Evaluating Broadscale Deep Learning for Maya Settlement Detection in G-LiHT Lidar

Authors:
Benjamin J. Britton[1] (ORCID: 0009-0006-7475-3795)
Alec McLellan[2] (ORCID: 0000-0002-4392-5391)
Jeffrey Brewer[3] (ORCID: 0000-0001-9724=2601)
Christopher Carr[4] (ORCID: 0000-0001-8891-5923)
Nicholas Dunning[5] (ORCID: 0000-0002-1843-3088)
Lin Liu[6] (ORCID: 0000-0002-7202-3418)

1 Department of Geography, College of Arts & Sciences, University of Cincinnati, Cincinnati, Ohio, USA
2 Department of Anthropology, College of Arts & Sciences, University of Cincinnati, Cincinnati, Ohio, USA
3 Department of Geography, College of Arts & Sciences, University of Cincinnati, Cincinnati, Ohio, USA
4 Department of Geography, College of Arts & Sciences, University of Cincinnati, Cincinnati, Ohio, USA
5 Department of Geography, College of Arts & Sciences, University of Cincinnati, Cincinnati, Ohio, USA
6 Department of Geography, College of Arts & Sciences, University of Cincinnati, Cincinnati, Ohio, USA

Corresponding Author:
Benjamin Britton
Department of Geography, College of Arts & Sciences, University of Cincinnati, Cincinnati, Ohio, USA
Email: brittobj@mail.uc.edu
Phone: +1 513-728-9852
ORCID iD: 0009-0006-7475-3795

Keywords: Lidar, Maya, archaeology, deep learning, object detection, Broadscale Model

A B S T R A C T

Examining Lidar data is an efficient way to detect ancient Maya features across the Yucatan Peninsula. Automated object detection powered by deep learning leverages Maya archaeologists' specialist knowledge in detecting the presence of ancient Maya settlements. By using a broadscale approach in its training, our new efficient multi-regional model Q2000 achieves comparable performance across a significantly broader and more diverse geographic region. This study addresses the current limitation of small-scale, area-specific models to generalize characteristics and properly detect a diverse range of target objects over a large area. This study introduces the foundational development of a broadscale, multi-region convolutional neural network (CNN) object detection model utilizing Lidar data across a significantly larger extent of the Maya area (approximately 35,584 km²). This model achieved accuracies comparable to previous local studies that relied on the annotation of a larger number of structures within smaller, more homogeneous areas. Comparative analysis of the model's test results indicates enhanced generalization across diverse topographic regions when trained on multi-area data, achieving a robust F1 Score of 0.89, even with a relatively small training sample set. Our results further indicate that a broadscale approach to deep learning is efficient, and that a pan-Yucatan model can be effective.

1. Introduction

The remains of ancient Maya civilization, hidden beneath the dense tropical forests of the Yucatan Peninsula, have long presented formidable challenges to archaeological research. Historically, scientists trudged through jungles, camped amid mosquitoes, dodged vipers, and hacked brush, mapping ruins with tapes and transits (Chase, 1988). The dense jungle environment significantly hampered early remote sensing methods, hiding sites and artifacts from view, making exhaustive ground-based surveys the dominant, albeit arduous, exploratory approach. While these efforts yielded valuable foundational knowledge, the scale and complexity of the Maya world demanded more efficient, comprehensive approaches to uncover the extent of their landscapes, infrastructure, and settlement patterns. The advent of Airborne Lidar Survey (ALS) technology marked a revolutionary turning point, enabling researchers to penetrate the forest canopy to reveal previously unimagined archaeological features across vast territories (Chase et al., 2011; Fernandez-Diaz et al., 2014; Inomata, 2017; Ruhl et al., 2017). This unprecedented wealth of data, however, brought forth a new challenge: the need for effective computational methods to process and analyze these extensive datasets to address complex questions about ancient Maya society and their interaction with the environment.

The primary objectives of this study are twofold: first, to evaluate the practicability of a broadscale deep learning model for the identification of archaeological features in unannotated Lidar imagery across the extensive Maya region to facilitate discovery and location of previously undocumented sites and features; and second, to develop a robust and replicable methodological framework for application of deep learning techniques in archaeological feature detection across diverse geographical contexts. Ultimately, this research aims to contribute to a collaborative framework that promotes a deeper understanding of ancient Maya culture and to further demonstrate the potential of the G-LiHT dataset for achieving this objective, offering a significant advancement over the limitations of geographically restricted models.

Deep learning architectures—including convolutional neural networks (CNNs) and emerging Transformer-based models—enable the rapid processing of large image datasets and identification of intricate patterns (Khan et al., 2023). This capability makes deep learning well-suited for detecting archaeological features across the varied landscapes of the Maya region—a crucial requirement for broadscale mapping. The increasing global adoption of machine learning in archaeology underscores its transformative potential (e.g., Argyrou & Agapiou, 2022; Bickler, 2021; Canedo et al., 2023; Canedo et al., 2024; Davis, 2020; Galwey et al., 2019; Marçal et al. 2024, Wang et al., 2024). Furthermore, a fundamental aspect of archaeological inquiry involves accurately differentiating between anthropogenic and natural features. In the Maya region, particularly in the southern lowlands, architectural practices often echoed natural landforms. As a result, many archaeological features—such as overgrown pyramids, terraces, and mounds—can resemble natural hills or ridges in lidar datasets (Fig. 1). Deep learning techniques offer valuable support in distinguishing these cultural features from natural terrain, enhancing the accuracy of remote sensing analyses.

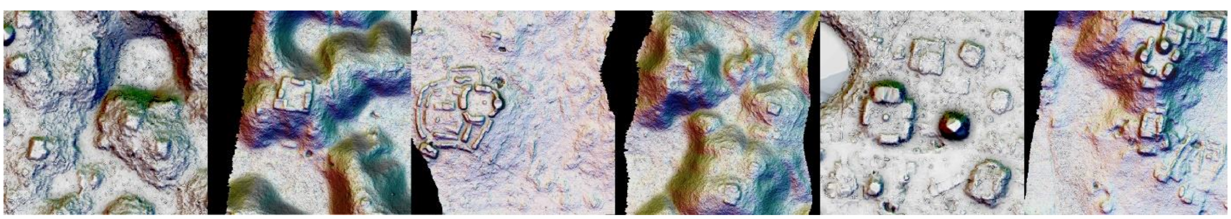


Fig. 1 *Examples of annotated training samples necessary for automated search of Lidar data for Maya structure; these training samples of target objects in the Maya lowlands were saved as 416x416 pixel, geospatially referenced TIFF images; in the North2 region, clusters of platforms around circular depressions in the limestone (rejolladas) are a common settlement pattern, in the South region, hilltop compounds with plazas and buildings are common*

Deep learning tools automate the search for transformations of the landscape that can be archaeological. The potential for using Lidar with deep learning tools to facilitate research has been a goal of Maya researchers for years (Beach et al., 2015; Bundzel et al., 2020; Character et al., 2024; Chase et al., 2017; Inomata, 2024; Jannat et al., 2023; Richards-Rissetto et al., 2024; Sevara et al., 2016; Somrak et al., 2020). This potential has driven the recognition of AI, machine learning principles, and deep learning methods applied to the analysis of Lidar data for Maya archaeology and has highlighted Lidar as an invaluable source for locating massive amounts of unrecorded Maya structures (Bundzel et al., 2020; Character et al. 2024; Jannat, 2023; Kokalj et al., 2023; Richards-Rissetto et al., 2021; Sevara et al., 2016; Somrak et al., 2020). Globally, since its adoption by archaeologists, there has been an explosion of studies around the world using Lidar with deep learning tools, including studies of the Pacific region (Bickler & Jones, 2021; Quintus et al., 2023), North America (Davis et al., 2021; der Vaart et al., 2023), Asia (Azarkhordad et al., 2025; Landauer et al., 2025; Wang et al., 2024), and Europe (Canedo et al., 2023; Canedo et al., 2024; Galwey et al., 2019; Marçal et al., 2024; Trier et al., 2021, Verschoof-van der Vaart & Landers, 2019).

Assessing population densities, unraveling the processes of urban growth, understanding their relationships with each other in space, and management of the region’s vital resources, like water, plants, and animals, can be effectively done using automated processes that can record and assess the data in vast amounts, commensurate with the vast quantities of data that Lidar provides. For this reason, archaeologists seek to use deep learning (DL) tools to identify, quantify, and analyze ancient Maya structures. A deep learning model designed to locate Maya ruins throughout the Maya lowlands could enable researchers to identify and analyze archaeological structures in days rather than decades, achieving a level of analysis that was previously unattainable.

While localized datasets and models have been developed for specific regions in the Maya area, there is increasing recognition of the value of sharing this data. This collaborative approach aims to address broader research questions, enabling a deeper and more illuminating understanding of the ancient Maya. For example, Maya researchers have already leveraged deep learning (Kokalj et al., 2023; Schroder et al., 2020) to analyze population trends, environmental impacts, paleo-climatological information, as well as the characteristics of ancient Maya population centers within specific regions. By combining data from diverse sources into prototype deep learning frameworks for archaeological data, this systematic cooperative information sharing can significantly enhance and expand our collective human knowledge.

To address the limitations inherent in localized models, this study investigates the potential of a broadscale deep learning approach—one that leverages geographically expansive and heterogeneous datasets as sources of training samples to enhance model practicability and performance. Specifically, we assess whether the advantages of this approach, stemming from its ability to incorporate increased terrain variability, diverse settlement morphologies and architectural types, and a broader pool of training samples, can be effectively realized.

In this study, we define 'broadscale' as referring to models trained across large spatial extents or heterogeneous datasets that span varied geographic, cultural, or environmental contexts. For example, rather than training a model exclusively on a contiguous block of 230 km² from Campeche, Mexico, we explore the feasibility of training on integrated datasets encompassing over 36,000 km² and spanning multiple regions across the Yucatán Peninsula.

This expansion introduces a broader range of geomorphological conditions, vegetation regimes, and settlement configurations, compelling the model to learn more transferable features and reducing the risk of it becoming too narrowly specialized to a single region, a limitation known as overfitting, as discussed by Goodfellow et al. (2016). Viewed through this broadscale lens, the model demonstrates improved capacity for identifying settlement indicators in previously unexamined landscapes, thereby advancing its potential to support site identification in environmentally and culturally diverse zones.

These potential benefits of broadscale deep learning are accompanied by challenges inherent in geographically extensive data. These include heightened computational demands due to dataset heterogeneity, the risk of overfitting due to data complexity, and hurdles in maintaining data consistency (see Section 5 for a detailed discussion). To navigate these complexities and realize the potential of broadscale models for generalizable applications, this study employs a Human-in-the-Loop (HITL) framework—a collaborative approach in which human expertise is integrated into machine learning workflows to improve accuracy, relevance, and transparency—as a key mechanism for managing data complexity, reducing bias, and optimizing model performance. For example, our team included archaeologists and geographers with expertise in quantitative spatial data analysis, lidar visualization of Maya structures and settlements, and ground-truthed Maya architecture. These domain experts consulted regularly to review proposed training samples, evaluate visualization strategies, and guide data analysis protocols. This perspective aligns with the general understanding that while deep learning models can automate certain tasks and improve efficiency, human interpretation and validation remain crucial for accurate and meaningful insights, as domain experts are essential to verify and interpret goals, targets, and AI-detected features to ensure their reliability and thorough analysis. A visualization of our HITL workflow is presented in Section 3. The application of HITL in this study, and its broader significance for deep learning, is further elaborated in Section 5.

Building upon the transformative capabilities of Lidar and the analytical power of deep learning, several key studies have paved the way for this research. Schroder and colleagues (2020) highlighted the significant value of the NASA Goddard Lidar, Hyperspectral, and Thermal (G-LiHT) dataset, the same dataset utilized in this study, for representing Maya architecture across the Yucatan Peninsula and advocated for collaborative landscape-level model development. Similarly, Ringle and colleagues (2021) underscored the potential of the G-LiHT dataset for a broader understanding of Maya structural diversity through point-cloud reprocessing. Furthering the collaborative trend, Kokalj and colleagues (2023) offered an annotated dataset to facilitate machine learning research in the Maya region. Zhang and colleagues (2024) highlighted the efficiency of deep learning models for segmenting archaeological features in Lidar imagery of the Puuc region in Yucatan, Mexico, but significantly cautioned that while these advancements show great promise, their research also confirmed the crucial importance of human expertise in validating AI-driven results. They emphasized that manual identification and confirmation of archaeological features are labor-intensive and require specialized knowledge that AI can complement but not entirely replace, asserting that the combination of AI and human review ensures both efficiency and accuracy in archaeological surveys.

Character and colleagues (2024) also strongly advocated for a collaborative approach combining human and machine learning in archaeological research. They proposed a comprehensive convolutional neural network (CNN) model for detecting and locating Maya structures across the entire peninsula, envisioning the potential to map archaeological areas in the Maya Lowlands within weeks or months, rather than decades. Their goal was to develop a broadscale model capable of analyzing unlabeled Lidar imagery from across the Maya region.

Although their data did not cover the entire Yucatan Peninsula (excluding the central and northern sections), they successfully developed a prototype for a pan-Yucatan deep learning model. This included creating robust methods of processing Aerial Lidar Surveys (ALS) as foundational steps toward a comprehensive collaborative study, This research aimed to identify and locate Maya archaeological structures using Lidar data, ultimately seeking to provide a deep learning model capable of recognizing Maya architecture across all areas of the Yucatan peninsula. Such a tool would enable researchers to efficiently uncover potential Maya ruins in Lidar data, thereby advancing knowledge of human culture and offering significant time and cost efficiencies.

