Archaeological Prospection

WILEY

RESEARCH ARTICLE OPEN ACCESS

Automated Detection of Hillforts in Remote Sensing Imagery With Deep Multimodal Segmentation

Daniel Canedo[1] | João Fonte[2,3,4] | Rita Dias[3,5] | Tiago do Pereiro[3] | Luís Gonçalves-Seco[6,7] | Marta Vázquez[6,8] | Petia Georgieva[1,9] | António J. R. Neves[1]

[1]IEETA/DETI, University of Aveiro, Aveiro, Portugal | [2]Department of Archaeology and History, University of Exeter, Exeter, UK | [3]ERA Arqueologia, Calçada de Santa Catarina, Cruz Quebrada, Portugal | [4]Centre for the Humanities (CHAM), Faculdade de Ciências Sociais e Humanas, Universidade NOVA de Lisboa, Lisboa, Portugal | [5]ICArEHB, Campus de Gambelas, Universidade do Algarve, Faro, Portugal | [6]UMAIA, University of Maia, Maia, Portugal | [7]INESC TEC–Institute for Systems and Computer Engineering, Technology and Science, University of Porto, Porto, Portugal | [8]N2i, Polytechnic Institute of Maia, Maia, Portugal | [9]Instituto de Telecomunicações, Universidade de Aveiro, Aveiro, Portugal

Correspondence: João Fonte (j.fonte3@exeter.ac.uk)

Received: 22 May 2024 | **Revised:** 17 July 2024 | **Accepted:** 16 August 2024

Funding: This research was supported by the project 'Odyssey: Platform for Automated Sensing in Archaeology' (Ref. ALG-01-0247-FEDER-070150), co-financed by COMPETE 2020 and Regional Operational Program Lisboa 2020, through Portugal 2020 and FEDER. This work was also supported by the European Union-NextGenerationEU, through the National Recovery and Resilience Plan of the Republic of Bulgaria, project N0 BG-RRP-2.004-0005, and the FCT/MCTES through national funds and when applicable co-funded EU funds under the project UIDB/50008/2020-UIDP/50008/2020. Additionally, this article had the support of CHAM (NOVA FCSH/UAc), through the strategic project sponsored by FCT (UIDB/04666/2020) – https://doi.org/10.54499/UIDB/04666/2020.

Keywords: computer vision | hillforts | LiDAR | multimodal semantic segmentation | orthoimagery | transformer

ABSTRACT

Recent advancements in remote sensing and artificial intelligence can potentially revolutionize the automated detection of archaeological sites. However, the challenging task of interpreting remote sensing imagery combined with the intricate shapes of archaeological sites can hinder the performance of computer vision systems. This work presents a computer vision system trained for efficient hillfort detection in remote sensing imagery. Equipped with an adapted multimodal semantic segmentation model, the system integrates LiDAR-derived LRM images and aerial orthoimages for feature fusion, generating a binary mask pinpointing detected hillforts. Post-processing includes margin and area filters to remove edge inferences and smaller anomalies. The resulting inferences are subjected to hard positive and negative mining, where expert archaeologists classify them to populate the training data with new samples for retraining the segmentation model. As the computer vision system is far more likely to encounter background images during its search, the training data are intentionally biased towards negative examples. This approach aims to reduce the number of false positives, typically seen when applying machine learning solutions to remote sensing imagery. Northwest Iberia experiments witnessed a drastic reduction in false positives, from 5678 to 40 after a single hard positive and negative mining iteration, yielding a 99.3% reduction, with a resulting F_1 score of 66%. In England experiments, the system achieved a 59% F_1 score when fine-tuned and deployed countrywide. Its scalability to diverse archaeological sites is demonstrated by successfully detecting hillforts and other types of enclosures despite their typical complex and varied shapes. Future work will explore archaeological predictive modelling to identify regions with higher archaeological potential to focus the search, addressing processing time challenges.

1 | Introduction

The application of machine learning for the detection of automated archaeological features in remote sensing data has a growing track record (Argyrou and Agapiou 2022; Câmara et al. 2023; Davis 2020; Fiorucci et al. 2020). Deep learning algorithms are currently at the forefront of this field, encompassing object detection and semantic segmentation tasks (Casini et al. 2023; Guyot et al. 2021). However, their effectiveness is dependent on large training datasets and significant computational resources (Fiorucci et al. 2022; Karamitrou et al. 2022; Kazimi and Sester 2023; Olivier and Verschoof-van der Vaart 2021). These techniques have successfully identified various archaeological features with varying morphologies, even considering ephemeral archaeological remains (Davis 2021). However, burial mounds are often the primary target due to their consistent rounded shape (Berganzo-Besga et al. 2021; Canedo et al. 2023; Cerrillo-Cuenca 2017; Orengo et al. 2020; Verschoof-van der Vaart and Lambers 2022). Furthermore, complex archaeological features with intricate shapes, such as enclosures, pose a significant challenge for these methods, particularly due to the prevalence of similar shapes in the landscape, increasing the likelihood of many false positives. In terms of this, promising research has already been conducted in hillforts, offering good insight into future advancements (Landauer and Verschoof-van der Vaart 2021; Landauer et al. 2023).

