A HIGH RESOLUTION CANOPY HEIGHT MODEL USING A DEEP LEARNING APPROACH

KEVIN RAFAEL ROA GARCIA

UNIVERSIDAD INDUSTRIAL DE SANTANDER
FACULTAD DE INGENIERÍAS FISICOMECÁNICAS
ESCUELA DE INGENIERÍA ELÉCTRICA, ELECTRÓNICA Y DE TELECOMUNICACIONES
BUCARAMANGA
2026

A HIGH RESOLUTION CANOPY HEIGHT MODEL USING A DEEP LEARNING APPROACH

KEVIN RAFAEL ROA GARCIA

Degree work presented as a requirement to qualify for the title of Electronic Engineer

Advisor:
Ana Beatriz Ramirez Silva.
PhD in Electrical Engineering

UNIVERSIDAD INDUSTRIAL DE SANTANDER
FACULTAD DE INGENIERÍAS FISICOMECÁNICAS
ESCUELA DE INGENIERÍA ELÉCTRICA, ELECTRÓNICA Y DE TELECOMUNICACIONES
BUCARAMANGA
2026

Dedicated to my family and friends.

ACKNOWLEDGEMENTS

I would like to extend my sincere gratitude to my thesis advisor, Ana Beatriz, for her patience, confidence, and expert guidance throughout this entire process. Her mentorship was fundamental to the completion of this work.

I am sincerely thankful to the Industrial University of Santander (UIS), which offered not only the academic framework for this research but also an environment that fostered my personal and intellectual development. My time at UIS helped me better understand myself and discover my interests and passions.

I also express my sincere appreciation to all the professors and faculty members for the invaluable lessons and examples they provided during this long journey, their mentorship has impacted my perspective and broadened my vision for the future.

Finally, my deepest love is reserved for my family, especially my grandparents, my aunts and uncles. The significant sacrifices you made to grant me the opportunity to study are the true foundation of this achievement. This accomplishment is as much yours as it is mine.

With appreciation, **Kevin Rafael Roa Garcia**

TABLE OF CONTENTS

Page.

LIST OF FIGURES

Page.

LIST OF TABLES

Page.

RESUMEN

TÍTULO: A HIGH RESOLUTION CANOPY HEIGHT MODEL USING A DEEP LEARNING APPROACH [*]

AUTOR KEVIN RAFAEL ROA GARCIA [**]

PALABRAS CLAVE: Modelo de Altura de Dosel (CHM), Aprendizaje Profundo (DL), Visión por Computadora, LiDAR, Teledetección, Procesamiento de Imágenes, Monitoreo Forestal.

DESCRIPCIÓN:

Este trabajo tiene como objetivo desarrollar un modelo de altura de dosel o por sus siglas en ingles (CHM) mediante la aplicación de algoritmos de aprendizaje profundo y técnicas de visión por computadora, a partir de la fusión de datos LiDAR e imágenes RGB de alta resolución.

El monitoreo y la gestión eficiente de los ecosistemas forestales requieren modelos de altura de dosel (CHM) de alta resolución; sin embargo, los productos globales existentes presentan limitaciones en resolución, precisión y aplicabilidad local. Esta falta de detalle afecta la estimación confiable de indicadores clave como el carbono almacenado y la biodiversidad. En este trabajo se desarrolló y evaluó un modelo de aprendizaje profundo basado en técnicas de visión por computadora para la estimación de la altura del dosel. El modelo fue entrenado utilizando imágenes RGB de alta resolución y datos de referencia CHM derivados de LiDAR aéreo en un área de estudio específica. Los resultados demuestran la capacidad de la metodología para generar mapas CHM predictivos de alta resolución y estimar la altura del dosel con alta precisión a partir únicamente de imágenes. De esta manera, se valida una solución práctica y escalable que contribuye como herramienta de apoyo para el monitoreo, la caracterización detallada de la estructura forestal y la gestión sostenible de los recursos forestales.

---

[*] Trabajo de Grado

[**] Facultad de Ingenierías Físico-Mecánicas. Escuela de Ingenierías Eléctrica, Electrónica y de Telecomunicaciones. Directora: Ana Beatriz Ramirez Silva

ABSTRACT

TITLE: A HIGH RESOLUTION CANOPY HEIGHT MODEL USING A DEEP LEARNING APPROACH [*]

AUTHORS: KEVIN RAFAEL ROA GARCIA [**]

Keywords: Canopy Height Model (CHM), Deep Learning (DL), Computer Vision, LiDAR, Remote Sensing, Image Processing, Forest Monitoring.

DESCRIPTION:

This work aims to develop a canopy height model (CHM) through the application of deep learning algorithms and computer vision techniques, based on the fusion of LiDAR data and high-resolution RGB imagery.

The monitoring and efficient management of forest ecosystems require high-resolution canopy height models (CHMs); however, existing global products present limitations in terms of resolution, accuracy, and local applicability. This lack of detail affects the reliable estimation of key indicators such as stored carbon and biodiversity. In this study, a deep learning model based on computer vision techniques was developed and evaluated for canopy height estimation. The model was trained using high-resolution RGB imagery and reference CHM data derived from airborne LiDAR within a specific study area. The results demonstrate the capability of the proposed methodology to generate high-resolution predictive CHM maps and accurately estimate canopy height using only imagery. In this way, a practical and scalable solution is validated, contributing as a supporting tool for forest monitoring, detailed structural characterization, and sustainable forest resource management.

---

[*] BSc Thesis

[**] Facultad de Ingenierías Físico-Mecánicas. Escuela de Ingenierías Eléctrica, Electrónica y de Telecomunicaciones. Advisor: Ana Beatriz Ramirez Silva

INTRODUCTION

The study and sustainable management of forest ecosystems fundamentally rely on accurate information about their structure, with the Canopy Height Model (CHM) being a critical parameter for estimating stored carbon and monitoring biodiversity.[1] Currently, obtaining highly detailed CHM models remains a persistent challenge in the field of remote sensing and engineering.

While data acquisition technologies have advanced rapidly, traditional methods like airborne LiDAR are costly and have limited coverage. Furthermore, global satellite-basedp products like GEDI[2] offer broad accessibility but often at low resolutions that are insufficient for detailed ecological studies at the local scale . This gap between accessibility and the need for high resolution (1 meter) is the central problem addressed by this work.

This project proposes the development and validation of a Canopy Height Model based on Deep Learning, utilizing a computer vision approach. The solution focuses on an advanced architecture (featuring a pre-trained Transformer backbone and an adapted U-Net header) to exploit the spatial richness of RGB imagery and train the model with high-quality reference CHM data. This methodological approach seeks to establish a practical and scalable solution for the scientific and forestry community.

The developed model demonstrates its capacity to generate predictive CHM maps with a 1-meter resolution, establishing a viable methodology for continuous monitoring using

---

[1] N. L. STEPHENSON et al. “Rate of tree carbon accumulation increases continuously with tree size”. In: *Nature* 507.7490 (2014), pp. 90–93. DOI: 10.1038/nature12914.

[2] R. DUBAYAH et al. *GEDI L2A Elevation and Height Metrics Data Global Footprint Level V002*. Fecha de acceso: 2025-07-28. 2021. DOI: 10.5067/GEDI/GEDI02_A.002.

only optical imagery. This breakthrough is key, as it provides a robust support tool for natural resource management, reducing the reliance on costly LiDAR acquisitions.

The preceding sections establish the context and justify the methodology utilized in this study. Following this Introduction, the document details the necessary theoretical framework, covering concepts in Remote Sensing, LiDAR, and Deep Learning architectures for computer vision. Subsequent sections rigorously describe the methodological flow, including data selection, architecture adaptation, and training strategies. The results section presents the quantitative and qualitative performance of the proposed architecture, including a comparative benchmark against a state-of-the-art model[3] to evaluate competitive standing within the field. Finally, conclusions and recommendations for future work are presented.

[3] Jamie TOLAN et al. "Very high resolution canopy height maps from RGB imagery using self-supervised vision transformer and convolutional decoder trained on aerial lidar". In: *Remote Sensing of Environment* 300 (2024), p. 113888.

OBJECTIVES

GENERAL OBJECTIVE

To develop a Canopy Height Model (CHM) by applying deep learning algorithms and computer vision techniques, based on the fusion of LiDAR data and high-resolution RGB imagery, with the purpose of generating a high-resolution geospatial product that can be used as a support tool in forest resource monitoring and management activities.

SPECIFIC OBJECTIVES

1. To review and select a deep learning architecture suitable for the task of canopy height regression from RGB imagery.
2. To develop, train, and optimize a deep learning model to generate high-resolution CHM maps from RGB imagery and reference CHM data derived from LiDAR data in a defined area.
3. To evaluate the proposed model using a reference dataset in defined regions, determining the accuracy of the tree canopy height estimations.
4. To generate and visualize high-resolution Canopy Height Maps (CHM) using only optical imagery in zones within the study area, comparing them with the reference CHMs derived from LiDAR.

1. BACKGROUND AND FOUNDATIONAL CONCEPTS

This chapter establishes the foundational context required to understand the methodology and results of this thesis. It begins with an overview of Remote Sensing and the role of the Canopy Height Model (CHM) in forest resource management and ecological monitoring. It then describes the key technologies involved in data acquisition, including LiDAR systems and high-resolution RGB imagery. Finally, the chapter introduces the core concepts of Deep Learning and Computer Vision architectures, with a focus on regression tasks and the specific requirements for predicting continuous spatial variables, thereby laying the groundwork for the model architecture presented in the subsequent chapter.

1.1. THE REMOTE SENSING ECOSYSTEM AND THE CHM

Remote sensing is a key discipline in engineering, defined as the process of obtaining information about the properties of an object without direct physical contact with it. This process involves measuring the energy emitted or reflected by the object, which is captured by sensors located on aerial or satellite platforms. In essence, remote sensing constitutes a chain of data processing that transforms the captured energy into useful information [4]. In the context of forest monitoring, remote sensing has become established as a highly useful tool for the observation and analysis of ecosystems at local, regional, and global scales. It enables the detection of changes in vegetation cover and forest structure over large areas, reducing or eliminating the need for extensive manual ground measurements. Forests play a critical role in the carbon cycle, biodiversity, and

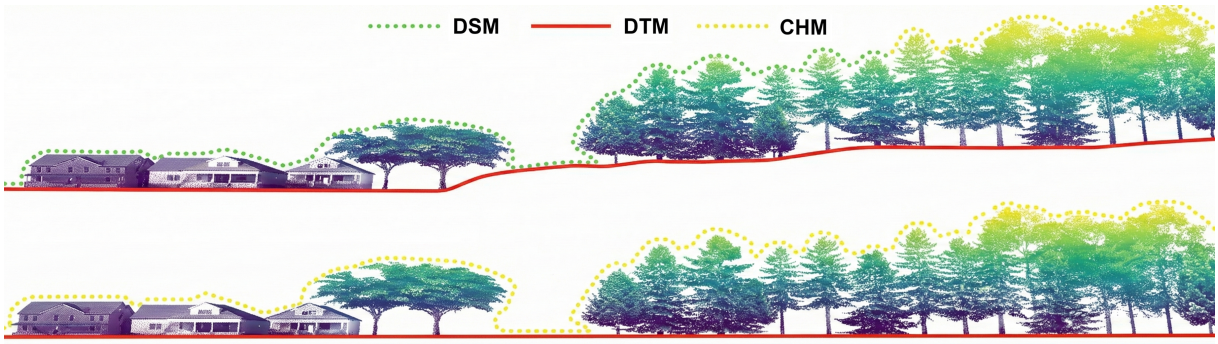
---

[4] Thomas LILLESAND, Ralph W KIEFER, and Jonathan CHIPMAN. *Remote sensing and image interpretation*. 7th. Hoboken, NJ: John Wiley & Sons, 2015.

climate regulation [5].