Building on these previous efforts and aiming to overcome the geographical limitations of prior work, this study uses data from the entire Yucatan Peninsula to replicate established methods, compare them with newer ones, and evaluate deep learning tools for analyzing ancient Maya architecture, settlement patterns, and terrain. This study assesses the feasibility of a comprehensive pan-Yucatan object detection tool to promote collaboration among researchers for a deeper understanding of ancient Maya culture and the potential of the G-LiHT dataset to contribute to that goal. Results demonstrate that G-LiHT point cloud data provides a strong foundation for creating regional deep learning models for the study of pan-Yucatan Maya features, effectively accounting for the diversity of settlement patterns, architecture styles, and environmental terrains. Furthermore, the study confirms that G-LiHT point clouds can be productively reprocessed to achieve improved spatial resolution.

The Yucatan Peninsula, approximately 190,000 km^2 in size, presents an immense archaeological landscape. Despite advancements in applying deep learning to Maya archaeology, including proposals for broader scale models, the current research landscape is predominantly characterized by models specifically designed for archaeological feature detection within limited geographical areas (typically <615 km^2). These existing models, while often achieving high accuracy within their relatively homogeneous input data, are tailored for specific sites. Consequently, they necessitate substantial retraining for deployment in new geographical contexts and do not adequately address the need for models applicable to the extensive unmapped regions of Mexico and Central America. Recognizing this critical gap and building upon the potential of landscape-level analysis, this study investigates the practical feasibility of developing a comprehensive Pan-Yucatan deep learning model for archaeological detection. The following sections will detail our methodology, the results of the broadscale model's validation, a discussion of the implications and challenges of this approach, and finally, directions for future work.

2. Background

The visionary work of Character and colleagues (2024), who proposed a broadscale CNN model for mapping archaeological areas across the Maya Lowlands, directly inspired this study. This research investigates the practicability of developing a deep learning model for Maya archaeology covering the entire Yucatan peninsula – a veritable Pan-Yucatan model with significant practical utility. Its design aims for broad applicability, allowing it to be effectively adapted through transfer learning to new, unanalyzed areas. This study presents the development and evaluation of such a model, specifically a convolutional neural network, for the automated detection of Maya archaeological structures. The model was trained and tested using Lidar data from the NASA Goddard G-LiHT dataset, encompassing two regions in the Yucatan Peninsula: South and North2, totaling an area of about 36,584 km^2

(Fig. 2). The model's capacity for generalization was assessed through examination of its detection performance on previously unanalyzed terrain.

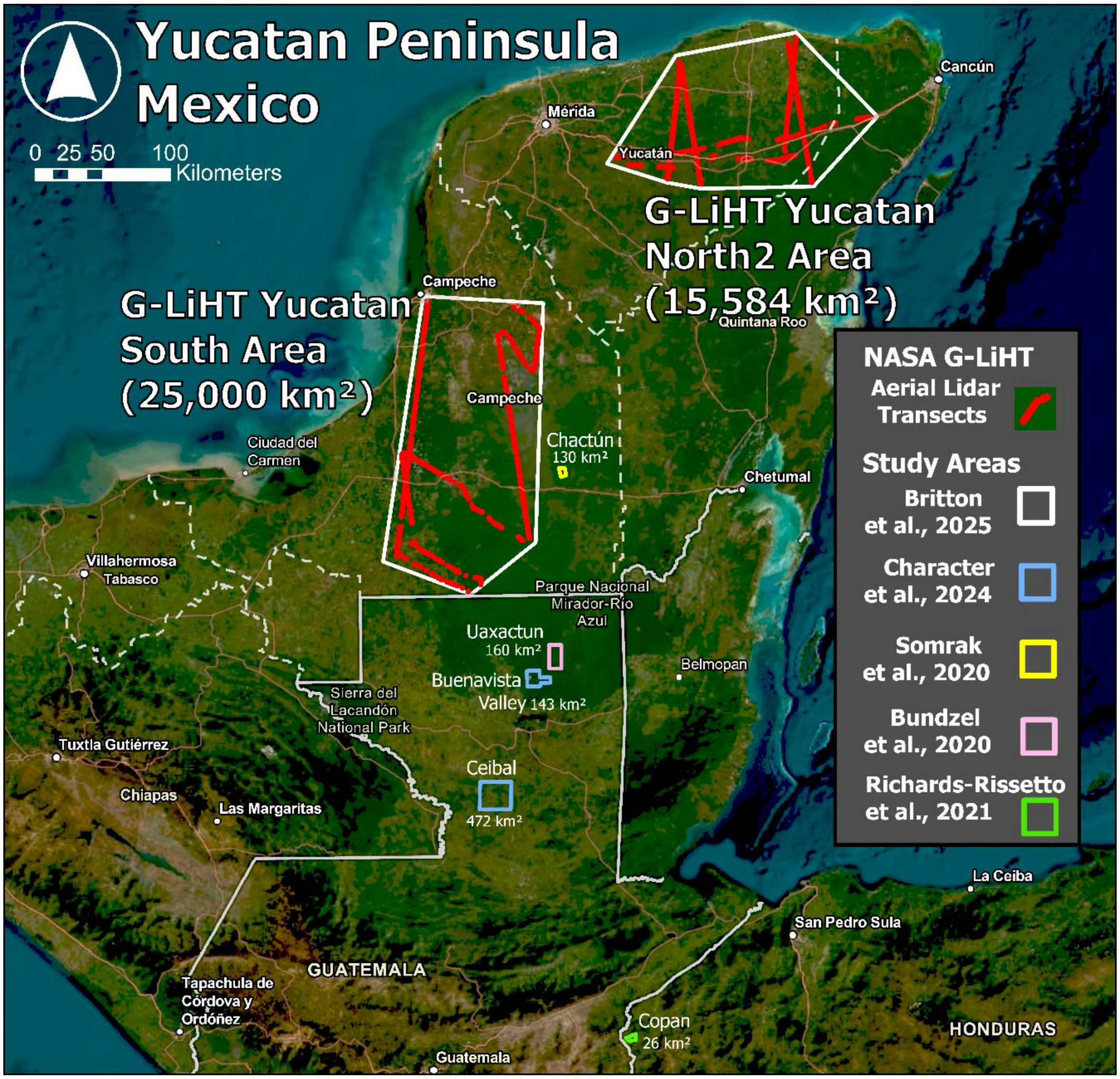


Fig. 2 *Map showing study areas in the Yucatan Peninsula, Mexico; NASA Goddard Lidar, Hyperspectral, and Thermal (G-LiHT) regions "South" & "North2"; red markings represent individual G-LiHT aerial transects*

The potential of the G-LiHT dataset for this purpose stems from its long-recognized value to contribute to a greater understanding of Maya archaeology (Golden et al., 2016). It serves as a crucial resource for collaborative efforts to gain insight into the natural and anthropological evolution of the Yucatan peninsula at a macro-regional level. The G-LiHT Yucatan dataset was initially created to examine forest cover and biomass in that region of Mexico, so the parameters of its flight plans were presumably influenced by this original motivation. This unbiased origin makes the data exceptionally valuable for archaeological study. Without pre-existing archaeological biases toward specific areas or sites of interest by subject area experts, and without inclinations toward regions historically favored by archaeological specialists, the dataset criss-crosses the region, impartially capturing a pseudo-random snapshot of the entire peninsula's terrain. Spatial analysis of this data is therefore uncorrupted by archaeological preconception. The data provided in these aerial transects lays bare a broad picture of ancient Maya settlement patterns across the peninsula. For pan-Yucatan analyses of spatial patterns in Maya archaeology, the G-LiHT dataset offers an invaluable snapshot of the peninsula, providing a robust basis for developing models capable of broad transfer learning applications across the region.

For this comparative analysis, we focus on four previously published deep learning models designed for archaeological feature detection within the Maya region. Each of these models was custom-developed for its respective archaeological site, trained to operate effectively over relatively small extents (typically <615 km²). In contrast to these geographically constrained approaches, the current study focuses on a significantly larger combined area of 36,584 km², encompassing two distinct archaeological zones of the Yucatan Peninsula.

It is important to clarify that our comparison does not involve re-testing these existing models on new, shared terrain, nor applying our model to their specific, limited study areas. Instead, we compare the performance of these existing localized models (which often show high accuracy due to their relatively homogeneous input data) with the results of the present study in Section 4 (Results) and summarized in Table 5. This will be done by examining their reported performance metrics (precision, recall, F1-score, and accuracy where available) from their original evaluations on previously unanalyzed data within their respective target regions, alongside our model's performance on previously unreviewed terrain from within its own distinct, broader target area across the Yucatan Peninsula. This comparative analysis aims to highlight the advantages and challenges inherent in a broadscale approach, specifically underscoring the inherent limitations of localized models in addressing the imperative need for models applicable across the extensive, unmapped regions of Mexico and Central America.

Character and colleagues (2024) applied a one-stage object detection architecture, YOLOv3, to Simple Local Relief Model (SLRM) Lidar visualizations from two archaeological sites in Petén, Guatemala: Ceibal (472 km²) and the Buenavista Valley (143 km²), together covering approximately 615 km². Their Lidar data from both Ceibal (Inomata et al., 2017) and Buenavista Valley (Canuto et al., 2018) was acquired using a Teledyne Optech Titan MW, typically emitting 15–19 laser shots per square meter. The combined Lidar dataset effectively yielded a density of 2.07 ground points per square meter, which was used to produce digital elevation models (DEMs) with 1 m/pixel spatial resolution for both areas, further visualized as SLRM imagery. The relatively high point density achieved in this study provides a useful benchmark for considering comparisons and potential improvements to other datasets like G-LiHT.

Their training dataset comprised 14,445 mounds from the Ceibal site and 1,300 mound structures from Buenavista Valley, resulting in 15,745 labeled mound objects across both areas. Character et al. produced two SLRM-trained models. Their Model 1 achieved test scores of 0.98 for precision, 0.61 for recall, and 0.76 for F1 Score. Their Model 2 achieved test scores of 1.0 for precision, 0.49 for recall, and 0.66 for F1 Score. Notably, unlike other deep learning projects reviewed in this section, specifically Bundzel et al., (2020), Somrak et al., (2020), and Richards-Rissetto et al., (2021)**,** Character et al. did not report an accuracy test score for their two models.

Bundzel and colleagues (Bundzel et al., 2020) conducted a comparative analysis of two distinct deep learning architectures, U-Net and Mask R-CNN, for the identification of ancient Maya structures across a 160 km² area around the Uaxactun archaeological site in Petén, Guatemala. The LiDAR data utilized in their research were acquired using a Teledyne Optech Titan MW system (Fernandez-Diaz et al., 2016), with average ground return densities ranging from 1.1 to 5.3 returns per square meter. This range highlights the variability encountered in LiDAR data and underscores the potential benefits of reprocessing datasets with lower densities. A digital elevation model (DEM) with a spatial resolution of 1 m/pixel was generated from the last return data and used as input for both the U-Net and Mask R-CNN models.

Their models—each implemented in both U-Net and Mask R-CNN variants—were trained to identify two classes of objects: “Mounds” (trained with 670 annotated target objects) and “Structures” (trained with approximately 600 annotated target objects). These annotated target objects were manually labeled by experts from the PARU project and validated through fieldwork in 2017 and 2019. The resulting models achieved overall accuracy scores of up to 0.99, an F1 score of 0.62, a recall (TPR) of 0.60, and a precision (PPV) of 0.72. U-Net consistently outperformed Mask R-CNN, particularly in the Mounds segmentation task.

Somrak and colleagues (Somrak et al., 2020) employed a VGG-19 convolutional neural network (CNN) to classify multiple types of Maya structures across a 130 km² area around the Chactún archaeological site in Campeche, Mexico. The LiDAR data were acquired independently by the National Center for Airborne Laser Mapping (NCALM) using a Teledyne Optech Titan MW system (Fernandez-Diaz et al., 2016), with an average classified ground return density of 14.7 points per square meter. This high density enabled the generation of a 0.5 m/pixel DEM, contributing to the detailed visualizations used in the study. Their training dataset included over 12,000 annotated anthropogenic structures. The researchers evaluated several DEM-derived visualizations, including four variations of Visualization for Archaeological Topography (VAT) (Kokalj et al., 2019), Red Relief Image Map (RRIM) (Chiba et al., 2008), and Local Dominance (LD) (Guyot et al., 2021). While VAT-HS and VAT-HS channels generally performed best, no single visualization method was found categorically superior across all scenarios. The models achieved overall accuracy scores of up to 0.99. Although the authors emphasized the importance of minimizing false negatives, they did not report recall or F1 scores, which would have provided a more direct assessment of the model's sensitivity to missed detections.

Richards-Rissetto and colleagues (2021) investigated the application of deep learning using two approaches: one using a 2D convolutional neural network (CNN) trained on hillshade visualizations, and another using a 3D point-based architecture (PointConv) trained on LiDAR point cloud data. Developed at the Copan archaeological site in Honduras, their target region covered an area of 24 km². The LiDAR data were acquired using a Leica ALS50 Phase II system (von Schwerin et al., 2016), with an average first-return point density of 21.57 points per square meter and an average ground return point density of 2.91 points per square meter. The contrast between first and ground return densities is an important consideration when reprocessing LiDAR data for archaeological feature detection. The point cloud data were annotated by classifying points into five categories: vegetation, default/unclassified, ground, archaeological features, and ruin grounds. A DEM with a 0.5 m/pixel spatial resolution was generated using the last return data. Derived from a shapefile containing approximately 3,500 structures, their training data included 142 annotated target objects for the 3D approach and 410 for the 2D approach. The hillshade-derived model achieved an overall accuracy of 0.88, a precision of 0.85, a recall of 0.89, and an F1 score of 0.87. The point cloud-derived model yielded higher performance metrics, with an overall accuracy of 0.95, a precision of 0.91, a recall of 0.94, and an F1 score of 0.92. Given that this study area was the smallest among the three discussed, higher accuracy metrics were not entirely unexpected, as smaller areas are likely to exhibit greater data uniformity in terrain and architectural characteristics.

