Remote Sensing
MDPI placeholder lo

Article

Attention-Based Enhancement of Airborne LiDAR Across Vegetated Landscapes Using SAR and Optical Imagery Fusion

Michael Marks [1,*], Daniel Sousa [1] and Janet Franklin [1,2]

[1] Department of Geography, San Diego State University, San Diego, CA 92182, USA; dan.sousa@sdsu.edu (D.S.); jfranklin2@sdsu.edu (J.F.)
[2] Center for Open Geographical Sciences, San Diego State University, San Diego, CA 92182, USA
* Correspondence: mmarks0561@sdsu.edu

Abstract: Accurate and timely 3D vegetation structure information is essential for ecological modeling and land management. However, these needs often cannot be met with existing airborne LiDAR surveys, whose broad-area coverage comes with trade-offs in point density and update frequency. To address these limitations, this study introduces a deep learning framework built on attention mechanisms, the fundamental building block of modern large language models. The framework upsamples sparse (<22 pt/m^2) airborne LiDAR point clouds by fusing them with stacks of multi-temporal optical (NAIP) and L-band quad-polarized Synthetic Aperture Radar (UAVSAR) imagery. Utilizing a novel Local–Global Point Attention Block (LG-PAB), our model directly enhances 3D point-cloud density and accuracy in vegetated landscapes by learning structure directly from the point cloud itself. Results in fire-prone Southern California foothill and montane ecosystems demonstrate that fusing both optical and radar imagery reduces reconstruction error (measured by Chamfer distance) compared to using LiDAR alone or with a single image modality. Notably, the fused model substantially mitigates errors arising from vegetation changes over time, particularly in areas of canopy loss, thereby increasing the utility of historical LiDAR archives. This research presents a novel approach for direct 3D point-cloud enhancement, moving beyond traditional raster-based methods and offering a pathway to more accurate and up-to-date vegetation structure assessments.

Keywords: LiDAR upsampling; point-cloud enhancement; multi-modal fusion; deep learning; attention mechanism; SAR; optical imagery; vegetation structure; transformers

Academic Editor: Pinliang Dong

Received: 15 June 2025
Revised: 22 August 2025
Accepted: 5 September 2025
Published:

Citation: Marks, M.; Sousa, D.; Franklin, J. Attention-Based Enhancement of Airborne LiDAR Across Vegetated Landscapes Using SAR and Optical Imagery Fusion. *Remote Sens.* **2025**, *1*, 0. https://doi.org/

1. Introduction

Accurate three-dimensional (3D) vegetation structure information at sub-meter spatial scales now plays a key role in applications ranging from wildfire risk modeling [1] to biodiversity and habitat assessment [2,3]. The way vegetation is arranged—its height, density, and continuity—directly influences both fire hazard and fire behavior and impacts how species use the landscape. Fuels, as opposed to topography and weather, are the only element of the fire behavior triangle that land managers can directly manipulate [4], making structural data vital for strategic interventions. Simultaneously, vegetation structure governs microclimate, resource availability, and landscape connectivity, making it a cornerstone of ecological monitoring and conservation planning [5–7]. Airborne light detection and ranging (LiDAR) has emerged as a premier tool for capturing this structural complexity, enabling detailed, landscape-scale mapping of vegetation structure that was previously unattainable with passive optical imagery [2,8,9].

National mapping programs such as the U.S. Geological Survey's 3D Elevation Program (3DEP) now collect LiDAR data over large areas, but these surveys have important limitations. Typical 3DEP acquisitions are performed at modest point densities (on the order of 0.5 pts/m^2 to 20 pts/m^2) [10], and with much of the national LiDAR baseline acquired over an extended period (e.g., roughly 2015–2023), a significant portion of this data is now several years old, a situation exacerbated by the lack of a guaranteed or universal update timeline [11]. Consequently, the available point clouds often reflect conditions from several years prior and are relatively sparse compared to those obtained from other platforms (e.g., uncrewed aerial vehicles—or UAVs; see Figure 1).

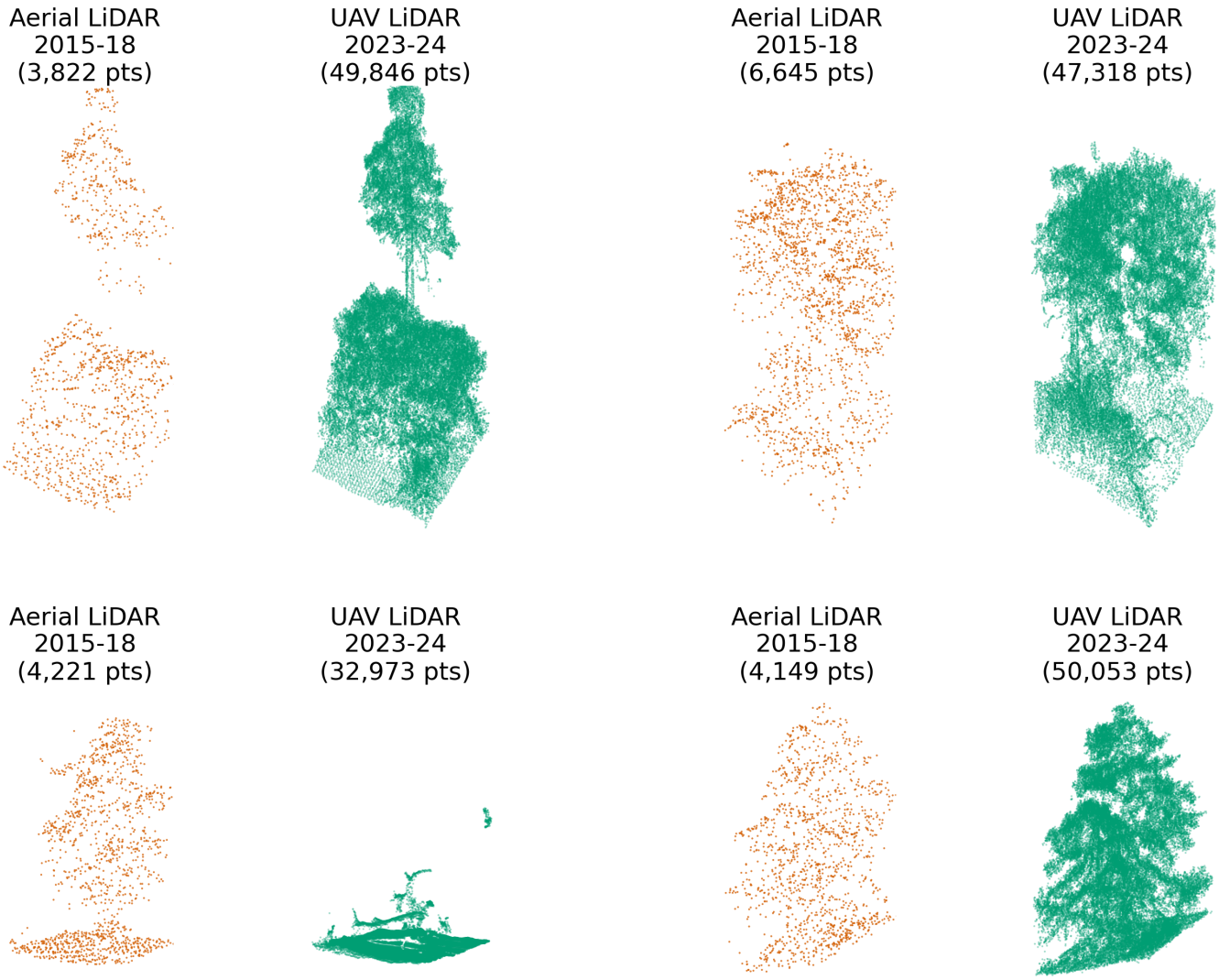

Figure 1. Comparison of point clouds with 10 m × 10 m footprints from USGS 3DEP aerial LiDAR (2015-2018, orange) and UAV LiDAR (2023-2024, green) over the same locations. UAV LiDAR captures significantly greater structural detail, especially in fine-scale canopy features. The bottom-left example shows clear canopy loss between surveys due to recent disturbance. The top left shows clear growth.

Critical changes in vegetation structure, such as disturbance-driven loss or ongoing vegetation growth, may go undetected between LiDAR acquisition cycles. This sparsity and temporal gaps limit the utility of national LiDAR datasets for applications that require up-to-date, high-resolution 3D information.

Given the limitations of national LiDAR in spatial and temporal resolution, one promising avenue is to enrich these sparse datasets using co-registered imagery from other remote sensing platforms that offer more frequent updates. Sub-meter-resolution

aerial imagery—such as orthophotos from the National Agriculture Imagery Program (NAIP) [12]—provides fine detail on canopy textures, gaps, and vegetation color, typically acquired at 2 year intervals. Complementing this, L-band synthetic aperture radar, known for its sensitivity to vegetation structure [13], can provide valuable multi-temporal data through repeat acquisitions. NASA's UAVSAR [14], for example, conducts roughly bi-annual L-band campaigns to monitor movement along the San Andreas fault in California, and the upcoming NISAR satellite mission will offer global L-band SAR at a 12-day revisit rate [15]. Fusing such temporally rich optical and radar imagery with existing LiDAR has the potential to produce a denser 3D point cloud reflecting more current vegetation conditions—a challenge well-suited to data-driven approaches such as deep learning. In particular, attention-based models offer a powerful way to integrate these diverse inputs by modeling their spatial and semantic relationships.

Attention mechanisms, first introduced for language translation in [16], enable a model to dynamically determine which parts of the input data are most relevant to each other, a capability crucial for understanding complex scenes. For example, in point clouds of vegetated landscapes, a point on a tree's leaf can learn, through self-attention, to connect more strongly with its own trunk or branches than with foliage from an adjacent, albeit closer, tree. This ability to discern intrinsic structural relationships could be particularly effective in natural vegetation, as its fractal and self-similar nature provides consistent patterns for self-attention to model across different scales [17,18]. When fusing data, cross-attention extends this by allowing features from one modality, such as a LiDAR point, to selectively query information from another modality, like relevant shadow patterns or canopy gaps identified in NAIP imagery or radar data. These powerful attention operations are the fundamental building blocks of the influential Transformer architecture [19], which serves as the foundation for nearly all large language model architectures in use today. Building on that success, Transformers have been adapted for vision tasks [20] and are now increasingly used across many remote sensing tasks [21]. While these advancements showcase their broad utility, their specific application and optimal adaptation for the enhancement of sparse airborne LiDAR in complex vegetated landscapes present unique challenges and open questions.

Consequently, key knowledge gaps remain. First, most prior work on data-driven LiDAR enhancement has focused on enhancing point cloud-derived metrics in raster form (e.g., canopy height [22,23], above-ground biomass [24], elevation [25], and other fuel/vegetation metrics [26,27]) rather than directly enhancing the point cloud itself. Although one recent study successfully upsampled mobile laser scanner point clouds in a forested environment using terrestrial LiDAR as the reference dataset [28], both sensors differ substantially from airborne systems in terms of scale and occlusion behavior. Second, existing deep learning frameworks for point-cloud upsampling have primarily been developed and tested on synthetic shapes or human-made objects, and their efficacy on the complex, irregular structures of natural vegetation is not well understood. Third, we found no studies that have attempted to leverage optical or radar imagery for enhancing point clouds in vegetated landscapes. Fourth, we found no studies that have analyzed model performance when the LiDAR input is temporally misaligned with the reference dataset, confounding performance metrics with real-world landscape changes. Deep models are typically trained and evaluated on static scenes, often using an artificially downsampled point cloud as the input. Thus, it remains unknown how upsampling errors behave in areas where substantial canopy growth or loss has occurred since the original LiDAR survey and whether multi-modal inputs can mitigate errors stemming from such changes.

1.1. Background and Related Work

1.1.1. Point-Cloud Upsampling with Deep Learning

In computer vision and graphics, a range of neural frameworks have been proposed to densify sparse point clouds. PU-Net [29] pioneered the task with multi-layer perceptron (MLP) feature extraction and a point-set expansion module, achieving good fidelity on synthetic computer-designed (CAD) objects. PU-GCN (Point Upsampling-Graph Convolution Network) later built upon this by replacing the expansion MLP with a graph-convolution upsampling unit called NodeShuffle and paired it with a GCN feature extractor [30]. Recently, PU-Transformer introduced the first Transformer-based upsampler to exploit both local and long-range relations [31]. While these methods deliver state-of-the-art results on synthetic shapes and man-made objects, their behavior on the irregular geometry of vegetation—and in LiDAR-derived point clouds more broadly—remains largely untested.

1.1.2. Upsampling in Vegetated Landscapes

Upsampling LiDAR data from forests and other natural vegetation introduces unique challenges. In natural vegetation, aerial LiDAR point clouds exhibit uneven densities—upper vegetation layers are well sampled due to their proximity to the sensor, while lower vegetation layers and the ground experience significantly reduced returns, introducing complexity that differs from human-made environments. Zhang and Filin [32] highlighted that most existing research had focused on upsampling point clouds of human-made objects, with little attention to natural scenes. They found that standard 3D interpolation or naïve point densification often leads to over-smoothed results in forests, since such methods ignore fine local variations in structure. To address this, Zhang and Filin proposed a graph convolutional network with a global attention mechanism that exploits vegetation's self-similar geometric patterns for superior vegetated landscape upsampling. Nevertheless, that work relied solely on the geometric information in the LiDAR point cloud, without incorporating external imagery or multi-modal data.

1.1.3. Utilizing Cross-Attention for Multi-Modal Fusion

Cross-attention mechanisms have proven valuable for multi-modal data fusion in remote sensing, though their application has largely centered on integrating various 2D raster datasets [33–36]. In remote sensing, the primary method for fusing 3D LiDAR data with imagery involves rasterizing the LiDAR information, most often by integrating digital surface models (DSMs) with hyperspectral data [37–39]. Consequently, the direct fusion of individual LiDAR point features with imagery using cross-attention represents a largely unexplored area in remote sensing research. Conversely, the broader computer vision community actively develops and utilizes such direct point-to-image cross-attention techniques to enhance detailed 3D scene perception [40–42].

Attention-Based Multi-Modal Upsampling

Building on these advances, we introduce an upsampling model that leverages attention mechanisms to capture both local and global context while fusing LiDAR with optical and radar inputs. Transformer-based architectures have recently shown promise in 3D point-cloud tasks by modeling long-range dependencies in point sets. Our network adopts a *Local–Global Point Attention* block structure inspired by this paradigm. At the local scale, a multi-head variant of a point Transformer architecture developed by Zhao et al. [43] applies self-attention within each point's neighborhood to learn fine-grained spatial details. This "multi-head" approach enables the model to learn multiple, distinct feature representations in parallel; for instance, one head may learn to model fine-scale canopy texture while another captures broader branch-level geometry. At the global scale, we

incorporate a position-aware multi-head attention mechanism over the entire point cloud to ensure structural coherence. To maintain computational efficiency, we implement this global attention with a FlashAttention [44] algorithm, allowing exact multi-head attention across thousands of points in a memory-efficient manner. By combining local and global attention pathways, the model preserves small-scale features (e.g., individual tree crown shapes) while enforcing consistency in larger-scale patterns (e.g., stand-level canopy height gradients). This architecture extends prior point upsampling networks but is uniquely tailored to handle multi-modal inputs and the complexities of natural scenes.