For the monitoring and management of these resources, it is crucial to obtain accurate information about the vertical structure of vegetation, which is a fundamental parameter for estimating biophysical indicators[6]. The Canopy Height Model (CHM) is a key geospatial product for characterizing this vertical structure, as it represents the absolute height of the vegetation canopy above the ground surface. Operationally, the CHM is derived by subtracting a Digital Terrain Model (DTM) — which depicts the elevation of the bare Earth surface from a Digital Surface Model (DSM), which captures the elevation of the uppermost surface, including tree canopies and other objects, as illustrated in Figure 1.1. This normalization process provides a direct and terrain-independent measure of vegetation height. The ability to generate high-resolution CHMs has become a priority in forest management and environmental policy-making, enabling more precise assessments of ecosystem health and supporting decisions related to conservation and sustainable use.

Figure 1.1. Conceptual derivation of the Canopy Height Model (CHM) from Digital Surface Model (DSM) and Digital Terrain Model (DTM).



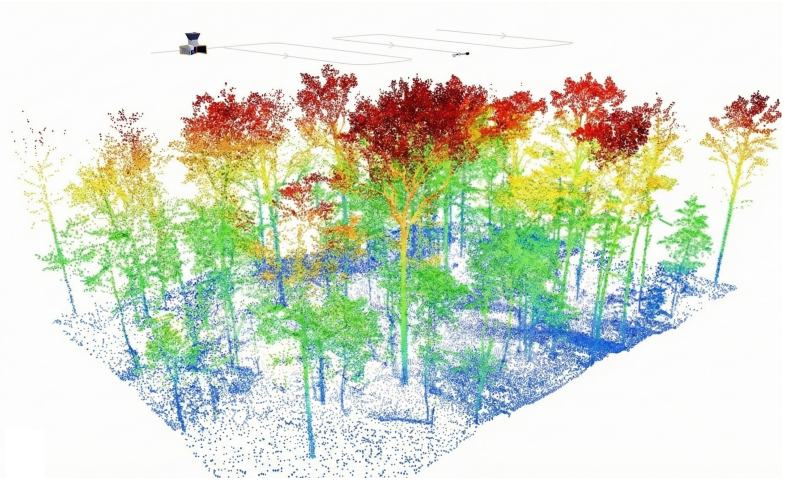
[5] Mirco MIGLIAVACCA et al. “The three major axes of terrestrial ecosystem function”. In: *Nature* 598.7881 (2021), pp. 468–472. DOI: 10.1038/s41586-021-03939-9.

[6] Andrew K SKIDMORE et al. “Priority list of biodiversity indicators to track progress towards the post-2020 global biodiversity framework”. In: *Environmental Research Letters* 16.9 (2021), p. 094049.

1.2. DATA ACQUISITION TECHNOLOGIES: LIDAR AND OPTICAL IMAGERY

The monitoring of forest vertical structure requires the use of technologies capable of capturing accurate three-dimensional data. Historically, LiDAR (Light Detection and Ranging) has been the reference technology for this task. This system operates by emitting laser pulses and measuring the time required for the signal to return to the sensor, thereby generating a three-dimensional point cloud (Figure 1.2). Because it's high density and good precision. Airborne LiDAR, in particular, is considered the ground truth for the generation of Canopy Height Models (CHMs)[7].

Figure 1.2. Example of an airborne LiDAR point cloud illustrating the three-dimensional vertical structure of forest vegetation.



However, despite its high accuracy, the acquisition of LiDAR data is costly and requires significant logistical efforts. These constraints limit both its spatial coverage and temporal resolution, restricting its applicability for frequent monitoring over large areas. To overcome these limitations, recent research has increasingly focused on the use of high-resolution optical imagery, such as RGB data, as a cost-effective alternative for

[7] Ralph O DUBAYAH and Jason B DRAKE. "Lidar remote sensing for forestry". In: *Journal of Forestry* 98.6 (2000), pp. 44–46.

continuous canopy height inference. Unlike LiDAR, optical imagery is widely available and easier to acquire, making it suitable for large-scale and frequent monitoring. Although optical images are inherently two-dimensional, the texture, color, and spatial context they contain provide valuable cues that advanced computer vision algorithms can exploit to infer the third dimension, namely canopy height, when appropriate reference data are available. Current methodological trends therefore rely on high-precision LiDAR data to establish ground truth and subsequently train image-based models to achieve broader spatial and temporal coverage.

1.3. DEEP LEARNING FOR PIXEL-WISE CANOPY HEIGHT ESTIMATION

To infer canopy height, a three-dimensional biophysical parameter, from two-dimensional data such as optical imagery, a modeling approach with a high capacity for complex feature extraction and large-scale data processing are required. Deep Learning (DL) provides such capability. Unlike traditional Machine Learning approaches, Deep Neural Networks (DNNs) enable the direct learning of data representations, which is essential for capturing the spatial and textural relationships between the visual appearance of tree canopies in imagery and their actual height.

Within the context of computer vision and geospatial data processing, it is essential to distinguish between different modeling tasks. While classification aims to assign a discrete label to each pixel (e.g., forest or bare ground), canopy height estimation constitutes a pixel-wise regression problem. In this case, the model must predict a continuous and spatially explicit value, namely canopy height in meters, for each pixel in the input image. This requirement to estimate continuous spatial variables directly influences the choice of the model architecture employed.

Figure 1.3. Conceptual representation of a deep learning model for pixel-wise canopy height estimation, where high-resolution RGB imagery is used as input to predict a continuous Canopy Height Model (CHM) as output.

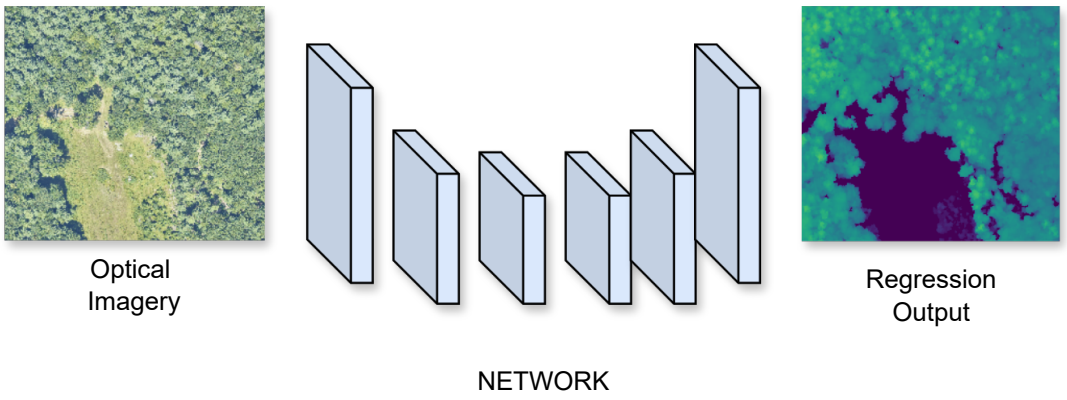


Figure 1.3 illustrates the general principle of pixel-wise canopy height estimation using deep learning, where high-resolution optical imagery is transformed into a continuous spatial prediction through a neural network.

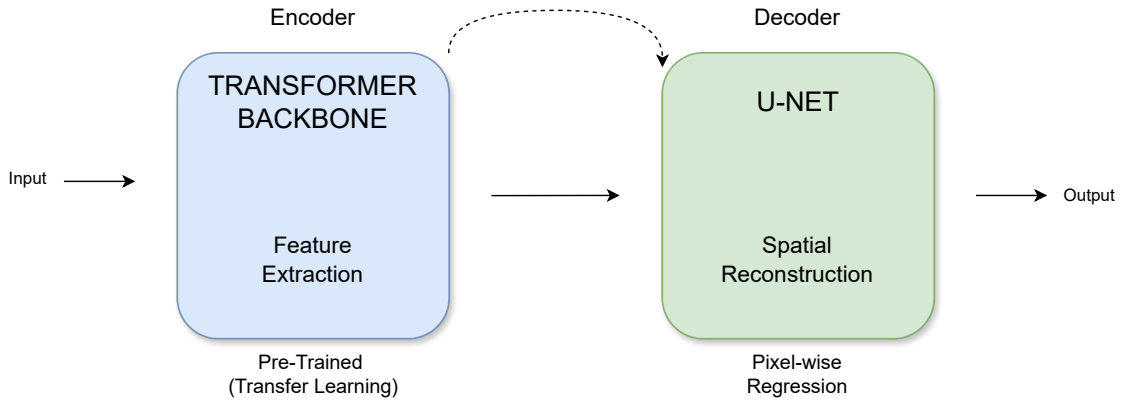
The neural network design is driven by the requirement to generate an output that preserves the high spatial resolution of the input imagery, namely a 1 m canopy height model. To meet this requirement, computer vision architectures commonly adopt an encoder–decoder structure, where a backbone is responsible for feature extraction and a decoder reconstructs the output at the original spatial resolution.

Within this paradigm, U-Net–based architectures, originally proposed for semantic segmentation [8], are particularly well suited for dense prediction tasks due to their ability to preserve fine spatial details through skip connections. In this work, an adapted U-Net was used as the decoder, reconstructing the CHM prediction from the features

[8] Olaf RONNEBERGER, Philipp FISCHER, and Thomas BROX. "U-net: Convolutional networks for biomedical image segmentation". In: *International Conference on Medical image computing and computer-assisted intervention* (2015), pp. 234–241.

extracted by the backbone. Complementarily, transformer-based models are employed as the backbone due to their ability to capture global contextual information through self-attention mechanisms[9], allowing the model to better represent long-range spatial and textural relationships that are critical for accurate canopy height estimation.

Figure 1.4. Conceptual architecture of a deep learning framework for pixel-wise canopy height estimation, consisting of a pre-trained Transformer backbone for feature extraction and a U-Net decoder for spatial reconstruction



Transfer learning has become a standard practice in computer vision, particularly when training deep models on limited domain-specific data[10]. The approach consists of leveraging models pre-trained on large-scale datasets to provide robust feature representations, which can then be fine-tuned for a specific task, making it especially useful for tasks such as canopy height estimation.

These principles provide the conceptual foundation for the methodology described in the next chapter, where the proposed framework for canopy height estimation is presented in detail.

---

[9] Ashish VASWANI et al. “Attention is all you need”. In: *Advances in neural information processing systems* 30 (2017).

[10] Lei MA et al. “Deep learning in remote sensing: A comprehensive review and list of resources”. In: *IEEE Geoscience and Remote Sensing Magazine* 7.2 (2019), pp. 67–105.

2. METHODOLOGY AND MODEL IMPLEMENTATION

This chapter presents a detailed description of the methodological workflow and the experimental design developed to achieve high-resolution canopy height estimation. It begins with an analysis of the study area and data sources, establishing the specifications for the LiDAR and RGB imagery used as the foundation for the model. The chapter then outlines the data pre-processing stage, which is critical for ensuring spatial alignment and data quality. Subsequently, the proposed deep learning architecture is described, focusing on the integration of the Transformer-based backbone and the adapted U-Net decoder. Finally, the experimental setup is presented, including the loss functions, optimization strategies, and training parameters necessary to ensure the model correctly learns the complex spatial relationships required to generate accurate Canopy Height Models (CHM).