The varying point densities reported in these four studies, describing the number of laser pulses that were emitted, reflected, and perceived by their respective Lidar systems, ranged from relatively low (Bundzel et al., 2020) to very high (Richards-Rissetto et al., 2021). Illustrating the spectrum of Lidar data quality available for archaeological research, this is particularly relevant to the current study's future work, which proposes reprocessing the G-LiHT dataset. Understanding the point densities of successful deep learning applications in the Maya area provides a valuable context for determining optimal reprocessing strategies for the G-LiHT data, which has its own unique point density characteristics.

Building upon the solid foundation established by prior research, the present study extends these efforts by developing a scalable model applicable across larger expanses of the Maya region. The significance of this work lies in providing empirical evidence for the practicability of such a broadscale model for Maya archaeological detection. Furthermore, it aims to produce a foundational model that can be immediately utilized for identifying structures. This model will offer a robust starting point for transfer learning in the development of more localized or specialized models, thus reducing the burden of extensive retraining. The ultimate objective of this work is to establish the practicability of producing a comprehensive model capable of offering a broad-scale alternative to numerous geographically constrained, area-specific models for detecting Maya structures in the Yucatan Peninsula.

For the northern area of our study, the G-LiHT Yucatan North2 dataset covers approximately 15,584 km² of the northern central part of the Yucatan peninsula, roughly west of Cancun and east of Merida. It was chosen for creation of the training samples because it covers a large area from west to east, presenting a large number of Maya structures that differ from those in the southern study area. This region contains numerous clusters of regular platforms surrounding *rejolladas* (generally round depressions in the limestone bedrock, often used by ancient Maya for planting crops). Otherwise, the terrain in this region is consistently flat.

In the Central Lowlands of the peninsula, the G-LiHT Yucatan South dataset covers an area in the Yucatan, roughly east of Ciudad Campeche and west of Calakmul, of approximately 21,000 km²; it was chosen for creation of the training samples because of its location roughly in the center of the Yucatan peninsula, covering a large area from north to south, presenting a large number of identifiable Maya structures of various types and styles on diverse, sometimes hilly terrain.

3. Methods

This study evaluated the feasibility of a pan-Yucatan deep learning model for Maya archaeology through experimentation with Lidar data of Yucatan archaeological features and data preparation for geospatial analysis. The development of this model followed a standard deep learning workflow applied to aerial laser survey data.

1. Data Acquisition, Preprocessing, and Conversion: Aerial Lidar Surveys of the Yucatan Peninsula are collected. These scans are converted into standard image files (TIFF format) with geospatial reference information.

2. Training Data Annotation: A subset of these images is imported into ESRI ArcGIS Pro, where the Deep Learning toolset—accessed through the Geoprocessing pane—is used to visually inspect and annotate ancient Maya structures. These structures are identified and outlined with bounding boxes. The annotated images are then exported as smaller image files to create a training dataset for model development.

3. Model Training and Architecture: This training dataset is used to train a deep learning model to recognize ancient Maya structures.

4. Model Application and Results Generation: The trained model is then fed the entire dataset to detect and locate ancient Maya structures. The model creates a table listing each detected structure with a unique ID, location, total area, and the source file.

5. Model Evaluation and Assessment: The results are reviewed, and standard analytical assessment metrics are generated.

Our Human-in-the-Loop (HITL) workflow (Table 1) was designed to incorporate iterative input from subject area experts, ensuring that domain expertise directly informed the technical development and evaluation processes.

Table 1 Human-in-the-Loop (HITL) Framework used in this study

	Stage	Automated Process	Human involvement
1.	Data Collection	Aggregation of Lidar datasets from multiple regions	Selection of representative sites; evaluation of coverage quality
2.	Preprocessing	Noise filtering, Lidar visualization, feature extraction	Review of visualization output; feedback on terrain and structure clarity
3.	Sample Labeling	Initial automated proposals using previous models	Manual annotation, correction and confirmation of sample labels
4.	Model Training	Deep learning model refinement on labeled samples	Regular review of intermediate results; adjustment of training parameters
5.	Model Evaluation	Accuracy, precision, recall, F1 score metrics	Review of object identifications; input on evaluation methods, cultural relevance
6.	Application and Discovery	Deployment on new regions for predictive mapping	Final review of model outputs; visual inspection and expert interpretation

3.1 Data Acquisition, Preprocessing, and Conversion:

The primary dataset for this research was the NASA Goddard's Lidar, Hyperspectral, and Thermal (G-LiHT) dataset, an aerial Lidar survey of the peninsula, chosen for its comprehensive and pseudo-random coverage of the Yucatan Peninsula and its high spatial resolution, as discussed in the Introduction and Background. Schroder and colleagues (Schroder et al., 2020) published an archaeological evaluation of 458 tiles from the NASA G-LIHT dataset, demonstrating the value of that Lidar dataset for representing the architecture of ancient Maya structures across the breadth and length of the peninsula. Schroder et al.'s study employed manual inspection and visual annotation of more than 60,000 ancient Maya structures within the same G-LiHT tiles utilized in this research.

The dataset used in the DL model by Character and colleagues (2024) originated from a region in the southern highlands of the Yucatan Peninsula. A comprehensive pan-Yucatan DL model would ideally be trained with architectural data encompassing the full range of ancient Maya architecture, settlement patterns, and terrain types across the Yucatan to optimize its understanding. This study considered the entire Yucatan peninsula as its target region. A dataset was sought that would provide representative samples of imagery from the entire region in a random or pseudo-random array with sufficiently high resolution to make accurate feature identification practicable to facilitate creation of effective broadscale deep learning models.

Several datasets were considered to produce such a pan-Yucatan data representation. The Lidar dataset from the Instituto Nacional de Estadística y Geografía (INEGI) provides 1.5m (150cm per pixel) spatial resolution, and the NASA Global Ecosystem Dynamics Investigation Shuttle Radar Topography Mission (GEDI SRTM) dataset provides imagery at 30m (3,000 cm per pixel) spatial resolution. Both of these datasets cover all areas of the peninsula, but a review of that data confirmed it was too low-resolution to accurately identify archaeological target features, in agreement with findings from Kokalj and Mask (2021). Consistent with Schroder and colleagues' (2020) analysis, the G-LiHT dataset was considered an excellent resource for analyzing the Yucatan, as it represents all areas of the peninsula in a pseudo-random array. The G-LiHT dataset consists of long narrow transects that cover thousands of kilometers in all the general areas of the target area, for example southern Mexico and the entire Yucatan Peninsula.

G-LiHT transects cross the peninsula in roughly vertical and horizontal strips, providing a pseudo-random collection of imagery from all areas of the entire target region. G-LiHT data offers fine spatial resolution (<1 m). The Lidar point clouds of the G-LiHT ALS system were recorded by a Riegl VQ-480 scanning Lidar system with an accuracy of 25mm, scanning rate of more than 100 khz, and as many as 10-20 returns/m², and augmented by a Riegl LD321-A40 profiling Lidar system to calibrate distance measurements. Theoretically optimal densities were determined by Liao et al. (2024), to be around 2.43 pts/m² for 0.2 m/pixel resolution, which suggests that the G-LiHT point clouds might be fruitfully processed at a significantly higher resolution than the published one meter (100 cm) per pixel versions. While the currently available DEMs do provide valuable data, the density and accuracy of the original point clouds suggests that a significant potential exists for generating higher spatial resolution digital elevation models (DEMs) through targeted reprocessing. Due to limitations in computational resources and time, we were unable to reprocess the transects for this study. However, we recognize a promising opportunity to generate enhanced visualizations and potentially more detailed feature detection capabilities through reprocessing at higher spatial resolutions, as further discussed in Section 6.3 (Discussion).

Table 2 List of G-LiHT transects used in this study

Map	Area	G-LiHTt datasets in the Yucatan Peninsula
A	Campeche	AMIGACarb Chiap Campeche Apr2013
B	South	AMIGACarb_Yuc_South_Apr2013
C	North	AMIGACarb Yuc North Apr2013
D	North2	AMIGACarb_Yuc North2_Apr2013
E	Out of the Yuc	AMIGACarb Out of the Yuc May2013
F	Centro	AMIGACarb_Yuc_Centro_Apr2013
G	Chiapas	AMIGACarb Chiaps1 Apr2013

The NASA G-LiHT online data site provides these datasets in their original form as LAS files and in processed format as single-channel Digital Elevation Model (DEM) files with a spatial resolution of one meter per pixel. For this study, both versions (LAS and DEM) of the following G-LiHT datasets were acquired and thoroughly reviewed: Norte, North2, Centro, South, Out-of-Yucatan, Chiapas, and Campeche (Table 2).

The names of these regions were originally assigned to the datasets by the NASA G-LiHT team. We have retained their original nomenclature for each region in this study to prevent confusion and facilitate direct analysis and comparison with the source data as published on the official G-LiHT online data download site (https://glihtdata.gsfc.nasa.gov/files/G-LiHT/). For instance, the dataset named "Norte" is the Spanish designation for one of the northern Yucatán transects, while "Out-of-Yucatan" refers to a G-LiHT transect that extends geographically beyond the generally accepted boundaries of the Yucatán Peninsula, despite still being within the broader study area of Mesoamerica (Fig. 3).

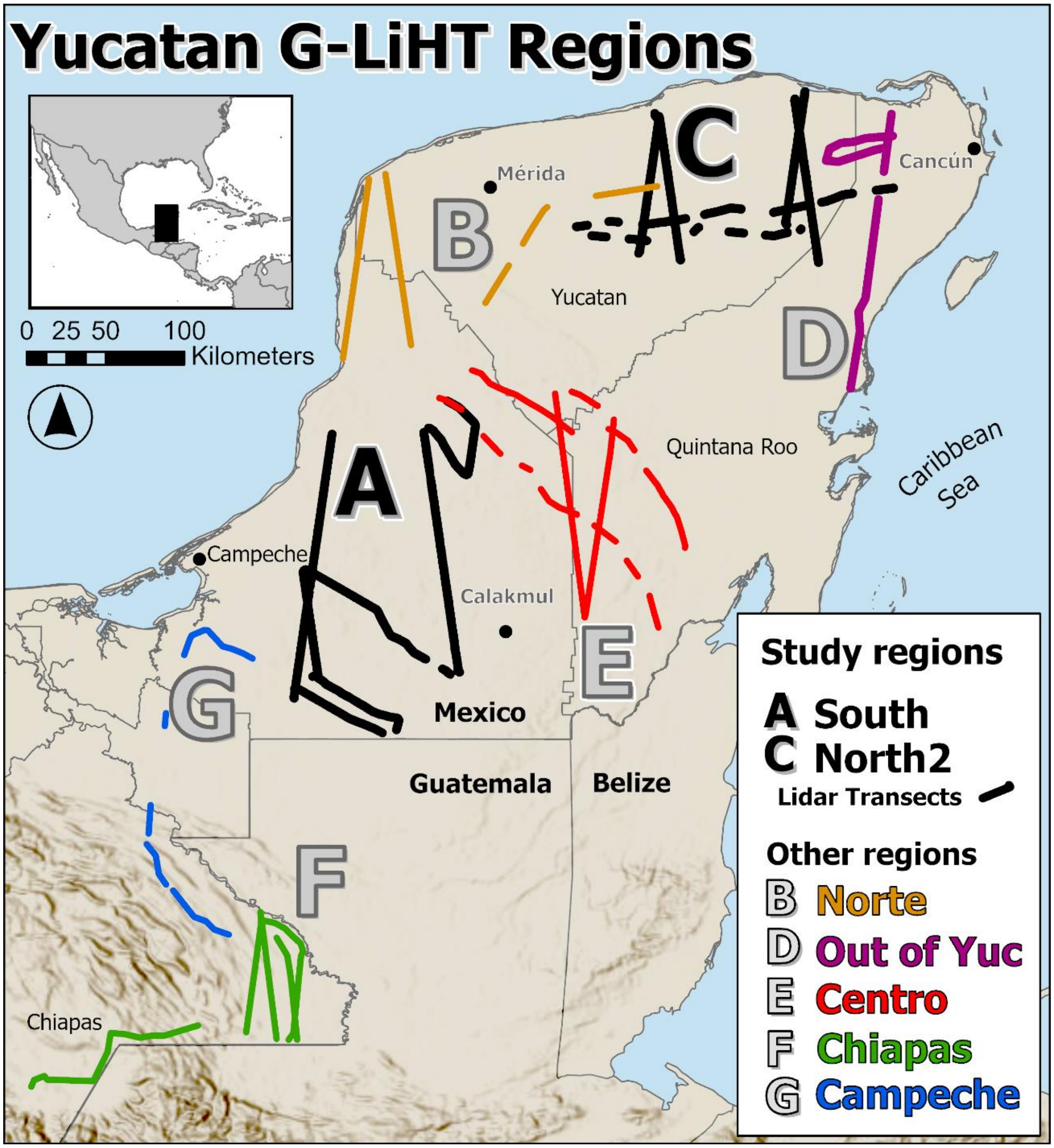


Fig. 3 *G-LiHT Yucatan Regions Aerial Lidar Survey transects, South and North2 were reviewed in this study; Transects of two regions were combined to create the Q2000 deep learning model: A (South) and C (North2)*

Each of these regional sets contained two subsets of data (except for Campeche which had only one), called GLAS and NFI. The GLAS files were generally north-south aerial surveys, whereas the NFI files were generally east-west surveys of the same general area. This patchwork of transects provided representative data for all regions of the

Mexican areas of the Yucatan peninsula, so it was chosen for this study. The North2 and South datasets were chosen to be combined into a multi-regional model for this study because of their large geographical area, their physical distance from each other, and for the contrast between their respective terrains, with the objective of examining if their diversity would make combining such related but different areas practically possible.

