Transformer-Based Tree Extraction from Remote Sensing Imagery: A Systematic Review

Attavit Wilaiwongsakul, Bin Liang, Bryan Zheng, Fang Chen

December 26, 2025

Abstract

Global sustainability commitments under the Paris Agreement and United Nations Sustainable Development Goal 15 require scalable tree-level monitoring to quantify forest carbon stocks and ecosystem services. While CNNs established strong baselines, their limited long-range dependency modeling constrains performance on complex canopy structures. Transformer architectures promise to address these limitations through self-attention mechanisms, yet their effectiveness remains unclear. This survey evaluates whether transformers meet three critical requirements: superior accuracy, cross-environment generalization, and data-efficient training. We synthesize 63 papers (2017-2025) documenting 550% growth, analyzing five architectural paradigms spanning pure Vision Transformers, CNN-transformer hybrids, hierarchical variants, Foundation Models, and Vision-Language Models across multiple continents and sensor modalities. Transformers achieve +2.0% to +5.5% median gains over CNN baselines depending on architecture, while Foundation Models with parameter-efficient fine-tuning match full fine-tuning using <1,000 samples. However, critical gaps persist: 28% average performance degradation in cross-environment transfer and 72% of studies lacking full reproducibility. We provide actionable deployment recommendations, identify research priorities including multi-biome benchmarking and domain adaptation, and release comprehensive supplementary materials at `https://github.com/attavit14203638/transformer-tree-survey`.

1 Introduction

Individual tree monitoring has emerged as a critical capability for achieving global sustainability targets. The United Nations (UN) 2030 Agenda for Sustainable Development [1], particularly Sustainable Development Goal (SDG) 13 (Climate Action) and SDG 15 (Life on Land), explicitly recognizes the role of forests in climate mitigation and biodiversity conservation, driving demand for scalable, accurate tree-level monitoring systems. Forests are central to meeting climate targets under the Paris Agreement, with credible mitigation requiring

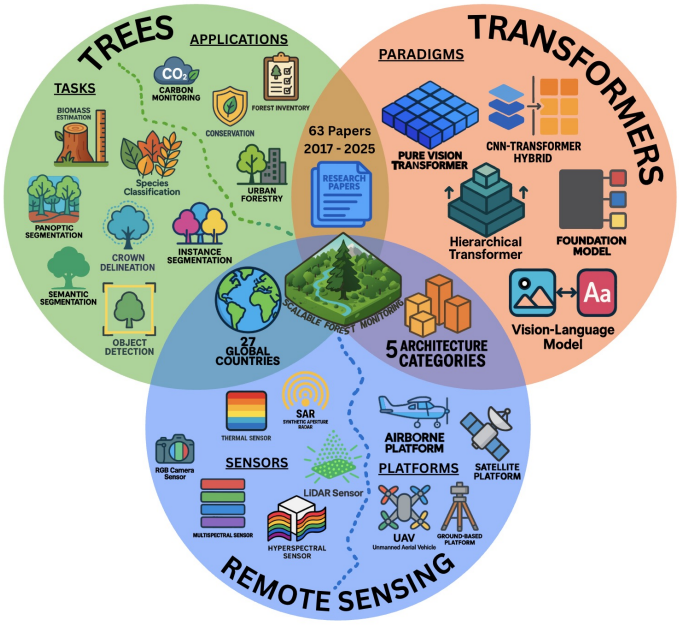


Figure 1: Graphical overview of survey scope.

precise quantification of forest carbon stocks [2], while global tree restoration potential analyses estimate that restoring degraded lands could capture substantial atmospheric carbon [3]. National and international frameworks including Reducing Emissions from Deforestation and Forest Degradation (REDD+) require monitoring at unprecedented spatial granularity. In Europe, the European Union (EU) Nature Restoration Law mandates ecosystem recovery across degraded landscapes, creating binding obligations for member states to restore forest ecosystems [4]. Urban greening initiatives similarly depend on individual tree inventories to model ecosystem services; tools such as i-Tree enable municipalities to quantify carbon sequestration, air pollution removal, and stormwater management benefits at the individual tree level [5]. These policy imperatives have accelerated research into automated tree extraction methods that can scale across diverse landscapes while maintaining the precision needed for regulatory compliance and carbon accounting.

Traditional Remote Sensing (RS) techniques, while valuable, often struggle with the scalability and feature engineering required for precise tree-level analysis across large, heterogeneous landscapes. Convolutional Neural Networks (CNNs) marked a significant leap forward, establishing strong baselines for automated tree detection and segmentation [6]. However, CNNs' inherent focus on local features limits their ability to model the complex structures and long-range spatial dependencies characteristic of diverse forest canopies and urban environments. Vision Transformers (ViTs) have emerged as a promising alternative, employing

self-attention mechanisms that directly model relationships between any two image regions regardless of spatial distance [7]. Their adoption in RS has been rapid [8], with particularly promising results in forestry applications including hierarchical transformers for tree detection [9] and Foundation Models (FMs) enabling tree crown segmentation with minimal labeled data [10].

This survey addresses a critical gap in the literature. While comprehensive reviews have documented CNN-based tree extraction [6, 11] and others have surveyed transformers in RS broadly [8, 12], no systematic synthesis exists at their intersection: transformer-based architectures for individual tree extraction across diverse environments. A rapidly growing body of research is applying transformers to tree extraction tasks, spanning instance segmentation, species classification, multimodal fusion, and FM adaptation, generating insights that require dedicated analysis. This survey systematically synthesizes this emerging literature, examining studies spanning multiple continents and biomes, varied sensor modalities, and a wide range of transformer architectures from pure ViTs to CNN-transformer hybrids and FM adaptations.

This survey addresses three research questions: (1) What are the performance trade-offs between transformer-based and CNN architectures for tree extraction? (2) How well do these models generalize across biomes, sensors, and resolutions? (3) What training strategies enable effective extraction with limited labeled data? Figure 1 provides a graphical overview of the survey scope, illustrating the intersection of tree extraction tasks, transformer architectures, and remote sensing modalities. To answer these questions, we contribute:

1. A data landscape analysis revealing geographic biases and reproducibility challenges across 63 papers;

2. An architecture taxonomy categorizing transformer approaches and documenting their temporal evolution;

3. A systematic performance comparison quantifying transformer advantages over CNN baselines; and

4. An analysis of FM adaptation strategies identifying when few-shot methods outperform fully-supervised specialists.

The remainder of this paper reviews related work (Section 2), analyzes datasets and reproducibility (Section 3), presents our architecture taxonomy (Section 4), addresses each research question through quantitative analysis (Section 5), identifies challenges and future directions (Section 6), and concludes with recommendations (Section 7).

2 Related Work

This survey addresses a critical gap at the intersection of transformer architectures and tree extraction applications. As shown in Table 1, existing literature addresses either tree extraction with CNNs or transformers in RS broadly, but

Table 1: Existing review paper coverage matrix.

Review Paper	Tree Extr.	Transf.	FMs	VLMs	Note
Zhao et al. [6]	●	○	○	○	CNN only
Khan et al. [7]	○	●	○	○	Computer vision
Aleissaee et al. [8]	○	●	◐	○	General RS
Zheng et al. [11]	●	◐	○	○	Traditional + CNN
Wang et al. [12]	○	●	◐	○	General RS
Abreu-Dias et al. [13]	●	○	○	○	Species classif.
Diez et al. [14]	●	○	○	○	UAV only
Zhong et al. [15]	●	○	○	○	Species classif.
Velasquez-Camacho et al. [16]	●	○	○	○	Urban trees
Li et al. [17]	○	●	○	○	Segmentation
Lu et al. [18]	○	◐	●	○	General RS
Chehreh et al. [19]	●	○	○	○	Agroforestry
Liu et al. [20]	○	●	○	○	Computer vision
Estrada et al. [21]	●	○	○	○	Forest health
Tao et al. [22]	○	◐	◐	●	General RS
THIS SURVEY	●	●	●	●	**Tree + Transformer**

Legend: ● = *Fully covered*; ◐ = *Partially covered*; ○ = *Not covered*
Abbreviations: Tree Extr. = Tree Extraction; Transf. = Transformers

not their intersection. While several comprehensive reviews have documented CNN-based tree extraction methods [6, 11, 13–16] and others have surveyed transformers in RS broadly [7, 8, 12, 17], no prior synthesis examines transformers specifically for individual tree extraction across diverse environments. Zhao et al. [6] systematically reviewed 35 papers (2019-2021) on CNN-based tree crown detection, establishing performance benchmarks but predating widespread transformer adoption. Similarly, Zheng et al. [11] analyzed approximately 150 papers (2000-2023) covering traditional to deep learning approaches for tree crown detection, while task-specific reviews addressed tree species identification [13, 15], Unmanned Aerial Vehicle (UAV)-based forestry applications [14], and urban tree characterization [16]. These reviews document CNN achievements but provide limited or no coverage of transformer architectures that have emerged prominently in recent years.

Conversely, transformer-focused surveys have examined these architectures across general RS applications without tree-level specificity. Aleissaee et al. [8] surveyed more than 60 papers (2017-2023) on transformers for scene classification, object detection, and semantic segmentation, while Wang et al. [12] extended this to 237 papers (2021-2023) with systematic architectural categorization. From the computer vision perspective, Khan et al. [7] and Liu et al. [20] established theoretical foundations of vision transformers, while Li et al. [17] focused on transformer-based visual segmentation. While these surveys confirm transformer success in general RS tasks, the unique challenges of individual tree extraction, including complex canopy structures, overlapping crowns, multi-scale detection requirements, and high variability across diverse biomes, warrant focused investigation beyond general applications.