The primary scientific contribution of our study is not merely a new network architecture but, rather, the exploration of a fused-modality approach to LiDAR point-cloud upsampling. In contrast to previous methods that input only sparse LiDAR points, we evaluate how incorporating additional imagery (optical NAIP and L-band SAR) can improve the reconstruction of vegetation structure. We also explicitly examine the temporal dimension by testing models in areas with known canopy growth or loss since the original LiDAR acquisition, an aspect largely overlooked in prior research.

1.2. Research Questions

- **RQ1**: To what extent does incorporating individual imagery modalities—(a) high-resolution optical imagery or (b) L-band Synthetic Aperture Radar (SAR) imagery—lower the point-cloud reconstruction error (measured by the Chamfer distance) compared to a baseline upsampling model that uses only the initial sparse LiDAR as input?
  - *Hypothesis*: Both modalities will reduce reconstruction error, but optical imagery will yield superior results.
  - *Reasoning*: The finer ground-sampling distance of optical imagery provides high-resolution texture essential for fine-scale detail. While L-band SAR is sensitive to volumetric structure, its coarser resolution is a limitation.
- **RQ2**: Does simultaneously fusing high-resolution optical and L-band SAR imagery yield additional reconstruction accuracy gains beyond the best single-modality model, indicating complementary rather than redundant information?
  - *Hypothesis*: The fused optical and SAR model will achieve the lowest reconstruction error, outperforming both single-sensor models.
  - *Reasoning*: Each sensor captures a different aspect of vegetation structure. The model's attention-based fusion is expected to leverage optical texture to define canopy boundaries and SAR backscatter to reconstruct internal volume.
- **RQ3**: How does reconstruction error change with net canopy-height gains and losses since the initial airborne LiDAR survey, and do the optical, SAR, and fused models mitigate these errors more effectively than a baseline upsampling model that uses only the initial sparse LiDAR as input?
  - *Hypothesis*: Errors will scale with canopy change and be greater for losses than gains. Model performance will stratify accordingly: the fused model will best mitigate these errors, followed by single-modality models, with the baseline performing most poorly, though it may capture some uniform growth.
  - *Reasoning*: The predicted error asymmetry stems from the differing nature of vegetation dynamics. Growth is often an incremental extrapolation of existing structure, whereas disturbance-driven loss (e.g., fire, treefall) can be abrupt and total. This creates a complete information void for the removed canopy in the legacy LiDAR, a more significant reconstruction challenge than modeling gradual growth.

2. Materials and Methods

The core challenge addressed by this research is the enhancement of sparse and outdated national airborne LiDAR datasets. To train and validate supervised upsampling models, we collected dense UAV LiDAR as a benchmark of actual vegetation structure. The sections below describe the study areas where this data was collected, the multi-modal input data sources, and steps taken to ensure spatial alignment and consistency across all modalities.

2.1. Study Area

The study area (Figure 2) consists of two separate sites in Southern California, USA. The first study area includes parts of the 24-square-kilometer Sedgwick Reserve, managed by the University of California, Santa Barbara, and the adjacent 11.5-square-kilometer Midland School property. Both sit within the San Rafael Mountain foothills of Santa Barbara County (Figure 3).

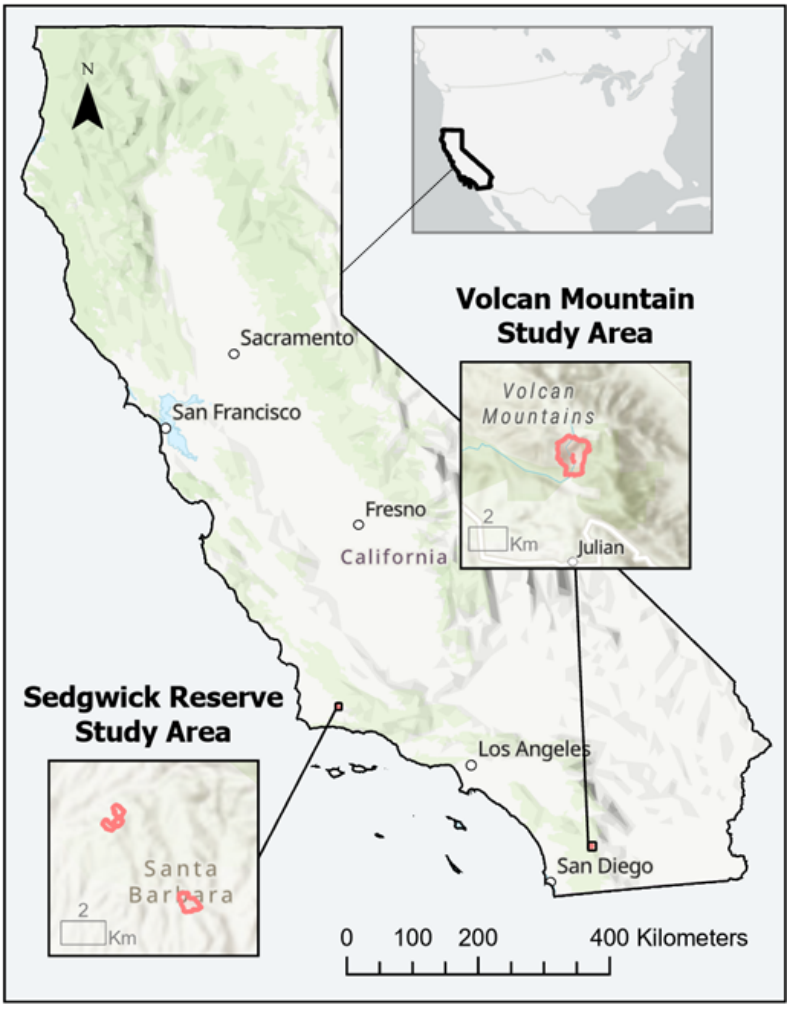

Figure 2. Locations of the Southern California UAV LiDAR surveys—Sedgwick Reserve–Midland School in the Santa Ynez Valley and Volcan Mountain Wilderness Preserve in the Peninsular Range. Red outlines indicate the extent of the UAV LiDAR surveys.

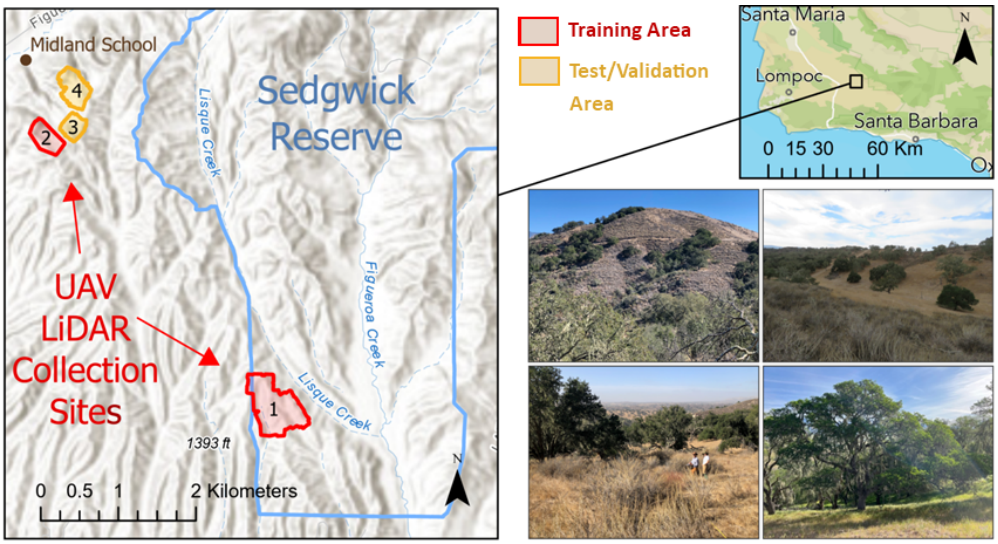

Figure 3. The first study area combines parts of the Sedgwick Reserve and Midland School property in the San Rafael Mountain foothills, Santa Barbara County. It features diverse vegetation, such as oak woodlands, chaparral, grasslands, and coastal sage scrub. Within this area, four UAV LiDAR sites were surveyed, covering a total of 70 hectares. The numbers indicate the UAV LiDAR survey sites which are detailed in Table 1.

The area spans elevations from 287 to 852 m and supports a mosaic of vegetation types with varying canopy architectures, including coastal sage scrub, native grasslands, chaparral (shrublands), coast live and blue oak woodlands, valley oak savanna, riparian habitats, and gray pine forests. UAV LiDAR was collected for four sites (totaling about 71 hectares) within the study area (Table 1). Given their similar terrain and vegetation, two sites were used for training, while the two smallest were withheld for independent model evaluation.

Table 1. UAV LiDAR sites used as ground truth for model development.

Site	Hectares	Location	Collection Date	Model Use
1	38	Sedgwick Reserve	30 June 2023	Training
2	12	Midland School	23 October 2023	Training
3	9	Midland School	23 October 2023	Test/Validation
4	11	Midland School	28 September 2023	Test/Validation
5	197	Volcan Mountain	25 October 2024	70% Training 30% Test/Validation

The second study area (Figure 4) comprises 197 hectares within and adjacent to the Volcan Mountain Wilderness Preserve in the Peninsular Range of Southern California. The reserve is managed by San Diego County and the Volcan Mountain Foundation and ranges in elevation from about 1220 m to over 1675 m. It hosts diverse plant communities, including oak woodlands, chaparral, mixed conifer forests, and grasslands. To ensure robust model evaluation across this ecological gradient, roughly 30 percent of the area (58 hectares) was reserved for testing and validation (Table 1). The three holdout zones (Figure 4) used for model evaluation were selected to reflect the site's vegetation diversity: the northernmost area includes chaparral that replaced forest following wildfires in 2003 and 2005; the central zone contains dense mixed-conifer and riparian vegetation interspersed with oak woodlands and chaparral; and the southernmost zone is predominantly semi-continuous oak canopy.

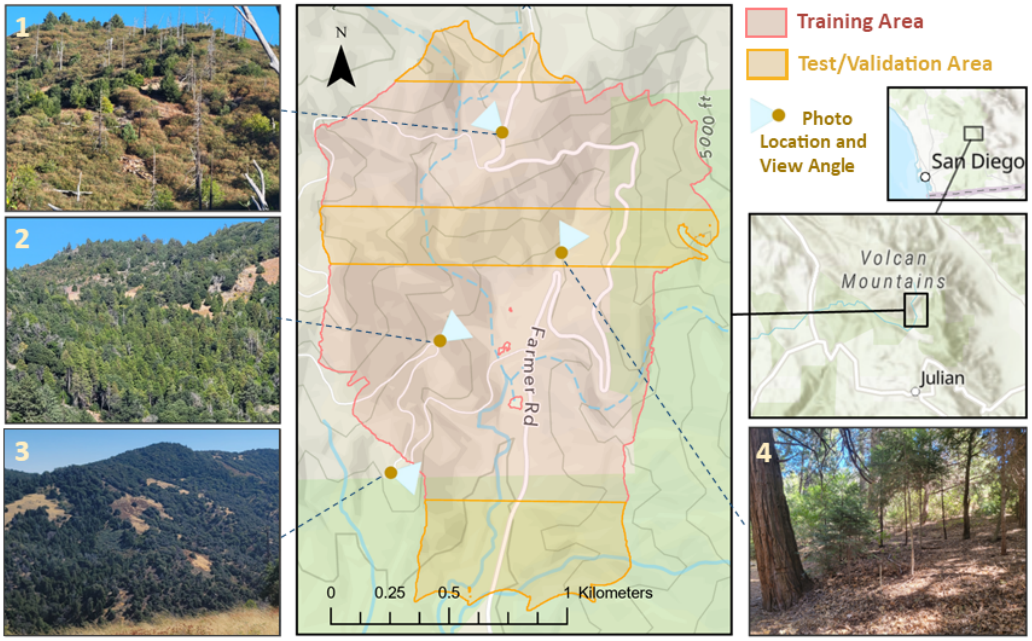

Figure 4. Volcan Mountain study area (197 hectares) in the Peninsular Range of Southern California, encompassing diverse vegetation types across an elevation range of 1220 to 1675 m. Thirty percent of the site (58 hectares) was set aside for testing/validation in three zones: post-fire chaparral in the north, mixed conifer and riparian vegetation with oak and chaparral in the center, and oak woodland with semi-continuous canopy in the south. Subfigures 1,2,3,4 show photos of the different vegetation types in the study area.

2.2. The Data

All remote-sensing assets were co-registered within a common tiling framework covering both study sites. First, the UAV-LiDAR acquisition footprints were tessellated into 10 m × 10 m analysis tiles with 15% overlap, yielding an initial set of 9800 tiles at Sedgwick–Midland and 26,557 tiles at Volcan Mountain. Each tile served as the spatial key for assembly of a four-layer data stack (Table 2).

Table 2. Summary of remote sensing datasets used in the data stack.

Data Source	Role in Study	Details	Revisit Rate
UAV LiDAR	Reference Data	>300 pts/m^2 *Acquired: 2023–2024*	N/A[1]
3DEP Airborne LiDAR	Sparse Input	Sedgwick: ~22 pts/m^2 (2018) Volcan: ~6 pts/m^2 (2015–2016)	N/A[1]
UAVSAR (L-band)	Ancillary Input	6.17 m GSD *Coverage: 2014–2024*	~2–3 years
NAIP (Optical)	Ancillary Input	0.6–1 m GSD *Coverage: 2014–2022*	~2 years

[1] Single acquisition; no revisit.

By standardizing on overlapping 10 m patches, we guarantee that each training example draws from the same footprint across sensors—ensuring the model learns consistent, co-registered features from LiDAR, SAR, and imagery.

2.2.1. UAV LiDAR Data

Sedgwick and Midland Sites—Between June and October 2023, UAV LiDAR data were collected by San Diego State Geography Department staff on the Sedgwick and Midland School Sites using a DJI Matrice 300 drone (DJI, Shenzhen, China) equipped with a TrueView 515 LiDAR instrument (GeoCue, Madison, AL, USA)/The drone was flown at an altitude

of approximately 60 m above ground level, achieving a point density of around 300 points per square meter.

Volcan Mountain Site—On 25 October 2024, the same Geography Department staff conducted flights using a DJI Matrice 350 drone (DJI, Shenzhen, China) equipped with a TrueView 540 LIDAR system (GeoCue, Madison, AL, USA). This newer drone and sensor were flown at a higher altitude of approximately 110 m above ground level and still achieved a point density of over 600 points per square meter.

2.2.2. Crewed Airborne LiDAR

The available crewed airborne LiDAR (C-ALS) data includes two separate 3DEP datasets. The Sedgwick 3DEP dataset, collected in 2018, has a point density of 22 pts/m^2. The C-ALS data for the Volcan Mountain site was collected between October 2015 and November 2016 and has a point density of 6.3 pts/m^2 [45]. Both datasets were obtained from Microsoft's Planetary Computer [46,47]. In addition to the point X, Y, and Z values we include three per-return attributes—intensity (16-bit integer 0–65,535 pulse magnitude), return number (ordinal of the return within its pulse), and number of returns (total returns from that pulse)—each standardized using the global mean and standard deviation computed across the full 3DEP dataset.

2.2.3. Synthetic Aperture Radar (SAR)

Fully polarimetric L-band imagery (23.84 cm wavelength) from NASA's UAVSAR system was obtained with the Alaska Satellite Facility's `asf_search` (version 8.1.1) Python API [48]. The UAVSAR flights were conducted from a Gulfstream-III platform with bidirectional acquisitions from opposite look directions at an average altitude of 12,495 m, providing multi-perspective radar coverage of the landscape at 6.17 m ground resolution. We utilized six fully polarimetric multi-look cross-product channels: `HHHH`, `HVHV`, `VVVV`, `HHHV`, `HHVV`, and `HVVV`. The specific campaigns and acquisition details for each study site are summarized in Table 3. For every 10 m × 10 m 3DEP tile, we extracted a co-centered 20 m × 20 m UAVSAR chip to accommodate layover and shadow extent, then bilinearly resampled each chip to 5 m GSD before fusion with NAIP imagery and LiDAR.