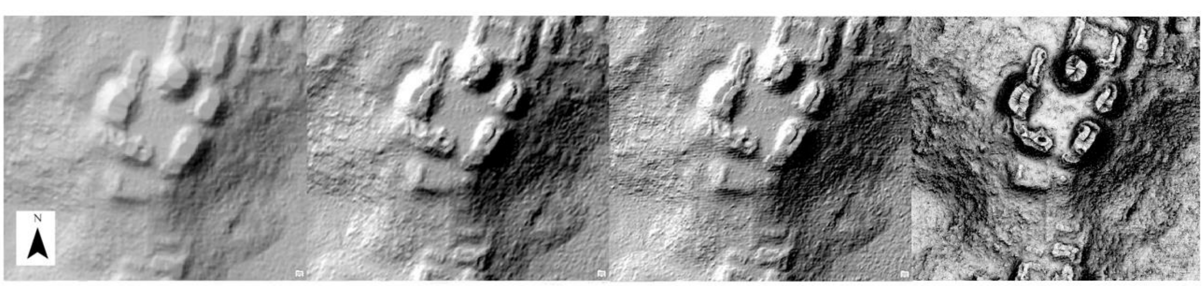
The two regions combined in this study are called North2 and South, respectively, reflecting the names assigned by the NASA Ames G-LiHT research team in 2013 for these discrete datasets of regionally collected transects.

The North2 region, covering approximately 15,584 km², is located in the north-central area of the Yucatan Peninsula, stretching generally east-west from east of Merida to west of Cancun, and extending from the northern coast southward to the vicinity of Chichen Itza. Geologically, this region is characterized by very flat terrain situated on Miocene-Pliocene limestone, colloquially known as the Carrillo Puerto Formation. This topography contains numerous clusters of rejolladas, which are circular limestone sinkholes presumably used for the cultivation of cacao and other plant products (Koby, 2012). The G-LiHT transects crisscross this area, providing Aerial Lidar Survey recordings that reveal ancient Maya structures, often appearing as rectangular mound clusters concentrated around these rejolladas.

In contrast, the South region, encompassing about 25,000 km², is located in the central-west area of the Yucatan Peninsula. It stretches generally north-south from the town of Campeche to the northern border of Guatemala, and reaches from the west coast eastward almost to the site of Calakmul in the center of the Yucatan's northern lowlands. Geologically distinct, the South region is situated on older, more weathered Eocene era sedimentary rock, resulting in a landscape that features many hilly areas with surrounding bajos (lowlands). The Lidar data for this region often shows that many ancient Maya structures were commonly located on these hills (Hansen et al., 2023). It's important to note that the North2 and South regions are physically separated at their nearest distance by approximately 150 kilometers, further emphasizing their distinct environmental and cultural contexts.

Data Pre-processing:

After downloading the data from the G-LiHT repository, the Digital Elevation Models (DEM) files derived from the G-LiHT Airborne Laser Scanning (ALS) data were inspected. Their published DEM files are processed at a 1 m/pixel spatial resolution by NASA, which offered a foundational layer for broad archaeological feature identification. However, the inherent limitations of this resolution became apparent when targeting smaller yet significant structures ubiquitous across the Yucatan landscape. Upon recognizing the untapped potential within the original G-LiHT point clouds, which had been saved in the LAS file format, our preliminary investigation explored the feasibility of generating higher resolution DEM files (0.75 m/px and 0.5 m/px). While these initial re-processed datasets revealed enhanced visibility of sub-meter features, practical constraints necessitated the utilization of the publicly available 1 m/pixel DTMs for the present study, ensuring direct comparability with existing regional analyses. The demonstrable gains in feature detection through higher resolution processing underscore the rich potential of the original G-LiHT LAS files, (e.g., clearly showing an abundance of looter’s excavations at the site of Pixoyal, Campeche archaeological site; cf. Merk et al. 2025). The prospect of reprocessing the G-LiHT transects to attain higher spatial resolution is a subject warranting exploration in forthcoming research. This could further optimize G-LiHT Lidar data for archaeological and broader environmental applications in the Yucatan Peninsula.



Original 100 cm/pixel 75 cm/pixel 50 cm/pixel 33 cm/pixel

Fig. 5 *NASA G-LiHT Yuc_South_GLAS_395 Lidar point cloud data reprocessed at higher spatial resolution showing part of the Pixoyal archaeological site in Campeche state, Mexico*

The DEM files were processed in Relief Visualization Toolkit (Kokalj & Somrak, 2019) to generate Multi-Hillshade, Slope, Open-Positive, and Skyview image files. We relied on the RVT graphical user interface to produce these layers. We were not time-constrained in our process, finally reaching a point of satisfaction that the combined layers produced a clear representation of the terrain.

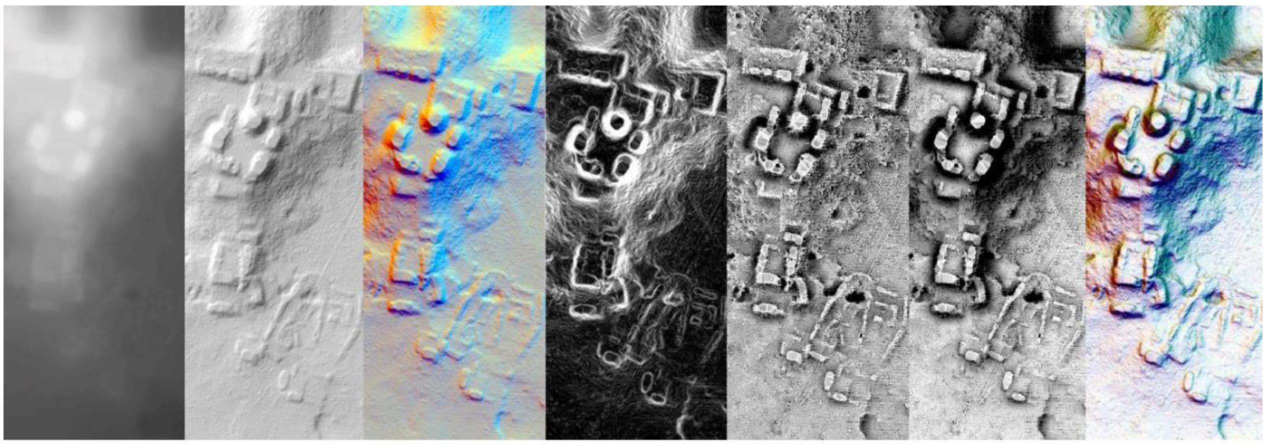


Fig. 4 *Examples of Lidar visualizations used in this study; the layers are combined to create a blended visualization*

These four types of visualizations were imported into ArcGIS Pro 3.3 as individual layers and combined using various layer blending modes (e.g., multiply, overlay) with specific weightings (detailed in Table 3) to create composite visualizations designed to enhance subtle topographic variations indicative of archaeological features in the Yucatan Peninsula. Skyview was used to show small-scale relief, depressions, and subtle elevations; Open-view Positive highlighted convex objects like ridges and crests; Slope indicated the steepness of inclines; and a colorized multidirectional hillshade served as a base to simultaneously bring out the subtle three-dimensional shape of the terrain and its associated characteristics. The image processing parameters, including a baseline formula (detailed in Table 2), were specifically developed for this study. This baseline proved effective for most G-LiHT transects, with occasional adjustments applied for specific layer combinations to optimize the visibility of archaeological features.

Table 3 Formula for combining visualizations to generate a three-channel Lidar visualization

Process Settings (all layers stretched to Min/Max)						
Layer	Visualization	Transparency	Blend Type	Brightness	Contrast	Gamma
4	Skyview	33	Multiply	1	30	.5
3	Open-Positive	33	Overlay	-30	55	1
2	Slope	33	Difference	10	20	3
1	Multi- hillshade		Normal	10	20	2

The layering and adjustment image processing parameters developed for this study to optimize the visibility of archaeological features in the Yucatan resulted in a combined image, which was then exported as a single-layer, three-channel RGB GEOTIFF image file (Table 4).

Table 4
Specifications for exported GEOTIFF files of combined Lidar visualizations

Share Map Frame as GEOTIFF
Image Compression: None
Write GEOTIFF Tag: On
Image Size: Slightly larger than original, using same H/W ratio, position Map in window

Embed Color Profile: On 24-Bit True Color Image

3.2 Training Data Annotation:

To train a deep learning model for pan-Yucatan detection of diverse ancient Maya structures, G-LiHT visualizations from the South and North2 regions were annotated in ArcGIS. The intention was to create a model that would recognize anthropogenic features consistent with those found in Maya archaeological sites. To achieve this, we prioritized fundamental geometric and topographic characteristics: rectilinear angles, specific slope values, and defined depressions. Unlike studies employing independent classification of architectural genres (e.g., Schroder et al., 2020; Characer et al., 2024), this research, consistent with the broad applicability sought by Ringle (Ringle et al., 2021), focused on these core features to develop a model robust to the Yucatan's varied terrain and settlement patterns. Identifiable features exhibiting these characteristics—platforms, buildings, annular structures, and mounds—were outlined with rectangular polygons under a unified "Maya structure" class. Presuming that the areas immediately surrounding these ruined structures are also inevitably transformed anthropogenically, it was deemed prudent to include a small surrounding area around each annotated target object during the annotation process to enhance generalizability and incorporate subtle cues in the terrain as training information. Given the study's extensive areal scope and the necessity of clearly identifiable training targets to ensure robust generalization, this single-class approach was deemed more practical for this broadscale study than delineating numerous specific architectural types across a limited, albeit unambiguous, training dataset.

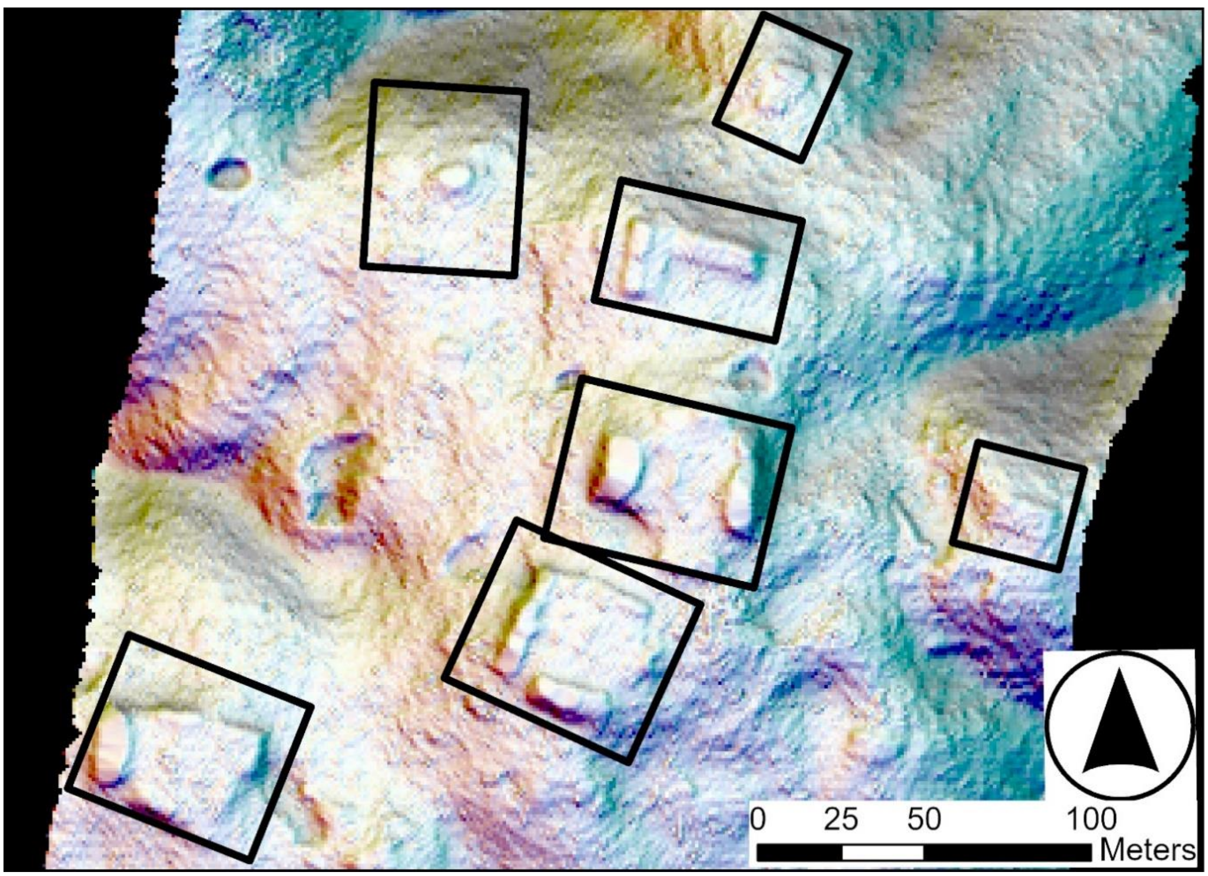


Fig. 6 *An example of* annotated training samples (dark rectangles) in South area; visualization is from 1 m/pixel G-LiHT DEM file; this cluster of structures is approximately 13.5 kilometers south of Pixoyal, Campeche*

To create the training dataset, visualizations of the G-LiHT transects for these two regions were visually reviewed. Objects exhibiting the characteristics of Maya structures were outlined with rectangular polygons using the Deep

Learning Tools in ArcGIS Pro, and their geospatial locations recorded. These annotations were rigorously reviewed in accordance with our Human-in-the-Loop (HITL) framework (see Sections 1, 3, and 5) by team members possessing extensive experience with both Maya archaeological sites and Lidar visualizations of such sites.

We deliberately omitted ambiguous examples, ignoring target objects near transect edges or those difficult to confidently identify. Consequently, our annotated sample set was purposely limited, comprising only Maya structures identified with high confidence. Given the vastness of our target area (over 36,000 km² spanning private and public lands), ground-truthing a statistically significant number of these objects was logistically impossible. Therefore, we relied upon the experience and expertise of team members highly familiar with the physical reality of Maya settlements and their representation in Lidar data. This desktop-based approach is consistent with established industry practice in large-scale Maya archaeology (Folan et al., 2009; Ringle et al., 2004; Chase et al., 2011; Garrison et al., 2019; Evans et al., 2021). In accordance with our HITL framework, we carefully selected 1,000 sample objects from each region (totaling 2,000 objects) to populate this study's inventory of annotated training samples for this multi-regional, broadscale deep learning model.