The emergence of FMs and Vision-Language Models (VLMs) has been docu-

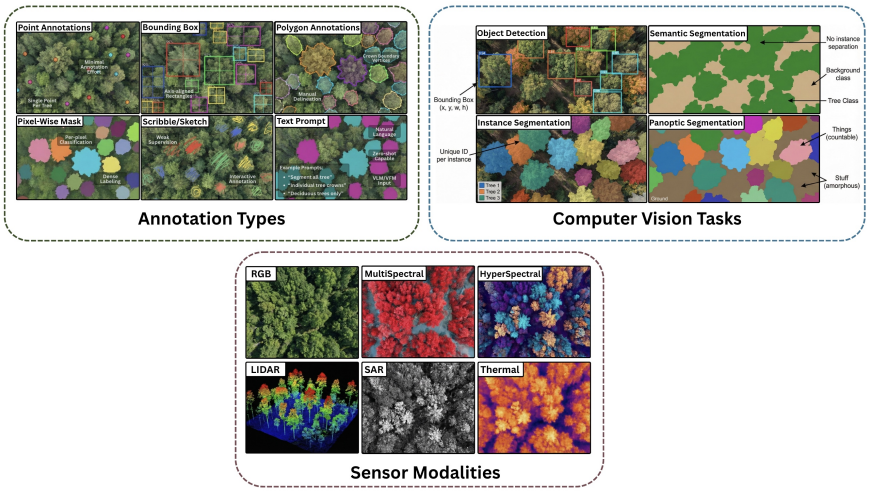


Figure 2: Overview of tree extraction research dimensions.

mented in recent surveys. Lu et al. [18] examined 58 FMs (2021-2024) analyzing models such as Segment Anything Model (SAM) and Contrastive Language-Image Pre-training (CLIP) for broad RS applications, while Tao et al. [22] reviewed approximately 90 papers (2020-2024) on VLMs for tasks including image captioning and visual question answering. Application-specific reviews have addressed agroforestry [19] and forest health assessment [21]. However, the adaptation of these powerful models to specialized tree extraction tasks, including domain shift challenges, fine-grained species discrimination, and detection of small or sparse trees in diverse environments, represents a nascent field requiring dedicated analysis.

Our survey fills this critical gap by systematically analyzing 63 primary research papers (2017-2025) that apply transformer-based models (including pure ViTs, hierarchical variants, CNN-transformer hybrids, Vision Foundation Model (VFM) adaptations, and VLMs) to individual tree extraction tasks across diverse biomes, sensor modalities, and environmental contexts. Unlike general transformer surveys [8, 12], we focus exclusively on tree-level applications. Unlike CNN-focused reviews [6, 11], we analyze the architectural shift to transformers and its implications for modeling complex canopy structures. Unlike FM surveys [18], we examine both custom-trained transformers and adapted FMs within a unified framework, providing practitioners with actionable guidance for model selection and deployment.

3 Data Landscape

3.1 Benchmarks and Reproducibility

The data landscape spans diverse annotation types (bounding boxes,

Table 2: Top public benchmarks for tree extraction with transformer models.

Dataset	Sensors	Location	Size	Res.	Used By
NeonTreeEval. [23]	RGB, LiDAR, HSI	22 US sites	31k trees	0.1m	[24–27]
Open-Canopy [30]	VHR RGB, S2	France (87k km^2)	1M+ samples	1.5m	[31]
FOR-instanceV2 [28]	LiDAR	9 countries	10.5k trees	0.2m	[29]
Quebec Plant. [32]	UAV RGB	Canada	19.5k trees	5mm	[10, 33, 34]
Detectree2 [39]	UAV RGB	Malaysia	3k crowns	0.05m	[26]
OAM-TCD [35]	RGB	Global	5k images	0.1m	[36]
DynamicEarthNet [40]	S2 Time-series	75 global sites	7k images	10m	[41]
Semantic3D [42]	Terr. LiDAR	Switzerland	4B points	0.03m	[43]
Landcover.AI [44]	Aerial RGB	Poland (216 km^2)	22k images	0.25m	[45]
Trento [46]	HSI + LiDAR	Italy (600×166 px)	99.9k pixels	1.0m	[47]

masks, points, polygons), computer vision tasks (object detection, semantic/instance/panoptic segmentation), and sensor modalities (Red-Green-Blue (RGB), Multispectral (MS), Light Detection and Ranging (LiDAR), Synthetic Aperture Radar (SAR), Hyperspectral (HS)). This is consolidated in Figure 2. Across 63 papers, 48.4% utilize public benchmarks (Table 2). Key datasets include NeonTreeEvaluation [23], the most widely adopted [24–27], providing 31k trees across 22 US sites with multi-modal data; FOR-instanceV2 [28] with 10.5k tree point clouds from 9 countries [29]; Open-Canopy [30] spanning 87k km^2 of France with 1M+ samples [31]; Quebec Plantations [32] at ultra-high 5mm Ground Sampling Distance (GSD) for Parameter-Efficient Fine-Tuning (PEFT) research [10, 33, 34]; and OAM-TCD [35] offering globally diverse tree cover maps [36]. FoMo-Bench [37] aggregates 15 datasets for FM evaluation, while large-scale pre-training resources include MillionST (1M Sentinel-2 (S2) images) [38] and GeoLangBind-2M (2M image-text pairs) [24].

Despite these resources, reproducibility challenges constrain progress. Across 63 papers, 64 distinct datasets are used, with 51.6% relying on private or unreported sources. Benchmark adoption remains at 25.8%, code release at 33.9% [10, 24, 26, 29, 31, 37, 38, 47–60], and no papers provide pre-trained transformer weights. Full reproducibility (public data and code) is achieved by only 27.4% [10, 26, 29, 31, 37, 38, 47, 48, 50–52, 54–58, 60]. The remaining 72.6% with private data or missing code inhibit comparative evaluation and limit translation to operational systems, particularly for computationally intensive FMs.

3.2 Acquisition Platforms and Sensors

Figure 3 summarizes the data acquisition landscape across 63 papers. Sensor modality distribution, shown in Figure 3(a), reveals that RGB sensors dominate at 42.7% [9, 10, 24–27, 33, 34, 36–38, 45, 48–51, 54–86], valued for low cost but limited in spectral discrimination. MS sensors constitute 28.2% [24, 25, 31, 37, 38, 41, 48, 49, 51, 53, 54, 57, 73, 75–78, 80, 81, 83, 87–92], enabling species discrimination and health assessment. LiDAR represents 16.4% [10, 24–26, 29, 34, 37, 43, 47, 52, 54, 58, 71, 74–76, 78, 81, 82, 86, 93], providing 3D

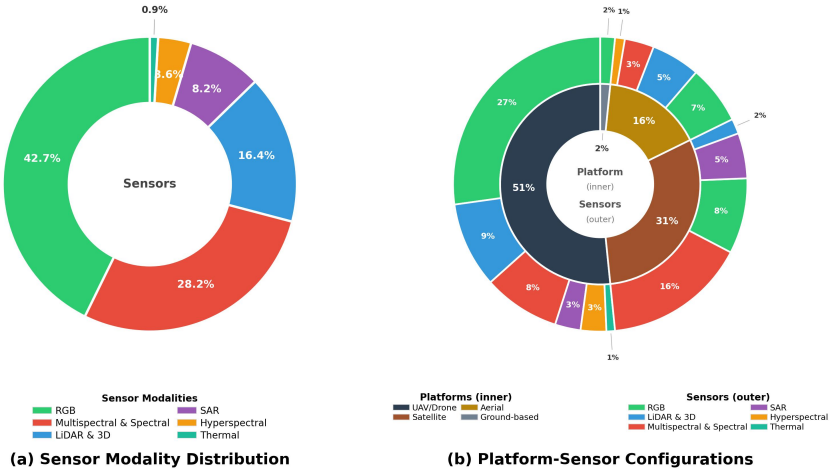


Figure 3: Data acquisition landscape.

structural information for biomass estimation. The remaining 12.7% comprises SAR (8.2%) [24, 37, 48, 53, 75, 78, 83, 91, 92] for weather-independent monitoring, HSI (3.6%) [24, 37, 47, 48] for fine-grained species classification, and thermal (0.9%) [48] for stress detection. Notably, 37.1% of papers employ RGB as their sole modality, while 62.9% leverage multi-modal fusion, a trend that directly motivates transformer cross-attention mechanisms.

Platform-sensor configurations are detailed in Figure 3(b), revealing a clear scale-resolution trade-off that shapes architectural decisions. UAV platforms dominate at 52%, predominantly deploying RGB sensors (52.5% of UAV studies) to capture sub-decimeter imagery (0.05-0.5m GSD) essential for individual crown delineation; this high resolution pairs naturally with hierarchical transformers (Swin, SegFormer) whose multi-scale feature extraction exploits fine spatial detail. Satellite platforms constitute 31%, primarily using MS (31.1%) and SAR for landscape-scale monitoring with consistent revisit times; coarser resolution motivates pure ViTs or CNN-transformer hybrids that process larger spatial contexts efficiently. Aerial platforms (15%) bridge this gap at 0.25-1.0m GSD for regional inventories, while ground-based terrestrial laser scanning (2%) [50] provides precise 3D structural characterization for height estimation. This platform-sensor distribution creates a methodological insight: multi-modal fusion tasks drive specialized cross-modal attention architectures that learn complementary representations (texture from RGB, structure from LiDAR) through attention-based alignment, a key transformer advantage over CNN feature concatenation, further examined in Section 4.

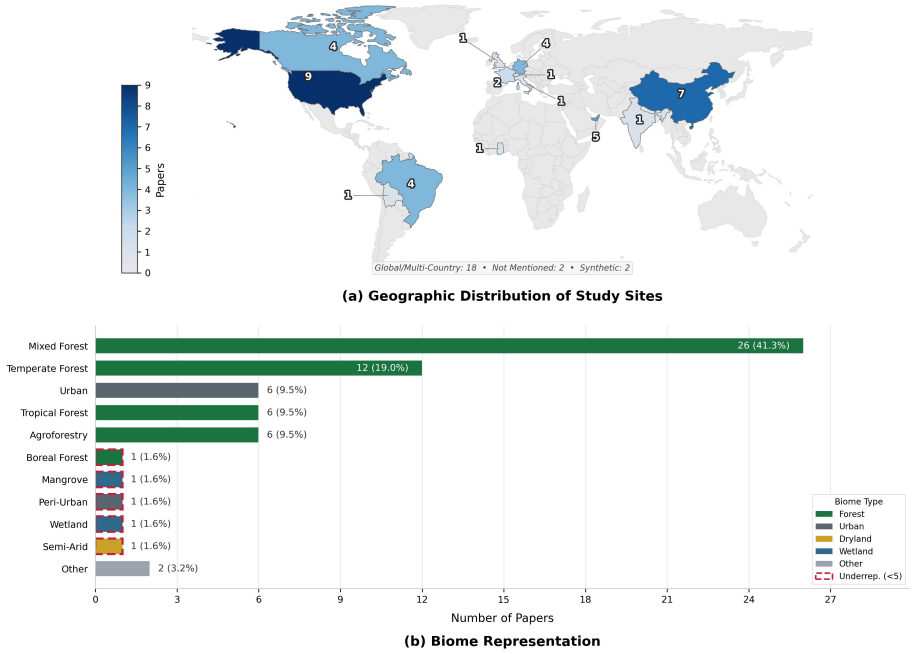


Figure 4: Study sites by country and biome.