Table 3. Timeline of NAIP and UAVSAR acquisitions for both study sites.

	Volcan Mountain		**Sedgwick Reserve**	
Year	**NAIP (GSD)**	**UAVSAR (# Looks[1])**	**NAIP (GSD)**	**UAVSAR (# Looks[2])**
2014	May (1 m)	June (3), October (2)	June (1 m)	June (8)
2016	April (60 cm)	—	June (60 cm)	April (6)
2018	August (60 cm)	October (2)	July (60 cm)	—
2020	April (60 cm)	—	May (60 cm)	—
2021	—	November (2)	—	—
2022	April (60 cm)	—	May (60 cm)	—
2023	—	—	—	September (6)[2]
2024	—	—	—	October (2)

[1] The count of UAVSAR acquisition passes that month. Each listed month includes at least two opposing look directions; counts >2 indicate additional passes that may be distinct or repeated look geometries.
[2] Part of the NASA FireSense initiative.

2.2.4. High-Resolution Aerial Imagery

We ingested NAIP imagery through the Microsoft Planetary Computer STAC API [47] for survey years 2014–2022. NAIP provides four-band (red, green, blue, and near-infrared)

orthoimagery of the conterminous United States, collected at peak green-up on a two- to three-year cycle. Prior to the 2016 flight season, data were delivered at 1 m ground-sample distance (GSD); since 2016, the native resolution has been 60 cm. A complete timeline of these acquisitions and their resolutions is provided in Table 3. For every 10 m × 10 m 3DEP tile, we extracted a 20 m × 20 m NAIP chip centered on the same point to accommodate viewing-geometry variance and to capture neighboring shadows. All NAIP scenes were then resampled to a common 50 cm grid.

2.3. Data Cleaning and Preprocessing

To reduce computational load and give the upsampling network a uniform-density target, we downsampled the UAV LiDAR for every tile with an adaptive anisotropic voxel grid. Each cloud was first voxelized at a 4 cm × 4 cm × 2 cm resolution; if more than 50,000 points remained, the horizontal voxel edge was incrementally enlarged (keeping the vertical edge at 50% of that value), and the filter was reapplied until the count fell below the limit. The resulting point sets preserve fine vertical structure while standardizing horizontal density.

To bound memory use and keep attention context sizes tractable, we cap the input 3DEP point cloud at $N_{\max} = 10{,}000$ points per 10×10 m tile. If a tile exceeds this count, we randomly subsample to $N_{\max}$ points; fewer than 1% of tiles were affected. This cap defines the maximum attention sequence length and thus the model's memory footprint.

The dataset was partitioned into training, validation, and test sets using reference polygons to ensure the holdout sets captured the full environmental gradients found in the training data. During supervised training/evaluation, we applied quality thresholds—UAV to 3DEP point count ratio > 2 and at least 16,000 UAV LiDAR points and 200 3DEP points per tile—to exclude edge-of-flight tiles and extremely sparse cases, ensuring a meaningfully denser UAV reference and minimally informative 3DEP input for stable supervision. These thresholds served only as dataset curation filters and are not required at inference.

2.4. Data Augmentation

To increase the model's robustness and prevent overfitting, we expanded the training dataset from 24,000 to 40,000 tiles via data augmentation [49,50]. First, we preferentially selected source tiles for this process, prioritizing those with high structural complexity (z-variance) and large vegetation changes. Each selected source tile was then used to generate a new, augmented sample by applying a random combination of transformations. These included geometric operations (rotations and reflections) applied to all data layers and point cloud-specific perturbations (random point removal and jittering) applied only to the input LiDAR data.

2.5. Model Architecture

Our multi-modal upsampling framework transforms a sparse 3DEP point cloud, plus co-registered NAIP and UAVSAR image chips, into a denser 3D point cloud (Figure 5). The network is built around the *Local–Global Point Attention Block* (LG-PAB; Section 2.6), which provides permutation-invariant feature learning, optional feature-guided upsampling, and long-range geometric context. Five macro components are arranged in a feed-forward sequence: (1) point feature extraction, (2) imagery encoding, (3) cross-attention fusion, (4) feature expansion and refinement, and (5) point decoding.

Notation Conventions

Throughout this section, we use the following notations for tensor dimensions:

- Counts: N_{pts} (number of input points), $N_{\text{curr\_pts}}$ (current points in LG-PAB), $N_{\text{pts\_up}} = R_{\text{up}} \cdot N_{\text{pts}}$ (upsampled points), N_{patch} (number of image patches, default 16), N_{time} (temporal stack length), and N_{looks} (maximum look angles, $\leq$2);
- Dimensions: $D_{\text{coord}} = 3$ (coordinate dimension), D_{attr} (attribute dimension), $D_{\text{p\_feat}} = 256$ (point feature dimension), $D_{\text{p\_in}}$ (input point feature dimension), $D_{\text{p\_out}}$ (output point feature dimension), and $D_{\text{token}} = 128$ (token dimension);
- Channels: $C_{\text{naip}} = 4$ (NAIP channels), and $C_{\text{uavsar}} = 6$ (UAVSAR channels);
- Image dimensions: $H_{\text{naip}} = W_{\text{naip}} = 40$ (NAIP height/width), $H_{\text{uavsar\_pr}} = W_{\text{uavsar\_pr}} = 4$ (UAVSAR patch region height/width).
- Other: R_{up} (upsampling ratio, default 2), $k_{\text{knn}} = 16$ (k-nearest neighbors).

For tensors, we use the notation TensorSymbol: (dim_1, dim_2, ..., dim_k) to describe their dimensions.

Figure notation bridge. In Fig. 6 the generic diagram labels `in_features` and `out_features` (and the shortened `out_feat` on two arrows) correspond to $D_{\text{p\_in}}$ and $D_{\text{p\_out}}$ in the above notation. (We used Mermaid for the schematic, which does not support mathematical subscripts, so plain text placeholders were used.)

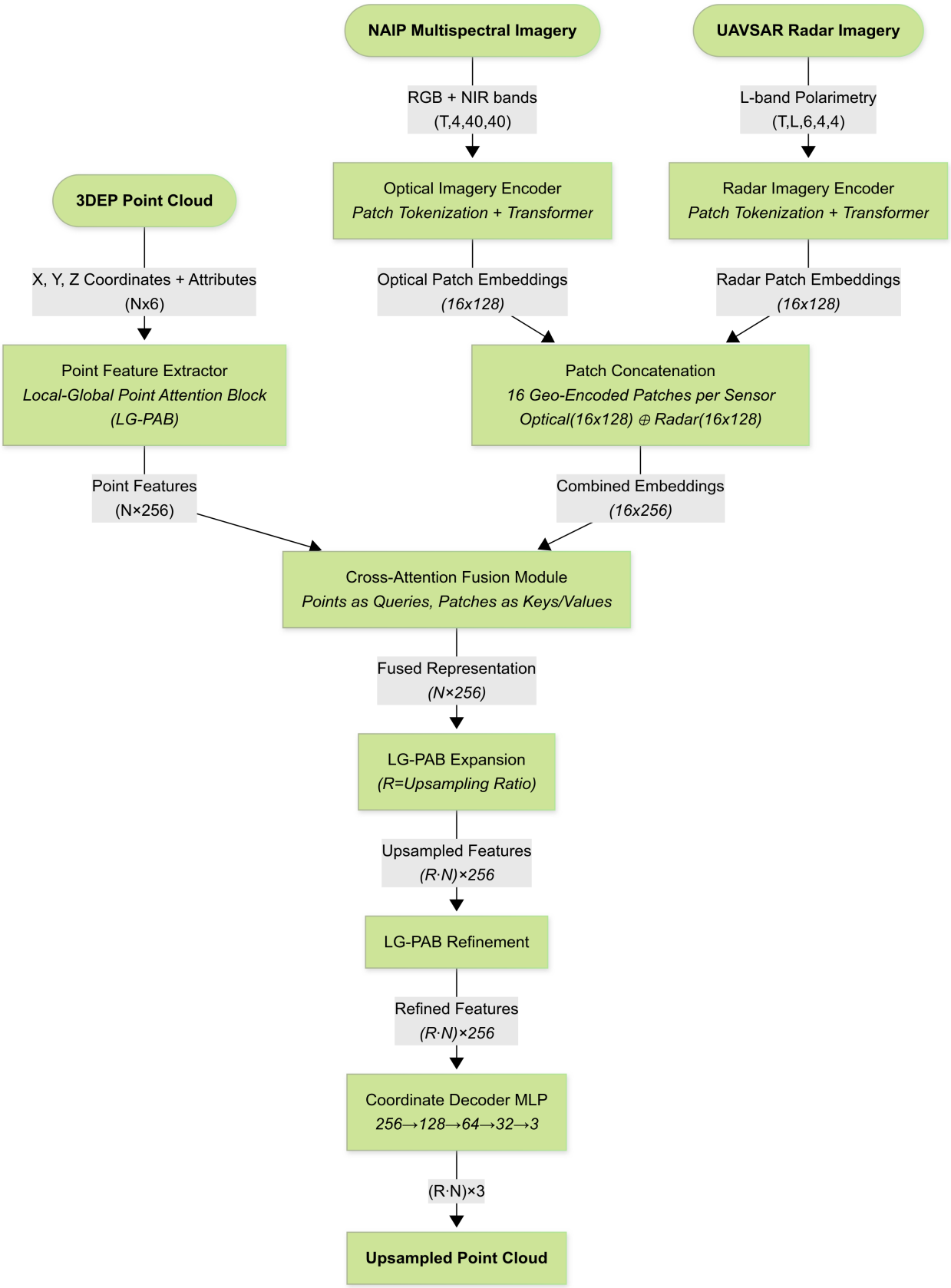

Figure 5. The overall multi-modal upsampling architecture. Key components include local–global point attention blocks, modality-specific imagery encoders for the processing of optical (NAIP) and radar (UAVSAR) data, and a cross-attention fusion module for combining imagery and LiDAR features.

Given an input cloud ($P_{\text{in}} : (N_{\text{pts}}, D_{\text{coord}})$) with points ($\{\mathbf{x}_i\}_{i=1}^{N_{\text{pts}}} \subset \mathbb{R}^{D_{\text{coord}}}$) and attributes ($\mathbf{a}_i : (D_{\text{attr}})$), the network outputs $P_{\text{out}} : (N_{\text{pts\_up}}, D_{\text{coord}})$ with points $\{\hat{\mathbf{x}}_j\}_{j=1}^{N_{\text{pts\_up}}}$, where $N_{\text{pts\_up}} = R_{\text{up}} \cdot N_{\text{pts}}$ (typically $R_{\text{up}} = 2$).

An overview is presented as follows:

1. **LG-PAB Extractor** $\mathcal{E}_{\text{pt}} \rightarrow$ local–global point features $F : (N_{\text{pts}}, D_{\text{p\_feat}})$;
2. **Imagery Encoders** $\mathcal{E}_{\text{opt}}, \mathcal{E}_{\text{rad}} \rightarrow$ NAIP and UAVSAR patch embeddings $E_{\text{opt}}, E_{\text{rad}} : (N_{\text{patch}}, D_{\text{token}})$;
3. **Cross-Attention Fusion** $\mathcal{F}_{\text{ca}} \rightarrow$ enriched point features $F_{\text{fused}} : (N_{\text{pts}}, D_{\text{p\_feat}})$;
4. **LG-PAB Expansion & Refinement** $\rightarrow$ upsampled features $F^{\uparrow} : (N_{\text{pts\_up}}, D_{\text{p\_feat}})$ and coordinates $P^{\uparrow} : (N_{\text{pts\_up}}, D_{\text{coord}})$;
5. **MLP Decoder** $\rightarrow$ residual offsets $\Delta P : (N_{\text{pts\_up}}, D_{\text{coord}})$ and final coordinates $P_{\text{out}} : (N_{\text{pts\_up}}, D_{\text{coord}})$.

Complexity, memory, and sequence length across stages.

Global multi-head attention remains quadratic in the number of points within a tile. The 10 m tiling and the input cap $N_{\text{max}} = 10{,}000$ fix the extractor's attention sequence length; with the feature-guided expansion set to $R_{\text{up}} = 2$, the largest sequence arises immediately after expansion and is at most $\approx 2 \times N_{\text{pts}} \leq 20{,}000$ before refinement and decoding. We implement global attention with FlashAttention, which reduces memory traffic while preserving exact attention. Under these bounds, we observed stable runtime and memory without requiring additional approximations or pruning.

2.6. The Local–Global Point Attention Block (LG-PAB)

Figure 6 presents a flow chart of the *Local–Global Point Attention Block*, the fundamental unit used three times in our architecture (extraction, expansion, and refinement stages). Each LG-PAB converts an input tuple $\langle X, P \rangle$ consisting of point features ($X : (N_{\text{curr\_pts}}, D_{\text{p\_in}})$) and 3D coordinates ($P : (N_{\text{curr\_pts}}, D_{\text{coord}})$) into refined (and optionally upsampled) features and positions. The block proceeds through the following stages that appear in the diagram:

1. **Local Attention Block:** A MULTI-HEAD POINT TRANSFORMER operates on a k_{knn}-nearest-neighbor graph ($k_{\text{knn}} = 16$) to capture fine-scale geometry. Its output passes through a two-layer FEED-FORWARD NETWORK (FFN) with GELU activation, producing an intermediate tensor ($Z : (N_{\text{curr\_pts}}, R_{\text{up}} \cdot D_{\text{p\_out}})$). When the upsampling ratio is $R_{\text{up}} = 1$, this step already delivers the final per-point features.

2 a **Feature-Guided Upsampling (optional):** If $R_{\text{up}} > 1$ (expansion stage), the intermediate tensor is reshaped to $[N_{\text{curr\_pts}}, R_{\text{up}}, D_{\text{p\_out}}]$, effectively cloning each feature vector R_{up} times. A small position-generator MLP then predicts a 3D offset for every clone, yielding new coordinates ($\hat{P} = P^{\text{rep}} + \Delta P : (N_{\text{curr\_pts}} \cdot R_{\text{up}}, D_{\text{coord}})$). The features are flattened back to $[N_{\text{curr\_pts}} \cdot R_{\text{up}}, D_{\text{p\_out}}]$.

2. **Global Attention Block:** To impose long-range coherence, the upsampled (or original) features are processed by POSITION-AWARE GLOBAL FLASH ATTENTION. Coordinates are first embedded by a two-layer MLP, concatenated to the features, and fed to a four-head FlashAttention layer that attends across *all* points in the tile. A second FFN refines the attended features, after which residual connections and LayerNorm complete the block.