In practical terms, after in-depth orientation regarding criteria for identifying Maya structures in Lidar data, our principal author processed the data, reviewed the transects, and selected sample objects. These selections were then reviewed with experienced team members to confirm the accuracy of the identifications for inclusion in the training sample inventory. When ambiguous settlement patterns or questionable target objects were encountered, multiple team members reviewed them to establish their nature and to include or exclude them from our inventory. This process generally resulted in a carefully selected set of training samples of a sufficient number to fulfill our goal of 1,000 samples of Maya structures from each of the two regions. This meticulous process was intended to ensure that the model would be provided with dependably representative examples of Maya architecture to enhance its capacity to recognize such objects.

Integrated complexes (temple plazas, marketplaces) were annotated either as a whole or with individual structures outlined. Water reservoirs (aguadas), stone quarries, and water storage pits (chultunes) were not individually annotated but were sometimes included within larger target structures. The emphasis of the annotations was on identifying the ruins of Maya buildings, platforms, and mounds, which collectively were classified in this study as "Maya structures" as a technical designator for the ArcGIS Deep Learning tools.

The desk-based identification of these features is grounded in established archaeological methodologies for interpreting Lidar data in the Maya Lowlands, where distinct anthropogenic modifications to terrain are recognizable even without ground-truthing. These interpretations rely on a comprehensive understanding of Maya architectural patterns and settlement organization across diverse environmental contexts (e.g., Folan et al., 2009; Ringle et al., 2021; Chase et al., 2011). Maya archaeological features, such as building platforms and mounds, exhibit characteristic rectilinear forms, precise orientations, and predictable arrangements that distinguish them from natural topography. This morphological consistency allows for their reliable identification in high-resolution Lidar-derived digital elevation models and various relief visualizations, a practice well-documented in the literature (e.g., Garrison et al., 2019; Evans et al., 2021). The shapes and spatial relationships of these features, even in a ruinous or heavily vegetated state, provide sufficient information for their classification by experienced interpreters.

Following target annotation with bounding boxes, visual representations of each identified feature were captured and exported as 416x416 pixel, 24-bit, GEOTIFF files, accompanied by KITTI metadata containing geospatial reference information, forming the training sample set. This dataset ultimately comprised images of 2,000 Maya structures identified through visual review of the G-LiHT transects, with 1,000 structures located in the South area and 1,000 of them from the North2 area.

3.3 Model Training and Architecture:

The annotated training samples served to train an original deep learning model based on the open-source YOLOv3 network architecture (Redmon and Farhadi, 2018) with a Darknet-53 backbone. The resulting model, packaged as an ESRI .dlpk file, was designed to identify objects within Lidar imagery, delineate their forms with polygon bounding boxes, assign unique IDs, record geospatial locations, and generate detection confidence scores.

We acknowledge that YOLOv3 (released in 2018) is an older architecture compared to more recent developments such as the current versions of Ultralytics YOLO (e.g., v12). Our selection of YOLOv3 was informed by several practical and strategic considerations during the model development process. Firstly, YOLOv3 was one of the deep learning architectures fully integrated and robustly supported within the ArcGIS Pro Deep Learning Tools at the time of this study's development, streamlining the workflow for data annotation, model training, and deployment within a GIS environment. Secondly, initial comparative experiments on this Lidar visualization dataset, which included Mask-RCNN, indicated that YOLOv3 detected a greater number of targets while achieving a comparable number of true positive detections. This empirical finding, coupled with the successful application of YOLOv3 by Character et al. (2024) in related archaeological research, reinforced our decision to utilize this architecture. Our choice was further driven by the intention to build upon their established work, thereby contributing to the broader progress of Maya archaeology.

For future work, we recognize the potential of examining newer versions of the Ultralytics YOLO package and exploring other contemporary network architectures. This may involve transitioning to more flexible training environments, such as a direct Python interface, to leverage the latest advancements in deep learning. However, for the scope and specific goals of this study, YOLOv3 provided a robust and well-supported framework.
The trained YOLOv3 deep learning model (Q2000) was designed with a primary objective of evaluating the viability of a broadscale approach to detection of Maya structures across diverse regions and terrain types. For training the deep learning model, the ArcGIS Pro 'Train Deep Learning Model' tool - part of the Spatial Analyst extension - was configured with the following parameters: a batch size of 16, a default learning rate of 0.10, and a maximum of 70 training epochs. 10% of the training data was reserved for validation, and ArcGIS Pro's default suite of geometric (crop, rotation, flip, zoom) and photometric (brightness, contrast) augmentations were applied to increase the dataset's variability. The training process was set to stop automatically when the model's performance on the validation set no longer improved. This training was conducted on a system equipped with one NVIDIA A4000 GPU and an Intel XEON W-11855 CPU with 64 GB of RAM. The NVIDA A4000 GPU was used for processing the training of the deep learning model.

3.4 Model Application and Results Generation:

Having been trained with selected samples from the South and North2 datasets, the Q2000 deep learning model was used to review all transects from both regions to assess its capacity to correctly locate Maya structures on previously unreviewed terrain, utilizing the ArcGIS Pro 'Detect Objects Using Deep Learning' tool. To handle overlapping detections, Non-maximum Suppression was set to On with a Max Overlap ratio of 0.5. Processor Type was set to GPU, using an NVIDIA A4000 GPU. The Q2000 deep learning model recorded the location, size, and confidence level of each detected object into a feature class stored within the project as a geodatabase feature and stored externally as a shapefile. Its review of the South and North2 regions resulted in thousands of feature detections across the G-LiHT transects of those regions. Our visual inspection revealed that detections frequently occurred in clusters, facilitating the identification of settlement areas and locations where anthropogenic terrain features were evident in abundance. The model often yielded positive detections for objects of ambiguous appearance; however, these rarely occurred in isolation, instead being more frequently observed in proximity to clearly identifiable structures, indicating the model's capacity to detect subtle anthropogenic terrain modifications.

The completeness and correctness of its detections are inherently imperfect, the extent of which can be indirectly measured by comparing its detections against the values of annotated test cells—discrete, square, grid-based spatial units used for evaluation—visually assessed (by HITL) as containing (Positive) or not containing (Negative) intersection with a Maya structure. After making detections, the performance metrics of the Q2000 model were analyzed across both regions of the Yucatan Peninsula by comparing its detections against the values of over 400 such units in each region created for external validation testing as described below in Section 3.5 (Model Evaluation and Assessment).

From the values of its detections relative to the values of the external validation test cells, the sums of True Positive (TP), False Positive (FP), True Negative (TN), and False Negative (FN) were generated and incorporated as a Confusion Matrix to calculate the value of the performance metrics, including Precision, Accuracy, Recall, and F1 Score. By visual notation, variations in effectiveness based on terrain, architectural styles, and settlement patterns were subjectively noted for comparison with the resulting objectively calculated performance metrics to evaluate the general performance characteristics of the model.

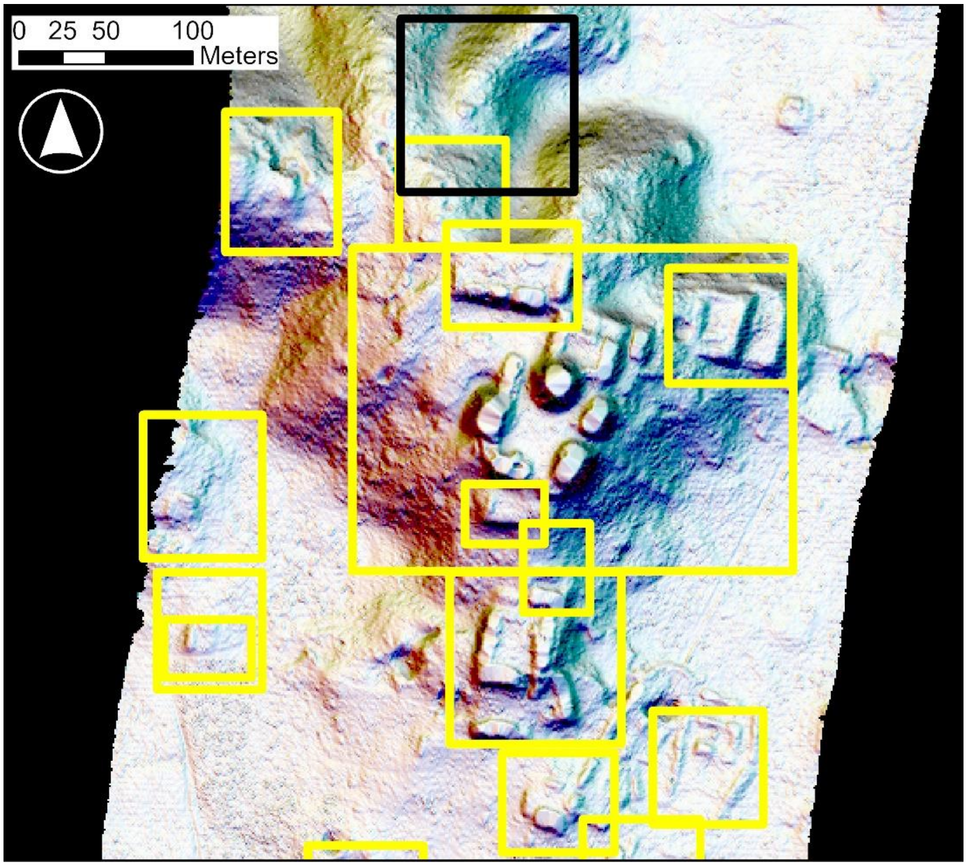


Fig. 7 *Pixoyal in the South area, G-LiHT visualization with detections (light);a test cell (dark) contains a structure, so it is valued as Positive, and since the model detected a structure intersecting the cell, the detection is counted as one True Positive detection.*

3.5 Model Evaluation and Assessment:

To evaluate the Q2000 deep learning model, a system for external validation testing was implemented. This validation process included the following processes:

- Generating all model detections across the two regions.
- Removing detections of objects used in training to create a validation set of only previously unseen data.
- Creating a grid of 100m x 100m cells over the terrain.

The Q2000 model was applied to the complete set of transects in the South and North2 regions using the ArcGIS Pro Detect Objects Using Deep Learning tool. To assess the model's performance on previously unseen terrain, a grid was superimposed over each region, and a subset of cells was selected for external validation (see Sections 3.5 and 3.8). Each test cell—defined as a square unit of terrain—was independently evaluated through Human-in-the-Loop (HITL) visual inspection and assigned a binary ground-truth label: Positive, if it contained or intersected one or more Maya structures, or Negative, if it contained none.

Following this annotation, each test cell was reviewed to determine whether it contained any detections produced by the model. For the purpose of evaluation, only model detections that intersected or were fully contained within these annotated test cells were considered. Subsequently, each test cell was classified to quantify overall model performance, including cases in which cells lacked detections, resulting in either False Negatives (for Positive cells) or True Negatives (for Negative cells). The classification of model performance was based on the following criteria:

True Positives (TP): One count per detection intersecting or contained within a Positive cell.
False Positives (FP): One count per detection intersecting or contained within a Negative cell.
False Negatives (FN): One count per Positive cell that contains no detections.
True Negatives (TN): One count per Negative cell that contains no detections.

This hybrid evaluation approach—counting individual detections for TP and FP, and counting cells for FN and TN—was adopted to reflect the spatial distribution of archaeological features and the diffuse anthropogenic modifications often surrounding Maya structures. It provides a rigorous and interpretable framework for quantitatively evaluating object detection performance at landscape-scale in archaeological settings.

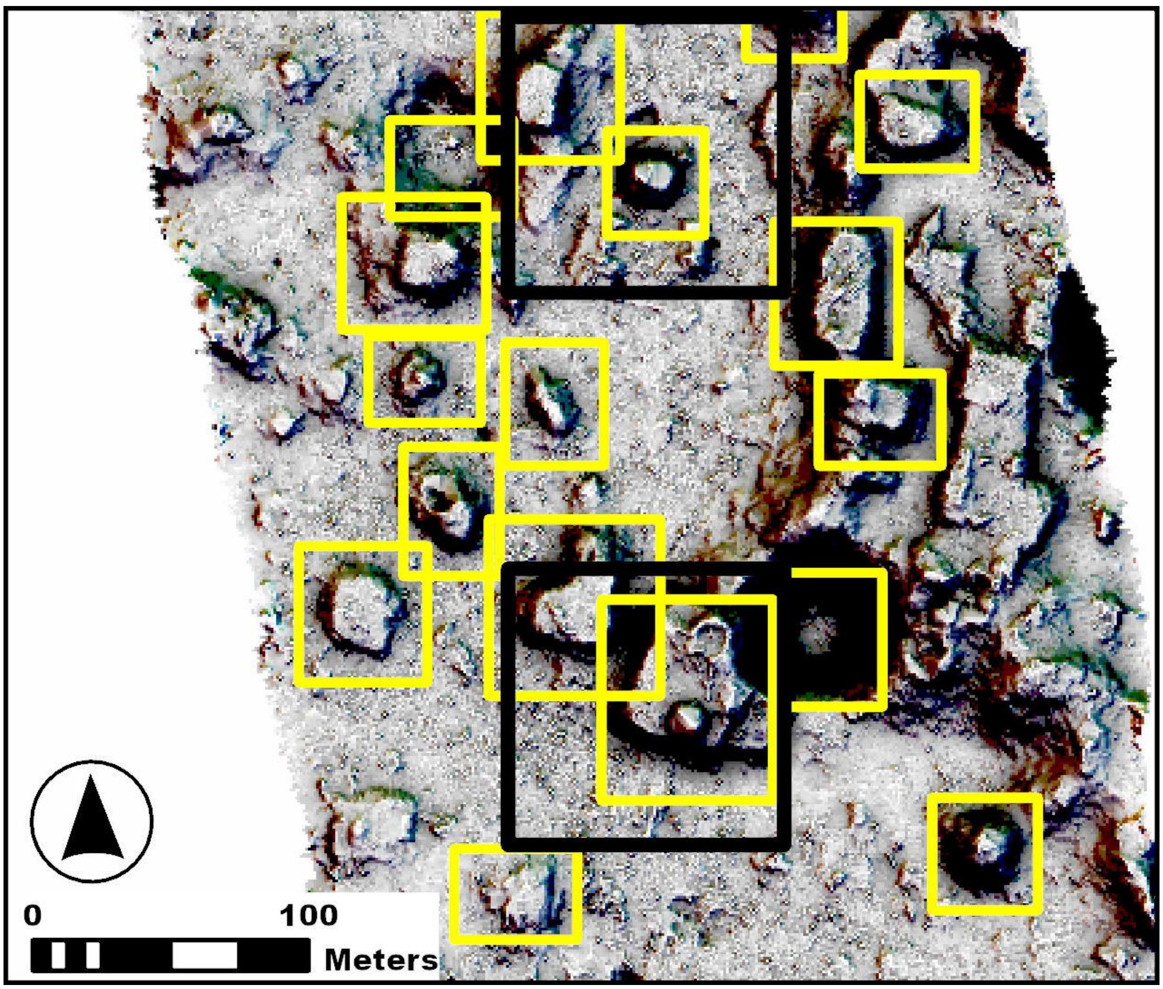


Fig. 8 *Chichen Itza area, North2 region; detections of structures (light rectangles), test cells (dark squares); target objects are clustered around rejolladas (karstic depressions above the water table containing well-drained soils and stable microclimates, archaeologically significant for agriculture, particularly moisture-sensitive crops like cacao, or other fruit and nut trees); note that the model did not trigger a detection of the largest structure in the image*