3.3 Geographic Distribution

Figure 4 summarizes study site distribution across geographic and ecological dimensions. Country-level analysis, presented in Figure 4(a), reveals severe geographic concentration. While 28.6% span multiple countries [10, 26, 36–38, 41, 45, 47–50, 60, 62, 72, 76, 83, 85], ∼85% of single-country studies concentrate in the Northern Hemisphere. The United States leads at 14.3% [25, 27, 51, 52, 55, 57, 67, 80, 93], followed by China (11.1%) [64, 69, 71, 74, 82, 86, 88], UAE (7.9%) [9, 61, 68, 70, 84], Canada/Brazil/Germany (each 6.3%) [33, 34, 53, 54, 59, 73, 77, 81, 87, 89–91], and France (3.2%) [29, 31]. Single studies cover Bolivia, Switzerland, Italy, Ghana, India, and the UK [43, 56, 65, 75, 78, 92]; synthetic (3.2%) [58, 66] and unspecified locations (3.2%) [63, 79] account for remaining studies.

Biome representation, illustrated in Figure 4(b), shows similar imbalance. Mixed forests dominate at 41.3% [10, 25–27, 29, 36–38, 41, 45, 47, 50, 58, 60, 62, 63, 65, 66, 72, 76, 78, 82, 83, 86, 90], followed by temperate (19.0%) [31, 33, 34, 54, 55, 59, 64, 71, 80, 81, 89, 93], then urban (9.5%) [43, 52, 61, 74, 75, 85], agroforestry (9.5%) [9, 49, 56, 69, 70, 84], tropical (9.5%) [51, 57, 77, 87, 91, 92], and other/unspecified (3.2%) [48, 79]. Critically underrepresented are boreal forests (1.6%) [73] despite comprising 29% of global forest area, mangroves (1.6%) [88], wetlands (1.6%) [53], peri-urban zones (1.6%) [67], and semi-arid drylands (1.6%) [68]. This imbalance has practical consequences: models trained

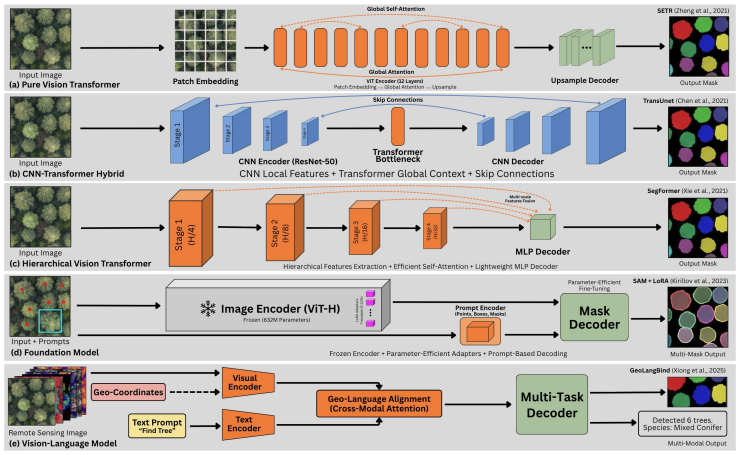


Figure 5: Transformer architectural categorization pipelines.

on temperate/mixed forests may degrade on tropical multi-layered canopies, spectrally ambiguous mangroves, seasonally variable boreal systems, or sparse savanna distributions. Cross-environment validation remains limited, with only 24% of papers testing across multiple biomes, and most of those within ecologically similar contexts.

4 Transformer Architectures

4.1 Architectural Taxonomy

Analysis of 63 papers reveals five primary architectural categories that trace the evolution from early ViT adaptations to contemporary FM approaches: Pure ViT, CNN-Transformer Hybrids, Hierarchical ViTs, FMs, and VLMs. An additional 28.6% employ custom or incompletely reported backbones and are grouped as Unspecified/Custom in Figure 6. The distribution reflects both the maturity of different approaches and their suitability for tree-level analysis in RS contexts. Conceptual processing pipelines for each category are illustrated in Figure 5, and temporal adoption trends are examined in Section 4.2.

4.1.1 Pure Vision Transformers

Pure ViT architectures apply standard ViT [94] and sequence-to-sequence formulations such as Segmentation Transformer (SETR) [95] to tree extraction (7.9%), treating inputs as sequences of fixed-size patches or tokens processed by global self-attention [55, 65, 71, 77, 92]. In this literature, pure ViT models are primarily used for classification-oriented tasks, including deforestation monitoring via multilabel classification [77] and tree species classification from UAV canopy imagery [55, 65] and LiDAR-derived inputs [71]. Sequence-based transformers

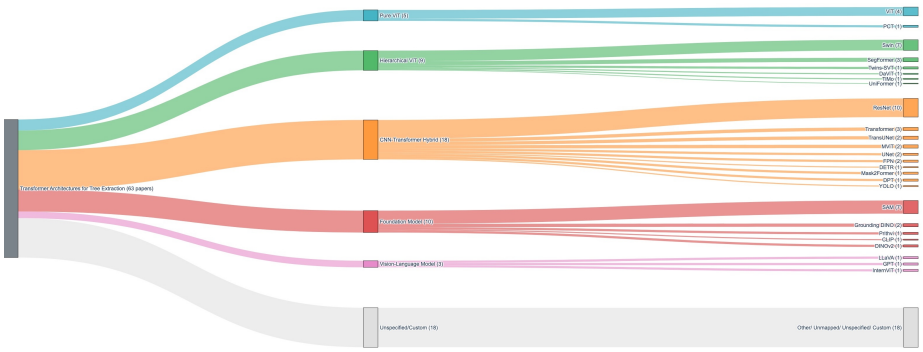


Figure 6: Transformer architecture taxonomy for tree extraction.

are also applied to seasonal disturbance detection from multi-sensor time series [92]. Overall adoption remains limited relative to hybrid and hierarchical designs, reflecting the data-hungriness and compute cost of training fully attention-based models in RS settings.

4.1.2 CNN-Transformer Hybrids

CNN-Transformer hybrids comprise 28.5% of the literature, pairing CNN encoders for efficient local feature extraction with transformer decoders or attention modules for global context modeling [29, 45, 54, 57, 62, 70, 81, 86, 89]. This hybrid design exploits complementary strengths: CNNs provide translation equivariance and hierarchical feature pyramids, while transformers contribute long-range dependency modeling that helps separate overlapping crowns and capture larger-scale spatial structure. ResNet-based feature pyramids remain the most common backbone pattern, often feeding transformer modules for delineation or dense prediction [62, 70]. The TransUNet family [96] adapts U-Net skip connections with transformer bottlenecks for RS segmentation [57]. Detection transformers (DETR) [97] replace hand-crafted pipelines with learned object queries for single-tree detection [81, 89], while Mask2Former [98] supports instance-aware crown delineation in high-resolution imagery [62]. Domain-specific hybrids address multimodal sensing and 3D structure: MTCDNet fuses RGB and LiDAR cues for crown detection [86], and ForestFormer3D extends hybrid modeling to point-cloud segmentation [29].

4.1.3 Hierarchical Vision Transformers

Hierarchical ViTs, particularly Swin Transformer [99] and SegFormer [100] (14.3%), employ hierarchical feature extraction without CNN backbones [38, 49, 63, 64, 67–69, 80, 84]. Unlike standard ViT's global self-attention with quadratic complexity, hierarchical variants compute attention within local windows, reducing complexity while maintaining long-range dependency modeling through shifted window mechanisms. Other hierarchical backbones that appear

in the broader vision literature include PVT [101], DaViT [102], UniFormer [103], and Twins-SVT [104]. SegFormer [100] combines hierarchical encoders with lightweight Multilayer Perceptron (MLP) decoders for applications including unhealthy tree crown detection [80] and invasive species mapping [67], while Twins-SVT [104] supports efficient inference for fine-grained vegetation monitoring [67]. As in Figure 6, papers that embed hierarchical transformers inside CNN-based pipelines (e.g., Mask R-CNN + Swin) are treated as hybrids to keep categories mutually exclusive.

4.1.4 Foundation Models

FM adaptations account for 15.9% and represent the fastest-growing direction, leveraging large pre-trained models such as SAM [105] and Grounding DINO [106] for tree segmentation and detection with reduced task-specific training [10, 26, 33, 34, 41, 50, 52, 56, 59, 66]. SAM-based studies explore a spectrum of adaptation strategies: zero-shot SAM2 evaluation without domain training yields low precision but moderate recall (0.41-0.65) depending on the benchmark [26], while parameter-efficient fine-tuning can reach near full fine-tuning performance with a tiny fraction of trainable parameters [33] and can match strong adapter baselines with substantially fewer parameters using Low-Rank Adaptation (LoRA) [59]. Fully supervised adaptation and prompt-learning variants further improve crown delineation but increase computational cost [10, 34, 50, 52]. Grounding DINO supports open-vocabulary or prompt-guided tree detection, enabling species-aware detection and downstream classification under weak supervision [33, 56]. Overall, FMs directly target a central bottleneck in tree extraction: the high cost of pixel-level annotation, by leveraging representations learned from large-scale natural-image or RS-specific pretraining corpora.

4.1.5 Vision-Language Models

VLMs represent an emerging frontier (4.8%), enabling interactive and multi-task capabilities through natural language interfaces [48, 74, 83]. EarthDial [48] turns multi-sensory earth observations into interactive dialogues for querying forest conditions and species composition, while Tree-GPT [74] implements a modular Large Language Model (LLM) expert system combining specialized visual encoders with LLM reasoning. REO-VLM [83] addresses regression challenges through language-guided visual grounding for quantitative attribute estimation (e.g., height, biomass, canopy cover). While VLMs promise more accessible forest monitoring through natural language interfaces, computational requirements and limitations in fine-grained spatial reasoning currently constrain operational deployment.

