

5. Conclusion

In this paper, we addressed the challenge of data scarcity in façade-view tree analysis by demonstrating that conditional diffusion models can successfully generate high-fidelity façade views from nadir-view LiDAR rasters. Our systematic evaluation of models with varying levels of conditioning revealed that while a nadir view alone is insufficient for this task, its combination with key ancillary attributes—namely tree height and species—provides the necessary constraints to synthesize structurally coherent and perceptually realistic outputs. The primary significance of this work lies in its potential to democratize the acquisition of detailed, 3D-aware forest data, transforming a logistical bottleneck into a solvable computational problem. Ultimately, this research validates a novel application of generative models in remote sensing and establishes a new baseline for cross-dimensional data synthesis, opening new frontiers for large-scale ecological analysis and forest management.

References

Anderson, B., Lottes, P., Stiller, C., 2024. SVDTree: Semantic Voxel-based Deep implicit-function-based Tree reconstruction from a single photograph. *arXiv preprint arXiv:2401.10874*.

Calders, K., Newnham, G., Burt, A., Murphy, S., Raumonen, P., Herold, M., Disney, M., 2015. Nondestructive estimates of above-ground biomass using terrestrial laser scanning. *Methods in Ecology and Evolution*, 6(2), 198–208.

Goodfellow, I., Pouget-Abadie, J., Mirza, M., Xu, B., Warde-Farley, D., Ozair, S., Bengio, Y., 2014. Generative adversarial nets. *Advances in Neural Information Processing Systems*, 27.

Ho, J., Jain, A., Abbeel, P., 2020. Denoising diffusion probabilistic models. *Advances in Neural Information Processing Systems*, 33, 6840–6851.

Kim, S., Park, J., Lee, J., Kim, D., 2023. MVDiffusion: Enabling Realistic Multi-view Image Generation with Correspondence-Aware Diffusion Models. *arXiv preprint arXiv:2307.01097*.

Kingma, D. P., Ba, J., 2014. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*.

Li, M., Li, Y., Feng, R., Li, B., 2023. Diff-tree: A Diffusion-based Model for Realistic 3D Tree Point Cloud Generation. *arXiv preprint arXiv:2308.14920*.

Paszke, A., Gross, S., Massa, F., Lerer, A., Bradbury, J., Chanan, G., Chintala, S., 2019. PyTorch: An imperative style, high-performance deep learning library. *Advances in Neural Information Processing Systems*, 32.

Puliti, S., Lines, E. R., Müllerová, J., Frey, J., Schindler, Z., Straker, A., Astrup, R., 2025. Benchmarking tree species classification from proximally sensed laser scanning data: Introducing the FOR-species20K dataset. *Methods in Ecology and Evolution*, 16(4), 801–818.

Rombach, R., Blattmann, A., Lorenz, D., Esser, P., Ommer, B., 2022. High-resolution image synthesis with latent diffusion models. *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 10684–10695.

Ronneberger, O., Fischer, P., Brox, T., 2015. U-net: Convolutional networks for biomedical image segmentation. *Proceedings of the International Conference on Medical Image Computing and Computer-Assisted Intervention*, 234–241.

Smith, L. N., 2018. A disciplined approach to neural network hyper-parameters: Part 1–learning rate, batch size, momentum, and weight decay. *arXiv preprint arXiv:1803.09820*.

Smith, L. N., Topin, N., 2017. Super-convergence: Very fast training of neural networks using large learning rates. *arXiv preprint arXiv:1708.07120*.

Song, Y., Sohl-Dickstein, J., Kingma, D. P., Kumar, A., Ermon, S., Poole, B., 2021. Score-based generative modeling through stochastic differential equations. *Proceedings of the International Conference on Learning Representations*.

Wang, G., Gu, J., He, K., Chen, X., Liu, Z., 2023. MultiDiff: Generating Geometrically Consistent and High-Quality Multiview Images from a Single Image. *arXiv preprint arXiv:2309.04424*.

Wang, Z., Bovik, A. C., Sheikh, H. R., Simoncelli, E. P., 2004. Image quality assessment: from error visibility to structural similarity. *IEEE Transactions on Image Processing*, 13(4), 600–612.

Zhang, R., Isola, P., Efros, A. A., Shechtman, E., Wang, O., 2018. The unreasonable effectiveness of deep features as a perceptual metric. *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 586–595.

Zhang, T., Zhang, S., Liu, C., Dong, J., 2023. GAvatar: A generalizable 3d-aware generative avatar model. *arXiv preprint arXiv:2307.13780*.