2.7. Imagery Encoders

Optical (NAIP) and radar (UAVSAR) image chips are processed by a *shared five-stage encoder* (with modality-specific weights) that converts each image stack into a fixed set of patch tokens. The encoder stages—illustrated in Figure 7—are outlined as follows:

1. **Patch Tokenization:** A two-layer Conv–GELU–Conv stem with a stride of 1 followed by average pooling (stride=`patch_size`) extracts features on a 4×4 grid. This design mirrors shifted patch tokenization, which adds local texture bias to Vision Transformers (ViTs) and improves sample efficiency on small datasets [51]. (Conv1: $C_{in} \to D_{\text{token}}/2$, 3×3, stride 1, pad 1; Conv2: $D_{\text{token}}/2 \to D_{\text{token}}$, 3×3, stride 1, pad 1; AvgPool: kernel=stride=`patch_size` $= 10$.) The $N_{\text{patch}} = 16$ patches are flattened to $Z_{\text{tok}} : (N_{\text{patch}}, D_{\text{token}})$ and normalized via LayerNorm.
2. **2D Positional Encoding:** Normalized patch centers are embedded via a two-layer MLP and added to the tokens: $Z_{\text{pos}} = Z_{\text{tok}} + \text{MLP}_{\text{pos}}(x, y)$.
3. **Transformer Encoder Block.** A 4-head self-attention layer is paired with LayerScale ($\gamma \approx 10^{-5}$), a learnable scalar that stabilizes deep Transformers [52]. The accompanying MLP includes a depth-wise 1D convolution, following the ConViT approach to introduce a soft convolutional inductive bias [53]. MLP: $D_{\text{token}} \to 4D_{\text{token}}$ (depth-wise 1-D conv $k = 3$) $\to D_{\text{token}}$. The block outputs $Z_{\text{enc}} : (N_{\text{patch}}, D_{\text{token}})$.
4. **Temporal GRU–Attention Head.** For inputs with a temporal length of N_{time}, a bidirectional Gated Recurrent Unit (GRU) with attention pooling aggregates tokens into patch descriptors ($E_{\text{patch}} : (N_{\text{patch}}, D_{\text{token}})$).

Full per-component layer specifications (including fusion projections, LG-PAB FFNs, upsampling position generator, and coordinate decoder) appear in Appendix Table A1 for transparency and reproducibility.

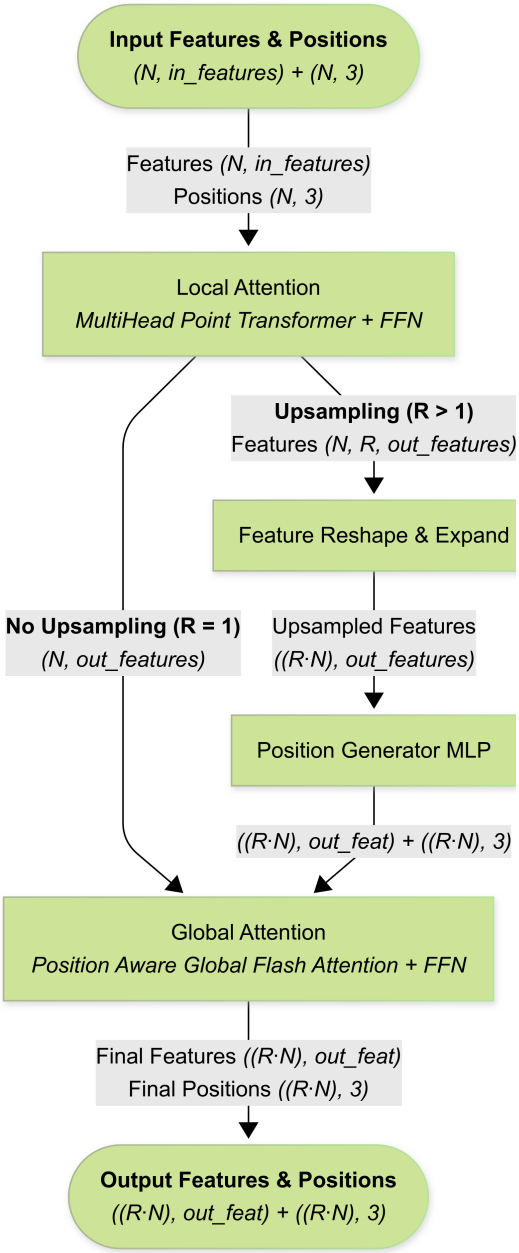

Figure 6. Flow chart of the Local–Global Point Attention Block (LG-PAB), the core computational unit used across the feature extraction, expansion, and refinement stages. The block applies local attention via a multi-head point Transformer, optional upsampling with learned position offsets, and global multi-head FlashAttentionfor broader spatial context. A feed-forward MLP follows both the local and global attention blocks.

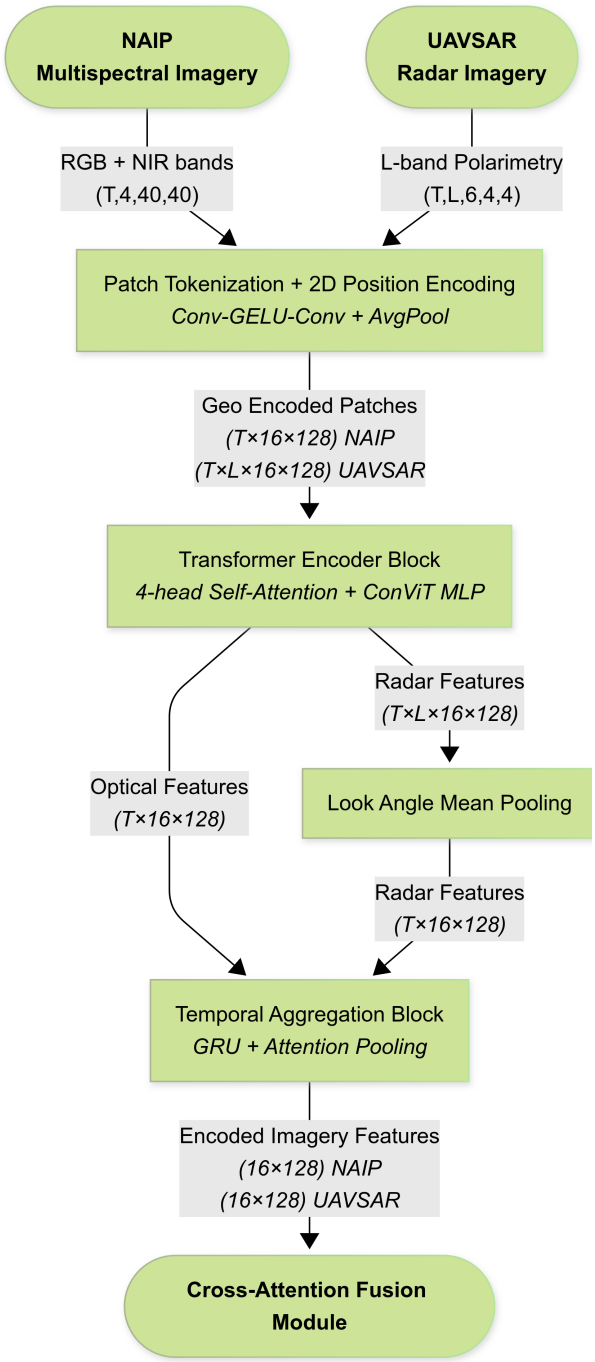

Figure 7. Architectural diagram of the imagery encoders. Image stacks from NAIP (optical) and UAVSAR (radar) pass sequentially through patch tokenization, Transformer encoder blocks for spatial context modeling, and a temporal GRU-Attention head for temporal aggregation.

UAVSAR Look-Angle Mean Pooling

- **UAVSAR.** Input $I_{\text{uavsar}} : (N_{\text{time}}, N_{\text{looks}}, C_{\text{uavsar}}, H_{\text{uavsar\_pr}}, W_{\text{uavsar\_pr}})$ with $N_{\text{looks}} \leq 2$. After the transformer encoder block, we average features across available look angles to produce $Z_{\text{avg}} : (N_{\text{time}}, N_{\text{patch}}, D_{\text{token}})$ before temporal processing. Otherwise, the pipeline matches NAIP.

Both encoders output token matrices (E_{opt} and $E_{\text{rad}} : (N_{\text{patch}}, D_{\text{token}})$) that are normalized and concatenated before fusion.

2.8. End-to-End Upsampling Pipeline

(1) Local–Global Feature Extraction.

The sparse 3DEP cloud—concatenated with intensity, return number, and number of returns (6 attributes in total)—is processed by the first *Local–Global Point Attention Block* (Section 2.6). The output is a set of point features ($F : (N_{\text{pts}}, D_{\text{p\_feat}})$) that encode both neighborhood morphology and tile-level context.

(2) Imagery Encoders.

Co-registered NAIP and UAVSAR chips are independently tokenized and fed to lightweight Transformer encoders (N_{patch} patches, embed dim D_{token}). The optical encoder (RGB + NIR) outputs patch embeddings ($E_{\text{opt}} : (N_{\text{patch}}, D_{\text{token}})$), whereas the radar encoder (L-band polarimetry) outputs $E_{\text{rad}} : (N_{\text{patch}}, D_{\text{token}})$. When multiple acquisition dates exist, per-modality tokens are fused temporally via a shared GRU–attention head.

(3) Cross-Attention Fusion.

Point features act as *queries*, and image patches act as *keys/values* in a four-head cross-attention block. Scaled dot-product scores are masked for patches with a centroid that is more than 8 m from the query point. The fused representation ($F_{\text{fused}} : (N_{\text{pts}}, D_{\text{p\_feat}})$) augments every point, with spectral texture and volumetric back-scatter cues improving discrimination of canopy surfaces versus gaps.

(4) Feature-Guided Upsampling.

A second LG-PAB with a ratio of $R_{\text{up}} = 2$ expands F_{fused} and predicts offsets (ΔP) that generate $N_{\text{pts\_up}}$ candidate coordinates. This stage doubles point density while maintaining local topology.

(5) Feature Refinement.

A third LG-PAB (ratio $R_{\text{up}} = 1$) re-computes local and global attention on the enlarged cloud, ironing out artifacts introduced during expansion and propagating context across newly formed neighborhoods.

(6) Coordinate Decoding.

Finally, a four-layer MLP ($D_{\text{p\_feat}} \to 128 \to 64 \to 32 \to D_{\text{coord}}$) regresses residual offsets that are added to the upsampled positions, producing the higher resolution prediction ($\hat{P} : (N_{\text{pts\_up}}, D_{\text{coord}})$).

2.9. Training Protocol and Analysis

To systematically evaluate our research questions, we established a rigorous training and analysis protocol. The full dataset, comprising tiles from both study areas, was partitioned into training, validation, and testing subsets. The training set consists of 24,000 original tiles and was expanded to 40,000 through the data augmentation process described previously. A separate set of 3792 tiles was reserved for validation during training, and a final hold-out set of 5688 tiles was used for testing and performance evaluation (Table 4).

Table 4. Dataset preparation

Subset	Tiles
Training	24,000 original + 16,000 augmented = 40,000
Validation	3792
Test	5688

To isolate the impact of each data modality, we trained four distinct model variants (Table 5). A baseline model was trained using only the sparse 3DEP LiDAR input. Two

single-modality models were trained by fusing the LiDAR with either NAIP optical imagery or UAVSAR radar data. Finally, a fully fused model was trained using all three data sources simultaneously.

Table 5. Model variants evaluated

oprule extbfVariant	**Active encoders**	**Cross-attention on**	**Parameters**
Baseline	LiDAR only	None	4.7M
+ NAIP	LiDAR, optical	NAIP tokens	5.9M
+ UAVSAR	LiDAR, radar	UAVSAR tokens	5.8M
Fused	LiDAR, optical, radar	Concatenated token	6.8M

Parameter differences relative to the baseline reflect only the additional modules enabled in each variant (e.g., imagery encoders and cross-attention fusion). The remaining architecture is unchanged to ensure a fair comparison.

All model variants were trained using an identical architecture, set of hyperparameters, and training protocol to ensure a fair comparison. Key model configuration parameters, including feature dimensions and attention head counts, are provided in Table 6. The models were trained for 100 epochs on four NVIDIA L40 GPUs using the ScheduleFreeAdamW optimizer [54] and a density-aware Chamfer distance loss function (Equation 9 in [55]). We set the loss hyperparameter α to 4, adapting the recommendation in [55] for our un-normalized, meter-scale data to prevent the loss function's exponential term from saturating and causing vanishing gradients. Further details on the training protocol and hardware are listed in Table 7.

Table 6. Model-configuration parameters

Parameter	**Value**	**Notes**
Core geometry		
Point-feature dimension	256	—
KNN neighbours (k)	16	Used in local attention graph
Upsampling ratio (R_{up})	2	Doubles point density per LG-PAB expansion
Point-attention dropout	0.02	Dropout inside global attention heads
Attention-head counts		
Extractor — local / global	8 / 4	Extra local heads help expand feature set
Expansion — local / global	8 / 4	Extra local heads aid point upsampling
Refinement — local / global	4 / 4	—
Imagery encoders		
Image-token dimension	128	Patch embeddings for NAIP & UAVSAR encoders
Cross-modality fusion		
Fusion heads	4	—
Fusion dropout	0.02	—
Positional-encoding dimension	36	—

We evaluated our research questions using non-parametric statistical tests to account for non-normality in the error distribution. For RQ1 and RQ2, we used Wilcoxon signed-rank tests to compare reconstruction error (measured by Chamfer distance) between models, with median percentage change and rank-biserial correlation as effect-size measures. For RQ3, we used Spearman rank correlations to analyze the relationship between reconstruction error and absolute change in canopy height across all models. We further split the dataset into canopy gains (N = 2423) and losses (N = 3264) to examine potential asymmetries in error patterns. Fisher r-to-z transformations were used to statistically compare correlation coefficients between models. All significance values are reported at α=0.05, with bold values indicating statistically significant results. For each tile, net canopy height change was calculated as the difference in the mean 95th percentile point height (z-value)

Table 7. Training protocol and hardware

Setting	Value
Hardware	4 × NVIDIA L40 (48 GB) GPUs under PyTorch DDP 2.5.1 (CUDA 12.4) DDP
Optimizer	*ScheduleFreeAdamW* [54]; base LR 5×10^{-4}, weight-decay 10^{-4}, $\beta_{1,2} = (0.9, 0.999)$; no external LR schedule
Loss function	Density-aware Chamfer distance (Equation 9 in [55]), $\alpha = 4$
Batch size	15 tiles per GPU
Epochs	100
Gradient clip	$\|g\|_2 \leq 10$
Training time	$\approx$ 7 h per model variant
Model selection	Epoch with lowest validation loss

across 2 m × 2 m grid cells between the UAV LiDAR and 3DEP clouds (N=5687; one tile removed due to a spurious below-ground return in 3DEP preventing canopy-change computation).

3. Results

A summary of the reconstruction performance, measured by Chamfer distance (CD), across all models is presented in Table 8.The 3DEP LiDAR-only baseline model yields a markedly lower CD and a tighter distribution than the raw, sparse input (Figure 8; Figure 9); however, because CD can vary with sampling density and point counts, we present this contrast as a qualitative soundness check rather than a normalized comparison. When comparing the performance across all models (Figure 10), the fused model consistently achieved the lowest median error and smallest interquartile range, indicating the most robust performance. To address our first research question (RQ1) on the impact of individual modalities, we compared the error distributions of the single-modality models against the baseline (Table 9).

Table 8. Descriptive statistics for Chamfer distance across all model variants (see Table 5).

Model	Mean CD (m)	Median CD (m)	Std Dev (m)	IQR (m)
Input	2.568	0.858	6.852	1.540
Baseline	1.043	0.340	5.717	0.465
NAIP	0.993	0.316	5.542	0.421
UAVSAR	0.924	0.331	5.505	0.437
Fused	0.965	0.298	5.753	0.393

Table 9. RQ1 and RQ2: Impact of single and fused modalities on reconstruction error.