4.2 Temporal Trends

Figure 7 traces the temporal evolution of transformer adoption in tree extraction, revealing over 550% growth from 2022 to 2025. Although our systematic review

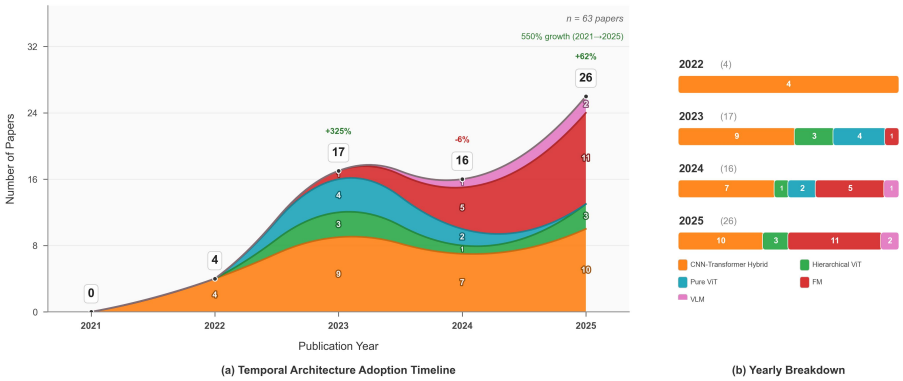


Figure 7: Temporal adoption of different transformer architectures.

scope spans 2017-2025 to capture the full transformer era following the original Transformer architecture [107], the first applications to tree extraction appeared in 2022, reflecting the time required for vision transformers to mature and for the RS community to adapt these architectures to forestry applications. Early adoption in 2022 established the viability of transformer approaches through pioneering work on Swin Transformer for date palm detection [9], Detection Transformer (DETR) for single-tree detection [89], 3D LiDAR transformers for urban vegetation [43], and individual tree detection from UAV imagery [85].

The field expanded rapidly in 2023, establishing CNN-Transformer hybrids as the dominant paradigm. Notable contributions include TransUNet variants for deforestation mapping [51, 53], canopy height estimation [78], species classification [64, 69, 71], optical-SAR fusion [91, 92], DETR-based detection refinement [81], cross-sensor benchmarking [76], and the first VLM application with TreeGPT [74]. The year 2024 marked a qualitative shift toward FMs, with SAM adaptations [50, 52, 54, 90] and distribution-shift benchmarking [56] emerging alongside continued hybrid refinement through e-TransUNet [57] and multi-modal fusion networks [58, 79, 88]. This FM wave accelerated dramatically in 2025, with SAM-based crown segmentation [10, 26, 34, 59], parameter-efficient fine-tuning frameworks [33], multi-modal FMs [24, 31, 37], and interactive VLMs for forest dialogue [48]. Hybrids remained substantial with innovations in multimodal fusion [25, 47, 82, 86] and 3D point cloud processing [29], while hierarchical transformers saw continued use for date palm detection [61, 68] and invasive species mapping [67, 80].

Three key trends emerge. First, CNN-Transformer hybrids maintain plurality throughout (28.6% of categorized papers), reflecting their pragmatic balance of leveraging established CNN infrastructure while incorporating transformer advantages. Second, FM adaptations show the steepest growth trajectory (absent in 2022 but representing the fastest-growing category by 2024-2025), driven by powerful pre-trained models and efficient adaptation techniques including LoRA and prompt tuning. Third, VLM applications remain nascent (4.8%), with

Table 3: Transformer architecture performance summary against CNN baselines.

Architecture	Papers	Median Gain	Primary Tasks	Parameters
Pure ViT	6	+5.5%	Cls.+Seg.	0.08-320M
Hierarchical ViT	8	+2.6%	Cls.+Seg.	3-91M
CNN-Transformer Hybrid	31	+3.4%	Seg.+Det.	1.3-116M
FM (PEFT)	12	+2.0%	Seg.+Det.	14-632M
VLM	4	—	Interactive	4B+

Note: Median Gains = over CNN baselines, Seg.=segmentation, Det.=detection, Cls.=classification, M = millions, B = billions, — = insufficient data.

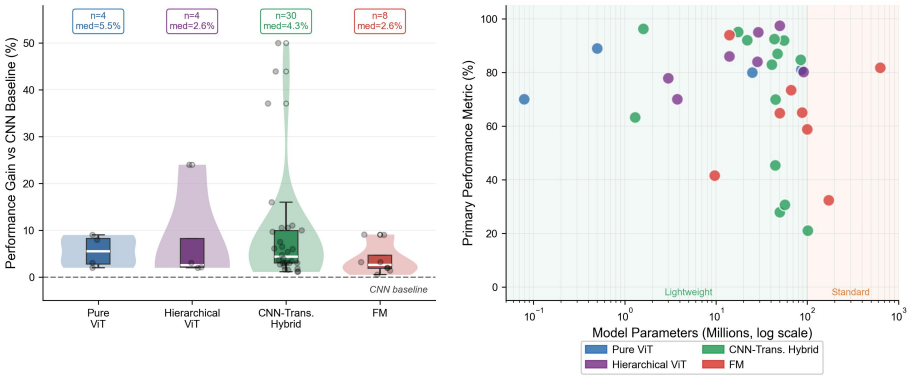
practical deployment constrained by computational requirements and limited fine-grained spatial reasoning. This rapid FM adoption signals a potential paradigm shift: rather than training task-specific architectures from scratch, the community increasingly adapts large-scale pre-trained models to tree extraction, promising reduced annotation requirements and improved cross-domain generalization. These themes are examined when addressing RQ3 on training strategies and data efficiency.

5 Discussion

5.1 RQ1: Performance Trade-offs

Transformer architectures consistently outperform CNN baselines across tree extraction tasks [9, 10, 24, 26, 34, 37, 38, 48–51, 53, 55–57, 63, 65, 67–71, 73, 74, 76, 77, 80, 81, 83, 84, 89, 92]. Table 3 summarizes architecture categories across 61 transformer-based studies (two non-transformer papers are excluded [64, 66]). Median gains are computed only when CNN baselines and percentage metrics (Intersection over Union (IoU), Average Precision (AP), F1, accuracy) are reported. Figure 8(a) reveals distinct gain distributions: pure ViTs achieve the highest median improvement (+5.5%) but with limited sample size; CNN-transformer hybrids deliver robust gains (+4.3%) with the largest representation; hierarchical ViTs show modest gains (+2.6%) while FMs exhibit the lowest median improvement (+2.0%) but trade raw performance for label efficiency via PEFT; VLMs remain early-stage for interactive analysis. Figure 8(b) plots model parameters against primary performance metrics, delineating efficiency frontiers: lightweight architectures (<100M parameters) achieve competitive performance (70–97%) while larger FMs (>100M) yield marginal accuracy gains at substantially higher computational cost.

The hybrid category [9, 25, 27, 29, 43, 45, 47, 51, 53, 57, 58, 60–62, 70, 72, 75, 76, 78, 79, 81, 82, 85–89, 91, 93], spanning TransUNet variants, Swin-backbone segmentation, DETR-based detection, and multi-modal fusion, leverages CNN encoders for local features and transformer decoders for global context, effective for crown delineation where boundary details and spatial relationships both matter. Pure ViTs [38, 49, 55, 63, 65, 67–69, 71, 73, 77, 80, 84, 92] concentrate in species classification, health assessment, invasive species mapping, and spa-



(a) Performance Gain Distribution by Architecture **(b) Performance vs. Computational Cost**

Figure 8: Performance-efficiency trade-offs across transformer architectures.

tiotemporal modeling, though their higher parameter requirements (up to 320M) and data hunger limit adoption for smaller datasets. FM adaptations [10, 24, 26, 31, 33, 34, 37, 41, 50, 52, 54, 56, 59, 90] represent the fastest-growing category: SAM-based approaches dominate, with PEFT methods (LoRA, adapters) achieving 90-95% of full fine-tuning performance while training only 5-10% of parameters, and zero-shot SAM2 reaching moderate recall (0.41-0.65) without domain-specific training. VLMs [24, 48, 74, 83] enable interactive forest analysis through natural language interfaces, though computational requirements and limited spatial reasoning constrain deployment.

Authors attribute transformer advantages to four mechanisms: (1) *long-range dependency modeling* for capturing spatial relationships between distant image regions, critical for distinguishing overlapping crowns [9, 81, 82]; (2) *multi-scale feature extraction* through hierarchical architectures [49, 55, 84]; (3) *adaptive receptive fields* via dynamic attention weights [62, 67]; and (4) *end-to-end learning* eliminating hand-crafted post-processing [81, 89]. However, transformers incur computational overhead (2.3× median parameter counts versus CNN baselines), though efficient variants narrow this gap: SegFormer-B0 achieves competitive performance with 3.75M parameters [80], WetMapFormer with 1.6M [53]. FMs present an efficiency paradox: massive pre-trained models (SAM: 632M) enable effective extraction with minimal labeled data, potentially reducing overall annotation costs despite high parameter counts [10, 50]. For resource-constrained deployment (UAV on-board processing), lightweight hierarchical transformers offer the best performance-to-efficiency ratio; for offline analysis, larger models justify their computational cost through improved accuracy.