Since the total number of cells on a grid covering a large area can be impracticably high, one can use a sampled subset of such a collection of cells, achieving a statistical level of significance by assessing a large enough number of samples. The formula for determining the minimum number of sampled cells required to achieve statistical confidence is calculated as follows:

$$n = \frac{Z^2 \cdot p \cdot (1 - p)}{E^2}$$

where n = required number of cells to achieve statistical significance
Z = Z-score corresponding to the desired confidence level (for 95% confidence, $Z \approx 1.96$)
p = estimated proportion of objects (commonly p=0.5 to maximize the required sample size)
E = desired margin of error (as a decimal, so 5% = 0.05)

Considering the number of cells to be in the tens of thousands, for an expected accuracy of 0.5 (50%), with a 95% confidence level and a +/- 5% margin of error, the minimum subsample size of cells we would need was approximately 385. To ensure robustness of measurement, we set a target of 400 cells in each area for our set of curated test samples. The shapefile containing training samples was overlain on the grid to ensure that no test cell would contain any structure that had been included in the training samples. By this means, we ensured that the test cells measured detections only against objects and areas that had been previously unreviewed and unseen.

In general, most of the target objects (Maya structures) are less than 1000 m^2 in area. They are often densely clustered strategically on advantageous terrain, often separated by other clustered settlements of structures at a distance of 10 km or so. The G-LIHT ALS transects collect Lidar recordings in long, thin strips, about 300-400 meters wide and often about 7-8 kilometers in length. This produces visual recordings of relatively small, clustered targets separated by long empty stretches of terrain, on long, thin strips of ALS Lidar data.

A random sample of 900 m^2 cells overlain on the transects of the South and North2 regions, generated millions of cells, requiring a minimum sampling of 385 cells in each of the two regions to obtain reliable test measurements to the level of statistical significance. This generated a set of test cells imbalanced by more than 10:1. To capture a larger number of positive detections, resizing the cells to 10,000 m^2 cells still generated a 7:1 imbalanced set. Further, cluster density weighting to emphasize settlement areas with 10,000 m^2 cells yielded a 5:1 imbalanced set. Increasing the size of the cells was not possible because of the narrow transects. Finally, in accordance with standard practice in dealing with an imbalanced test set, we employed curated sampling, also called stratified or targeted sampling. This approach has the benefit of providing an opportunity to balance the test set, for targeted evaluation and focus on critical cases specific to the discipline in which the model will be used; in our case this is archaeology in the Yucatan peninsula. Despite these challenges in achieving a naturally balanced sample, the standard practice in model validation emphasizes the importance of ensuring proper balance for reliable evaluation.

A balanced (1:1 Neg:Pos) test set is generally considered crucial for reliable detection model validation (TP, FP, TN, FN), as imbalanced sets can skew accuracy, providing a misleading sense of performance, particularly for the minority class. Such balancing ensures a fair evaluation of the model's ability to detect both Positive and Negative instances, leading to more trustworthy precision, recall, and F1 Score metrics, and facilitating meaningful comparisons between models while mitigating the risk of masked biases.

However, for specific applications like archaeological detection, the relative importance of these metrics can be consciously weighted, directly influencing the composition of the test set. Although prioritized values could vary depending on project objectives, we established as our primary goal to maximize True Positives (TP), ensuring the reliable discovery of archaeological features. Next in priority is minimizing False Positives (FP) to avoid wasting resources on fruitless ground-truthing efforts. While False Negatives (FN) mean overlooking potential sites, this concern can be somewhat mitigated if the model exhibits a high True Positive (TP) capacity, as the discovery of objects through successful TP detections can inspire ground-truth expeditions where clustered, initially undetected targets might also be found. True Negatives (TN) are often considered the least critical for direct archaeological impact. This prioritization thus provides a rational basis for deviating from a strict 50/50 split between Negative and Positive test cells. For instance, to more rigorously assess the model's ability to find actual structures (TP) and minimize wasted resources (FP), one might adjust the balance of the test set; a process referred to in machine learning as Model Calibration.

Drawing upon this prioritization, our analysis established a relative weighting that guided the selection of test cells. For each set of 400 curated cells, this resulted in an allocation of 240 Positive cells and 160 Negative cells (a 3:2 ratio of Positive to Negative cells). For this reason, we employed a curated set of over 400 cells per area (more than 800 test cells) with a balance strategically weighted (240 Positive to 160 Negative per 400 cells) to achieve optimal test results, reflecting the inherent prioritization of maximizing object detection (TP) and minimizing wasted resources on false leads (FP) within an archaeological research context.

After the Q2000 model reviewed both regions and made its predictions, each of the 400 cells was visually reviewed to compare its value with the model's prediction for that spot on the terrain, categorizing the outcomes for each cell as True Positive (TP), False Negative (FN), False Positive (FP), or True Negative (TN) based on whether a detection occurred and the cell's actual classification.

The sums of each outcome (TP, FP, TN, FN) from the Q2000 model on the test cells were used for calculating externally validated Accuracy, Precision, Recall, and F1 Scores using a Confusion Matrix populated with the counts of TP, FP, TN, and FN, as follows:

1. **Accuracy**:
   - Measures the overall correctness of the model.
   - $Accuracy = \frac{TP + TN}{TP + FP + TN + FN}$
2. **Precision** (Positive Predictive Value):
   - Indicates the accuracy of positive predictions.
   - $Precision = \frac{TP}{TP + FP}$
3. **Recall** (Sensitivity or True Positive Rate):
   - Measures the ability to identify all positive instances.
   - $Recall = \frac{TP}{TP + FN}$
4. **F1 Score**:
   - Harmonic mean of precision and recall, useful for imbalanced datasets.
   - $F1\ Score = 2 \times \frac{Precision \times Recall}{Precision + Recall}$

4. Results

The model's performance was assessed through external validation on previously unseen, unreviewed terrain from two G-LiHT regions in the Yucatan Peninsula (South and North2). Observations regarding the occurrence of TP, FP, TN, and FN in both regions were recorded for analysis of the model's performance to populate a confusion matrix from which were derived externally validated values for Precision, Accuracy, Recall and F1 Score. The Q2000 model reviewed the South region, detecting the presence of 1,917 objects, and those detections were evaluated by comparing with a set of 400 External Validation Test cells. The Q2000 model also reviewed the North2 region, detecting the presence of 6,913 objects, and was also evaluated on this region with a set of 400 External Validation Test cells. The external validation test results from each area yielded the results shown in Table 5, and Table 6:

Table 5
Q2000 model external validation scores for regions South and North2, and combined regions

	Detections	Precision	Accuracy	Recall	F1 Score	TP	FP	TN	FN
South	1917	0.9167	0.7773	0.8073	0.8586	352	32	53	84
North2	6913	0.9089	0.8568	0.9205	0.9135	718	72	84	62
Combined	8830	0.91141	0.8284	0.8799	0.8944	1070	104	137	146

Table 6
Q2000 Model Validation Compared with other Deep Learning Studies of the Maya Area

Project	Study Area km²	Training Dataset	Precision	Accuracy	Recall	F-1 Score	Visualization	Model Type
This paper	35,584	2,000	0.91	0.82	0.88	0.89	Custom	YOLOv3
Character 2024	615	15,765	0.98	Not given	0.67	0.80	SLRM	YOLOv3
Richards-Rissetto 2021	26	410	0.85	0.85	0.89	0.87	Hillshade	Inception-v3 CNN
Richards-Rissetto 2021	26	142	0.91	0.95	0.94	0.92	3D Point Cloud	3D PointConv
Somrak 2020	130	8,687	Not given	0.99	Not given	Not given	VAT	VGG-19
Bundzel 2020	160	670	0.72	0.99	0.6	0.62	DEM	U-Net/ Mask-RCNN

5. Discussion

Our Q2000 model produced test results comparable to other projects employing a broadscale approach. Significantly, these results were derived from a small, cautiously curated set of training samples distributed relatively evenly across an extremely large and varied combination of two regions. This approach, which inherently prioritizes robust generalizability, relies on the diversity of input data, although caveats are warranted regarding the potential for excessive data complexity to create overfitting.

This study specifically examined the practicability of a broadscale approach across the entire Maya region. The generalizability of our proposed methodology is strongly supported by the considerably larger study area of 35,584 km², a spatial extent vastly exceeding that of previous research (e.g., Character et al., 2024; Somrak et al., 2020; Bundzel et al., 2020; Richards-Rissetto et al., 2021), which respectively covered less than 615 km². This substantial increase in coverage highlights the potential for wider application of this methodology.

A large target area provides both greater diversity of environments and target objects, and potentially a larger population of appropriate target objects for creating training samples. Despite our limitation of selecting only clearly identifiable Maya structures, drawing samples from two diverse regions provided both greater diversity and a sufficient number of appropriate samples. Our experiment, utilizing a very large target area represented by regionally organized, interwoven ALS transects in a pseudo-random pattern and a small, yet representative and conservatively curated training set, successfully tested the potential of a broadscale approach. The resulting model effectively detected Maya structures by synthesizing information from a variety of terrains and structure types, proving effective in its detections as measured by standard metrics of Precision, Accuracy, Recall, and F1 Score. These scores were competitive compared to similar experiments using much smaller areas and larger datasets, as described in Section 4. This experiment thus supports the contention of Character and colleagues that a broadscale approach can be effective (Character et al., 2024). We propose that the diversity within a representative training set of target objects and the variety of terrain types across an expansive area of the Yucatan Peninsula can meaningfully enhance a deep learning model's capacity to generalize effectively.

5.1 Curating Training Samples for a Broadscale Model

To examine the potential for broad generalization by leveraging the inherent terrain and object diversity of this expansive target area, this model employed a small, conservatively selected number of training samples, each chosen based on high confidence in their identification as valid Maya structures. The logistical challenges of comprehensive ground-truthing across such a vast and varied landscape, encompassing both public and private lands, made it essential to prioritize these dependable examples. This cautious approach aimed to guide the model towards learning accurate characteristics, minimizing the risk of mistaken identifications arising from uncertain or false positives. The expectation was that the rich environmental and archaeological diversity of this encompassing region would provide the necessary variation in terrain and target characteristics. This diversity enabled the model to learn generalizable features, even from this focused initial dataset. For a broadscale approach, this is valuable particularly given the infeasibility of extensive ground-truthing to verify more ambiguous examples across such a broad and complex study area. This strategy sought to assess the practicability of leveraging broad geographic diversity to achieve a foundational level of generalizability from a small, but highly reliable, set of training data, potentially paving the way for its use as a pre-trained model for transfer learning. This curation process was informed by HITL, as described in 5.3, ensuring high-confidence samples.

The model achieved strong performance with this limited yet carefully curated dataset, suggesting that its architecture possesses the capacity to learn the fundamental underlying patterns in the data relevant to the target objects. These findings are further supported by similar successes in the field. For instance, Davis et al. (2021) demonstrated that deep learning models could effectively identify archaic shell rings in LiDAR data, even when trained on a relatively small number of original annotated target objects (e.g., 18 shell rings). Their work highlights how strategic data augmentation and transfer learning can enable robust generalizability on previously unseen terrain, making the most of limited original data and leveraging pre-existing knowledge within the model. The encouragingly high scores achieved by our Q2000 model indicate that the features the model has discerned from the small sample are demonstrably discriminative of the target structures against the background and the diverse array of terrain present across the combined regions. This outcome suggests that the inherent diversity of this extensive region is indeed providing a valuable signal, even with sparse labeling.

This promising initial performance provides a robust foundation upon which to build. With the incorporation of a more substantial volume of labeled data, the model will have the opportunity to learn more nuanced features, become increasingly resilient to variations in terrain and object appearance, and potentially achieve even higher levels of accuracy and generalization across the entire expansive target area. This strong initial result can readily justify the potentially significant effort and resources required to label a much larger dataset, offering compelling evidence that the approach has considerable merit and is likely to yield further substantial improvements with more data. For this reason, pursuit of the goal of producing a broadscale deep learning model to detect Maya structures in Lidar data is endorsed.