Comparison	Median Change (%)	Effect Size
RQ1: Single Modality vs. Baseline		
NAIP vs. Baseline	**0.5** ($p < 0.001$)	**0.088**
UAVSAR vs. Baseline	**0.3** ($p < 0.001$)	**0.062**
RQ2: Fused Modality vs. Best Single Modality		
Fused vs. NAIP	**0.7** ($p < 0.001$)	**0.133**

Note: Bold values indicate statistical significance at $p \leq 0.05$

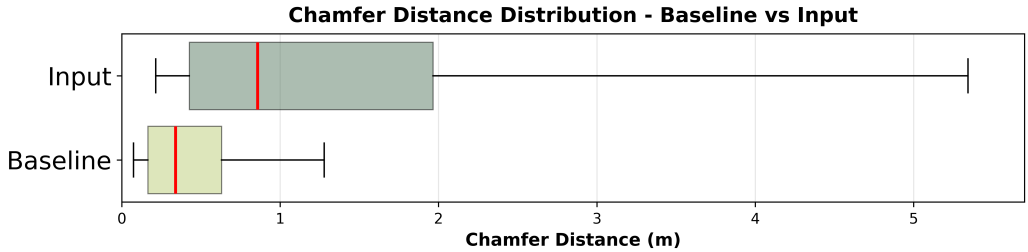

Figure 8. Distribution of Chamfer distance (CD) reconstruction errors comparing raw input data to the 3DEP LiDAR-only baseline model, each evaluated against the reference. Horizontal boxplots show the median (red line), interquartile range (colored boxes), and 10th–90th percentile whiskers; outliers beyond the 90th percentile are excluded. The baseline model shows a lower median CD (0.340 m vs. 0.858 m) with a tighter distribution. (Note: CD can vary with sampling density and point counts; since the input and predictions (2× upsampled) differ in size relative to the reference, these magnitudes are not a normalized comparison and are shown for qualitative context.)

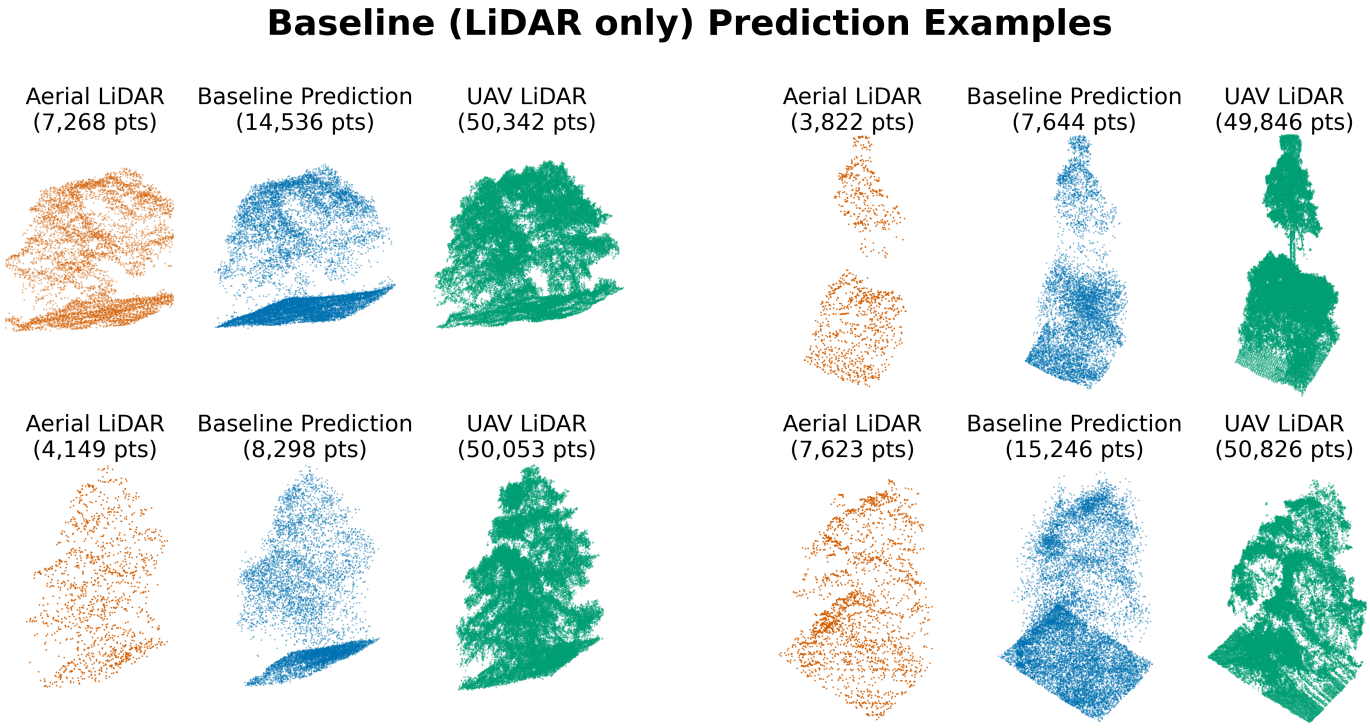

Figure 9. A comparison of baseline (3DEP LiDAR only) model output vs. input.

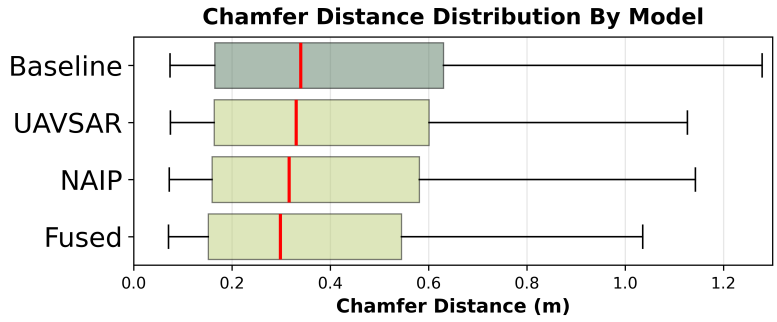

Figure 10. Distribution of Chamfer distance reconstruction errors across point-cloud upsampling models. Horizontal box plots show median (red line), interquartile range (colored boxes), and 10th–90th percentile whiskers for models trained with different input modalities. All models demonstrate substantial improvement over the LiDAR-only baseline, with the fused model achieving the lowest median error (0.298 m) and tightest distribution. Outliers beyond the 90th percentile are excluded for clarity.

Both high-resolution optical imagery (NAIP) and L-band SAR imagery significantly reduced reconstruction error compared to the LiDAR-only baseline, with optical imagery providing slightly larger improvements (0.5% vs. 0.3%). While statistically significant, the modest effect sizes (0.088 and 0.062 respectively) suggest that single-modality improvements over the baseline upsampling approach are limited in magnitude or may be concentrated in a limited number of tiles. Next, to evaluate our second research question (RQ2), we assessed whether fusing both imagery types yielded additional benefits over the best single-modality model (Table 9).

The fusion of both optical and SAR imagery yielded additional reconstruction accuracy gains (0.7% median reduction) beyond using NAIP alone, with a stronger effect size (0.133) than either individual modality achieved in RQ1. This statistically significant improvement supports our hypothesis that the two modalities contain complementary information that can be effectively combined through attention-based fusion. The results demonstrate that multi-modality approaches can leverage different sensing capabilities to achieve superior point-cloud reconstruction (Figure 11).

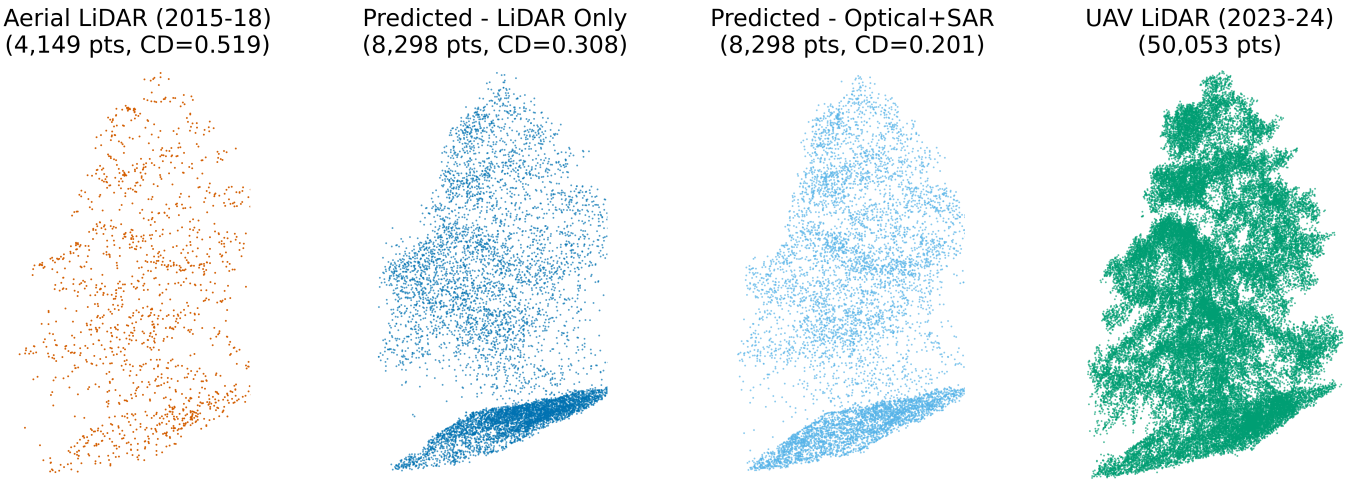

Figure 11. Example tile where no major vegetation change occurred between the 3DEP aerial LiDAR (2015-2018) and UAV LiDAR (2023-2024). The optical + SAR fusion model more accurately recovers fine-scale canopy structure compared to the LiDAR-only model, producing a point cloud that more closely matches the UAV LiDAR reference (lower Chamfer distance).

Finally, to evaluate our third research question (RQ3) on the impact of vegetation change, we correlated model error with the magnitude of canopy height changes between the legacy 3DEP and recent UAV LiDAR surveys. Table 10 shows the overall Spearman rank correlations. To investigate this relationship further, we split the dataset into areas of net canopy gain and net canopy loss, with the results of this extended analysis presented in Table 11.

Table 10. RQ3: Correlation between reconstruction error and canopy height change.

Model	Spearman ρ	p-Value
Baseline	**0.650**	$p < 0.001$
NAIP	**0.612**	$p < 0.001$
UAVSAR	**0.628**	$p < 0.001$
Fused	**0.582**	$p < 0.001$
Baseline vs. NAIP (z)	**3.377**	$p < 0.001$
Baseline vs. UAVSAR (z)	**1.999**	$p = 0.046$
Baseline vs. Fused (z)	**5.868**	$p < 0.001$

Note: Bold values indicate statistical significance at $p \leq 0.05$.

Table 11. RQ3 extended: Correlation between reconstruction error and canopy height changes (gains vs. losses).

Model	Canopy Gains (N = 2423)		Canopy Losses (N = 3264)	
	Spearman ρ	p-Value	Spearman ρ	p-Value
Baseline	**0.601**	$p < 0.001$	**0.671**	$p < 0.001$
NAIP	**0.586**	$p < 0.001$	**0.621**	$p < 0.001$
UAVSAR	**0.597**	$p < 0.001$	**0.637**	$p < 0.001$
Fused	**0.587**	$p < 0.001$	**0.580**	$p < 0.001$
Baseline vs. NAIP (z)	0.825	$p = 0.409$	**3.440**	$p < 0.001$
Baseline vs. UAVSAR (z)	0.233	$p = 0.816$	**2.406**	$p = 0.016$
Baseline vs. Fused (z)	0.768	$p = 0.442$	**6.074**	$p < 0.001$

Note: Bold values indicate statistical significance at $p \leq 0.05$.

All models showed strong correlations with absolute canopy height change, confirming that reconstruction error systematically increases with vegetation structure changes since to the original LiDAR collection. The baseline model exhibited the strongest correlation with canopy change ($\rho = 0.650$), while the fused model showed the weakest correlation ($\rho = 0.582$), with this difference being statistically significant ($z = 5.868$, $p < 0.001$). Importantly, the extended analysis revealed that this pattern was driven primarily by canopy losses, where the baseline model performed significantly worse than all other models, particularly the fusion approach ($z = 6.074$, $p < 0.001$), while for canopy gains, all models performed similarly, without statistically significant differences. These findings partially support our hypothesis that advanced models better mitigate error from canopy changes—specifically for canopy removal scenarios, where legacy LiDAR contains no information about the removed vegetation (Figures 12 and 13). For canopy gains the limited separation is consistent with our original reasoning: most growth cases represent incremental vertical or volumetric accretion within an existing structural envelope already partly encoded in the sparse legacy LiDAR, allowing even the baseline to extrapolate plausibly. Ancillary optical and SAR cues add comparatively little discriminative signal for these modest positive changes, whereas abrupt canopy removal creates a true information void that fused imagery helps fill.

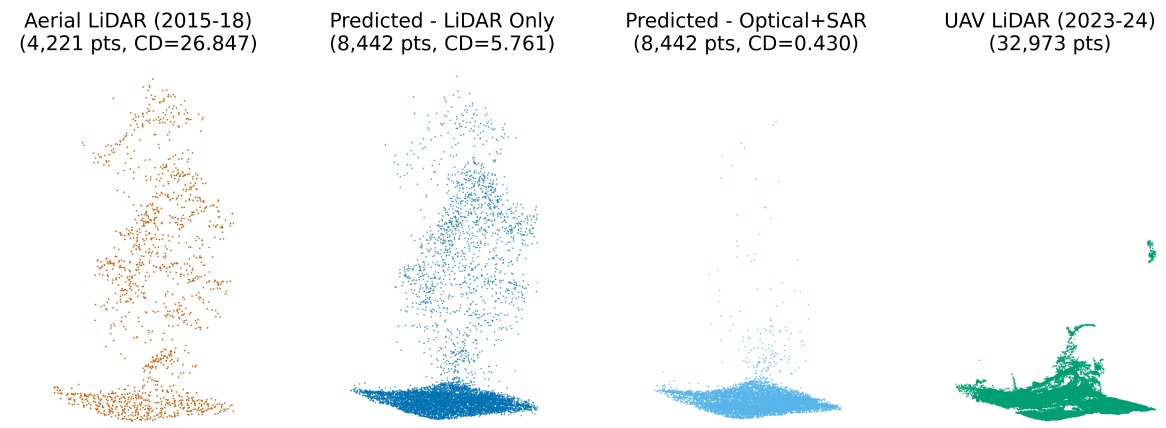

Figure 12. Example tile illustrating vegetation structure change between legacy aerial LiDAR (2015-2018) and recent UAV LiDAR (2023-2024). The optical + SAR fusion model accurately reconstructs the canopy loss visible in the UAV LiDAR reference, whereas the LiDAR-only model retains outdated structure from the earlier survey. This highlights the value of multi-modal imagery in correcting legacy LiDAR and detecting structural change.