5.2 RQ2: Cross-Environment Generalization

Cross-environment generalization emerges as a critical challenge [10, 24–26, 29, 34, 37, 38, 45, 50, 51, 53–57, 60, 65, 70, 73, 75–77, 81, 83, 86, 88–91]. Figure 9(a)

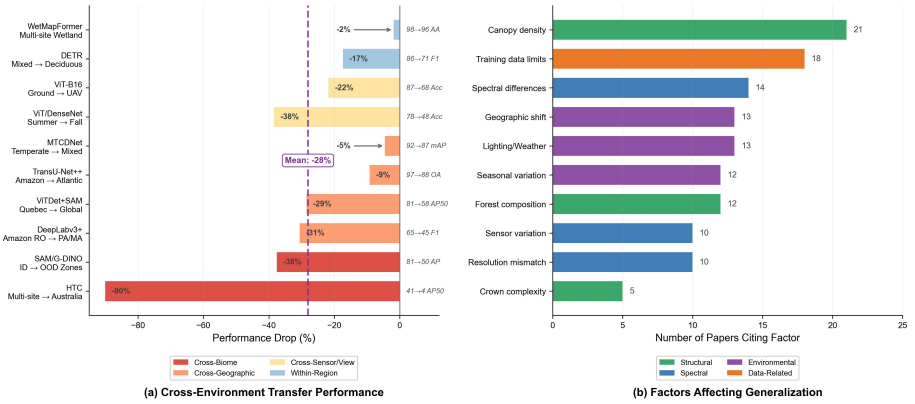


Figure 9: Cross-environment generalization.

synthesizes transfer performance across 10 representative experiments, revealing an average 28% performance drop with severity ranging from severe (cross-biome: mean 64% drop) to mild (within-region: mean 10%). Cross-biome transfers exhibit the most severe degradation: the HTC model [62] trained on 9 globally distributed sites achieves 41% AP50 on average but plummets to 4% on unseen Australian savanna, while SAM with Grounding-DINO shows 38% drops on out-of-distribution agro-climatic zones [56]. Cross-geographic transfers within similar biomes show more favorable results (mean 18% drop): TransU-Net++ degrades only 9% from Amazon to Atlantic Forest [51, 57], MTCDNet maintains performance within 5% between Chinese temperate and mixed forests [86], and ViTDet+SAM trained on globally diverse FLICKRTREE data degrades only 3.8% on Quebec versus 23% when trained on Quebec alone [50], demonstrating that diverse training exposure enables environment-invariant feature learning.

Cross-sensor and viewpoint shifts present distinct challenges (mean 30% drop) [24, 55, 58, 65, 69, 75, 79, 88, 91, 92]. Ground-level classifiers trained on iNaturalist suffer 22% degradation on UAV aerial imagery [65], while seasonal variation proves particularly challenging, models trained on summer imagery achieve only 48% accuracy on fall data [55, 92]. SAM2 demonstrates notable viewpoint invariance, generalizing from California temperate to Malaysian tropical forests without fine-tuning [26], though with moderate recall (41-65%). Within-region transfers show the mildest degradation: DETR-based detection drops only 17% from mixed to deciduous stands within the same German site [81, 89], while WetMapFormer maintains 96% accuracy across New Brunswick wetland sites [53]. These patterns indicate that architectural sophistication alone cannot overcome fundamental domain mismatches, but FMs pre-trained on diverse imagery offer promising generalization [10, 24, 26, 34, 37, 38].

Figure 9(b) categorizes factors affecting generalization based on citation frequency. Structural factors dominate: canopy density [9, 27, 29, 70, 73, 76, 81, 84, 85, 87, 93] emerges as the primary challenge since models trained on

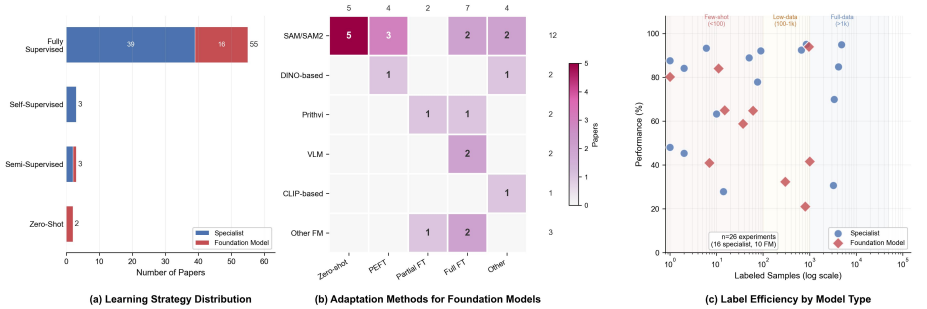


Figure 10: Training strategy and data efficiency analysis.

open woodland fail on dense tropical canopy, while forest composition and crown complexity also impact performance significantly [43, 45, 47, 62, 82]. Spectral factors, including signature differences, resolution mismatch [56, 60, 72, 78], and sensor variation [26, 58, 75, 79, 88, 91], affect models relying on specific band combinations. Environmental factors (geographic shift, lighting/weather, seasonal variation) [25, 49, 55, 71, 92] introduce systematic appearance differences. Critically, training data diversity limitations [37, 45, 50, 59, 66, 90] are addressable through careful dataset curation and FM pre-training on globally diverse imagery.

5.3 RQ3: Data-Efficient Training

Training strategy selection significantly impacts model performance and deployment feasibility, particularly with limited labeled data [10, 24, 26, 31, 33, 34, 37, 41, 48, 50, 52, 54, 56, 59, 60, 74, 83, 90]. Figure 10(a) shows fully supervised learning dominates (87%), with FMs concentrated in emerging strategies: self-supervised (5%), semi-supervised (5%), and zero-shot (3%). The relationship between FM types and tuning methods is examined in Figure 10(b), where SAM/SAM2 leads (12 papers) across zero-shot (5), PEFT (3), partial (2), and full fine-tuning (2). To assess data efficiency empirically, Figure 10(c) plots labeled sample count against performance for 26 experiments, revealing high variance (20-95% at similar counts) due to heterogeneous evaluation settings.

FM adaptation represents the fastest-growing category, with SAM-based approaches leading [10, 26, 33, 34, 50, 52, 54, 59, 90]. Table 4 summarizes five adaptation strategies: zero-shot SAM/SAM2 achieves moderate recall (41-65%) without labeled data [26, 34], sufficient for prototyping but below operational thresholds; PEFT with LoRA reaches approximately 99% of full fine-tuning with <1,000 samples [59]; ASCS adapters on Grounding-DINO achieve 60.6% AP50 with 0.28% trainable parameters [33]; Prithvi-100M outperforms from-scratch models (+2.9% ACC); and DOFA-CLIP achieves +6.7% mIoU on large datasets. Prompt-based approaches (RSPrompter [10], Digital Surface Model (DSM)-enhanced prompting [54]) leverage SAM's capabilities with minimal adaptation, while full fine-tuning maximizes performance at higher cost [50, 52, 90].

Table 4: FM Adaptation Strategies for Tree Extraction

Model	Tuning	Papers	Samples	Result
SAM/SAM2	Zero-shot	5	0	Recall: 41-65%
SAM (ViT-H)	PEFT (LoRA)	3	<1k	∼99% of adapter FT
G-DINO	PEFT (ASCS)	1	<500	AP50: 60.6%
Prithvi-100M	Full FT	1	Med.	+2.9% ACC
DOFA-CLIP	Full FT	1	Large	+6.7% mIoU

Note: PEFT = Parameter-Efficient Fine-Tuning. FT = Fine-Tuning.

Controlled experiments reveal that FMs with PEFT outperform specialist models trained from scratch when labeled data is limited [10, 34, 59], while fully-supervised specialists achieve higher peak performance with abundant data [50, 56]. This suggests a decision framework: for rapid deployment with limited labels, FMs with PEFT offer the best path; for maximum accuracy with abundant data, invest in specialist architectures. Semi-supervised approaches provide a middle ground, with TreeFormer achieving 96% of supervised performance using only 30% labeled data [60].

Pre-training strategy significantly affects downstream performance [9, 24, 31, 37, 41, 51, 55, 84]. While ImageNet pre-training dominates, RS-specific pre-training shows advantages: FoMo-Net achieves competitive performance across tasks without task-specific tuning [37], DOFA-CLIP enables cross-modal transfer through vision-language alignment [24], and Prithvi-100M's Harmonized Landsat Sentinel (HLS) pre-training outperforms from-scratch models on change detection [41]. Self-supervised pre-training via Masked Autoencoders (MAEs) on unlabeled RS imagery shows particular promise [37, 90]. A surprising finding: head-only fine-tuning of FMs outperforms full fine-tuning for out-of-distribution generalization [56], suggesting that minimal adaptation preserves generalizable pre-trained features. VLMs extend capabilities through natural language interfaces [48, 74] and language-guided regression [83], though computational requirements constrain deployment.

6 Challenges and Future Directions

6.1 Current Challenges

Three interconnected challenges limit operational deployment. First, the *capability-efficiency tension* remains unresolved: transformers achieve superior accuracy but impose 2.3× higher parameter counts than CNN baselines, with FMs like SAM ViT-H demanding 632M parameters [10, 26]. While lightweight variants (SegFormer-B0: 3.75M; WetMapFormer: 1.6M) demonstrate feasibility [53, 80], systematic accuracy-efficiency benchmarking is lacking. Practitioners face trade-offs between cloud processing (latency, connectivity) and edge deployment (reduced accuracy).

Second, *cross-environment generalization* presents the most consequential limitation. Our analysis reveals 28% average performance degradation outside

training distributions, with cross-biome transfers causing severe failures [50, 56, 62]. Root causes are structural: 85% of studies concentrate in the Northern Hemisphere, boreal forests (29% of global area) receive 1.6% of attention, and only 24% validate across biomes. This bias means published metrics may overestimate real-world capability for most global forests.

Third, the *reproducibility crisis* impedes progress. Only 27.4% of studies are fully reproducible, with 51.6% using private datasets, benchmark adoption at 25.8%, code availability at 33.9%, and no pre-trained weights released [37, 45]. Despite FMs enabling few-shot learning, 87% of papers require fully-supervised training [10, 60]. While PEFT achieves ∼99% of full fine-tuning with <1,000 samples [33, 59], guidance on sample selection and active learning remains absent.

6.2 Future Research Directions

For practitioners, we recommend architecture selection based on data availability: with limited labels (<100), FMs with PEFT achieve near full fine-tuning accuracy at reduced cost; with moderate data (100-1,000), semi-supervised approaches offer favorable trade-offs; with abundant data (>1,000), specialist hybrids maximize accuracy. For edge deployment, prioritize lightweight hierarchical transformers; for cloud processing, larger FMs.

For the research community, four priorities emerge: (1) establishing globally representative benchmarks spanning 8+ biomes with standardized multi-modal acquisition; (2) systematic domain adaptation research including unsupervised methods, continual learning, and meta-learning; (3) model compression for UAV and portable deployment; and (4) advancing multi-modal fusion through cross-modal attention and self-supervised pre-training.

Sustained progress requires reproducibility standards: validation across ecologically distinct environments, code and weight release, and computational reporting. A centralized model repository would reduce duplication and accelerate operational translation.