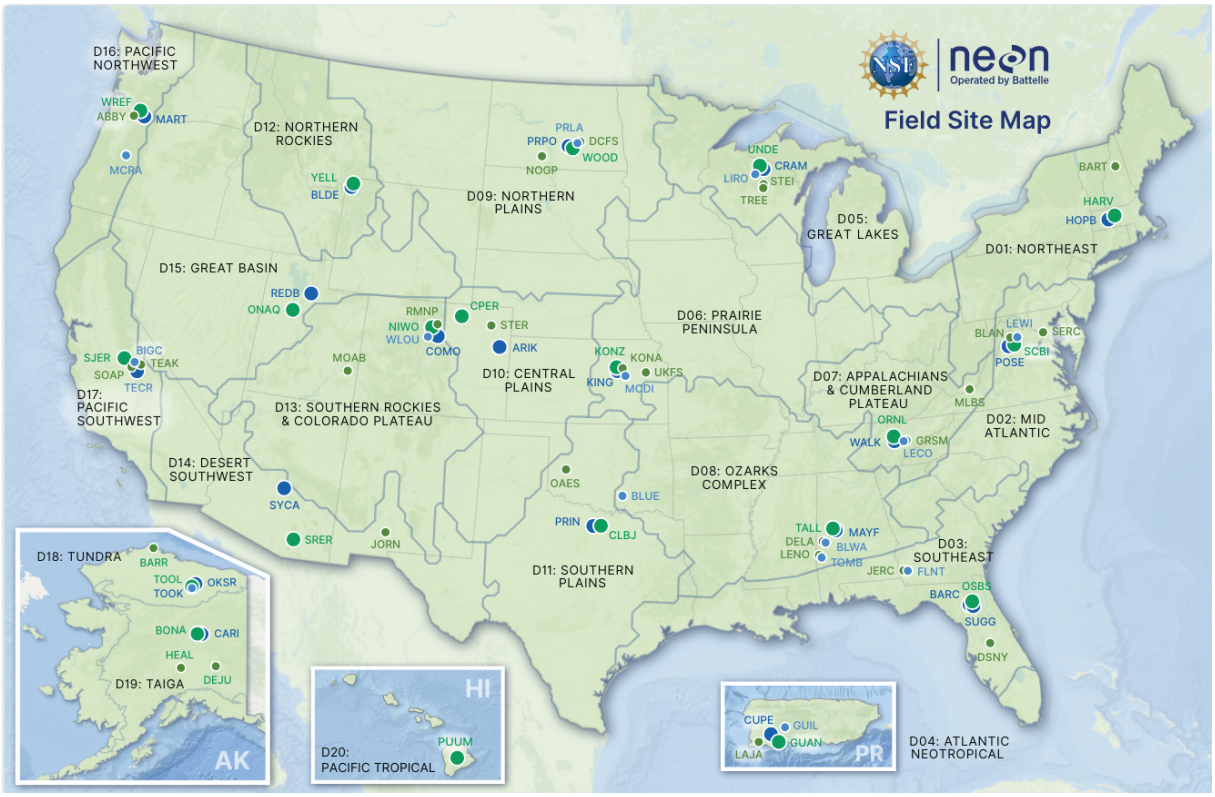
2.1. STUDY AREA AND DATA SOURCES

This study utilizes publicly available data from the National Ecological Observatory Network (NEON) Airborne Observation Platform (AOP)[11], incorporating a selection of forested sites across the continental United States (figure 2.1). These sites were chosen to provide a diverse range of forest structures and canopy height distributions, aiming to develop a model with sufficient generalization capability for canopy height estimation from optical imagery. By including various ecological domains, the dataset captures a wide range of canopy structural complexities, vegetation densities, and varying phenological conditions. This diversity ensures that the model is exposed to different spectral and textural patterns, promoting a more robust learning process that is

[11] Thomas KAMPE et al. “NEON: the first continental-scale ecological observatory with airborne remote sensing of vegetation canopy biochemistry and structure”. In: *Proc. SPIE* (2009).

less dependent on specific local characteristics.

Figure 2.1. Distribution of the National Ecological Observatory Network (NEON) field sites. Source: Official NEON Data Portal.



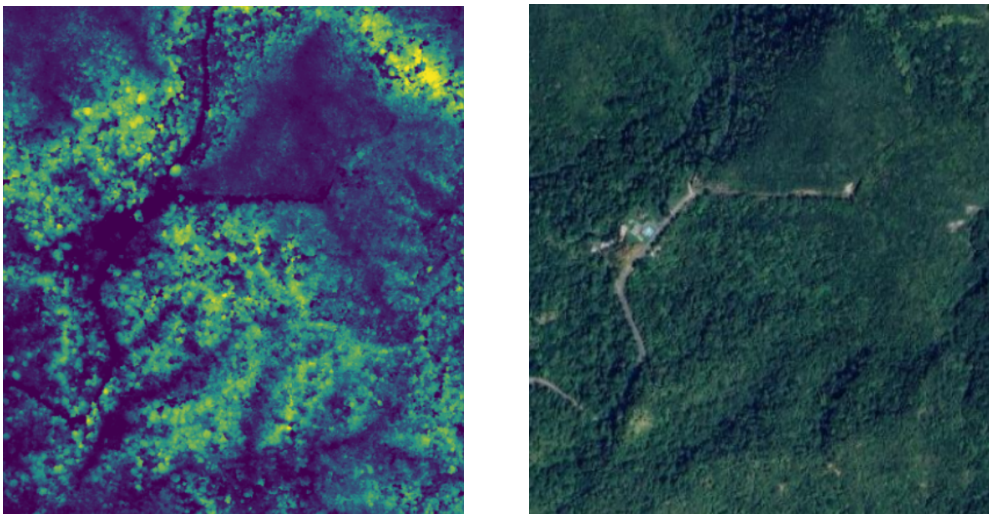
2.1.1. Reference Canopy Height Data. The reference data for this study consist of the NEON Ecosystem Structure data product (DP3.30015.001: Ecosystem Structure)[12], which provides pre-computed CHMs at a 1 m spatial resolution in 1 km×1 km GeoTIFF tiles. These CHMs are derived from discrete-return airborne LiDAR point clouds using NEON's standardized processing pipeline, which includes ground classification, height normalization, and the subtraction of the Digital Terrain Model (DTM)

[12] NATIONAL ECOLOGICAL OBSERVATORY NETWORK (NEON). *Ecosystem structure (DP3.30015.001)*. 2025. DOI: 10.48443/JQQD-1N30.

from the Digital Surface Model (DSM).

2.1.2. RGB Aerial Imagery. The optical inputs are high-resolution RGB orthomosaics from the NEON data product (DP3.30010.001: High-resolution orthorectified camera imagery)[13]. These images are captured simultaneously with the LiDAR flights using a digital frame camera, ensuring temporal consistency between the optical features and the structural ground truth. To optimize canopy visibility and spectral response, these acquisitions are performed during peak vegetation periods. Although originally captured at a higher spatial resolution (0.1 m), the imagery is resampled to 1 m to ensure strict pixel-wise correspondence with the LiDAR-derived CHM during the training and inference stages. The spatial correlation between both products is exemplified in Figure 2.2.

Figure 2.2. Comparison between (a) the reference Canopy Height Model and (b) the corresponding RGB aerial imagery for a selected NEON forest site.



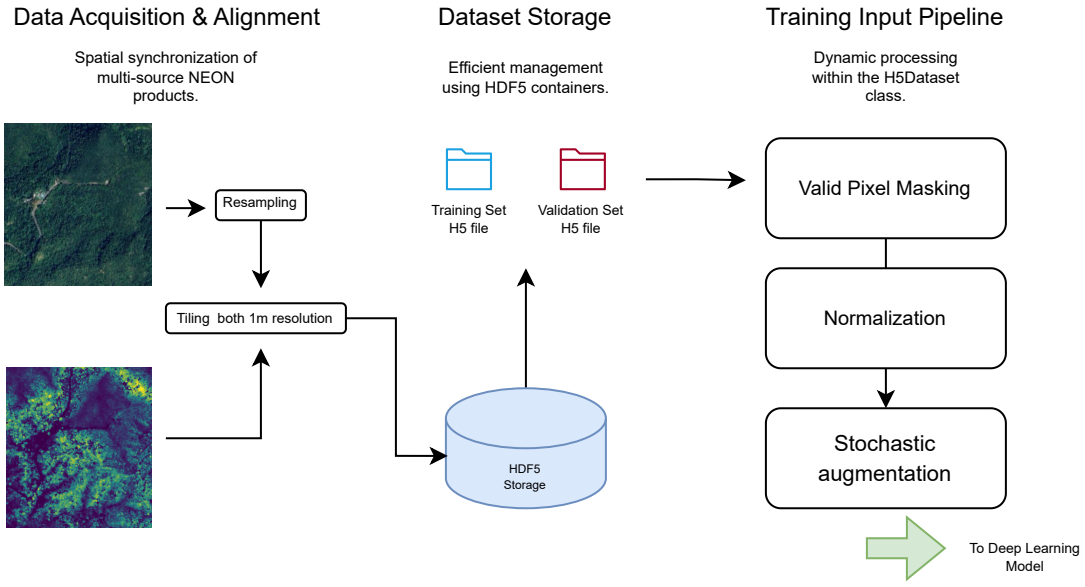
(a) LiDAR-derived CHM (b) High-resolution RGB

[13] NATIONAL ECOLOGICAL OBSERVATORY NETWORK (NEON). *High-resolution orthorectified camera imagery (DP1.30010.001)*. 2025. DOI: 10.48443/F8DG-8W18.

2.2. DATA PREPROCESSING

The raw NEON data products, although high-quality and spatially co-registered, require pre-processing steps to generate a dataset suitable for supervised deep learning. These steps ensure spatial consistency, remove unreliable or noisy areas, normalize input values, and create training samples that maximize the model's ability to learn robust mappings between RGB imagery and canopy height. Therefore, a pre-processing pipeline was implemented to improve data quality and computational efficiency. All processing was carried out using open-source Python libraries, primarily GDAL, Rasterio, NumPy, HDF5, and PyTorch.

Figure 2.3. Methodological workflow of the data preprocessing pipeline. The diagram illustrates the transition from raw NEON tiles to synchronized HDF5 datasets, including all the stages



2.2.1. Tiling, resampling, and alignment. The original 1km×1km consumes a considerable amount of memory. Consequently, each tile was subdivided into non-overlapping patches of 256×256 pixels (256 m × 256 m on the ground). Since RGB orthophotos are provided at 0.1 m resolution, they were resampled to 1 m using bilinear interpolation to match the Canopy Height data resolution, ensuring exact spatial correspondence.

2.2.2. Dataset storage and HDF5 integration. Processed patches were stored in HDF5 (.h5) files to enable fast random access and efficient input/output operations during training. This storage strategy also facilitates experimental flexibility, as additional processing steps (e.g., masking, normalization, or augmentation) can be applied dynamically at load time without regenerating the dataset.

The dataset was partitioned into two independent subsets to ensure unbiased evaluation. One HDF5 file was created for the training data, while a separate file was generated for validation data, allowing a clear separation between model training and performance assessment.

Table 2.1. Composition and spatial coverage of the generated HDF5 datasets.

Dataset	**Number of Patches**	**Percentage**	**Approx. Coverage (km^2)**
Training	20,629	$\approx$ 90%	1,351.9
Validation	2,579	$\approx$ 10%	169.0
Total	23,208	100%	1,520.9

2.2.3. Dynamic data processing and masking strategy. To maximize training efficiency and ensure numerical stability, several critical cleaning and transformation operations were implemented within the custom class created to manage the data (H5Dataset). These operations are applied dynamically during each training iteration, ensuring that the model receives optimized tensors without the need for large pre-

processed storage.

- **Value Clipping and Noise Filtering:** The LiDAR-derived CHM data was analyzed to identify and mitigate sensor noise. Values below 0m were treated as invalid (NaN), and a maximum height threshold of 60 m was established to cap unrealistic outliers that do not correspond to the typical forest structure of the selected sites. This prevents extreme values from disproportionately affecting the gradient during backpropagation.

- **Dynamic Valid Pixel Masking:** A fundamental challenge in pixel-wise regression is the presence of "NoData" areas. As observed in the training pipeline, a binary validity mask is computed as the logical *AND* of: (i) valid CHM pixels (non-NaN) and (ii) non-black RGB pixels (where all spectral channels > 0). This mask is utilized to ensure that the loss function only evaluates pixels with reliable ground truth, effectively ignoring sensor gaps or edge artifacts in the flight lines.

- **Statistical Normalization:** Input and target variables were scaled to facilitate faster convergence and numerical stability. The RGB channels were standardized to a standard normal distribution ($\mu = 0, \sigma = 1$) using ImageNet[14] statistics: $\mu = [0.485, 0.456, 0.406]$ and $\sigma = [0.229, 0.224, 0.225]$. This specific normalization is critical as the model utilizes a pre-trained MiT-B5[15] backbone via transfer learning; by aligning the input distribution with the data the encoder was originally trained on, we ensure that the pre-learned spatial features (textures and edges) are correctly extracted from the aerial imagery. Simultaneously, the CHM target was normalized to a bounded $[0, 1]$ range

---

[14] Jia DENG et al. “ImageNet: A Large-Scale Hierarchical Image Database”. In: *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. 2009, pp. 248–255. DOI: 10.1109/CVPR.2009.5206848.