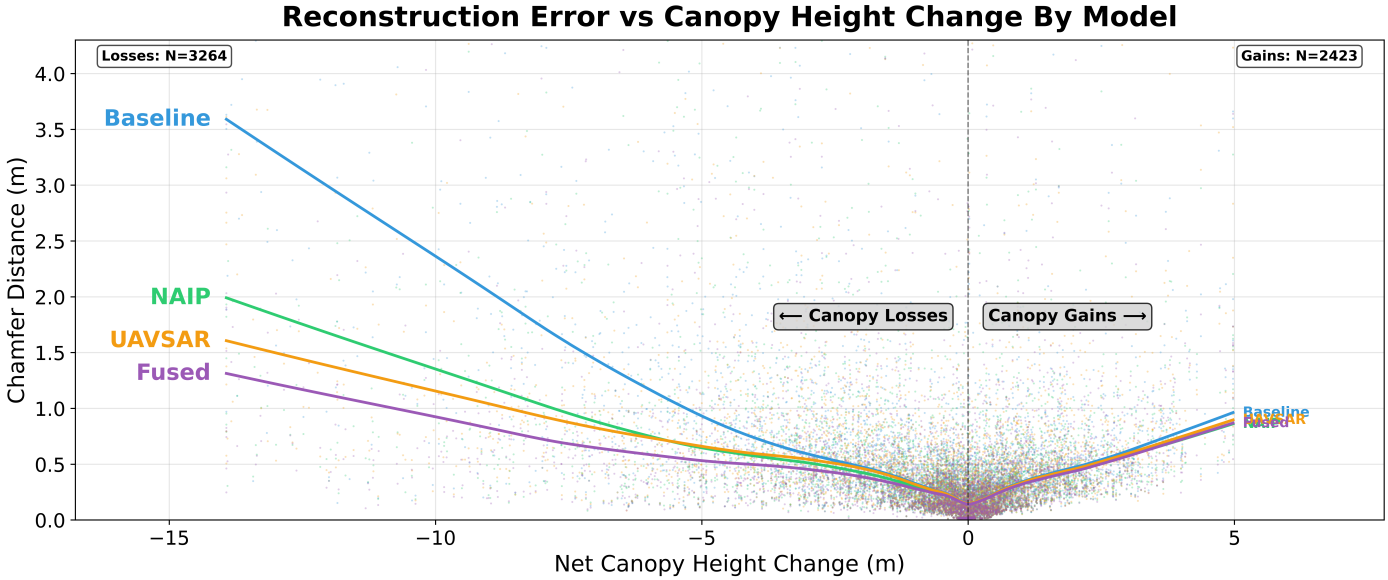

Figure 13. Relationship between point-cloud reconstruction error (Chamfer distance) and net canopy height change since the original LiDAR survey. Scatter points show individual sample tiles (N = 5687) with LOWESS trend lines for each model; outliers above the 99.5th percentile for both error and height change metrics were excluded for visual clarity. Negative values represent canopy losses, while positive values represent gains. The baseline model (blue) shows substantially higher error rates for large canopy losses compared to models incorporating additional remote sensing data (NAIP optical, UAVSAR radar, and fused approaches), while all models perform similarly for canopy gains.

4. Discussion

This study demonstrates the significant potential of attention-based deep learning models to enhance sparse and outdated airborne LiDAR point clouds by fusing them with more frequently acquired optical (NAIP) and synthetic aperture radar (UAVSAR) imagery. Our findings directly address the critical need for up-to-date, high-resolution 3D vegetation structure information in applications like wildfire risk modeling and ecological monitoring.

The core success of our approach lies in the effective fusion of multi-modal data. As hypothesized (RQ1), both NAIP optical imagery and L-band UAVSAR imagery, when individually integrated, improved point-cloud reconstruction accuracy compared to a

baseline model relying solely on sparse LiDAR. NAIP, with its finer spatial resolution, offered a slightly greater enhancement, likely due to its ability to delineate canopy edges and small gaps with high fidelity. However, the true advancement was observed when these modalities were combined (RQ2). The fused model, leveraging both NAIP's textural detail and UAVSAR's structural sensitivity, outperformed single-modality enhancements. This confirms our hypothesis that these sensors provide complementary rather than redundant information and that the cross-attention mechanisms within our architecture can effectively identify and leverage these synergistic relationships. This synergy is particularly valuable for capturing the complex, heterogeneous nature of vegetation.

Our investigation into temporal dynamics (RQ3) revealed that all models, including the baseline, exhibited increased reconstruction error in areas with substantial canopy change since the initial LiDAR survey. This is expected, as the input LiDAR reflects a past state. However, the fused model demonstrated the most robust performance, showing a significantly weaker correlation between error and the magnitude of canopy change, especially in areas of canopy loss. This is a crucial finding: by incorporating more recent imagery, particularly optical data that clearly depicts vegetation absence, the model can more effectively correct for outdated LiDAR information. The baseline model, lacking this current-state information, struggled most in loss scenarios. Even in areas of canopy growth, while not as pronounced as with loss, the image-informed models offered an advantage by providing cues about new or denser vegetation that the original sparse LiDAR could not capture. This capacity to "update" historical LiDAR datasets significantly enhances their utility for long-term monitoring and management, especially in landscapes prone to rapid changes from disturbances or growth.

The developed Local–Global Point Attention Block (LG-PAB) proved to be a robust architectural component. Its ability to capture both fine-grained local details through neighborhood-level self-attention and broader structural coherence via global attention across the entire point patch is central to its success. This hierarchical attention is well-suited to the fractal-like patterns often observed in natural vegetation.

Despite these promising results, some limitations exist. The improvements, while statistically significant, were modest in terms of percentage change in Chamfer distance. This suggests that while fusion helps, there might be inherent limits to the upsampling of very sparse LiDAR or that the current metrics may not fully capture all aspects of structural improvement relevant to ecological applications.

We acknowledge that perfect cross-sensor alignment is not guaranteed in practice. We deliberately selected established, well-orthorectified public datasets (3DEP, NAIP, UAVSAR) to minimize gross misregistration, and we chose point-to-image cross-attention (with a modest spatial proximity mask) rather than direct concatenation of co-located patch features partly because attention is more tolerant to small horizontal offsets and timing differences among inputs: it allows each point to attend to the most informative nearby patch rather than a single fixed location. We did not conduct a formal misregistration sensitivity study, which we identify as future work.

More broadly, our experiments and training were conducted in Southern California's Mediterranean ecosystems; model performance—and the relative contributions of optical versus radar—may differ in other biomes with distinct vegetation structure and phenology. Accordingly, out-of-domain application without additional high-density LiDAR may degrade performance; however, incorporating UAV LiDAR from new biomes into training is a straightforward path to broader generalization.

Beyond the primary 2× upsampling task, we also conducted a preliminary investigation into the model's potential for higher ratio densification (8×). This exploratory model was scaled up significantly, with larger feature dimensions (512 points and 192 images),

more local attention heads for extraction and expansion (16), and additional LG-PAB layers (six total) , and a total of 125 M parameters. With our hardware, this increased model size necessitated reducing the batch size to one per GPU. The results (Figure 14) show that the architecture can, indeed, produce highly dense outputs that qualitatively approach the reference data. Appendix A provides further qualitative examples from both the standard 2× and experimental 8× models, illustrating their performance across different vegetation change scenarios (Figures A1–A3).

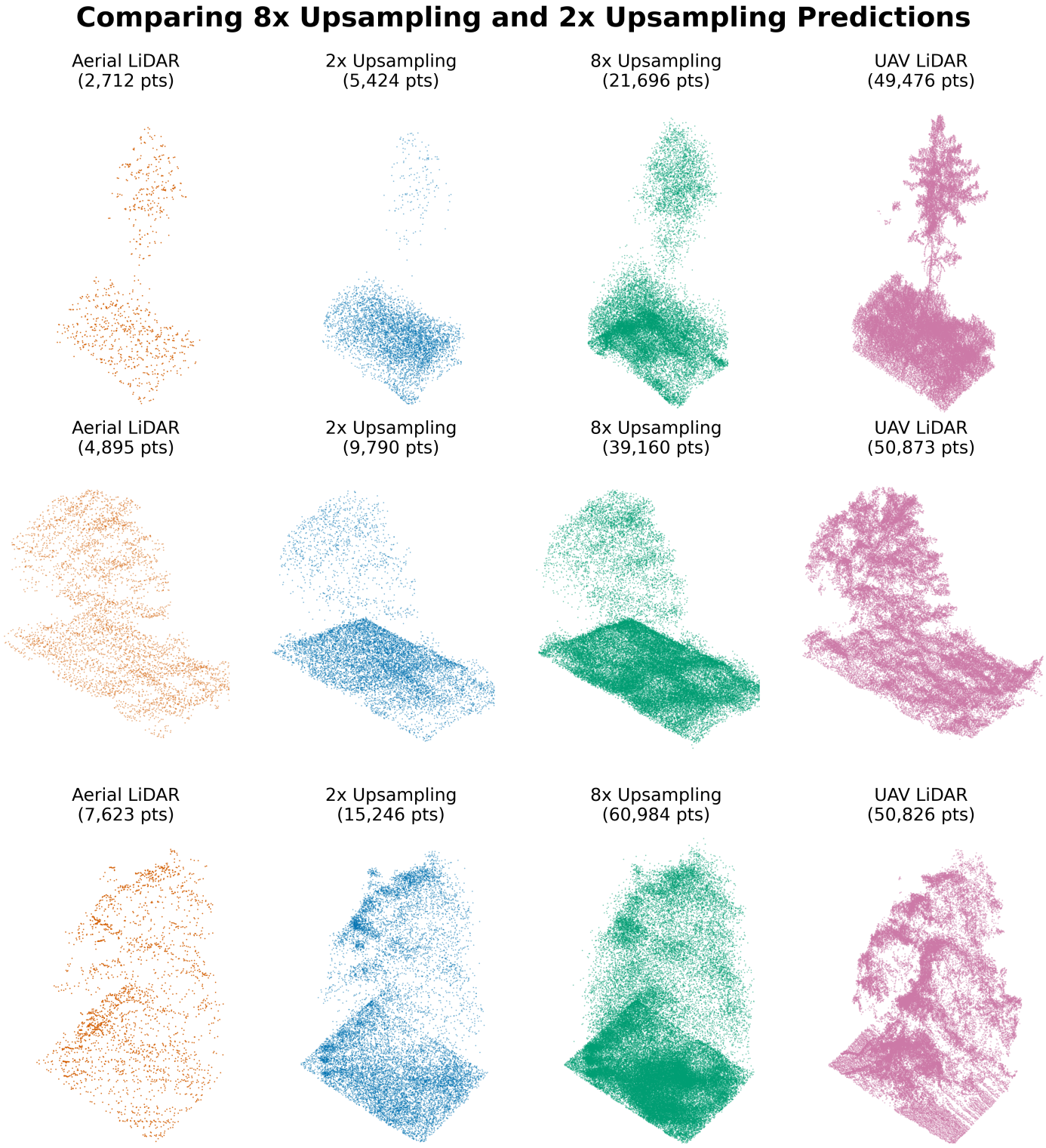

Figure 14. Comparison of the standard 2× upsampling model (optical + SAR, 6.8 million parameters) output versus a preliminary high-density 8× model (optical + SAR, 125 million parameters).

A recommended area for future research involves adapting the validated multi-modal feature extraction pipeline for direct prediction of key vegetation structure rasters, such as Canopy Height Models (CHMs), canopy cover, Above-Ground Biomass (AGB), and fuel types. Such an adaptation would entail replacing the current point-cloud generation head with task-specific regression or classification heads, potentially broadening the practical applicability of this work. Integrating geometric priors like Digital Surface Models (DSMs) and Digital Terrain Models (DTMs) into the loss function also represents a valuable direction. This could not only enforce greater structural realism but also enable the calculation of reconstruction error at various canopy strata, offering deeper insights into model performance. Furthermore, fine-tuning emerging foundation-model vision Transformers, such as the Clay model [56], as shared encoders for NAIP and UAVSAR imagery warrants exploration to leverage large-scale pretraining for enhanced feature representation. Complementary investigations could include a thorough evaluation of the UAVSAR encoder, particularly with respect to the optimization of multi-look fusion beyond simple averaging and assessment of resampling impacts, alongside ablation studies on the LG-PAB and imagery encoders to pinpoint key architectural contributions and guide further optimization. Future architectural research could also explore simplifying the Local–Global Point Attention Block into a Pure Point Attention Block by replacing the k-NN local attention with a global attention module. For greater scalability, this could be paired with a latent attention strategy [57] to bypass the quadratic complexity inherent to self-attention.

5. Conclusions

This research successfully demonstrates that attention-based deep learning, leveraging our novel Local–Global Point Attention Block, can significantly enhance sparse airborne LiDAR point clouds in vegetated landscapes through the fusion of more recent optical and radar imagery. We have shown that while individual imagery modalities provide benefits, their combined use yields superior reconstruction accuracy, particularly in mitigating errors arising from vegetation changes over time. Specifically, high-resolution optical imagery (NAIP) proved slightly more effective as a standalone ancillary dataset than L-band SAR within our framework, but the fusion of both offered the best performance, validating the complementary nature of these sensors. A key contribution is the model's ability to substantially reduce reconstruction degradation in areas of vegetation loss, thereby increasing the utility of historical LiDAR archives. This study presents a novel approach to direct 3D point-cloud upsampling using multi-modal fusion in complex natural environments, moving beyond prevalent raster-based enhancement techniques and paving the way for more accurate and timely assessments of vegetation structure.

Author Contributions: Conceptualization, M.M. and D.S.; methodology, M.M.; software, M.M.; validation, M.M.; formal analysis, M.M.; investigation, M.M. and D.S.; resources, D.S. and J.F.; data curation, M.M.; writing—original draft preparation, M.M.; writing—review and editing, D.S. and J.F.; visualization, M.M.; supervision, D.S. and J.F.; project administration, D.S. and J.F.; funding acquisition, D.S. and J.F. All authors have read and agreed to the published version of the manuscript.

Funding: Funding for the research was provided by California Climate Action Matching Grants, University of California, Office of the President (grant #R02CM708 to M. Jennings, SDSU (PI) and JF). DS gratefully acknowledges funding from the NASA FireSense program (Grant # 80NSSC24K0145), the NASA FireSense Implementation Team (Grants #80NSSC24K1320),the NASA Land-Cover/Land Use Change program (Grant #NNH21ZDA001N-LCLUC), the EMIT Science and Applications Team program (Grant #80NSSC24K0861), the NASA Remote Sensing of Water Quality program (Grant #80NSSC22K0907), the NASA Applications-Oriented Augmentations for Research and Analysis Program (Grant #80NSSC23K1460), the NASA Commercial Smallsat Data Analysis Program (Grant

#80NSSC24K0052), the USDA NIFA Sustainable Agroecosystems program (Grant #2022-67019-36397), the USDA AFRI Rapid Response to Extreme Weather Events Across Food and Agricultural Systems program (Grant #2023-68016-40683), the California Climate Action Seed Award Program, and the NSF Signals in the Soil program (Award #2226649).

Data Availability Statement: All Python code used for data collection, preprocessing, model training, and analysis is publicly available on GitHub at https://github.com/mmarks13/geoai_veg_map The repository uses the Poetry package manager for Python dependency management; all required libraries and their exact versions are listed in the root-level pyproject.toml and locked in 'poetry.lock'' for full reproducibility. The 3DEP airborne LiDAR and NAIP optical imagery used as model inputs were sourced via Microsoft's Planetary Computer (https://planetarycomputer.microsoft.com/) using the provided code. The UAVSAR radar imagery was sourced via the Alaska Satellite Facility (https://search.asf.alaska.edu/) using the provided code. The UAV LiDAR point clouds used as reference data for model training and evaluation are available through OpenTopography (https://opentopography.org/); specific DOIs will be provided upon acceptance or can be requested from the authors. The exact data stacks (input features and reference point clouds) used for model training and evaluation are available from the corresponding author upon reasonable request.