7 Conclusion

This survey systematically analyzes transformer-based approaches for tree extraction, reviewing 63 papers (2017-2025). The field has experienced 550% growth from 2022 to 2025, evolving from pure ViTs through hybrids to FM adaptations. Key findings: transformers consistently outperform CNN baselines (+3.4% median IoU for hybrids) with architectural choices mattering more than scale; models exhibit 28% average performance drops across domain shifts, with FMs pre-trained on diverse data demonstrating better robustness; and FMs with PEFT offer the best path for limited-label scenarios while specialist architectures maximize accuracy with abundant data.

This survey contributes a data landscape analysis revealing geographic bias (85% Northern Hemisphere, 27.4% fully reproducible), an architecture taxonomy documenting the shift toward FMs, quantified performance comparisons,

and practical deployment recommendations. Priority directions include multi-biome benchmarks, domain adaptation research, efficient edge architectures, and reproducibility standards. Realizing transformer potential for global forest monitoring requires addressing limitations in generalization, efficiency, and open science to accelerate progress toward scalable tree extraction systems supporting sustainability goals.

References

[1] B. X. Lee et al., "Transforming our world: Implementing the 2030 agenda through sustainable development goal indicators," *Journal of public health policy*, vol. 37, no. Suppl 1, pp. 13–31, 2016, ISBN: 0197-5897 Publisher: Springer.

[2] G. Grassi, J. House, F. Dentener, S. Federici, M. Den Elzen, and J. Penman, "The key role of forests in meeting climate targets requires science for credible mitigation," *Nature Climate Change*, vol. 7, no. 3, pp. 220–226, 2017, ISBN: 1758-678X Publisher: Nature Publishing Group UK London.

[3] J.-F. Bastin et al., "The global tree restoration potential," *Science*, vol. 365, no. 6448, pp. 76–79, 2019, ISBN: 0036-8075 Publisher: American Association for the Advancement of Science.

[4] J. Penca and M. Tănăsescu, "The transformative potential of the EU's Nature Restoration Law," *Sustainability Science*, vol. 20, no. 2, pp. 643–647, 2025, ISBN: 1862-4065 Publisher: Springer.

[5] D. J. Nowak, "Understanding i-Tree: 2023 summary of programs and methods," *General Technical Report NRS-200-2023. Madison, WI: US Department of Agriculture, Forest Service, Northern Research Station. 103 p.[plus 14 appendixes].*, vol. 200, 2024.

[6] H. Zhao, J. Morgenroth, G. Pearse, and J. Schindler, "A Systematic Review of Individual Tree Crown Detection and Delineation with Convolutional Neural Networks (CNN)," en, *Current Forestry Reports*, vol. 9, no. 3, pp. 149–170, Apr. 2023, ISSN: 2198-6436. DOI: 10.1007/s40725-023-00184-3. [Online]. Available: http://dx.doi.org/10.1007/s40725-023-00184-3.

[7] S. Khan, M. Naseer, M. Hayat, S. W. Zamir, F. S. Khan, and M. Shah, "Transformers in vision: A survey," *ACM computing surveys (CSUR)*, vol. 54, no. 10s, pp. 1–41, 2022, ISBN: 0360-0300 Publisher: ACM New York, NY.

[8] A. A. Aleissaee et al., "Transformers in remote sensing: A survey," *Remote Sensing*, vol. 15, no. 7, p. 1860, 2023, ISBN: 2072-4292 Publisher: MDPI.

[9] M. B. A. Gibril et al., "Deep convolutional neural networks and Swin transformer-based frameworks for individual date palm tree detection and mapping from large-scale UAV images," *Geocarto International*, vol. 37, no. 27, pp. 18 569–18 599, 2022, ISBN: 1010-6049 Publisher: Taylor & Francis.

[10] M. Teng, A. Ouaknine, E. Laliberté, Y. Bengio, D. Rolnick, and H. Larochelle, "Bringing SAM to new heights: Leveraging elevation data for tree crown segmentation from drone imagery," *arXiv preprint arXiv:2506.04970*, 2025.

[11] J. Zheng, S. Yuan, W. Li, H. Fu, L. Yu, and J. Huang, "A Review of Individual Tree Crown Detection and Delineation From Optical Remote Sensing Images: Current progress and future," *IEEE Geoscience and Remote Sensing Magazine*, vol. 13, no. 1, pp. 209–236, Mar. 2025, ISSN: 2168-6831. DOI: 10.1109/MGRS.2024.3479871. [Online]. Available: http://dx.doi.org/10.1109/MGRS.2024.3479871.

[12] R. Wang et al., "Transformers for remote sensing: A systematic review and analysis," *Sensors*, vol. 24, no. 11, p. 3495, 2024, ISBN: 1424-8220 Publisher: MDPI.

[13] R. Abreu-Dias, J. M. Santos-Gago, F. Martín-Rodríguez, and L. M. Álvarez-Sabucedo, "Advances in the Automated Identification of Individual Tree Species: A Systematic Review of Drone- and AI-Based Methods in Forest Environments," en, *Technologies*, vol. 13, no. 5, p. 187, May 2025, ISSN: 2227-7080. DOI: 10.3390/technologies13050187. [Online]. Available: http://dx.doi.org/10.3390/technologies13050187.

[14] Y. Diez, S. Kentsch, M. Fukuda, M. L. L. Caceres, K. Moritake, and M. Cabezas, "Deep learning in forestry using uav-acquired rgb data: A practical review," *Remote Sensing*, vol. 13, no. 14, p. 2837, 2021, ISBN: 2072-4292 Publisher: MDPI.

[15] L. Zhong, Z. Dai, P. Fang, Y. Cao, and L. Wang, "A review: Tree species classification based on remote sensing data and classic deep learning-based methods," *Forests*, vol. 15, no. 5, p. 852, 2024, ISBN: 1999-4907 Publisher: MDPI.

[16] L. Velasquez-Camacho, A. Cardil, M. Mohan, M. Etxegarai, G. Anzaldi, and S. de-Miguel, "Remotely sensed tree characterization in urban areas: A review," *Remote Sensing*, vol. 13, no. 23, p. 4889, 2021, ISBN: 2072-4292 Publisher: MDPI.

[17] X. Li et al., "Transformer-based visual segmentation: A survey," *IEEE transactions on pattern analysis and machine intelligence*, 2024, ISBN: 0162-8828 Publisher: IEEE.

[18] S. Lu et al., "Vision foundation models in remote sensing: A survey," *IEEE Geoscience and Remote Sensing Magazine*, 2025, ISBN: 2168-6831 Publisher: IEEE.

[19] B. Chehreh, A. Moutinho, and C. Viegas, "Latest trends on tree classification and segmentation using UAV data—A review of agroforestry applications," *Remote sensing*, vol. 15, no. 9, p. 2263, 2023, ISBN: 2072-4292 Publisher: MDPI.

[20] Y. Liu et al., "A survey of visual transformers," *IEEE transactions on neural networks and learning systems*, vol. 35, no. 6, pp. 7478–7498, 2023, ISBN: 2162-237X Publisher: IEEE.

[21] J. S. Estrada, A. Fuentes, P. Reszka, and F. Auat Cheein, "Machine learning assisted remote forestry health assessment: A comprehensive state of the art review," *Frontiers in plant science*, vol. 14, p. 1 139 232, 2023, ISBN: 1664-462X Publisher: Frontiers Media SA.

[22] L. Tao et al., "Advancements in vision–language models for remote sensing: Datasets, capabilities, and enhancement techniques," *Remote Sensing*, vol. 17, no. 1, p. 162, 2025, ISBN: 2072-4292 Publisher: MDPI.

[23] B. G. Weinstein et al., "A benchmark dataset for canopy crown detection and delineation in co-registered airborne RGB, LiDAR and hyperspectral imagery from the National Ecological Observation Network," *PLoS computational biology*, vol. 17, no. 7, e1009180, 2021, ISBN: 1553-734X Publisher: Public Library of Science San Francisco, CA USA.

[24] Z. Xiong et al., "Geolangbind: Unifying earth observation with agglomerative vision-language foundation models," *arXiv preprint arXiv:2503.06312*, 2025.

[25] T. Chang et al., "VibrantVS: A High-Resolution Vision Transformer for Forest Canopy Height Estimation," *Remote Sensing*, vol. 17, no. 6, p. 1017, 2025, ISBN: 2072-4292 Publisher: MDPI.

[26] M. Chen, D. Russell, A. Pallavoor, D. Young, and J. Wu, "Zero-Shot Tree Detection and Segmentation from Aerial Forest Imagery," *arXiv preprint arXiv:2506.03114*, 2025.

[27] Q. Liang, "Application of the vision transformer and mask r-cnn joint algorithm to assist forest decisions," in *2023 5th International Conference on Geoscience and Remote Sensing Mapping (GRSM)*, IEEE, 2023, pp. 127–131, ISBN: 979-8-3503-2955-1.

[28] S. Puliti et al., "For-instance: A uav laser scanning benchmark dataset for semantic and instance segmentation of individual trees," *arXiv preprint arXiv:2309.01279*, 2023.

[29] B. Xiang et al., "ForestFormer3D: A Unified Framework for End-to-End Segmentation of Forest LiDAR 3D Point Clouds," *arXiv preprint arXiv:2506.16991*, 2025.

[30] F. Fogel et al., "Open-Canopy: Towards Very High Resolution Forest Monitoring," in *Proceedings of the Computer Vision and Pattern Recognition Conference*, 2025, pp. 1395–1406.

[31] Z. Yu, M. Idris, and P. Wang, "Satellitecalculator: A multi-task vision foundation model for quantitative remote sensing inversion," *arXiv preprint arXiv:2504.13442*, 2025.

[32] M. Cloutier, M. Germain, and E. Laliberté, "Influence of temperate forest autumn leaf phenology on segmentation of tree species from UAV imagery using deep learning," *Remote Sensing of Environment*, vol. 311, p. 114 283, 2024, ISBN: 0034-4257 Publisher: Elsevier.

[33] J. Zhang, F. Lei, and X. Fan, "Parameter-Efficient Fine-Tuning for Individual Tree Crown Detection and Species Classification Using UAV-Acquired Imagery," *Remote Sensing*, vol. 17, no. 7, p. 1272, 2025, ISBN: 2072-4292 Publisher: MDPI.