5.2 Challenges of a Broadscale Approach

While the integration of broadscale and diverse datasets holds significant promise for enhancing the generalization capabilities of deep learning models, it is crucial to acknowledge several potential counterarguments and inherent considerations that warrant careful attention in research and implementation. Despite these challenges, this study provides a sound foundation for understanding the potential effectiveness of a broadscale approach to deep learning for object detection.

Firstly, the increased complexity associated with training models on such expansive and varied datasets presents a considerable challenge. For example, Lidar data and feature visibility are affected by factors such as vegetation cover, terrain complexity, settlement patterns influencing characteristic significance, target object properties (size, relief, materials, preservation), sensor characteristics impacting point clouds, and point cloud processing affecting DEM quality. For a broadscale approach to work, the inherent heterogeneity of these data sources necessitates increased computational demands, requiring substantial resources and longer training periods. To quantify such requirements, this study required a dedicated workstation properly equipped with GPU, RAM, and storage for a period of about one year, equipped with a wide variety of Lidar processing, programming, and GIS tools, a substantial but not an insurmountably unavailable resource. Human visual review of the Lidar data, which accounted for ~70% of the research time, was invested to ensure training sample accuracy. The inherent need for significant coordination with local and subject area experts when scaling HITL across a very large area presented a practical limitation as it would for any broadscale projects. Obtaining Lidar data of a large area is challenging; the high-quality NASA G-LIHT Lidar data was expensive and bureaucratically challenging to produce but is available as a public resource, supporting the development of deep learning models for archaeology.

Secondly, data heterogeneity challenges pose a significant risk to model performance. A naive aggregation of data from disparate regions and sources may prove suboptimal, potentially leading to the model learning spurious correlations that are specific to certain sub-populations or failing to capture nuanced, region-specific characteristics (Klie et al, 2024). To mitigate this, advanced techniques such as domain adaptation, transfer learning, and the strategic incorporation of regional contextual information as input features may be necessary to ensure robust and generalizable learning. A cautious approach to collection of training samples also mitigates spurious correlations. Creation of derivative models, using a more cautiously developed foundation model for pre-training is one approach to integration of diverse datasets without risk of diminishing the capacity of the foundation model. One expects that a good foundation model would be adopted as a base for researchers with new primary source data, and the foundation model could serve as a point of integration with new source data to create effective tools. For this reason, the NASA G-LIHT is an invaluable asset for Maya archaeology in the search to build a broadscale model.

Thirdly, excessive data complexity, sometimes called "curse of dimensionality", remains a pertinent concern. While a larger volume of data is generally beneficial for training deep learning models, the introduction of excessive irrelevant features or an increase in the data's complexity without a commensurate increase in meaningful examples for each distinct sub-population can paradoxically hinder the learning process. Careful feature selection, dimensionality reduction techniques, and a thorough understanding of the underlying data structure are essential to navigate this challenge.

Finally, ensuring data quality and consistency across a broad geographic area represents a substantial logistical and methodological hurdle. Variations in data collection methodologies, sensor specifications, and labeling protocols across different regions can introduce significant noise and systematic biases into the training dataset. Rigorous data cleaning, standardization, and harmonization procedures are therefore paramount to ensure the integrity and reliability of the data used for training broadscale deep learning models. Effective integration of edge cases, ambiguous identifications, or less distinct examples of target objects needs, which were not integrated in this study, is also an important area of study to be explored. This might be effectively accomplished by integrating Lidar visualizations of physically ground-truthed targets, and by confirmation with subject area experts to achieve confidence that edge cases proposed for integration are in correctly identified target objects.

5.3 Human in the Loop

The inherent complexities of broadscale deep learning models—such as increased heterogeneity and potential biases—demand robust safeguards and intelligent oversight. A key strategy for navigating these challenges is the Human-in-the-Loop (HITL) framework, which integrates domain expertise into the data management and model training process (Wu et al., 2022).

In our study, a team of geographers and Maya archaeologists provided ongoing guidance to ensure the project's relevance, rigor, and adherence to professional standards. Their subject area expertise played a critical role in refining the training dataset, identifying valid targets, and excluding irrelevant features. For instance, in the central northern lowlands, a distinctive settlement pattern emerged: clusters of rectangular mounds surrounding circular depressions in karstic limestone. Initially flagged as anomalous, these features were validated through specialist review as documented examples of rejolladas—agricultural depressions around which communities built structures, such as those found at Xuenkal (Koby, 2021). The confirmed structures were then integrated into the training sample.

HITL also facilitated the exclusion of non-Maya features, such as modern constructions, by cross-referencing Lidar imagery with archaeological records and satellite data. This iterative, domain-informed review process ensured consistency, reduced bias, and improved the model's generalizability (Canedo et al., 2024; Klie et al., 2024; Wu et al., 2022; Character et al., 2024).

Broadscale models offer the potential to detect novel patterns across diverse environments, but their success hinges on effective HITL implementation. When thoughtfully applied, HITL enhances model robustness, fosters collaboration with local communities, and ensures ethical and reliable outcomes. Compared to geographically constrained approaches, broadscale modeling—guided by expert human oversight—offers a more scalable and impactful approach to pattern detection across diverse contexts.

6. Future Work

As much as this study suggests the pursuit of a broadscale deep learning model is fruitful, it also raises important questions regarding the confirmation of its results and the potential for improvement of the source data. The study indicates that a broadscale approach can draw meaning from an inherently heterogeneous dataset across a large area using a modest sample set selected from a pseudo-random, relatively uniformly distributed area across two different regions. Building upon these findings, several key areas warrant future investigation.

6.1 Developing model enhancement

Future efforts will focus on enhancing the model to increase its capacity to serve as a foundation model for archaeological use in the Maya area. Opportunities for development include adding more areas, using more samples, incorporating more diverse and potentially edge-case training samples to address the limitations of the current curated dataset, and implementing multi-class object detection (e.g., platforms, pyramids, annular features, water reservoirs (*aguadas*), etc.). How well can it recognize unusually shaped target objects (e.g., in Figure 7, note that the model did not detect the largest structure in the image)? How well can it integrate data made with other sensors from other projects, considering potential differences in data format and resolution, by enlarging the set of training samples or by using the model as a pre-trained "foundation model" for creating new models?

6.2 Testing the model on other areas

A crucial next step is to test the model on regions other than those from which its training samples were drawn would be illuminating, directly assessing its broadscale generalizability. Would the Q2000 model work well on the G-LiHT Yucatan-Centro area, or Chiapas, or the Campeche, or on the Out-of-the-Yucatan regions? A related question is how well would it work on Lidar data derived from other sources besides the G-LiHT dataset? How well might it work on data from the Amazon or from Southeast Asia where a similar forest cover hides archaeological sites from similar periods?

6.3 Potential for Higher Resolution G-LiHT DEM files

A key avenue for future research is the potential to significantly enhance the visualization of Maya archaeological features within the G-LiHT dataset. We found that reprocessing NASA's LAS point cloud recordings to generate higher resolution DEM files (potentially reaching 30 cm/pixel, a nine-fold increase over the current 1 m/pixel resolution) would dramatically improve visual imagery. This enhanced detail would provide a far better source for developing deep learning models and allow for the clearer identification of Maya structures.

7. Limitations

For an initial survey of the prospect of developing an effective deep learning model for broadscale detection of Maya structures in the Yucatan by using a broadscale approach in its development, we selected the G-LIHT transects, a pseudo-random array effectively crisscrossing the region. This dataset was originally generated to estimate lumber resources in the region, so no bias regarding archaeological features or sites influenced the survey's design. This study included the area of two bounded regions (South and North2) totaling together roughly the broader study area, and the total area of the all the G-LiHT ALS transects of these two regions totals 421 km².

The included area contained by these Lidar transects provided an amount of areal coverage within its bounds to satisfy the theoretical requirement for 95% confidence level and 5% margin of error; specifically, for the overall region, at least 384 sq km would be required, assuming a random distribution of features for a basic areal proportion estimate.

Our selection of subsampling included transects exceeding that amount, totaling 421 sq km across this extensive region. This selection, meeting theoretical criteria for an initial areal survey at those statistical standards, was logistically possible because of the existence of the G-LiHT data, a publicly available resource. Without this publicly available G-LIHT data, a random collection of areal information at this scale would be practically impossible. The resources and logistical effort required to create a comparable, randomly collected areal dataset at this scale would be so substantial as to render it practically infeasible for most research projects, especially as an *initial* exploratory step.

While the G-LIHT Lidar transect survey provides a spatially unbiased and systematic sample of the north central Yucatan and western southern lowlands, representing approximately 1.2% of the total area, several inherent limitations must be acknowledged concerning its utility for evaluating a broadscale deep learning approach for Maya structure detection. Firstly, extrapolating the observed density and characteristics of detected features from the sampled 421 km² to the entirety of the expansive and topographically diverse study region for the purpose of assessing the potential effectiveness of a deep learning model across this scale relies on the assumption that the characteristics of Maya structures (e.g., their variability in form, preservation, and surrounding environment) observable in Lidar visualizations are consistent enough for a model trained on a subset to generalize effectively across the region.

Secondly, while spatial autocorrelation might theoretically reduce the effective diversity of training examples captured within the transects, the robust performance of a deep learning model trained on a carefully curated set of 2000 visually assessed targets (achieving an F1 score of .89 on a test set) suggests that the provided sample was indeed sufficient for the model to learn discriminatory features and generalize effectively to unseen data. This empirical evidence of strong model performance indicates that the diversity within the training set was likely adequate for the task.

Thirdly, while the G-LIHT survey offers a broad regional perspective, the 1.2% areal coverage inherently limits the scale at which the model's broad applicability can be directly evaluated across the entire study area. The model's performance on this subsample provides a promising indication of its potential, but its effectiveness across the full study area will ultimately depend on the consistency of the Lidar data quality and, more importantly, the characteristics of Maya structures in the unsampled regions and their differences compared to those included as training samples.

Finally, the sheer scale of the study area made comprehensive physical ground-truthing of all training and test cells logistically prohibitive, necessitating the reliance on HITL and existing authoritative data. Because of these practical logistics, selection of samples and valuation of test cells was not confirmed by physical, ground-truthed observation, so HITL selection bias and/or cases of mistaken identification may have occurred, potentially leading to an overestimation or underestimation of the model's true performance. Visual identification, even by experts, can be subjective and prone to human error or interpretation bias. While this project aimed for a broadscale model by encompassing two regions of the Yucatan, the very nature of such an approach across a diverse landscape means that the natural diversity of its terrain and settlement types is not entirely encapsulated in the training of the current model, potentially leading to lower performance or missed detections in areas with significantly different environmental or archaeological characteristics. Systematically evaluating model bias, misinterpretation, and false negatives through rigorous validation is essential for understanding the model's robustness and limitations. Further examination and development of the model with data from other areas could evaluate and ameliorate deficiencies in that regard.

8. Conclusions:

This study examines the practicability of a more efficient broadscale approach to creating deep learning models for Maya archaeology that are effective across the length and breadth of the whole Maya region. The generalizability of the proposed broadscale approach is supported by the considerably larger study area of this research, covering 35,584 km², a spatial extent vastly exceeding that of previous studies which covered <615 km² (Bundzel et al., 2020; Character et al., 2024; Richards-Rissetto et al., 2021; Somrak et al., 2020). This substantial increase in coverage highlights the potential for wider application of this deep learning methodology to Mayan archeology and more broadly to object detection models of target objects with a irregular, diverse, and relatively ambiguous nature.

Our Q2000 deep learning model produced test results comparable to the other projects which have intended to expand the idea of a broadscale approach. Significantly, our results were derived from a small, cautiously curated set of training samples distributed relatively evenly across an extremely large and varied combination of two areas, lending credence to the notion that a broadscale approach can succeed in deriving generalizable characteristics by virtue of the diversity of its input. Future work should focus on testing the model's applicability across the entire Maya Lowlands region, beyond the Yucatan Peninsula, to fully realize the benefits of this broadscale strategy for uncovering ancient Maya infrastructure and other ancient civilizations in tropical forest regions.

9. Acknowledgements:
The NASA G-LiHT data team is acknowledged for creating and processing the dataset and making their dataset publicly available.

10. Declaration of Competing Interests:

The authors declare that they have no known competing financial interests or personal relationships that have or could have appeared to influence the work reported in this paper.

11. Data availability:

Project data including the deep learning model and training data is available by contacting the corresponding author at brittobj@mail.uc.edu.
The deep learning model, Q2000 is available for open access, public domain download at: https://doi.org/10.7945/tbvg-8n31
The original G-LiHT data is available at: https://gliht.gsfc.nasa.gov/index.php?section=34

12. References:

Argyrou, A. and Agapiou, A., 2022. A review of artificial intelligence and remote sensing for archaeological research. *Remote Sensing*, *14*(23), p.6000. https:// 10.3390/rs14236000

Azarkhordad, F., Zarajabad, H.H., Taghavi, A. and Kherad, M., 2025. Detecting archaeological phenomena using deep learning in the study of the old aerial images of historical city of Zuzan. *Archaeological Prospection*, *32*(2), pp.409-418. https://doi.org/10.1002/arp.1967

Beach, T., Luzzadder-Beach, S., Cook, D., Dunning, N., Kennett, D.J., Krause, S., Terry, R., Trein, D. and Valdez, F., 2015. Ancient Maya impacts on the Earth's surface: An Early Anthropocene analog?. Quaternary Science Reviews, 124, pp.1-30. https://doi.org/10.1016/j.quascirev.2015.05.028

Bickler, S.H. and Jones, B., 2021. Scaling up deep learning to identify earthwork sites in Te Tai Tokerau, Northland, New Zealand. *Archaeology*, *16*(1).