[15] Enze XIE et al. “SegFormer: Simple and Efficient Design for Semantic Segmentation with Transformers”. In: *arXiv preprint arXiv:2105.15203* (2021).

using a min-max scaling approach based on a 60 m height threshold. This transformation maps the physical height values to the activation range of the model's output layer, ensuring consistent gradient flow during the training process.

2.2.4. Data augmentation strategy. To enhance the model's generalization and mitigate overfitting, a stochastic data augmentation pipeline was implemented for the training set by applying geometric transformations that simulate variations in sensor perspective and flight line orientations. Crucially, these transformations are applied synchronously to the RGB input, the CHM target, and the validity mask, ensuring that the pixel-wise spatial correspondence and structural ground truth remain perfectly aligned even after transformation as shown in figure 2.4.
During each training iteration, every patch has a 50% probability of undergoing the following operations:

- **Horizontal and Vertical Flips:** The tensors are reflected along their axes to account for different orientations of forest structures and topography.
- **Discrete Rotations:** Random rotations of 90°, 180°, or 270° are applied to ensure the model learns rotation-invariant textural features of the canopy.

The complete set of specifications for the data pre-processing pipeline, including patch dimensions, normalization strategies, and augmentation parameters, is summarized in the table 2.2.

Figure 2.4. Visual representation of the stochastic data augmentation process. The top row displays the input RGB patches, while the bottom row shows the corresponding LiDAR-derived CHM targets.

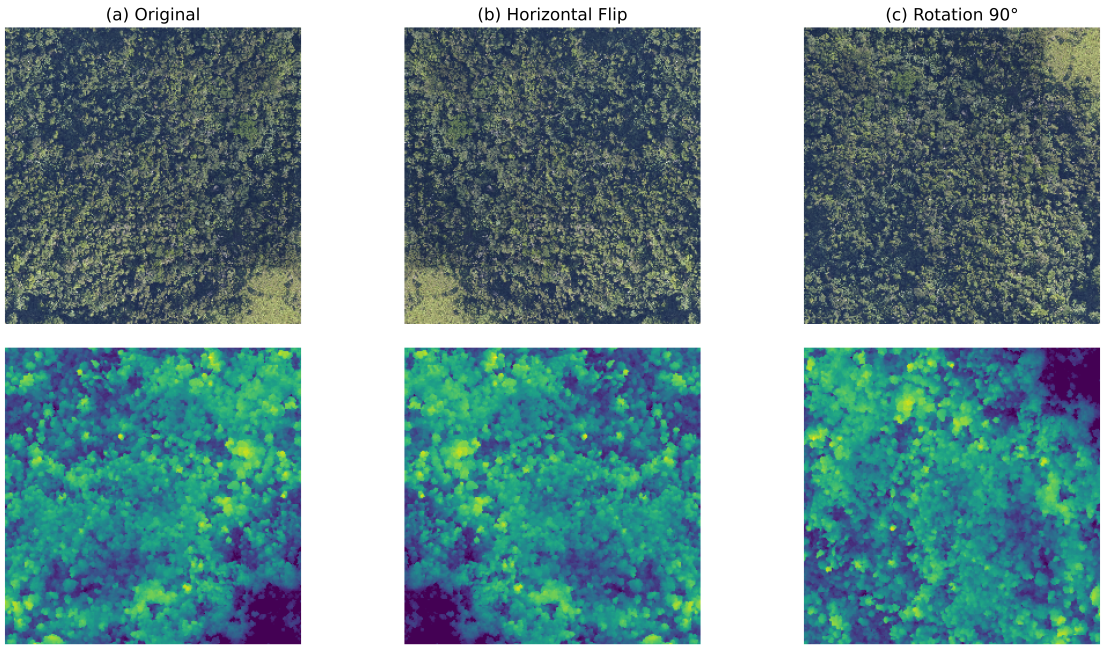


Table 2.2. Summary of pre-processing parameters and dataset specifications.

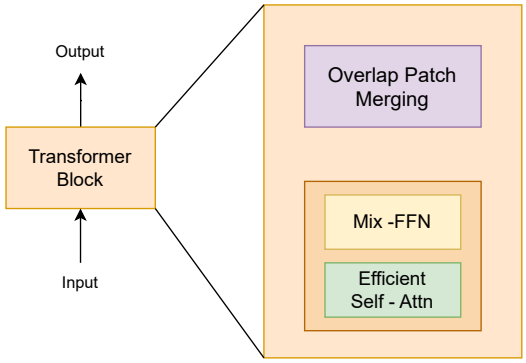
Parameter	**Value**
Patch Size	256×256 pixels
Spatial Resolution	1 m
CHM Height Threshold (H_{max})	60 m
RGB Normalization Strategy	ImageNet parameters
CHM Scaling Range	$[0, 1]$
Augmentation Probability	0.5
Resampling Method	Bilinear

2.3. PROPOSED DEEP LEARNING ARCHITECTURE

The proposed model for pixel-wise canopy height estimation is based on a U-Net architecture adapted for regression tasks, implemented using the segmentation models

pytorch (SMP) library [16]. This library provides modular and efficient implementations of state-of-the-art semantic segmentation models, which are inherently suitable for dense prediction problems due to their encoder-decoder structure with skip connections. Specifically, a U-Net with a Mix Transformer (MiT-B5) encoder from the SegFormer family was selected. The MiT-B5 encoder is a hierarchical transformer consisting of four stages that progressively downsample the spatial resolution (to 1/4, 1/8, 1/16, and 1/32 of the input) while increasing channel depth (64, 128, 320, and 512 channels, respectively) and incorporating efficient self-attention mechanisms. Each stage includes overlapping patch merging for spatial reduction and a series of Mix-FFN blocks with efficient self-attention layers (Figure 2.5). This design enables the capture of both local details and long-range contextual relationships, which are essential for interpreting complex canopy textures, shadows, and structural patterns visible in high-resolution RGB aerial imagery.

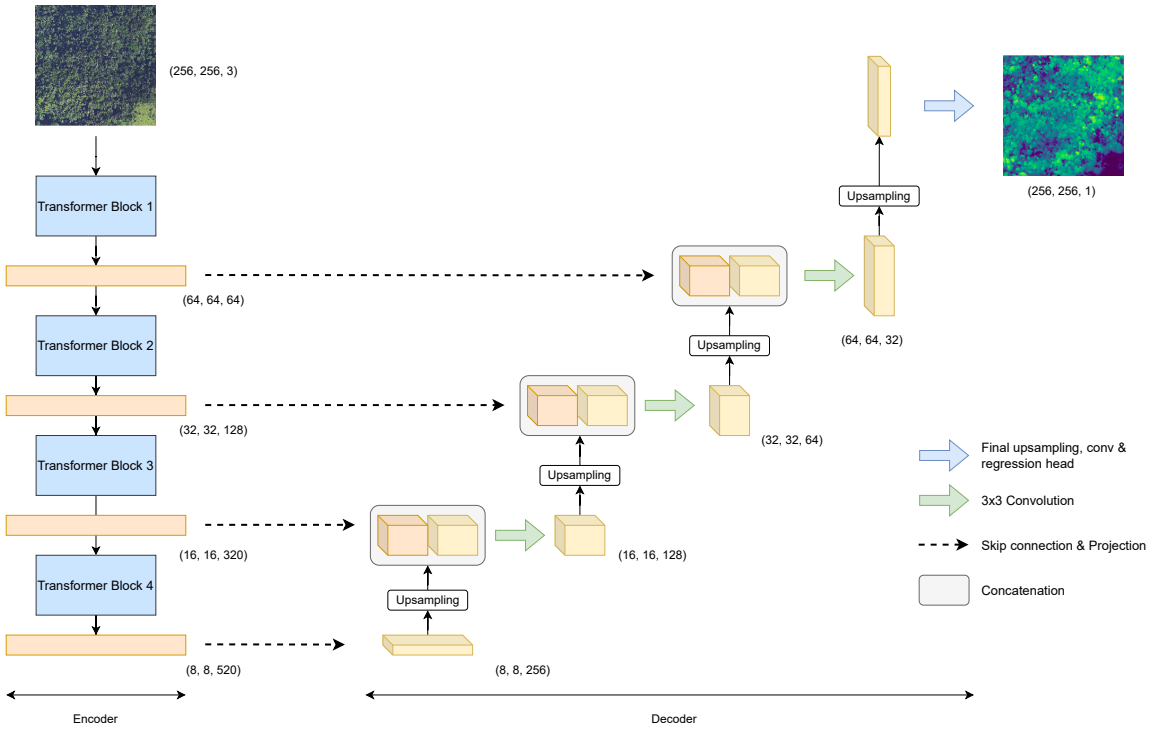
Figure 2.5. The internal architecture consists of an Efficient Self-Attention layer and a Mix-FFN module for global and local feature processing. The Overlapping Patch Merging stage ensures spatial consistency while preparing the feature maps for the subsequent hierarchical level.



[16] Pavel IAKUBOVSKII. *Segmentation Models Pytorch*. `https://github.com/qubvel/segmentation_models.pytorch`. 2019.

The decoder follows the classic U-Net design, upsampling encoded features and concatenating them with high-resolution features from the encoder via skip connections. These connections include a 1×1 convolution projection to align encoder channel dimensions with the fixed decoder channels—256, 128, 64, and 32, respectively. This mechanism preserves fine spatial details critical for accurate delineation of tree crowns and canopy edges, as shown in the general schematic in Figure 2.6.

Figure 2.6. Overview of the Architecture. The model integrates a hierarchical MiT-B5 transformer encoder for multi-scale feature extraction with a symmetric U-Net decoder that recovers spatial resolution through skip connections.



The segmentation head was modified to output a single channel with a linear activation, producing normalized height predictions in the range [0, 1]. These values are subsequently denormalized to meters during inference and evaluation using the fixed scaling factor (maximum height of 60 m). Transfer learning was leveraged by

initializing the MiT-B5 encoder with weights pre-trained on ImageNet-1k, providing robust low-level feature representations. This hybrid convolutional-transformer architecture balances expressive power with computational efficiency, making it well-suited for the challenging task of regressing continuous three-dimensional canopy structure from two-dimensional optical data across diverse forest ecosystems.

2.3.1. Justification. The choice of a hybrid Transformer-U-Net architecture is driven by the necessity to bridge the gap between local spatial precision and global contextual understanding. Recent benchmarks in forest structural modeling have demonstrated that hierarchical transformers significantly outperform traditional CNN-based architectures and standard isotropic transformers in canopy height estimation[3] [17].

- Overcoming the Receptive Field Limitation: Standard CNNs rely on local kernels that often fail to capture the long-range dependencies required to interpret large-scale forest patterns and crown shadows. The self-attention mechanism in the MiT-B5 encoder provides a global receptive field from the earliest stages, a feature identified as critical for handling the high variance in forest density.

- Multiscale Feature Hierarchies: Unlike flat transformers (like ViT), the hierarchical nature of the MiT-B5 allows the model to process features at multiple resolutions (from H/4 to H/32). This aligns with the findings that multi-scale representations are essential for delineating individual tree crowns while simultaneously accounting for regional topographic variations.

- Preservation of Spatial Detail through U-Net: While transformers excel at encoding semantic information, the U-Net decoder ensures that the regression output maintains the pixel-level fidelity of the input RGB imagery. The use of skip connections

[17] Fajwel FOGEL et al. “Open-Canopy: A Country-Scale Benchmark for Canopy Height Estimation at Very High Resolution”. In: (2024). eprint: 2407.09392.

effectively reintroduces high-frequency spatial information that is typically lost in pure transformer-based segmentation models (like the original SegFormer All-MLP head), making it more suitable for precise height regression.

2.4. TRAINING AND EXPERIMENTAL SETUP

This section describes the training and evaluation protocols used to train the proposed U-Net model with MiT-B5 encoder. The implementation was carried out in PyTorch, leveraging mixed precision training for efficiency and a custom training loop to handle masked loss computation on valid pixels only.