Acknowledgments: The authors wish to express their sincere gratitude to Lloyd L. ("Pete") Coulter (Center for Earth Systems Analysis Research, Department of Geography, SDSU) for his skillful piloting and management of all UAV LiDAR data acquisition campaigns; this data was foundational to the research presented. The authors also thank two anonymous reviewers for constructive comments which helped improve the clarity of the final manuscript. Separately, during the preparation of this manuscript, generative AI models (including Google's Gemini, Anthropic's Claude, and OpenAI's ChatGPT) were utilized to assist with language refinement, conceptual brainstorming, generating preliminary code structures, and code debugging. The authors have reviewed and edited all AI-generated suggestions and take full responsibility for the content of this publication.

Conflicts of Interest: The authors declare no conflicts of interest. The funders had no role in the design of the study; in the collection, analyses, or interpretation of data; in the writing of the manuscript; or in the decision to publish the results.

Abbreviations

The following abbreviations are used in this manuscript:

3DEP	3D Elevation Program
C-ALS	Crewed Airborne LiDAR
CAD	Computer-Aided Design
CD	Chamfer Distance
DSM	Digital Surface Model
FFN	Feed-Forward Network
GSD	Ground-Sample Distance
GRU	Gated Recurrent Unit
LG-PAB	Local–Global Point Attention Block
LiDAR	Light Detection and Ranging
MLP	Multi-Layer Perceptron
NAIP	National Agriculture Imagery Program
RQ1	Research Question 1
SAR	Synthetic Aperture Radar
UAV	Unmanned Aerial Vehicle
UAVSAR	Uninhabited Aerial Vehicle Synthetic Aperture Radar
USGS	U.S. Geological Survey
ViT	Vision Transformer

Appendix A

Table A1. Detailed layer specifications not fully enumerated in the main text. All experiments use $D_{\text{p\_feat}} = 256$, $D_{\text{token}} = 128$, upsampling ratio $R_{\text{up}} = 2$, and KNN $k = 16$.

Component	Specification
Patch Conv Stem	Conv1 $C_{in} \to D_{tok}/2$ (3×3, s=1,p=1) – GELU – Conv2 $D_{tok}/2 \to D_{tok}$ (3×3, s=1,p=1) – AvgPool (k=stride=10) – LayerNorm.
Transformer Encoder MLP	$D_{tok} \to 4D_{tok}$ – depth-wise 1-D conv (k=3) – GELU – $\to D_{tok}$ (LayerScale $\gamma \approx 10^{-5}$; 4 heads self-attn).
Temporal GRU Head	Bi-GRU hidden=D_{tok} (fwd+bwd) – attention pooling (Linear $2D_{tok} \to 1$).
Cross-Attention Fusion	Point query proj: Linear $256 \to 256$; NAIP/UAVSAR key value proj: Linear $(128 + pos_{patch}) \to 256$ (pos enc dim 36); multi-head (4) scaled dot-product; post-concat MLP: Linear $C \to C$ – GELU – Linear $C \to 256$ + residual + LayerNorm.
LG-PAB Local Attention	Multi-head PointTransformerConv (8 heads extractor, 8 expansion, 4 refinement) over KNN graph ($k = 16$); per-head out dim 256/heads; FFN after local attn: $256 \to 512 \to 256$ (GELU).
LG-PAB Global Attention	Position MLP $3 \to 32 \to 32$; concatenate with features; multi-head global FlashAttention (4 heads all stages); FFN: $256 \to 512 \to 256$ (GELU).
Feature-Guided Upsampling	Intermediate reshape $[N, R, 256]$; Position-Generator MLP $256 \to 64 \to 32 \to 3$ (GELU); feature clones flattened to $(R \cdot N, 256)$.
Coordinate Decoder	MLP $256 \to 128 \to 64 \to 32 \to 3$ (GELU) producing residual offsets.
Normalization	LayerNorm after major residual joins (fusion output; after local and global attention outputs).
Dropout	Fusion attention: 0.02; point global attention: 0.02 (no dropout in local PointTransformerConv).
Positional Encodings	Global point attention position MLP (32 dims); patch sinusoidal encoding dim 36.
Optimizer	ScheduleFreeAdamW (shared across variants; see main text Table 7).

Appendix visualization note: All quantitative metrics in the manuscript are derived from the production $2\times$ upsampling models. The two vegetation growth examples below use an experimental $8\times$ model *only* to improve visual legibility of subtle new canopy structure (higher point density reveals incremental infill); these outputs are illustrative and not part of reported statistics. The vegetation loss example is shown with the production $2\times$ fusion model to reflect the evaluated configuration.

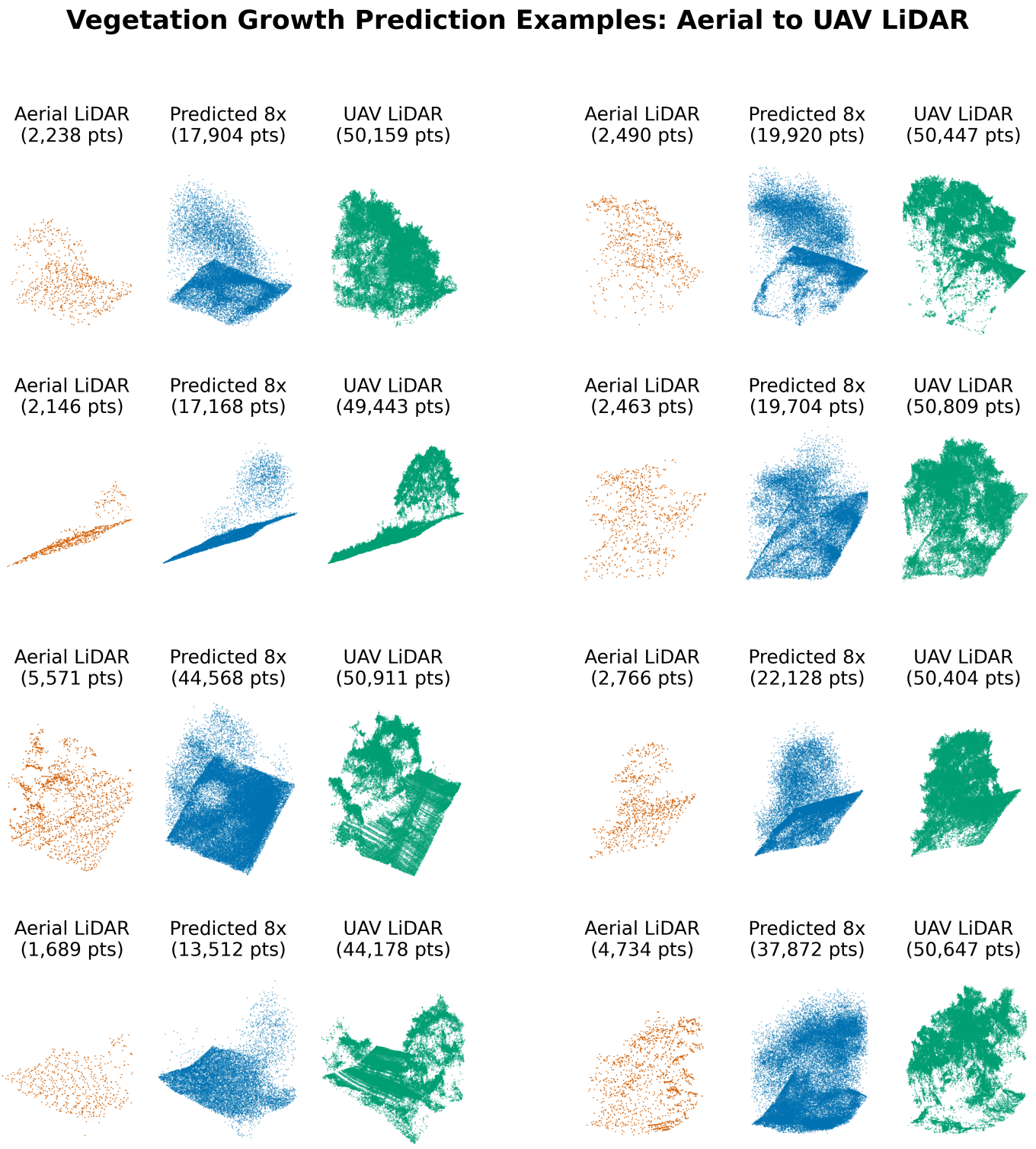

Figure A1. Example of vegetation growth reconstruction (experimental 8× model for illustrative visual clarity only; all reported metrics use the production 2× model). The model successfully infers new canopy structure (center) that is absent in the input 3DEP LiDAR but present in the recent UAV LiDAR reference.

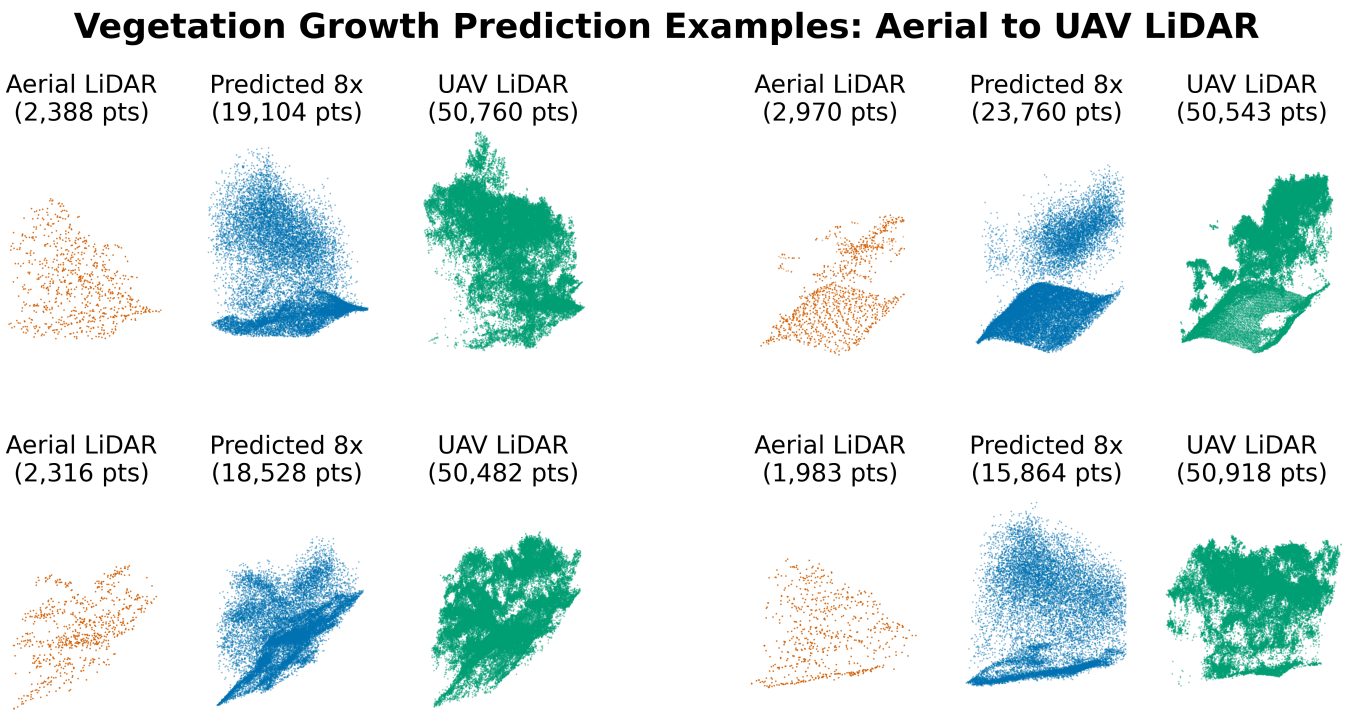

Figure A2. Second vegetation growth example (experimental 8× visualization only; quantitative results use the 2× model)

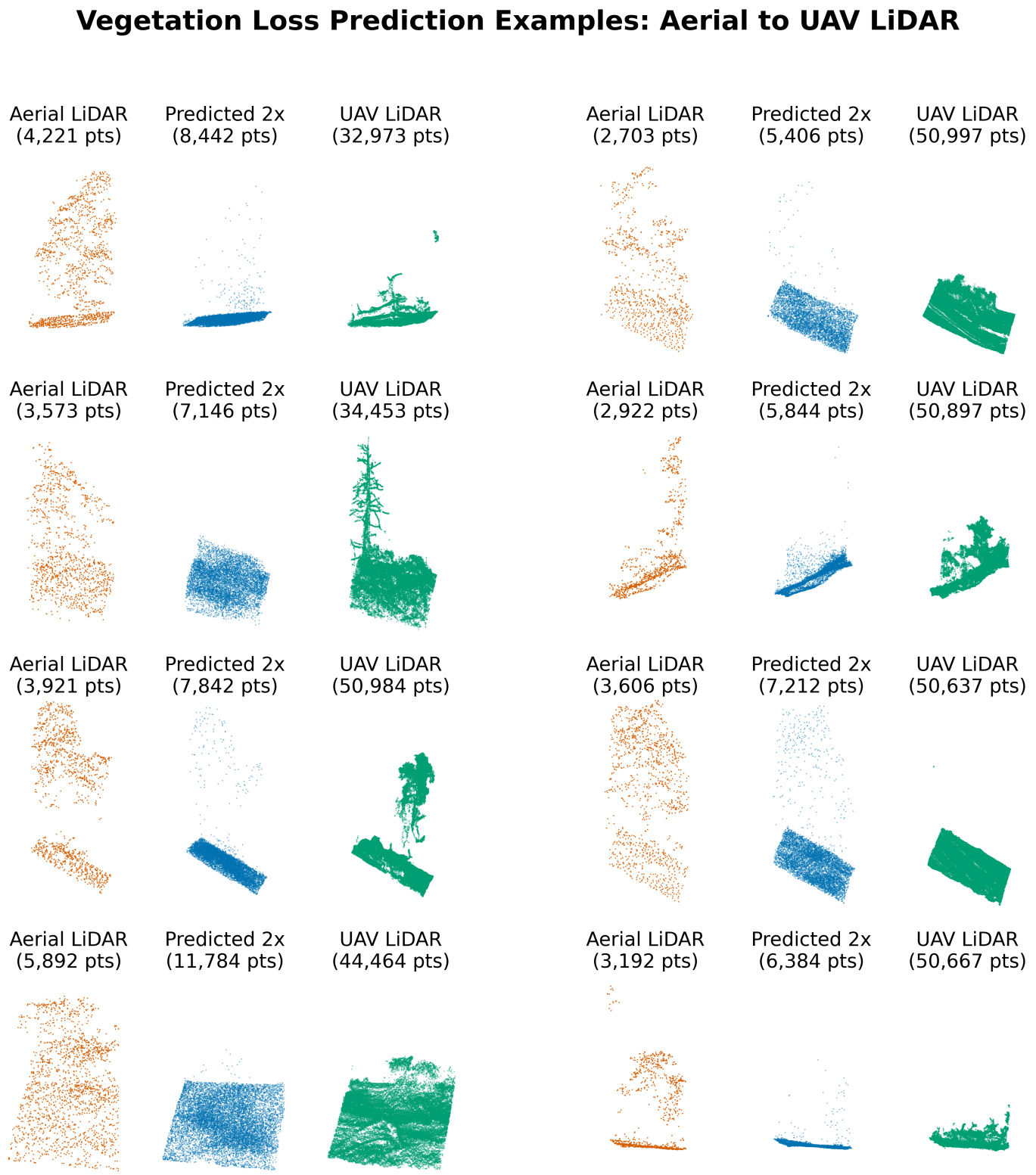

Figure A3. Vegetation loss example using the production 2× fusion model (configuration used for all quantitative evaluation). The model removes vegetation present in the outdated input LiDAR (left) to match the state shown in the recent UAV LiDAR reference (right), highlighting the value of multi-modal fusion for change detection.