Bickler, S.H., 2021. Machine Learning Arrives in Archaeology. *Advances in Archaeological Practice. 2021; 9: 186–191* [online] https://10.1017/aap.2021.6

Bundzel, M., Jašćur, M., Kováčs, M., Lieskovský, T., Sinčák, P., Tkáčik, T., 2020. Semantic segmentation of airborne LiDAR data in Maya archaeology. Rem. Sens. 12 (22), 3685. https://doi.org/10.3390/rs12223685

Canedo, D., Fonte, J., Seco, L.G., Vázquez, M., Dias, R., Do Pereiro, T., Hipólito, J., Menéndez-Marsh, F., Georgieva, P. and Neves, A.J., 2023. Uncovering archaeological sites in airborne LiDAR data with data-centric artificial intelligence. *IEEE Access*. https://10.1109/ACCESS.2023.3290305

Canedo, D., Hipólito, J., Fonte, J., Dias, R., do Pereiro, T., Georgieva, P., Gonçalves-Seco, L., Vázquez, M., Pires, N., Fábrega-Álvarez, P. and Menéndez-Marsh, F., 2024. The Synergy between Artificial Intelligence, Remote Sensing, and Archaeological Fieldwork Validation. *Remote Sensing*, *16*(11), p.1933. https://doi.org/10.3390/rs16111933

Character, L., Beach, T., Inomata, T., Garrison, T.G., Luzzadder-Beach, S., Baldwin, J.D., Cambranes, R., Pinzón, F. and Ranchos, J.L., 2024. Broadscale deep learning model for archaeological feature detection across the Maya area. *Journal of Archaeological Science*, *169*, p.106022 https://doi.org/10.1016/j.jas.2024.106022

Chase, A.F. and Chase, D.Z., 2017. Detection of Maya ruins by LiDAR: Applications, case study, and issues. *Sensing the Past: From artifact to historical site*, pp.455-468. https://doi.org/10.1007/978-3-319-50518-3_22

Chase, A.F., Chase, D.Z., Weishampel, J.F., Drake, J.B., Shrestha, R.L., Slatton, K.C., Awe, J.J. and Carter, W.E., 2011. Airborne LiDAR, archaeology, and the ancient Maya landscape at Caracol, Belize. *Journal of Archaeological Science*, *38*(2), pp.387-398. https://doi.org/10.1016/j.jas.2010.09.018

Chase, A.F., 1988. Jungle surveying: mapping the archaeological site of caracol, Belize. *POB (Point of Beginning)*, *13*(3), p.10.

Davis, D.S., 2020. Geographic disparity in machine intelligence approaches for archaeological remote sensing research. *Remote Sensing*, *12*(6), p.921. https://doi.org/10.3390/rs12060921

Davis, D. S., Caspari, G., Lipo, C. P., & Sanger, M. C. (2021). Deep learning reveals extent of Archaic Native American shell-ring building practices. Journal of Archaeological Science, 132, 105433.

der Vaart, W. V., Bonhage, A., Schneider, A., Ouimet, W., & Raab, T. (2023). Automated large-scale mapping and analysis of relict charcoal hearths in Connecticut (USA) using a Deep Learning YOLOv4 framework. Archaeological Prospection, 30(3), 251–266. https://doi.org/10.1002/arp.1889

Esri. (2024). *ArcGIS Pro* (Version 3.3) [Computer software]. Redlands, CA: Environmental Systems Research Institute.
Fernandez-Diaz, Juan Carlos, 2014. Now You See It… Now You Don't: Understanding Airborne Mapping LiDAR Collection and Data Product Generation for Archaeological Research in Mesoamerica. https://doi.org/10.3390/rs6109951

Evans, T. L., Fisher, C. T., & Arroyo, E. (2021). *Airborne Lidar in Archeology: A Guide for Practitioners*. Springer.

Folan, W. J., Fletcher, L. A., & Kintz, E. R. (2009). *Coba: A Classic Maya Metropolis*. Academic Press.

Folan, W.J., Marcus, J., Pincemin, S., Carrasco, M.D.R.D., Fletcher, L. and López, A.M., 1995. Calakmul: New data from an ancient Maya capital in Campeche, Mexico. *Latin American Antiquity*, *6*(4), pp.310-334.

https://doi.org/10.2307/971834

G-LiHT Data Products and File Formats (n.d.) *G-LiHT Data Products and File Formats*. Available at: https://gliht.gsfc.nasa.gov/index.php?section=34 (Accessed: 29 December 2024).

Gallwey, J., Eyre, M., Tonkins, M. and Coggan, J., 2019. Bringing lunar LiDAR back down to earth: Mapping our industrial heritage through deep transfer learning. Remote Sensing, 11(17), p.1994. https://doi.org/10.3390/rs11171994

Garrison, T. G., Houston, S. D., & Inomata, T. (2019). Lidar and the future of Maya archaeology. Latin American Antiquity, 30(4), 793-802.

GEDI Ecosystem LiDAR. (n.d.) *GEDI Ecosystem LiDAR*. Available at: https://gedi.umd.edu/ (Accessed: 29 December 2024).

Golden, C., Murtha, T., Cook, B., Shaffer, D.S., Schroder, W., Hermitt, E.J., Firpi, O.A. and Scherer, A.K., 2016. Reanalyzing environmental lidar data for archaeology: Mesoamerican applications and implications. *Journal of Archaeological Science: Reports*, *9*, pp.293-308. https://doi.org/10.1016/j.jasrep.2016.07.029

Goodfellow, I., Bengio, Y., Courville, A. and Bengio, Y., 2016. *Deep learning* (Vol. 1, No. 2). Cambridge: MIT press.

Hansen, R.D., Morales-Aguilar, C., Thompson, J., Ensley, R., Hernández, E., Schreiner, T., Suyuc-Ley, E. and Martínez, G., 2023. LiDAR analyses in the contiguous Mirador-Calakmul Karst Basin, Guatemala: an introduction to new perspectives on regional early Maya socioeconomic and political organization. *Ancient Mesoamerica*, *34*(3), pp.587-626. https://doi.org/10.1017/S0956536122000244

Inomata, T., Pinzón, F., Ranchos, J.L., Haraguchi, T., Nasu, H., Fernandez-Diaz, J.C., Aoyama, K., Yonenobu, H., 2017. Archaeological application of airborne LiDAR with object-based vegetation classification and visualization techniques at the Lowland Maya site of Ceibal, Guatemala. Rem. Sens. 9 (6), 563. https://doi.org/10.3390/rs9060563

Inomata, T., 2024. Lidar, Space, and Time in Archaeology: Promises and Challenges. Annual Review of Anthropology, 53. https://doi.org/10.1146/annurev-anthro-041222-093758

Jannat, F.E., Zhang, J., Willis, A. and Ringle, W., 2023, April. Extracting Ancient Maya Structures from Aerial LiDAR Data using Deep Learning. In *SoutheastCon 2023* (pp. 295-302). IEEE. https://doi.org/10.1109/SoutheastCon51012.2023.10115095

Khan, A., Rauf, Z., Sohail, A., Khan, A.R., Asif, H., Asif, A. and Farooq, U., 2023. A survey of the vision transformers and their CNN-transformer based variants. *Artificial Intelligence Review*, *56*(Suppl 3), pp.2917-2970. https://doi.org/10.1007/s10462-023-10595-0

Klie, J.C., Castilho, R.E.D. and Gurevych, I., 2024. Analyzing dataset annotation quality management in the wild. Computational Linguistics, 50(3), pp.817-866. https://doi.org/10.1162/coli_a_00516

Koby, P.J., 2012. Spatial analysis of ancient maya settlement near karst sinkholes at Xuenkal, Yucatan, Mexico (Bachelor's thesis, Kent State University).

Kokalj, Ž., & Somrak, M. (2019). R.V.T. (Relief Visualization Toolbox) version 2.21 (Software). https://iaps.zrc-sazu.si/en/rvt#v

Kokalj, Ž. and Mast, J., 2021. Space lidar for archaeology? Reanalyzing GEDI data for detection of ancient Maya buildings. *Journal of Archaeological Science: Reports*, *36*, p.102811.
https://doi.org/10.1016/j.jasrep.2021.102811

Kokalj, Ž., Džeroski, S., Šprajc, I., Štajdohar, J., Draksler, A. and Somrak, M., 2023. Machine learning-ready remote sensing data for Maya archaeology. *Scientific Data*, *10*(1), p.558.
https://doi.org/10.1038/s41597-023-02455-x
https://storymaps.arcgis.com/stories/b1ca4149cd074b5f8b728ab7fcc171aa

Landauer, J., Klassen, S., Wijker, A. P., van der Kroon, J., Jaszkowski, A., & Verschoof-van der Vaart, W. B. (2025). Beyond the Greater Angkor Region: Automatic large-scale mapping of Angkorian-period reservoirs in satellite imagery using deep learning. PloS one, 20(3), e0320452.

Liao, Z., Dong, X. and He, Q., 2024. Calculating the Optimal Point Cloud Density for Airborne LiDAR Landslide Investigation: An Adaptive Approach. *Remote Sensing*, *16*(23), p.4563.
https://www.mdpi.com/2072-4292/16/23/4563

Marçal, D., Câmara, A., Oliveira, J. and de Almeida, A., 2024, June. Evaluating R-CNN and YOLO V8 for Megalithic Monument Detection in Satellite Images. In *International Conference on Computational Science* (pp. 162-170). Cham: Springer Nature Switzerland.
https://doi.org/10.1007/978-3-031-63759-9_20

Merk, S., Huacuja Gallman, T., Del Gatto, R., Marti, K.. 2025 (in press). A brief report on the ancient Maya ruins of Pixoyal, Campeche. *Mexicon*.

Quintus, S., Davis, D. S., & Cochrane, E. E. (2023). Evaluating Mask R-CNN Models to Extract Terracing across Oceanic High Islands: An example from Sāmoa. Archaeological Prospection, 30(4), 477–492.

Redmon, J. and Farhadi, A., 2018. Yolov3: An incremental improvement. *arXiv preprint arXiv:1804.02767*.
https://doi.org/10.48550/arXiv.1804.02767

Richards-Rissetto, H., Newton, D., Al Zadjali, A., 2021. A 3D point cloud deep learning approach using lidar to identify ancient Maya archaeological sites. *ISPRS Annals of the Photogrammetry, Remote Sensing and Spatial Information Sciences*, Volume VIII-M-1-2021, 28 August-1 September. In: 28th CIPA Symposium "Great Learning and Digital Emotion. Beijing, China.
https://digitalcommons.unl.edu/anthropologyfacpub/183/

Ringle, W.M., Gallareta Negrón, T., May Ciau, R., Seligson, K.E., Fernandez-Diaz, J.C. and Ortegón Zapata, D., 2021. Lidar survey of ancient Maya settlement in the Puuc region of Yucatan, Mexico. *PLoS One*, *16*(4), p.e0249314.
https://doi.org/10.1371/journal.pone.0249314

Ruhl, T., Dunning, N.P. and Carr, C., 2018. Lidar reveals possible network of ancient Maya marketplaces in southwestern Campeche, Mexico. *Mexicon*, *40*(3), pp.83-91.
https://www.researchgate.net/publication/326403117_Lidar_reveals_possible_network_of_ancient_Maya_marketplaces_in_southwestern_Campeche_Mexico

Sheets, P., Hoopes, J., Melson, W., McKee, B., Sever, T., Mueller, M., Chenault, M. and Bradley, J., 1991. Prehistory and volcanism in the Arenal area, Costa Rica. Journal of Field Archaeology, 18(4), pp.445-465.
https://doi.org/10.1179/009346991791549112

Somrak, M., Džeroski, S., Kokalj, Ž., 2020. Learning to classify structures in ALS-derived visualizations of ancient Maya settlements with CNN. Rem. Sens. 12 (14), 2215.
https://doi.org/10.3390/rs12142215

Sevara, C., Pregesbauer, M., Doneus, M., Verhoeven, G., Trinks, I., 2016. Pixel versus object–a comparison of strategies for the semi-automated mapping of archaeological features using airborne laser scanning data. J. Archaeol. Sci.: Reports 5, 485–498.
https://doi.org/10.1016/j.jasrep.2015.12.023

Schroder, W., Murtha, T., Golden, C., Anaya Hernández, A., Scherer, A., Morell-Hart, S., Almeyda Zambrano, A., Broadbent, E., Brown, M., 2020. The lowland Maya settlement landscape: environmental LiDAR and ecology. J. Archaeol. Sci.: Reports
https://doi.org/10.1016/j.jasrep.2020.102543

Trier, Ø. D., Reksten, J. H., & Løseth, K. (2021). Automated mapping of cultural heritage in Norway from airborne lidar data using faster R-CNN. International Journal of Applied Earth Observation and Geoinformation, 95, 102241.
https://doi.org/10.1016/j.jag.2020.102241

Verschoof-van der Vaart, W. B., & Lambers, K. (2019). Learning to Look at LiDAR: The Use of R-CNN in the Automated Detection of Archaeological Objects in LiDAR Data from the Netherlands. Journal of Computer Applications in Archaeology, 2(1), 31–40. https://doi.org/10.5334/jcaa.32

Wang, S., Hu, Q., Wang, S., Ai, M. and Zhao, P., 2024. Archaeological site segmentation of ancient city walls based on deep learning and LiDAR remote sensing. *Journal of Cultural Heritage*, *66*, pp.117-131.
https://doi.org/10.1016/j.culher.2023.11.005

Why data annotation precision is critical important and how to provide 100% quality?. (2023) Keylabs. Available at:
https://keylabs.ai/blog/why-data-annotation-precision-is-critical-important-and-how-to-provide-100-quality/
(Accessed: 29 December 2024).

Wu, X., Xiao, L., Sun, Y., Zhang, J., Ma, T. and He, L., 2022. A survey of human-in-the-loop for machine learning. *Future Generation Computer Systems*, *135*, pp.364-381.

Zhang, J., Ringle, W. and Willis, A.R., 2024. Unveiling Ancient Maya Settlements Using Aerial LiDAR Image Segmentation. *arXiv preprint arXiv:2403.05773*.
https://doi.org/10.48550/arXiv.2403.05773