2.4.1. Comparative analysis of loss functions. While the architecture defines the model's capacity to learn, the selection of an appropriate objective function is critical for translating visual features into accurate vertical measurements. This section involves a systematic evaluation of five different loss functions, aiming to identify the optimal balance between outlier robustness, spatial smoothness, and edge sharpness.

- **Mean Absolute Error ($L1$):** This objective function is defined by the absolute difference between the predicted and ground-truth values:

$$\ell_{L1} = |\hat{h}_i - h_i|$$

  Unlike quadratic losses, $L1$ provides a constant gradient for all error magnitudes, making it inherently more robust to outliers and potential noise in the LiDAR reference data.

- **Mean Squared Error (MSE):** The loss term is defined by the squared difference:

$$\ell_{MSE} = (\hat{h}_i - h_i)^2$$

Due to its quadratic nature, this term heavily penalizes larger residuals, forcing the model to prioritize the reconstruction of emergent tree structures.

- **Huber Loss:** This term acts as a hybrid between $L1$ and MSE based on a threshold δ:

$$\ell_{Huber} = \begin{cases} \frac{1}{2}(\hat{h}_i - h_i)^2 & \text{if } |\hat{h}_i - h_i| \leq \delta \\ \delta(|\hat{h}_i - h_i| - \frac{1}{2}\delta) & \text{otherwise} \end{cases}$$

  It provides the stability of MSE for small errors and the robustness of L1 for large deviations.

- **Log-Cosh Loss:** A smoother approximation of the $L1$ loss:

$$\ell_{LC} = \log(\cosh(\hat{h}_i - h_i))$$

  This function is twice-differentiable everywhere, which can lead to more stable convergence compared to the non-smooth $L1$ derivative[18].

- **Gradient Difference Loss (GDL):** Unlike pixel-wise losses, the GDL term $\mathcal{L}_{reg}$ in the general equation accounts for the spatial relationship between neighboring pixels (i, j):

$$\ell_{GDL} = \sum_{i,j} \left| |h_{i,j} - h_{i-1,j}| - |\hat{h}_{i,j} - \hat{h}_{i-1,j}| \right| + \left| |h_{i,j-1} - h_{i,j}| - |\hat{h}_{i,j-1} - \hat{h}_{i,j}| \right|$$

  This term is combined with $L1$ to penalize the loss of sharpness at crown boundaries and high-gradient transitions[19].

---

[18] J. CHEN et al. "Log-Cosh Loss Function for Deep Learning Regression". In: *arXiv preprint arXiv:2210.03210* (2022).

[19] Michael MATHIEU, Camille COUPRIE, and Yann LECUN. "Deep multi-scale video prediction beyond mean square error". In: *arXiv preprint arXiv:1511.05440* (2015).

To ensure the model only learns from reliable data, all objective functions are computed exclusively over the valid pixels defined by the mask. The general masked loss ($\mathcal{L}$) is formulated as:

$$\mathcal{L} = \frac{1}{N_v} \sum_{i \in \mathcal{V}} \ell(\hat{h}_i, h_i) + \lambda \mathcal{L}_{reg} \tag{1}$$

where $\mathcal{V}$ represents the set of valid pixels, N_v denotes the total number of valid pixels, and ℓ is the specific pixel-wise loss criterion ($L1$, $LogCosh$, $Huber$, $L1+GDL$ or MSE), and $\lambda \mathcal{L}_{reg}$ represents an optional structural regularization term, such as GDL, weighted by the hyperparameter λ.

The quantitative performance and the impact of each objective function on the predicted canopy structure are detailed and discussed in the next chapter.

2.4.2. Optimization strategy: scheduling and early stopping. The models were optimized using the Adam optimizer[20] with an initial learning rate of 5×10^{-4}. To ensure stable convergence and prevent the training process from stalling at local minima, we implemented an adaptive learning rate decay strategy based on validation performance. This mechanism was configured based on empirical performance during the tuning phase to monitor the loss function, reducing the learning rate by a factor of 0.5 after 5 consecutive epochs without significant improvement.

Furthermore, an early stopping mechanism was integrated to safeguard against overfitting and guarantee optimal generalization. Based on the observation of the validation loss dynamics, a patience of 15 epochs was established, ensuring that the training process terminated once the model reached its maximum representative capacity without compromising its performance on unseen data.

---

[20] Diederik P. KINGMA and Jimmy BA. “Adam: A Method for Stochastic Optimization”. In: *International Conference on Learning Representations (ICLR)*. 2015.

This configuration allowed the model sufficient time to converge and help that the training terminated once the loss function performance reached a stable asymptote. These parameters provided a good convergence across all five loss function experiments, the results are detailed in the results section.

TRAINING HYPERPARAMETERS. The key hyperparameters used for the final training are summarized in table 2.3.

Table 2.3. Training hyperparameters.

Hyperparameter	Value
Batch size	32
Number of epochs	Up to 150
Optimizer	Adam
Initial learning rate	5×10^{-4}
Scheduler	ReduceLROnPlateau (factor=0.5, patience=5)
Early stopping patience	15 epochs
Precision	Mixed (FP16)

Evaluation Metrics. Model performance was evaluated using three standard regression metrics, computed only on valid pixels and reported in meters after denormalization:

- **Mean Absolute Error (MAE)**: $\frac{1}{N}\sum_{i=1}^{N}|\hat{h}_i - h_i|$, which provides an intuitive average of the absolute residuals.
- **Root Mean Square Error (RMSE)**: $\sqrt{\frac{1}{N}\sum_{i=1}^{N}(\hat{h}_i - h_i)^2}$, which penalizes larger deviations more heavily.
- **Coefficient of Determination (R^2)**: $1 - \frac{\sum(\hat{h}_i - h_i)^2}{\sum(h_i - \bar{h})^2}$, used to quantify the proportion of the height variance explained by the model.

These metrics provide a comprehensive evaluation of the model, quantifying the error magnitude in meters (MAE, RMSE) and the overall predictive fit relative to the LiDAR ground truth (R2), in accordance with standard spatial regression protocols [21]. This experimental configuration ensures robust convergence and enables the learning of accurate mappings from RGB imagery to canopy height across diverse forest sites, the results of which are analyzed in the subsequent chapter.

[21] Christopher M. BISHOP. *Pattern Recognition and Machine Learning*. New York: Springer-Verlag, 2006.

3. RESULTS AND DISCUSSION

This chapter presents the experimental results obtained from the proposed Hybrid Transformer-U-Net architecture for canopy height estimation. The evaluation begins with an analysis of the training dynamics, focusing on the convergence of the loss function and the effectiveness of the optimization strategies described in the previous methodology. Subsequently, the model's performance is quantified using three standard regression metrics: Mean Absolute Error (MAE), Root Mean Square Error (RMSE), and the Coefficient of Determination (R^2), to establish its precision and explanatory power across diverse forest structures. Finally, a qualitative and quantitative assessment is conducted on both internal validation sets and unseen geographic sites, providing insights into the model's ability to recover complex spatial features, its capacity for spatial generalization, and its inherent limitations.

3.1. IMPACT OF LOSS FUNCTIONS ON MODEL PERFORMANCE

To determine the most effective optimization strategy for canopy height regression, a systematic comparison was conducted among the five objective functions (MSE, $L1$, $LogCosh$, $Huber$, $L1 + GDL$). This analysis evaluates which formulation provides the best balance between pixel-wise accuracy and the preservation of the forest's structural integrity.

Preliminary quantitative results, summarized in Table 3.1, reveal that Mean Squared Error (MSE) achieved the highest overall performance, particularly in terms of R^2 and RMSE. This suggests that the quadratic penalization of large residuals is highly effective for capturing the vertical variation of the canopy. However, a deeper examination through scatter plots and visual patches indicates that while MSE dominates numerically, composite functions like L1+GDL offer advantages in preserving the sharpness

of tree crown boundaries.

3.1.1. Quantitative results. Table 3.1 summarizes the performance of the MiT-B5 U-Net architecture evaluated on the independent validation set. Metrics were calculated after denormalizing the predictions to meters to ensure a direct comparison with the LiDAR ground truth.

Table 3.1. Quantitative comparison of objective functions for canopy height estimation.

Objective Function	**MAE (m)** $\downarrow$	**RMSE (m)** $\downarrow$	$R^2 \uparrow$
L1	1.3025	2.5976	0.9086
Huber	1.3711	2.5209	0.9139
Log-Cosh	1.3663	2.5072	0.9148
L1 + Gradient (GDL)	1.3942	2.5514	0.9118
MSE	1.1728	2.2189	0.9333

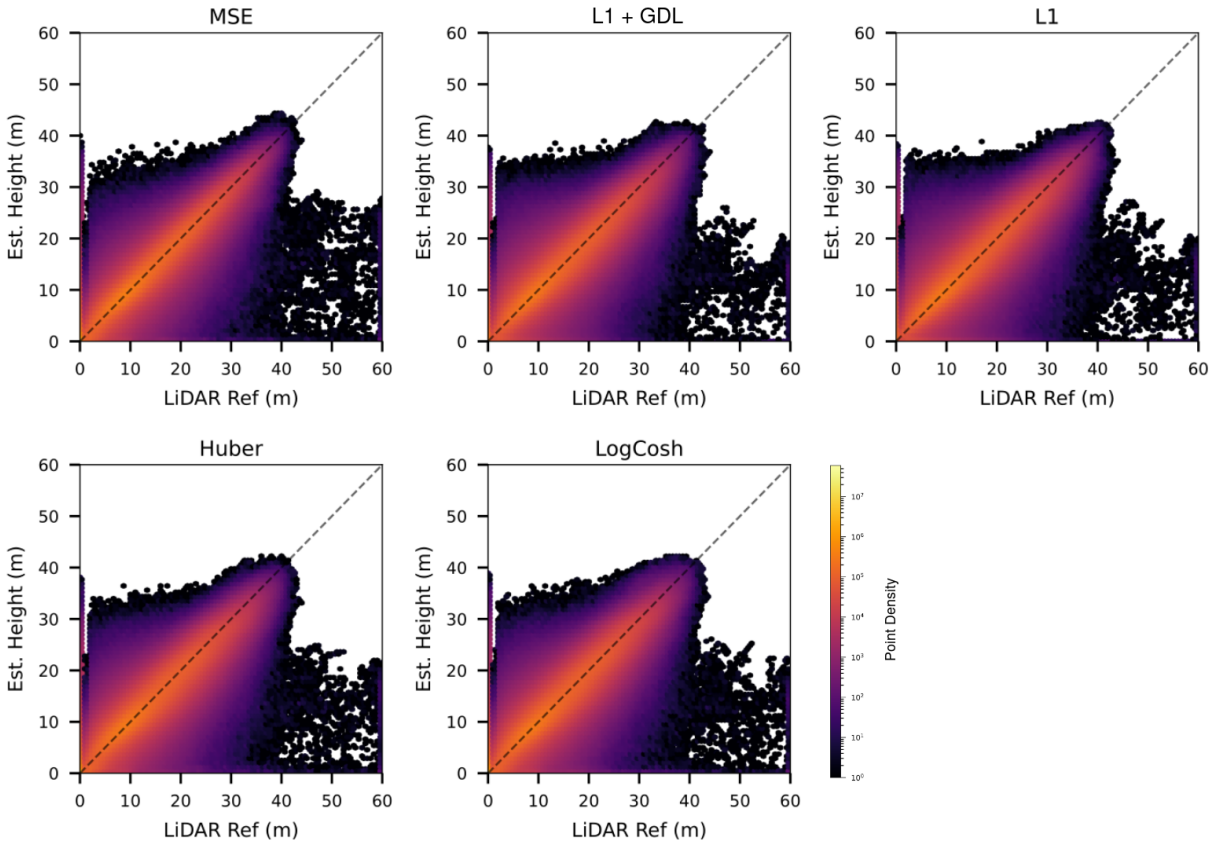
The quantitative results demonstrate that the Mean Squared Error (MSE) outperformed all other objective functions across every evaluated metric. This superior performance suggests that penalizing larger residuals quadratically is particularly effective for canopy height regression, as it forces the model to better capture the vertical extremes of the forest structure.