[34] M. Teng, A. Ouaknine, E. Laliberté, Y. Bengio, D. Rolnick, and H. Larochelle, "Assessing SAM for Tree Crown Instance Segmentation from Drone Imagery," *arXiv preprint arXiv:2503.20199*, 2025.

[35] J. Veitch-Michaelis et al., "OAM-TCD: A globally diverse dataset of high-resolution tree cover maps," *Advances in neural information processing systems*, vol. 37, pp. 49 749–49 767, 2024.

[36] A. Wilaiwongsakul, B. Liang, W. Jia, B. Zheng, and F. Chen, "BARE: Boundary-Aware with Resolution Enhancement for Tree Crown Delineation," in *AusDM'25*, 2025.

[37] N. I. Bountos, A. Ouaknine, I. Papoutsis, and D. Rolnick, "Fomo: Multi-modal, multi-scale and multi-task remote sensing foundation models for forest monitoring," in *Proceedings of the AAAI Conference on Artificial Intelligence*, Issue: 27, vol. 39, 2025, pp. 27 858–27 868, ISBN: 2374-3468.

[38] X. Qin et al., "TiMo: Spatiotemporal Foundation Model for Satellite Image Time Series," *arXiv preprint arXiv:2505.08723*, 2025.

[39] J. G. Ball et al., "Accurate delineation of individual tree crowns in tropical forests from aerial RGB imagery using Mask R-CNN," *Remote Sensing in Ecology and Conservation*, vol. 9, no. 5, pp. 641–655, 2023, ISBN: 2056-3485 Publisher: Wiley Online Library.

[40] A. Toker et al., "Dynamicearthnet: Daily multi-spectral satellite dataset for semantic change segmentation," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2022, pp. 21 158–21 167.

[41] J. Sadel, L. Tulczyjew, A. M. Wijata, M. Przeliorz, and J. Nalepa, "Monitoring Forest Changes with Foundation Models and Sentinel-2 Time Series," *IEEE Geoscience and Remote Sensing Letters*, 2025, ISBN: 1545-598X Publisher: IEEE.

[42] T. Hackel, N. Savinov, L. Ladicky, J. D. Wegner, K. Schindler, and M. Pollefeys, "Semantic3D.Net: A new large-scale point cloud classification benchmark," *arXiv preprint arXiv:1704.03847*, 2017.

[43] A. Khan, W. Asim, M. Ibrahim, and A. Ulhaq, "3D LiDAR transformer for city-scale vegetation segmentation and biomass estimation," in *2022 International Conference on Digital Image Computing: Techniques and Applications (DICTA)*, IEEE, 2022, pp. 1–7, ISBN: 1-6654-5642-6.

[44] A. Boguszewski, D. Batorski, N. Ziemba-Jankowska, T. Dziedzic, and A. Zambrzycka, "LandCover.AI: Dataset for automatic mapping of buildings, woodlands, water and roads from aerial imagery," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2021, pp. 1102–1110.

[45] G. Voulgaris, "Bridging Classical and Modern Computer Vision: PerceptiveNet for Tree Crown Semantic Segmentation," in *Proceedings of the Computer Vision and Pattern Recognition Conference*, 2025, pp. 2215–2224.

[46] P. Ghamisi, J. A. Benediktsson, and S. Phinn, "Land-cover classification using both hyperspectral and LiDAR data," *International Journal of Image and Data Fusion*, vol. 6, no. 3, pp. 189–215, 2015, ISBN: 1947-9832 Publisher: Taylor & Francis.

[47] X. Shu, L. Ma, and F. Chang, "Integrating Hyperspectral Images and LiDAR Data Using Vision Transformers for Enhanced Vegetation Classification," *Forests*, vol. 16, no. 4, p. 620, 2025, ISBN: 1999-4907 Publisher: MDPI.

[48] S. Soni et al., "Earthdial: Turning multi-sensory earth observations to interactive dialogues," in *Proceedings of the Computer Vision and Pattern Recognition Conference*, 2025, pp. 14 303–14 313.

[49] R. Al-Ruzouq et al., "Spectral–Spatial transformer-based semantic segmentation for large-scale mapping of individual date palm trees using very high-resolution satellite data," en, *Ecological Indicators*, vol. 163, p. 112 110, Jun. 2024, ISSN: 1470-160X. DOI: 10.1016/j.ecolind.2024.112110. [Online]. Available: http://dx.doi.org/10.1016/j.ecolind.2024.112110.

[50] V. Grondin, P. Massicotte, M. Gaha, F. Pomerleau, and P. Giguère, "Leveraging Prompt-Based Segmentation Models and Large Dataset to Improve Detection of Trees," in *Proceedings of the Conference on Robots and Vision*, PubPub, 2024.

[51] A. Jamali, S. K. Roy, J. Li, and P. Ghamisi, "TransU-Net++: Rethinking attention gated TransU-Net for deforestation mapping," *International Journal of Applied Earth Observation and Geoinformation*, vol. 120, p. 103 332, 2023, ISBN: 1569-8432 Publisher: Elsevier.

[52] H. Liu, C. Mou, J. Yuan, Z. Chen, L. Zhong, and X. Cui, "Estimating urban forests biomass with LiDAR by using deep learning foundation models," *Remote Sensing*, vol. 16, no. 9, p. 1643, 2024, ISBN: 2072-4292 Publisher: MDPI.

[53] A. Jamali, S. K. Roy, and P. Ghamisi, "WetMapFormer: A unified deep CNN and vision transformer for complex wetland mapping," *International Journal of Applied Earth Observation and Geoinformation*, vol. 120, p. 103 333, 2023, ISBN: 1569-8432 Publisher: Elsevier.

[54] S. Speckenwirth, M. Brandmeier, and S. Paczkowski, "TreeSeg—A toolbox for fully automated tree crown segmentation based on high-resolution multispectral UAV data," *remote sensing*, vol. 16, no. 19, p. 3660, 2024, ISBN: 2072-4292 Publisher: MDPI.

[55] Y. Huang et al., "Tree species classification from UAV canopy images with deep learning models," *Remote Sensing*, vol. 16, no. 20, p. 3836, 2024, ISBN: 2072-4292 Publisher: MDPI.

[56] S. Sachdeva, I. Lopez, C. Biradar, and D. Lobell, "A distribution shift benchmark for smallholder agroforestry: Do foundation models improve geographic generalization?" In *The Twelfth International Conference on Learning Representations*, 2024.

[57] A. Jamali, S. K. Roy, and B. Pradhan, "E-TransUNet: TransUNet provides a strong spatial transformation for precise deforestation mapping," *Remote Sensing Applications: Society and Environment*, vol. 35, p. 101 221, 2024, ISBN: 2352-9385 Publisher: Elsevier.

[58] Y. Lu et al., "M2fnet: Multi-modal forest monitoring network on large-scale virtual dataset," in *2024 IEEE Conference on Virtual Reality and 3D User Interfaces Abstracts and Workshops (VRW)*, IEEE, 2024, pp. 539–543, ISBN: 979-8-3503-7449-0.

[59] M. Wasil et al., "Parameter-Efficient Fine-Tuning of Vision Foundation Model for Forest Floor Segmentation from UAV Imagery," *arXiv preprint arXiv:2505.08932*, 2025.

[60] H. A. Amirkolaee, M. Shi, and M. Mulligan, "TreeFormer: A semi-supervised transformer-based framework for tree counting from a single high-resolution image," *IEEE transactions on geoscience and remote sensing*, vol. 61, pp. 1–15, 2023, ISBN: 0196-2892 Publisher: IEEE.

[61] H. Shanableh et al., "A Comparative Analysis of Deep Learning Methods for Ghaf Tree Detection and Segmentation from UAV-based Images," *ISPRS Annals of the Photogrammetry, Remote Sensing and Spatial Information Sciences*, pp. 805–811, 2025, ISBN: 2194-9050 Publisher: Copernicus Publications Göttingen, Germany.

[62] Y. Wang, X. Dou, and X. Liang, "Fine-grained individual tree crown segmentation based on high-resolution images," *The International Archives of the Photogrammetry, Remote Sensing and Spatial Information Sciences*, vol. 48, pp. 1529–1535, 2025, ISBN: 1682-1750 Publisher: Copernicus GmbH.

[63] S. R. Begum, V. Mishra, and D. Saha, "Sustainable Forestry: AI-Driven Forest Health Diagnostics using Swin Transformer," in *2025 International Conference on Advances in Modern Age Technologies for Health and Engineering Science (AMATHE)*, IEEE, 2025, pp. 1–8, ISBN: 979-8-3315-0103-7.

[64] N. Wang, T. Pu, Y. Zhang, Y. Liu, and Z. Zhang, "More appropriate DenseNetBL classifier for small sample tree species classification using UAV-based RGB imagery," *Heliyon*, vol. 9, no. 10, 2023, ISBN: 2405-8440 Publisher: Elsevier.

[65] R. Pierdicca, L. Nepi, A. Mancini, E. S. Malinverni, and M. Balestra, "UAV4TREE: Deep Learning-based system for automatic classification of tree species using RGB optical images obtained by an unmanned aerial vehicle," *ISPRS Annals of the Photogrammetry, Remote Sensing and Spatial Information Sciences*, vol. 10, pp. 1089–1096, 2023, ISBN: 2194-9050 Publisher: Copernicus Publications Göttingen, Germany.

[66] J. Jia, J. Kang, L. Chen, X. Gao, B. Zhang, and G. Yang, "A Comprehensive Evaluation of Monocular Depth Estimation Methods in Low-Altitude Forest Environment," *Remote Sensing*, vol. 17, no. 4, p. 717, 2025, ISBN: 2072-4292 Publisher: MDPI.

[67] S. K. Valicharla, R. Karimzadeh, X. Li, and Y.-L. Park, "Transformer-Based Semantic Segmentation of Japanese Knotweed in High-Resolution UAV Imagery Using Twins-SVT," *Information*, vol. 16, no. 9, p. 741, 2025, ISBN: 2078-2489 Publisher: Multidisciplinary Digital Publishing Institute.

[68] M. B. A. Gibril et al., "Efficient Large-scale Mapping of Acacia Tortilis Trees Using UAV-based Images and Transformer-based Semantic Segmentation Architectures," *ISPRS Annals of the Photogrammetry, Remote Sensing and Spatial Information Sciences*, pp. 285–290, 2025, ISBN: 2194-9050 Publisher: Copernicus Publications Göttingen, Germany.