References

1. Martin-Ducup, O.; Dupuy, J.L.; Soma, M.; Guerra-Hernandez, J.; Marino, E.; Fernandes, P.M.; Just, A.; Corbera, J.; Toutchkov, M.; Sorribas, C.; et al. Unlocking the potential of Airborne LiDAR for direct assessment of fuel bulk density and load distributions for wildfire hazard mapping. *Agric. For. Meteorol.* **2025**, *362*, 110341. https://doi.org/10.1016/j.agrformet.2024.110341.

2. Merrick, M.J.; Koprowski, J.L.; Wilcox, C. Into the Third Dimension: Benefits of Incorporating LiDAR Data in Wildlife Habitat Models. In *Merging Science and Management in a Rapidly Changing World: Biodiversity and Management of the Madrean Archipelago III*; USDA Forest Service Proceedings RMRS-P-67; U.S. Department of Agriculture, Forest Service, Rocky Mountain Research Station: Fort Collins, CO, USA, 2013; pp. 389–395.
3. Moudrỳ, V.; Cord, A.F.; Gábor, L.; Laurin, G.V.; Barták, V.; Gdulová, K.; Malavasi, M.; Rocchini, D.; Stereńczak, K.; Prošek, J.; et al. Vegetation structure derived from airborne laser scanning to assess species distribution and habitat suitability: The way forward. *Divers. Distrib.* **2023**, *29*, 39–50.
4. Agee, J.K. The Influence of Forest Structure on Fire Behavior. In Proceedings of the 17th Annual Forest Vegetation Management Conference, Redding, CA, USA, 16–18 January 1996; pp. 52–68.
5. Guo, X.; Coops, N.C.; Gergel, S.E.; Bater, C.W.; Nielsen, S.E.; Stadt, J.J.; Drever, M. Integrating airborne lidar and satellite imagery to model habitat connectivity dynamics for spatial conservation prioritization. *Landsc. Ecol.* **2018**, *33*, 491–511.
6. Mahata, A.; Panda, R.M.; Dash, P.; Naik, A.; Naik, A.K.; Palita, S.K. Microclimate and vegetation structure significantly affect butterfly assemblages in a tropical dry forest. *Climate* **2023**, *11*, 220.
7. Ustin, S.L.; Middleton, E.M. Current and near-term advances in Earth observation for ecological applications. *Ecol. Processes* **2021**, *10*, 1.
8. Belov, M.; Belov, A.; Gorodnichev, V.; Alkov, S. Capabilities analysis of lidar and passive optical methods for remote vegetation monitoring. *J. Phys. Conf. Ser.* **2019**, *1399*, 055024.
9. Guo, Q.; Su, Y.; Hu, T.; Guan, H.; Jin, S.; Zhang, J.; Zhao, X.; Xu, K.; Wei, D.; Kelly, M.; et al. Lidar boosts 3D ecological observations and modelings: A review and perspective. *IEEE Geosci. Remote Sens. Mag.* **2020**, *9*, 232–257.
10. Wu, Z.; Dye, D.; Stoker, J.; Vogel, J.; Velasco, M.; Middleton, B. Evaluating LiDAR point densities for effective estimation of aboveground biomass. *Int. J. Adv. Remote Sens. GIS* **2016**, *5*, 1483–1499.
11. USGS. *What Is 3DEP?*; U.S. Geological Survey: Reston, VA, USA, 2019.
12. USDA. *National Agriculture Imagery Program—NAIP Hub Site*; USDA: Washington, DC, USA, 2024.
13. Wang, C.; Song, C.; Schroeder, T.A.; Woodcock, C.E.; Pavelsky, T.M.; Han, Q.; Yao, F. Interpretable Multi-Sensor Fusion of Optical and SAR Data for GEDI-Based Canopy Height Mapping in Southeastern North Carolina. *Remote Sens.* **2025**, *17*, 1536.
14. Rosen, P.A.; Hensley, S.; Wheeler, K.; Sadowy, G.; Miller, T.; Shaffer, S.; Muellerschoen, R.; Jones, C.; Zebker, H.; Madsen, S. UAVSAR: A new NASA airborne SAR system for science and technology research. In Proceedings of the 2006 IEEE Conference on Radar, Verona, NY, USA, 24–27 April 2006; IEEE: Piscataway, NJ, USA, 2006; p. 8.
15. Kellogg, K.; Hoffman, P.; Standley, S.; Shaffer, S.; Rosen, P.; Edelstein, W.; Dunn, C.; Baker, C.; Barela, P.; Shen, Y.; et al. NASA-ISRO synthetic aperture radar (NISAR) mission. In Proceedings of the 2020 IEEE Aerospace Conference, Big Sky, MT, USA, 7–14 March 2020; IEEE: Piscataway, NJ, USA, 2020; pp. 1–21.
16. Bahdanau, D.; Cho, K.; Bengio, Y. Neural machine translation by jointly learning to align and translate. *arXiv* **2014**, arXiv:1409.0473.
17. Scheuring, I.; Riedi, R.H. Application of multifractals to the analysis of vegetation pattern. *J. Veg. Sci.* **1994**, *5*, 489–496.
18. Yang, H.; Chen, W.; Qian, T.; Shen, D.; Wang, J. The extraction of vegetation points from LiDAR using 3D fractal dimension analyses. *Remote Sens.* **2015**, *7*, 10815–10831.
19. Vaswani, A.; Shazeer, N.; Parmar, N.; Uszkoreit, J.; Jones, L.; Gomez, A.N.; Kaiser, Ł.; Polosukhin, I. Attention is all you need. In *Advances in Neural Information Processing Systems*; Neural Information Processing Systems Foundation, Inc.: San Diego, CA, USA, 2017; Volume 30.
20. Dosovitskiy, A.; Beyer, L.; Kolesnikov, A.; Weissenborn, D.; Zhai, X.; Unterthiner, T.; Dehghani, M.; Minderer, M.; Heigold, G.; Gelly, S.; et al. An image is worth 16 × 16 words: Transformers for image recognition at scale. *arXiv* **2020**, arXiv:2010.11929.
21. Aleissaee, A.A.; Kumar, A.; Anwer, R.M.; Khan, S.; Cholakkal, H.; Xia, G.S.; Khan, F.S. Transformers in remote sensing: A survey. *Remote Sens.* **2023**, *15*, 1860.
22. Wilkes, P.; Jones, S.D.; Suarez, L.; Mellor, A.; Woodgate, W.; Soto-Berelov, M.; Haywood, A.; Skidmore, A.K. Mapping Forest Canopy Height Across Large Areas by Upscaling ALS Estimates with Freely Available Satellite Data. *Remote Sens.* **2015**, *7*, 12563–12587. https://doi.org/10.3390/rs70912563.
23. Wagner, F.H.; Roberts, S.; Ritz, A.L.; Carter, G.; Dalagnol, R.; Favrichon, S.; Hirye, M.C.M.; Brandt, M.; Ciais, P.; Saatchi, S. Sub-meter tree height mapping of California using aerial images and LiDAR-informed U-Net model. *Remote Sens. Environ.* **2024**, *305*, 114099. https://doi.org/10.1016/j.rse.2024.114099.
24. Shendryk, Y. Fusing GEDI with earth observation data for large area aboveground biomass mapping. *Int. J. Appl. Earth Obs. Geoinf.* **2022**, *115*, 103108.
25. Li, Z.; Zhu, X.; Yao, S.; Yue, Y.; García-Fernández, Á.F.; Lim, E.G.; Levers, A. A large scale Digital Elevation Model super-resolution Transformer. *Int. J. Appl. Earth Obs. Geoinf.* **2023**, *124*, 103496.
26. Taneja, R.; Wallace, L.; Hillman, S.; Reinke, K.; Hilton, J.; Jones, S.; Hally, B. Up-scaling fuel hazard metrics derived from terrestrial laser scanning using a machine learning model. *Remote Sens.* **2023**, *15*, 1273.

27. Gazzea, M.; Solheim, A.; Arghandeh, R. High-resolution mapping of forest structure from integrated SAR and optical images using an enhanced U-net method. *Sci. Remote Sens.* **2023**, *8*, 100093.
28. Remijnse, T. Upsampling LiDAR Point Clouds of Forest Environments Using Deep Learning. Master's Thesis, Wageningen University and Research, Wageningen, The Netherlands, 2024.
29. Yu, L.; Li, X.; Fu, C.W.; Cohen-Or, D.; Heng, P.A. Pu-net: Point cloud upsampling network. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Salt Lake City, UT, USA, 18–23 June 2018; pp. 2790–2799.
30. Qian, G.; Abualshour, A.; Li, G.; Thabet, A.; Ghanem, B. Pu-gcn: Point cloud upsampling using graph convolutional networks. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Nashville, TN, USA, 20–25 June 2021; pp. 11683–11692.
31. Qiu, S.; Anwar, S.; Barnes, N. Pu-transformer: Point cloud upsampling transformer. In Proceedings of the Asian Conference on Computer Vision, Macao, China, 4–8 December 2022; pp. 2475–2493.
32. Zhang, T.; Filin, S. Deep-Learning-Based Point Cloud Upsampling of Natural Entities and Scenes. *Int. Arch. Photogramm. Remote Sens. Spat. Inf. Sci.* **2022**, *43*, 321–327.
33. Yan, W.; Cao, L.; Yan, P.; Zhu, C.; Wang, M. Remote sensing image change detection based on swin transformer and cross-attention mechanism. *Earth Sci. Inform.* **2025**, *18*, 106.
34. Ma, X.; Zhang, X.; Pun, M.O. A crossmodal multiscale fusion network for semantic segmentation of remote sensing data. *IEEE J. Sel. Top. Appl. Earth Obs. Remote Sens.* **2022**, *15*, 3463–3474.
35. Qingyun, F.; Zhaokui, W. Cross-modality attentive feature fusion for object detection in multispectral remote sensing imagery. *Pattern Recognit.* **2022**, *130*, 108786.
36. Li, K.; Xue, Y.; Zhao, J.; Li, H.; Zhang, S. A cross-attention integrated shifted window transformer for remote sensing image scene recognition with limited data. *J. Appl. Remote Sens.* **2024**, *18*, 036506.
37. Yu, W.; Huang, F. DMSCA: Deep multiscale cross-modal attention network for hyperspectral and light detection and ranging data fusion and joint classification. *J. Appl. Remote Sens.* **2024**, *18*, 036505.
38. Li, Z.; Liu, R.; Sun, L.; Zheng, Y. Multi-Feature Cross Attention-Induced Transformer Network for Hyperspectral and LiDAR Data Classification. *Remote Sens.* **2024**, *16*, 2775.
39. Yang, J.X.; Zhou, J.; Wang, J.; Tian, H.; Liew, A.W.C. LiDAR-guided cross-attention fusion for hyperspectral band selection and image classification. *IEEE Trans. Geosci. Remote Sens.* **2024**, *62*, 5515815.
40. Zhu, W.; Wang, H.; Hou, H.; Yu, W. CAMS: A Cross Attention Based Multi-Scale LiDAR-Camera Fusion Framework for 3D Object Detection. In *Advances in Guidance, Navigation and Control, Proceedings of the International Conference on Guidance, Navigation and Control*; Springer: Berlin/Heidelberg, Germany, 2024; pp. 533–542.
41. Yoo, J.H.; Kim, Y.; Kim, J.; Choi, J.W. 3d-cvf: Generating joint camera and lidar features using cross-view spatial feature fusion for 3d object detection. In Proceedings of the Computer Vision—ECCV 2020: 16th European Conference, Glasgow, UK, 23–28 August 2020; Proceedings, Part XXVII 16; Springer: Berlin/Heidelberg, Germany, 2020; pp. 720–736.
42. Wu, H.; Miao, Y.; Fu, R. Point cloud completion using multiscale feature fusion and cross-regional attention. *Image Vis. Comput.* **2021**, *111*, 104193.
43. Zhao, H.; Jiang, L.; Jia, J.; Torr, P.H.; Koltun, V. Point transformer. In Proceedings of the IEEE/CVF International Conference on Computer Vision, Montreal, QC, Canada, 10–17 October 2021; pp. 16259–16268.
44. Dao, T.; Fu, D.; Ermon, S.; Rudra, A.; Ré, C. Flashattention: Fast and memory-efficient exact attention with io-awareness. *Adv. Neural Inf. Process. Syst.* **2022**, *35*, 16344–16359.
45. USGS. *USGS 3DEP—CA SanDiego 2015 C17 1*; USGS: Reston, VA, USA, 2016.
46. U.S. Geological Survey, 3D Elevation Program. *USGS 3DEP Lidar Point Cloud (COPC)*; Microsoft Planetary Computer, Redmond, WA, USA, 2023. Available online: https://planetarycomputer.microsoft.com/dataset/usgs-3dep-lidar-copc .
47. Microsoft. *Microsoft Planetary Computer*; Microsoft: Redmond, WA, USA, 2025.
48. Alaska Satellite Facility. *ASF Search Python API*; Alaska Satellite Facility: Fairbanks, AK, USA, 2024.
49. Zhu, Q.; Fan, L.; Weng, N. Advancements in point cloud data augmentation for deep learning: A survey. *Pattern Recognit.* **2024**, *153*, 110532.
50. Shorten, C.; Khoshgoftaar, T.M. A survey on image data augmentation for deep learning. *J. Big Data* **2019**, *6*, 60.
51. Lee, S.H.; Lee, S.; Song, B.C. Vision Transformer for Small-Size Datasets: SPT + LSA. *arXiv* **2021**, arXiv:2112.13492.
52. Touvron, H.; Cord, M.; Sablayrolles, A.; Synnaeve, G.; Jégou, H. Going Deeper with Image Transformers. In Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), Montreal, QC, Canada, 10–17 October 2021; pp. 32–42.
53. d'Ascoli, S.; Touvron, H.; Leavitt, M.L.; Morcos, A.S.; Biroli, G.; Sagun, L. ConViT: Improving Vision Transformers with Soft Convolutional Inductive Biases. In Proceedings of the 38th International Conference on Machine Learning (ICML), Virtual, 18–24 July 2021; Volume 139, pp. 2286–2296.
54. Defazio, A.; Yang, X.A.; Mehta, H.; Mishchenko, K.; Khaled, A.; Cutkosky, A. The Road Less Scheduled. *arXiv* **2024**. https://doi.org/10.48550/arXiv.2405.15682.

55. Wu, T.; Pan, L.; Zhang, J.; Wang, T.; Liu, Z.; Lin, D. Density-Aware Chamfer Distance as a Comprehensive Metric for Point Cloud Completion. *arXiv* **2021**. https://doi.org/10.48550/arXiv.2111.12702.
56. Clay Foundation Team. Clay Foundation Model: An Open-Source AI Foundation Model for Earth Observation. Software Version 1.5 (Commit <hash>), Apache-2.0 License. 2024. Available online: https://github.com/Clay-foundation/model (accessed on 20 November 2025).
57. Liu, A.; Feng, B.; Wang, B.; Wang, B.; Liu, B.; Zhao, C.; Dengr, C.; Ruan, C.; Dai, D.; Guo, D.; et al. Deepseek-v2: A strong, economical, and efficient mixture-of-experts language model. *arXiv* **2024**, arXiv:2405.04434.