To further investigate these numerical differences, Figure 3.1 presents the density scatter plots for each experiment. These plots reveal clear differences in how each loss function handles height distribution; specifically, the MSE model shows the most compact alignment with the 1:1 diagonal, indicating that its predictions are more consistent and less dispersed than those of the other functions.

However, a common issue across all models is the underestimation of trees taller than 45 meters, where data points consistently fall below the reference line. This behavior is attributed to the limited number of high-altitude samples, as emergent trees are ecologically rare, making it difficult for the models to characterize these extremes. Additionally, some values reaching the 60-meter threshold in the LiDAR reference may rep-

resent signal noise or artifacts rather than actual vegetation structure. At lower height intervals, the observed vertical artifacts near the origin are influenced by NEON's sensor processing policy, which, depending on the sensor generation used, sets returns below 0.7 m or 2 m to zero to eliminate ground-level noise, this thresholding explains the sparse zones and vertical densities observed near the zero-height axis.

Figure 3.1. Density scatter plots of predicted versus LiDAR-derived canopy height for the five objective functions. The color intensity (log-scale) represents point density, and the dashed red line indicates the 1:1 ideal correlation.



3.1.2. Qualitative visual comparison. Beyond numerical metrics, a qualitative evaluation allows direct comparison of loss functions in reconstructing forest canopy structure from RGB imagery. Figure 3.2 shows predictions for two representative val-

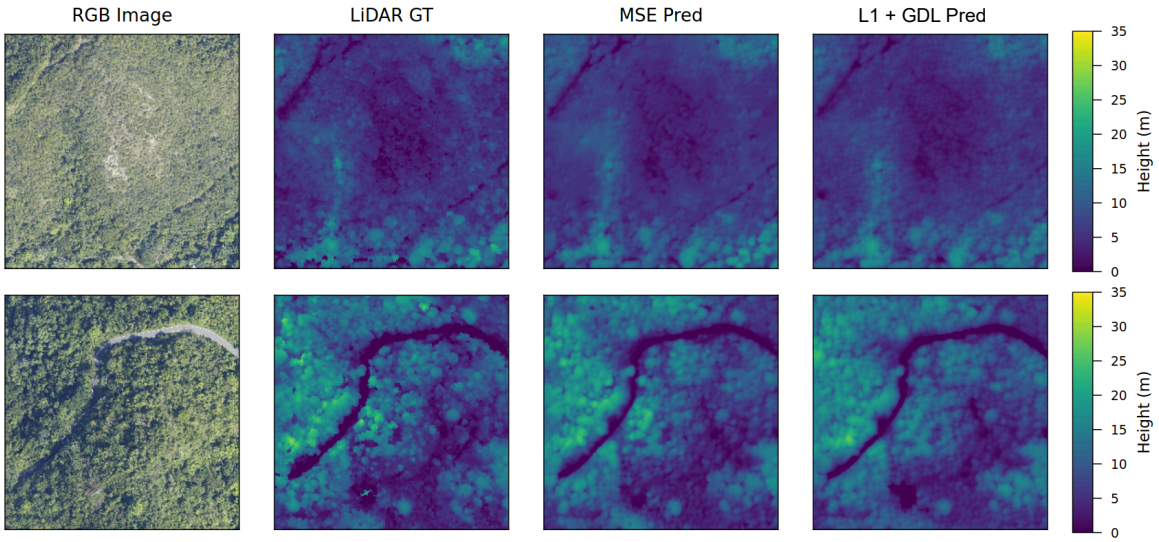
idation patches. The comparison focuses on the baseline MSE and the specialized L1 + Gradient Difference Loss (GDL) variant to assess the effect of explicit gradient penalization.

Visual results align with Table 3.1, where MSE showed the strongest performance in capturing canopy volumetric complexity, producing spatially consistent predictions of dominant crown heights and canopy gaps.

The L1 + GDL variant yields comparable visual quality, with slight enhancements in edge definition in some areas, but marginally noisier height distributions. Given its slightly lower numerical performance and the potential for training instability when increasing the GDL weight, it offers limited additional benefits in this canopy height regression setting.

Consequently, the MSE-based model is selected as the preferred configuration for the final inference pipeline, providing the best balance of pixel-wise accuracy, structural fidelity, and training stability in this context.

Figure 3.2. Visual assessment of canopy height estimation for different forest structures. The rows display the RGB input, LiDAR ground truth, and predictions for the evaluated loss functions.

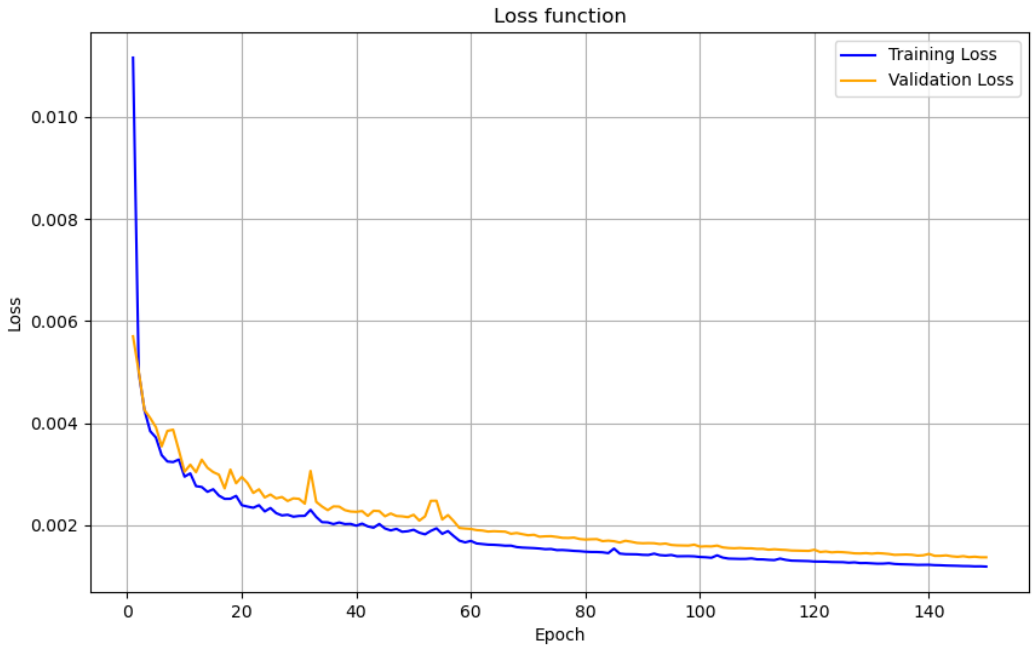


Furthermore, hybrid loss configurations involving MSE and GDL were explored during the experimental phase. However, these combinations exhibited significant training instability. It was observed that higher GDL coefficients led poor convergence, while lower coefficients rendered the gradient penalization negligible, yielding results nearly identical to the standard MSE.

3.2. TRAINING DYNAMICS OF THE OPTIMAL MODEL

To understand the learning process of the selected architecture, this section analyzes the training evolution of the MSE-optimized model, which demonstrated the best performance. The training was conducted over a maximum of 150 epochs, utilizing mixed-precision to optimize computational efficiency while maintaining numerical stability.

Figure 3.3. Loss Function evolution during training



The convergence behavior, illustrated in Figure 3.3, reveals a stable and consistent

reduction in both training and validation errors across the 150 epochs. Several key observations can be made regarding the optimization strategy and the model's learning dynamics:

- **Generalization Stability:** The training and validation loss curves exhibit a high degree of synchronization, with a minimal and constant gap between them towards the end of the process. This behavior confirms that the model effectively generalizes the learned spatial features showing no signs of overfitting.
- **Scheduler Impact:** A notable improvement in performance is observed around epoch 60, where the scheduler adjusted the learning rate. leading to a visible drop in error and a more stable optimization path in the subsequent epochs.
- **Convergence Stability and Optimization:** After epoch 120, the loss curves demonstrate a clear stabilization in their learning trajectory. This behavior indicates that the model has successfully captured the predominant structural patterns within the training data, entering a phase where further iterations yield only marginal improvements in error reduction. Continuing the optimization beyond this point would offer negligible gains in predictive performance while unnecessarily increasing the computational overhead and the risk of over-calibrating the model to specific noise in the training set.

The observed training dynamics reinforce the effectiveness of MSE as a primary objective function for this architecture. The results demonstrate that the quadratic penalization successfully guided the transformer-based encoder to converge toward a highly representative of the vertical canopy structure.

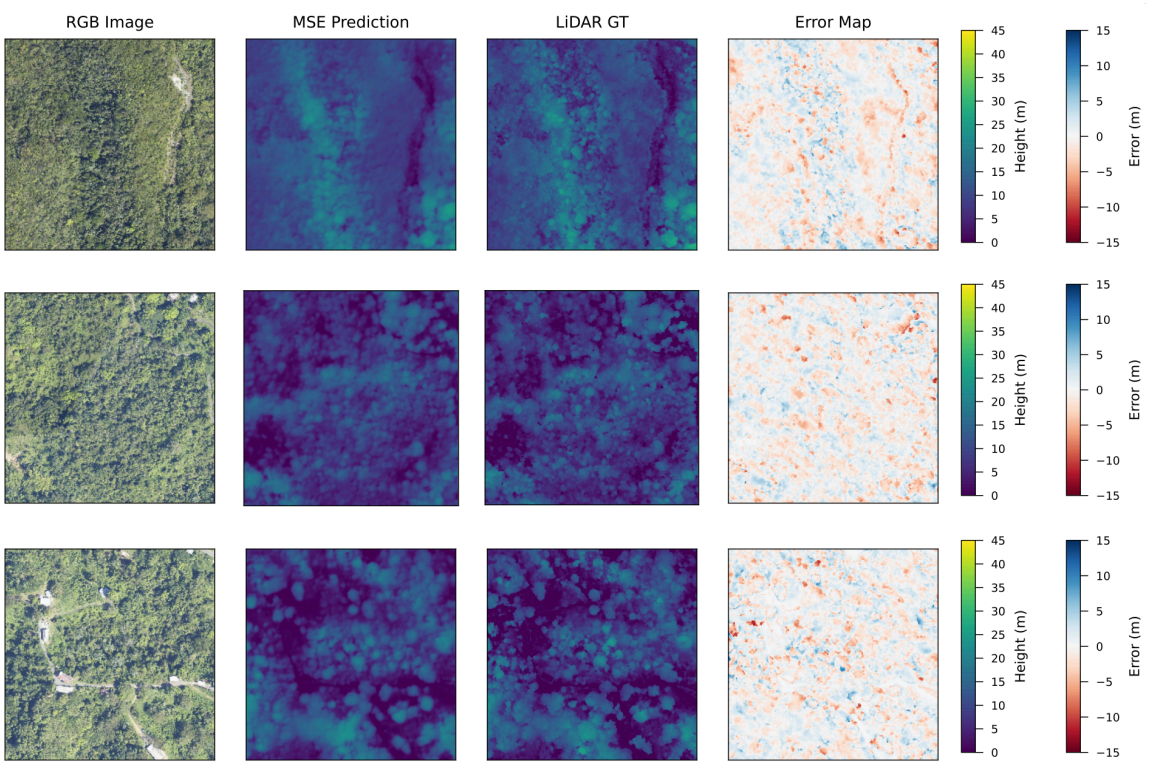
3.3. EXTENDED QUALITATIVE ANALYSIS

To further analyze the spatial reliability of the final MSE-optimized model, Figure 3.4 presents the absolute error distribution for representative test patches. This visualiza-

tion illustrates the pixel-wise difference between the LiDAR reference and the model's predictions.

The error maps reveal that discrepancies are predominantly localized at the boundaries of tree crowns and in areas with extreme vertical gradients. This suggests that while the model captures the overall volumetric structure with high precision, fine-scale alignment remains a challenge in dense canopy environments. However, for the vast majority of the forest area, the error remains consistently low, confirming the model's robustness. Furthermore, a closer inspection of the larger tree structures shows a slight increase in error variance compared to lower-stature vegetation.