[69] Y. Huang, X. Wen, Y. Gao, Y. Zhang, and G. Lin, "Tree species classification in UAV remote sensing images based on super-resolution reconstruction and deep learning," *Remote Sensing*, vol. 15, no. 11, p. 2942, 2023, ISBN: 2072-4292 Publisher: MDPI.

[70] M. B. A. Gibril et al., "Large-scale assessment of date palm plantations based on UAV remote sensing and multiscale vision transformer," en, *Remote Sensing Applications: Society and Environment*, vol. 34, p. 101 195, Apr. 2024, ISSN: 2352-9385. DOI: 10.1016/j.rsase.2024.101195. [Online]. Available: http://dx.doi.org/10.1016/j.rsase.2024.101195.

[71] P. Sun, X. Yuan, and D. Li, "Classification of individual tree species using UAV LiDAR based on transformer," *Forests*, vol. 14, no. 3, p. 484, 2023, ISBN: 1999-4907 Publisher: MDPI.

[72] M. N. Ton-That, T. V. Le, N. H. Truong, A. D. Le, A.-D. Pham, and H. B. Vo, "Expanding Vision in Tree Counting: Novel Ground Truth Generation and Deep Learning Model," in *2024 Tenth International Conference on Communications and Electronics (ICCE)*, IEEE, 2024, pp. 409–414, ISBN: 979-8-3503-7979-2.

[73] P. Perbet, L. Guindon, J.-F. Côté, and M. Béland, "Evaluating deep learning methods applied to Landsat time series subsequences to detect and classify boreal forest disturbances events: The challenge of partial and progressive disturbances," *Remote Sensing of Environment*, vol. 306, p. 114 107, 2024, ISBN: 0034-4257 Publisher: Elsevier.

[74] S. Du, S. Tang, W. Wang, X. Li, and R. Guo, "Tree-gpt: Modular large language model expert system for forest remote sensing image understanding and interactive analysis," *arXiv preprint arXiv:2310.04698*, 2023.

[75] Y. Gui et al., "Multi-modal Uncertainty Robust Tree Cover Segmentation For High-Resolution Remote Sensing Images," *arXiv preprint arXiv:2509.04870*, 2025.

[76] D. Gominski et al., "Benchmarking individual tree mapping with sub-meter imagery," *arXiv preprint arXiv:2311.07981*, 2023.

[77] M. Kaselimi, A. Voulodimos, I. Daskalopoulos, N. Doulamis, and A. Doulamis, "A vision transformer model for convolution-free multilabel classification of satellite imagery in deforestation monitoring," *IEEE Transactions on Neural Networks and Learning Systems*, vol. 34, no. 7, pp. 3299–3307, 2022, ISBN: 2162-237X Publisher: IEEE.

[78] I. Fayad et al., "Vision transformers, a new approach for high-resolution and large-scale mapping of canopy heights," *arXiv preprint arXiv:2304.11487*, 2023.

[79] L. Zhu, Y. Lin, and C.-W. Lin, "Transformer-based Instance Segmentation with Multi-Scale Spectrum-Averaging Blend Queries," in *2024 9th International Symposium on Computer and Information Processing Technology (ISCIPT)*, IEEE, 2024, pp. 515–518, ISBN: 979-8-3503-8840-4.

[80] D. Joshi and C. Witharana, "Vision transformer-based unhealthy tree crown detection in mixed northeastern us forests and evaluation of annotation uncertainty," *Remote Sensing*, vol. 17, no. 6, p. 1066, 2025, ISBN: 2072-4292 Publisher: MDPI.

[81] S. Dersch, A. Schoettl, P. Krzystek, and M. Heurich, "Towards complete tree crown delineation by instance segmentation with Mask R–CNN and DETR using UAV-based multispectral imagery and lidar data," *ISPRS Open Journal of Photogrammetry and Remote Sensing*, vol. 8, p. 100 037, 2023, ISBN: 2667-3932 Publisher: Elsevier.

[82] F. Zhou et al., "Semantic-Aware Cross-Modal Transfer for UAV-LiDAR Individual Tree Segmentation," *Remote Sensing*, vol. 17, no. 16, p. 2805, 2025, ISBN: 2072-4292 Publisher: MDPI.

[83] X. Xue et al., "Reo-vlm: Transforming vlm to meet regression challenges in earth observation," *arXiv preprint arXiv:2412.16583*, 2024.

[84] M. B. A. Gibril, H. Z. M. Shafri, R. Al-Ruzouq, A. Shanableh, F. Nahas, and S. Al Mansoori, "Large-scale date palm tree segmentation from multiscale uav-based and aerial images using deep vision transformers," *Drones*, vol. 7, no. 2, p. 93, 2023, ISBN: 2504-446X Publisher: MDPI.

[85] L. Zhang, H. Lin, and F. Wang, "Individual tree detection based on high-resolution RGB images for urban forestry applications," *IEEE Access*, vol. 10, pp. 46 589–46 598, 2022, ISBN: 2169-3536 Publisher: IEEE.

[86] H. Zhang, C. Yang, and X. Fan, "MTCDNet: Multimodal Feature Fusion-Based Tree Crown Detection Network Using UAV-Acquired Optical Imagery and LiDAR Data," *Remote Sensing*, vol. 17, no. 12, p. 1996, 2025, ISBN: 2072-4292 Publisher: MDPI.

[87] P. J. S. Vega, D. L. Torres, G. X. Andrade-Miranda, and R. Q. Feitosa, "Assessing the Generalization Capacity of Convolutional Neural Networks and Vision Transformers for Deforestation Detection in Tropical Biomes," in *ISPRS Archives of the Photogrammetry, Remote Sensing and Spatial Information Sciences*, vol. 48, 2024, pp. 519–525.

[88] Z. Wang, J. Yang, C. Dong, X. Zhang, C. Yi, and J. Sun, "SSMM-DS: A semantic segmentation model for mangroves based on Deeplabv3+ with swin transformer.," *Electronic Research Archive*, vol. 32, no. 10, 2024, ISBN: 2688-1594.

[89] S. Dersch, A. Schöttl, P. Krzystek, and M. Heurich, "Novel single tree detection by transformers using uav-based multispectral imagery," *The International Archives of the Photogrammetry, Remote Sensing and Spatial Information Sciences*, vol. 43, pp. 981–988, 2022, ISBN: 1682-1750 Publisher: Copernicus GmbH.

[90] M. Muszynski et al., "Fine-tuning of geospatial foundation models for aboveground biomass estimation," *arXiv preprint arXiv:2406.19888*, 2024.

[91] F. Ferrari, M. P. Ferreira, and R. Q. Feitosa, "Fusing SENTINEL-1 and SENTINEL-2 images with transformer-based network for deforestation detection in the Brazilian Amazon under diverse cloud conditions," *ISPRS Annals of the Photogrammetry, Remote Sensing and Spatial Information Sciences*, vol. 10, pp. 999–1006, 2023, ISBN: 2194-9042 Publisher: Copernicus GmbH.

[92] A. Mullissa, J. Reiche, and S. Saatchi, "Seasonal Forest Disturbance Detection Using Sentinel-1 SAR & Sentinel-2 Optical Timeseries Data and Transformers," in *IGARSS 2023-2023 IEEE International Geoscience and Remote Sensing Symposium*, IEEE, 2023, pp. 3122–3124, ISBN: 979-8-3503-2010-7.

[93] J. Yang, L. El Mendili, Y. Khayer, S. McArdle, and L. Hashemi Beni, "Instance Segmentation of LiDAR Data with Vision Transformer Model in Support Inundation Mapping under Forest Canopy Environment," 2023.

[94] A. Dosovitskiy et al., "An image is worth 16x16 words: Transformers for image recognition at scale," *arXiv preprint arXiv:2010.11929*, 2020.

[95] S. Zheng et al., "Rethinking semantic segmentation from a sequence-to-sequence perspective with transformers," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2021, pp. 6881–6890.

[96] J. Chen et al., "Transunet: Transformers make strong encoders for medical image segmentation," *arXiv preprint arXiv:2102.04306*, 2021.

[97] N. Carion, F. Massa, G. Synnaeve, N. Usunier, A. Kirillov, and S. Zagoruyko, "End-to-end object detection with transformers," in *European conference on computer vision*, Springer, 2020, pp. 213–229.

[98] B. Cheng, I. Misra, A. G. Schwing, A. Kirillov, and R. Girdhar, "Masked-attention mask transformer for universal image segmentation," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2022, pp. 1290–1299.

[99] Z. Liu et al., "Swin transformer: Hierarchical vision transformer using shifted windows," in *Proceedings of the IEEE/CVF international conference on computer vision*, 2021, pp. 10 012–10 022.

[100] E. Xie, W. Wang, Z. Yu, A. Anandkumar, J. M. Alvarez, and P. Luo, "SegFormer: Simple and efficient design for semantic segmentation with transformers," *Advances in neural information processing systems*, vol. 34, pp. 12 077–12 090, 2021.

[101] W. Wang et al., "Pyramid vision transformer: A versatile backbone for dense prediction without convolutions," in *Proceedings of the IEEE/CVF international conference on computer vision*, 2021, pp. 568–578.

[102] M. Ding, B. Xiao, N. Codella, P. Luo, J. Wang, and L. Yuan, "Davit: Dual attention vision transformers," in *European conference on computer vision*, Springer, 2022, pp. 74–92.

[103] K. Li et al., "Uniformer: Unifying convolution and self-attention for visual recognition," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 45, no. 10, pp. 12 581–12 600, 2023, ISBN: 0162-8828 Publisher: IEEE.

[104] X. Chu et al., "Twins: Revisiting the design of spatial attention in vision transformers," *Advances in neural information processing systems*, vol. 34, pp. 9355–9366, 2021.

[105] A. Kirillov et al., "Segment anything," in *Proceedings of the IEEE/CVF international conference on computer vision*, 2023, pp. 4015–4026.

[106] S. Liu et al., "Grounding dino: Marrying dino with grounded pre-training for open-set object detection," in *European conference on computer vision*, Springer, 2024, pp. 38–55.

[107] A. Vaswani et al., "Attention is all you need," *Advances in neural information processing systems*, vol. 30, 2017.