Figure 3.4. Spatial distribution of errors for the final MSE model. The absolute error maps highlight that most inaccuracies occur at canopy edges and extreme heights.



3.4. SPATIAL GENERALIZATION PERFORMANCE

To evaluate the robustness and transferability of the MiT-B5 U-Net, the model was tested on independent geographic locations within the NEON network that were completely excluded from both the training and internal validation phases. Notably, we selected the same geographic sites characterized in recent global benchmarks [3], ensuring that the model is evaluated against realistic forest heterogeneity previously identified as challenging in the literature. While these locations align with established benchmarks, the evaluation data in this section were generated using our internal processing workflow to ensure consistency. This analysis is fundamental as it addresses the impact of domain shift, where the model must face significant environmental and biological variations, such as changes in composition, canopy density, and species that define the forest's appearance. Furthermore, external sites could introduce diverse lighting conditions, solar angles, and soil background reflectance that differ from the training distribution, requiring the architecture to rely on fundamental structural patterns rather than site-specific characteristics. Successfully performing across these diverse scenarios demonstrates the model's ability to maintain its representational capacity and precision in ecological and atmospheric variability inherent to large-scale forest monitoring.

3.4.1. Numerical performance and domain shift. Table 3.2 summarizes the numerical performance on 4753 patches (256x256) of the NEON test sites. While an increase in the Mean Absolute Error (MAE) and Root Mean Square Error (RMSE) is observed, the results remain highly competitive.

This performance differential represents a standard behavior during the model evaluation phase on an independent test set, where results naturally fluctuate across different geographic regions. Such variations are expected as the model encounters a diverse range of compositions, canopy densities, and solar illumination conditions that

vary from one site to another. Despite this inherent environmental diversity, the model maintains a solid correlation with the LiDAR reference throughout the test. This stability confirms that the features extracted by the hierarchical Transformer are capable of capturing fundamental structural patterns of the forest, allowing the height estimations to remain consistent and reliable across multiple ecological contexts.

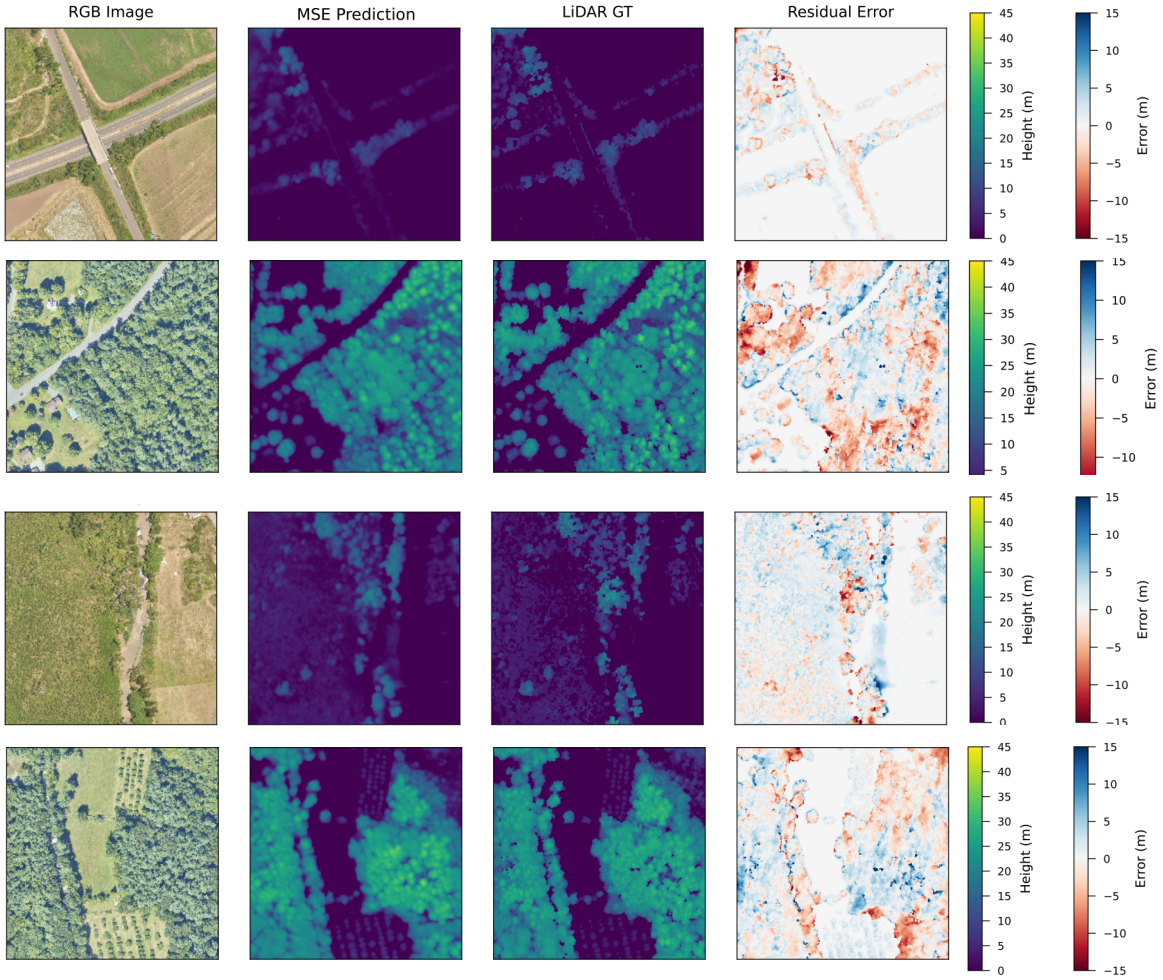
Table 3.2. Model performance on unseen geographic NEON sites

	MAE (m) $\downarrow$	**RMSE (m)** $\downarrow$	$R^2 \uparrow$
NEON Test	2.2540	3.9499	0.6951

3.4.2. Qualitative assessment of generalization. The qualitative results, illustrated in Figure 3.5, confirm that the architecture successfully recovers the vertical complexity of the canopy even in forest types not encountered during training. As observed in the generated Canopy Height Models (CHM), the model maintains consistent crown delineation and height.

A detailed inspection of the predictions highlights the model's generalization, showing a clear differentiation between dense forest clusters and open areas that replicates the LiDAR spatial distribution. The model successfully identifies isolated crowns in transition zones, preserving the structure of the canopy even in new geographic contexts. While the structural logic remains consistent, the residual map reveals localized height underestimations, particularly at crown edges and in areas where complex textures can be deceptive. This suggest that while the MiT-B5 U-Net effectively captures robust textural relationships across diverse forest sites, the inherent variability of vegetation structure still presents challenges for maintaining absolute precision in height magnitude when moving beyond the original training distribution.

Figure 3.5. Qualitative analysis of the model's performance on the NEON test. From left to right: RGB input imagery, Predicted CHM, LiDAR Ground Truth, and Residual Error map.



3.5. BENCHMARKING: COMPARISON WITH STATE-OF-THE-ART

To evaluate the competitive standing of the proposed architecture, a direct comparison was conducted against the SSL-Large and SSL-Aerial models developed by Meta AI[3]. Unlike the previous spatial generalization analysis, this evaluation was performed strictly using the official Meta AI benchmark dataset and their established inference pro-

tocols, including the specific normalization strategies required for their architectures. While the SSL-Large represents the general-purpose baseline, the SSL-Aerial model is specifically optimized for high-resolution aerial imagery, providing a more rigorous benchmark for our proposed model.

3.5.1. Numerical performance Table 3.3 summarizes the performance of both models on the external test set. These results provide a comparative overview of the models performance when evaluated under the same data distribution and processing constraints.

Table 3.3. Comparative performance results on the Meta AI NEON test set.

Model	**MAE (m)** $\downarrow$	**RMSE (m)** $\downarrow$	R^2 **Block** $\uparrow$
Proposed Model	2.51	4.26	0.69
SSL-Large (Meta AI)	3.31	5.41	0.37
SSL-Aerial (Meta AI)	2.50	3.86	0.70

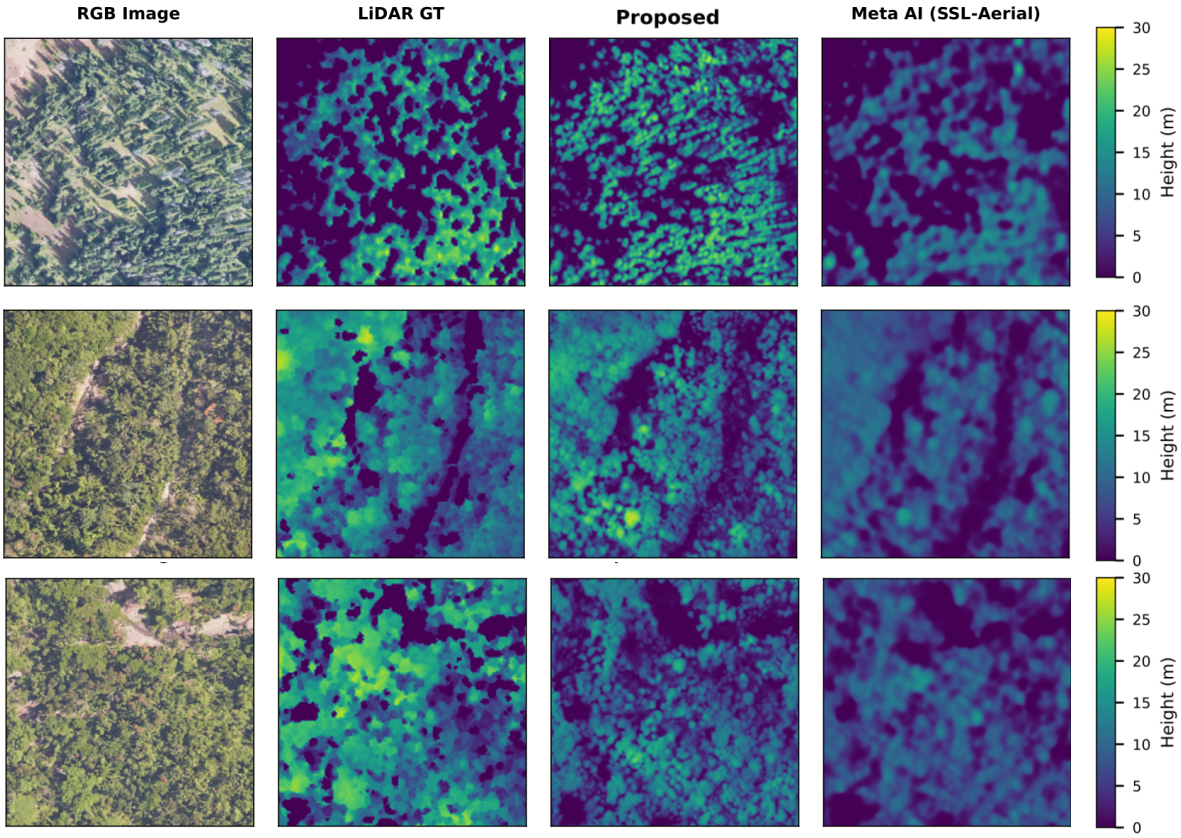
Metrics obtained through local replication of the inference protocol and verified against the official project repository.

The R^2 values presented here correspond to the "spatial block R^2" metric. This methodological approach involves aggregating height values over spatial blocks to evaluate the correlation of canopy height trends at a broader scale, rather than a pixel-by-pixel basis.

3.5.2. Qualitative analysis and structural fidelity Despite the competitive numerical performance of the SSL-Aerial model, which represents the highest performance in this benchmark, the qualitative assessment reveals fundamental differences in the practical utility of the predictions. As illustrated in Figure 3.6, while the SSL-Aerial achieves high statistical correlation, the Proposed Model demonstrates a superior capacity to preserve the forest's morphological complexity. This distinction is particularly evident in the delineation of individual crowns and the clear identification of canopy

gaps, where the reference models including the specialized Aerial version tend to produce smoothed height surfaces that obscure fine-grained structural details.

Figure 3.6. Visual comparison between MiT-B5 U-Net and SSL-Aerial



The following observations characterize the models' performance:

- **Individual Tree Crown (ITC) Delineation:** The MiT-B5 U-Net demonstrates a superior capacity to segment discrete crowns, which is essential for precision tasks such as tree counting. In contrast, the SSL-Aerial model yields a more continuous and smoothed surface, where individual tree structures become less distinct.
- **Identification of Non-Vegetated Areas:** A significant difference is observed in the detection of open spaces and soil background. The proposed architecture effectively identifies non-vegetated zones, maintaining sharp transitions between the canopy

and the ground. Conversely, the reference model tends to predict residual elevation values in these areas, resulting in a "height haze" where zero-height values are not clearly reached in the absence of trees.

- **Structural Precision and Detail Retrieval:** While both models are evaluated on the same task of canopy height estimation, the MiT-B5 U-Net recovers higher spatial frequency details. This results in a more granular representation of the forest canopy, whereas the reference model's output exhibits a smoothed morphological signature across different forest densities.

- **Ground Truth Nature:** Visual inspection of the benchmark's LiDAR reference reveals a low-frequency morphological signature. Despite this characteristic, the proposed model identifies high-frequency structural details present in the RGB imagery, demonstrating its robustness in maintaining spatial definition even when such features have been attenuated or smoothed in the reference labels.

In summary, the benchmark comparison demonstrates that the MiT-B5 U-Net is not only numerically competitive with the state-of-the-art but also offers superior structural definition. While the SSL-Large model captures the general vertical distribution of the canopy, the proposed architecture excels in retrieving its fine-grained geometry and accurately isolating non-vegetated areas. These results confirm that the MiT-B5 U-Net provides a more detailed representation of forest structure.

3.6. ANALYSIS OF SYSTEMATIC ERRORS AND LIMITATIONS

Despite the high overall accuracy achieved by the MSE-optimized model , a detailed analysis of the residuals reveals certain systematic limitations. As illustrated in the density scatter plots Figure 3.1, all evaluated models exhibit a consistent underestimation of canopy heights exceeding 45 meters.
This saturation effect can be explained by two primary factors:

- **Data Imbalance and Ecological Distribution:** The dataset exhibits a lower frequency of samples in the 45-60 meter range, primarily because trees of such height are biologically rare and less common in the natural forest structure compared to the dominant 10–30 meter group. Consequently, the hierarchical transformer has fewer opportunities to characterize the specific textural features of these emergent individuals, leading to the observed saturation effect where the model tends to regress towards the more frequently represented height intervals.

- **LiDAR Reference Noise:** In several test patches, the LiDAR ground truth reports heights near 60 meters that do not correspond to clear tree structures in the RGB imagery. These points, visible as isolated outliers in the bottom-right of the scatter plots, likely represent signal noise or artifacts in the LiDAR CHM. The model's tendency to predict lower values in these areas suggests it acts as a spatial regularizer, ignoring non-representative extreme values in favor of the most probable canopy height.

These limitations highlight that while RGB-based estimation is highly effective for the majority of the forest structure, accurate mapping of extreme vertical outliers may require the integration of multi-modal data or a higher density of samples for the tallest vegetation classes.

3.7. CHAPTER SUMMARY

The experimental results presented in this chapter demonstrate that the proposed MiT-B5 U-Net architecture, optimized with a Mean Squared Error (MSE) loss function, is highly effective for canopy height mapping from RGB imagery. The model achieves a performance that is competitive with current state-of-the-art benchmarks within the NEON ecosystem.

While a systematic underestimation occurs in emergent trees exceeding 45 meters—a limitation linked to the natural ecological scarcity of such individuals and the resulting

data imbalance—the model's robust ability to generalize to unseen geographic sites confirms its potential as a scalable tool for forest monitoring. These findings underscore the effectiveness of transformer-based encoders for structural characterization and lay the foundation for future work involving multi-modal data integration, such as the fusion of optical imagery with Radar or NIR sensors, to further refine height estimation in extreme vertical structures.

4. CONCLUSIONS AND FUTURE WORK

The development and evaluation of the Hybrid Transformer-U-Net architecture for canopy height estimation demonstrate the potential of hierarchical attention mechanisms in remote sensing. This research successfully addressed the challenge of extracting complex vertical forest structures from high-resolution RGB imagery, achieving a performance that is competitive with current state of the art models. While there is still room for improvement, the model exhibited robust efficiency and precision, establishing a new high-performance baseline within the NEON ecosystem. Our findings show the potential of transformer-based architectures to serve as a scalable and accessible alternative for large-scale forest monitoring, bridging the gap between passive optical sensors and high-fidelity structural characterization.

4.1. CONLCUSIONS

The experimental results obtained in this study lead to the following key conclusions:

- **Optimal Optimization Strategy:** The Mean Squared Error (MSE) was identified as the most effective objective function for canopy height regression. The quadratic penalization of MSE proved superior in capturing vertical extremes and maintaining numerical stability. Unlike robust or gradient-based losses—which in this study exhibited lower convergence precision or instability—the MSE effectively prioritized the correction of large residual errors, leading to a more accurate representation of the tallest canopy structures.
- **Structural Generalization:** The model demonstrated remarkable robustness when tested on unseen geographic sites. The ability to maintain structural fidelity in areas previously used as global benchmarks confirms that the hierarchical transformer

encoder captures fundamental textural features that are transferable across different sites.

- **Saturation and Height Limitations:** A systematic saturation effect was identified for emergent trees exceeding 45 meters. This limitation is attributed to the inherent loss of textural variance in passive optical data at extreme heights and the natural ecological scarcity of such individuals, leading to a data imbalance in the training distribution. Despite this, the model demonstrated a significant capacity for spatial regularization, providing consistent canopy representations and effectively filtering artifacts and noise present in the LiDAR reference data.

- **Benchmarking and High-Resolution Fidelity:** Comparative analysis against the state-of-the-art confirmed that the MiT-B5 U-Net is numerically competitive while offering superior structural definition. The model's ability to accurately discriminate between individual crowns and non-vegetated areas highlights its effectiveness for precision forestry applications.

4.2. FUTURE WORK

To overcome the identified limitations and expand the operational scope of this research, the following directions are proposed:

- **Multispectral Data Fusion:** Incorporating the Near-Infrared (NIR) band represents a critical next step. Previous studies have demonstrated that the NIR spectrum significantly enhances height estimation by providing detailed information on leaf biomass and canopy density, which is particularly useful in multi-layered environments. While the limited availability of high-resolution datasets containing both NIR and precise LiDAR ground truth remains a challenge, future efforts should prioritize the acquisition of such multispectral data to further refine the model's structural characterization.

- **Active Sensor Integration:** Future work should explore the integration of Synthetic Aperture Radar (SAR) data to complement optical imagery. The ability of radar to penetrate the canopy and interact directly with woody structures could effectively break saturation ceiling observed in RGB-only models. Furthermore, SAR data could prove essential in mitigating the impact of complex terrain irregularities and slope-induced shadows that often distort passive optical signals. Although recent studies have implemented SAR for height estimation, the primary challenge remains the lower spatial resolution of current radar products, suggesting that a super-resolution or multi-stream fusion approach would be necessary to preserve the fine-scale detail achieved in this research.
- **Dataset Enrichment:** Expanding the training set to include a broader range of climatic zones and sensor specifications would enhance global applicability. Prioritizing the inclusion of emergent tree samples and diverse atmospheric conditions will further reduce the impact of domain shift improving the scalability of the model.

This work provides a scalable and efficient methodology for forest structure characterization, contributing to the development of accessible tools for large-scale environmental monitoring through the integration of computer vision and remote sensing imagery.

BIBLIOGRAPHY

BISHOP, Christopher M. *Pattern Recognition and Machine Learning*. New York: Springer-Verlag, 2006 (cit. on p. 35).

CHEN, J. et al. “Log-Cosh Loss Function for Deep Learning Regression”. In: *arXiv preprint arXiv:2210.03210* (2022) (cit. on p. 32).

DENG, Jia et al. “ImageNet: A Large-Scale Hierarchical Image Database”. In: *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. 2009, pp. 248–255. DOI: 10.1109/CVPR.2009.5206848 (cit. on p. 25).

DUBAYAH, R. et al. *GEDI L2A Elevation and Height Metrics Data Global Footprint Level V002*. Fecha de acceso: 2025-07-28. 2021. DOI: 10.5067/GEDI/GEDI02_A.002 (cit. on p. 11).

DUBAYAH, Ralph O and Jason B DRAKE. “Lidar remote sensing for forestry”. In: *Journal of Forestry* 98.6 (2000), pp. 44–46 (cit. on p. 16).

FOGEL, Fajwel et al. “Open-Canopy: A Country-Scale Benchmark for Canopy Height Estimation at Very High Resolution”. In: (2024). eprint: 2407.09392 (cit. on p. 30).

IAKUBOVSKII, Pavel. *Segmentation Models Pytorch*. https://github.com/qubvel/segmentation_models.pytorch. 2019 (cit. on p. 28).

KAMPE, Thomas et al. “NEON: the first continental-scale ecological observatory with airborne remote sensing of vegetation canopy biochemistry and structure”. In: *Proc. SPIE* (2009) (cit. on p. 20).

KINGMA, Diederik P. and Jimmy BA. “Adam: A Method for Stochastic Optimization”. In: *International Conference on Learning Representations (ICLR)*. 2015 (cit. on p. 33).

LILLESAND, Thomas, Ralph W KIEFER, and Jonathan CHIPMAN. *Remote sensing and image interpretation*. 7th. Hoboken, NJ: John Wiley & Sons, 2015 (cit. on p. 14).

MA, Lei et al. “Deep learning in remote sensing: A comprehensive review and list of resources”. In: *IEEE Geoscience and Remote Sensing Magazine* 7.2 (2019), pp. 67–105 (cit. on p. 19).

MATHIEU, Michael, Camille COUPRIE, and Yann LECUN. “Deep multi-scale video prediction beyond mean square error”. In: *arXiv preprint arXiv:1511.05440* (2015) (cit. on p. 32).

MIGLIAVACCA, Mirco et al. “The three major axes of terrestrial ecosystem function”. In: *Nature* 598.7881 (2021), pp. 468–472. DOI: 10.1038/s41586-021-03939-9 (cit. on p. 15).

NATIONAL ECOLOGICAL OBSERVATORY NETWORK (NEON). *Ecosystem structure (DP3.30015.001)*. 2025. DOI: 10.48443/JQQD-1N30 (cit. on p. 21).

— *High-resolution orthorectified camera imagery (DP1.30010.001)*. 2025. DOI: 10.48443/F8DG-8W18 (cit. on p. 22).

RONNEBERGER, Olaf, Philipp FISCHER, and Thomas BROX. “U-net: Convolutional networks for biomedical image segmentation”. In: *International Conference on Medical image computing and computer-assisted intervention* (2015), pp. 234–241 (cit. on p. 18).

SKIDMORE, Andrew K et al. “Priority list of biodiversity indicators to track progress towards the post-2020 global biodiversity framework”. In: *Environmental Research Letters* 16.9 (2021), p. 094049 (cit. on p. 15).

STEPHENSON, N. L. et al. “Rate of tree carbon accumulation increases continuously with tree size”. In: *Nature* 507.7490 (2014), pp. 90–93. DOI: 10.1038/nature12914 (cit. on p. 11).

TOLAN, Jamie et al. “Very high resolution canopy height maps from RGB imagery using self-supervised vision transformer and convolutional decoder trained on aerial lidar”. In: *Remote Sensing of Environment* 300 (2024), p. 113888 (cit. on pp. 12, 30, 43, 45).

VASWANI, Ashish et al. “Attention is all you need”. In: *Advances in neural information processing systems* 30 (2017) (cit. on p. 19).

XIE, Enze et al. “SegFormer: Simple and Efficient Design for Semantic Segmentation with Transformers”. In: *arXiv preprint arXiv:2105.15203* (2021) (cit. on p. 25).