Landauer and Verschoof-van der Vaart (2021) introduced a computer vision system that consists of a convolutional neural network (CNN) for the automated detection of complex archaeological features, focusing on hillforts from England as a case study. The authors selected light detection and ranging (LiDAR)-derived digital surface models of England as their primary dataset. An interesting aspect of their approach lies in the building of their training dataset. Specifically, the authors intentionally biased the training dataset towards negative samples (non-hillforts) due to the low probability of encountering them when scanning through the data. This deliberate bias aims to align CNN more closely with real-world scenarios. Additionally, the authors employed data augmentation to mitigate centre bias, ensuring that CNN does not expect hillforts to always be in the centre of the image, as their dataset comprised cropped images centred around hillforts. The CNN was then trained for image classification, assessing the likelihood that each cropped image contained a hillfort. Three different regions with varying typologies in England were fed to CNN, resulting in 306 inferences out of a total of 20 600 images. This highlights that the CNN makes approximately one inference per 67 images, a notable achievement considering the challenges posed by aerial imagery and hillforts' diverse and intricate shapes. However, the authors note that a significant number of these 306 inferences were ultimately identified as false positives during remote verification. This not only highlights the difficulty of detecting these complex archaeological sites in aerial imagery but also illustrates the challenges facing CNN in detecting hillforts with diverse typologies across different landscapes. In a following work, Landauer et al. (2023) introduced the concept of fine-tuning CNN before deploying it in other countries with distinct site typologies and landscapes. Furthermore, the authors proposed a web application named *HillfortFinderApp* (Landauer) to disseminate their findings and approach to the scientific community.

This work builds on existing research and proposes an alternative approach for automatically detecting hillforts in remote sensing imagery. The approach utilizes a deep multimodal segmentation model and was tested using archaeological and geospatial data from Alto Minho (Portugal), Galicia (Spain) and England (United Kingdom). In contrast to the image classification method proposed by Landauer and Verschoof-van der Vaart (2021) and Landauer et al. (2023), or even object detection approaches, this work suggests addressing the task of detecting hillforts through segmentation. This suggestion is mainly related to accurately delineating irregular shapes like hillforts during annotation, and the consequent impact on the model's learning process (Fontana 2022; Murrugarra-Llerena, Kirsten, and Jung 2022). Annotating with masks allows for a precise understanding of the exact shape of hillforts, which will better guide the model during training. Additionally, segmentation reduces ambiguity compared with image classification and object detection methods. Cropped images and bounding boxes might encompass a large area of background information, particularly when the annotated instance has an irregular shape, which can negatively impact the model's learning process. Masks, however, address this issue by accurately delineating hillforts without introducing unnecessary information to the annotation.

The main novelty of this work lies in its comprehensive methodology that is discussed in Section 2. Starting with how the datasets are pre-processed and generated, followed by the early fusion of local relief models (LRMs) and orthoimages. This approach aims to replicate the multi-data analysis of expert archaeologists during digital surveys. The methodology also introduces margin and area filters to reduce false positives and bridges archaeological expertise and artificial intelligence through hard positive and negative mining. Additionally, a major contribution of this study is its large scope. Experiments were conducted in three different areas, each with its own site typologies and landscapes, requiring adaptations to address these differences. The conclusions derived from these varied contexts further enhance the study's impact on the field.

This document is organized as follows: Section 2 outlines the methodology. Section 3 presents the results. Section 4 discusses the results and provides additional information about the conducted work, and Section 5 formulates the concluding remarks.

2 | Methodology

Figure 1 summarizes the computer vision system proposed that will be discussed in this section.

Based on findings from the literature (Section 1) and previous experiments, many false positives are inevitable when machine learning algorithms are implemented on remote sensing images to detect archaeological sites. This is particularly true in cases where archaeological sites of interest exhibit complex and varying shapes, as is often the case with hillforts (Fontana 2022; Landauer et al. 2023; Landauer and Verschoof-van der Vaart 2021). Similar morphologies, natural or artificial, are commonly found in the landscape and are not necessarily hillforts. This is an additional difficulty, as machine learning models will struggle to distinguish hillforts from other similar landscape shapes. These structures often have common features, which

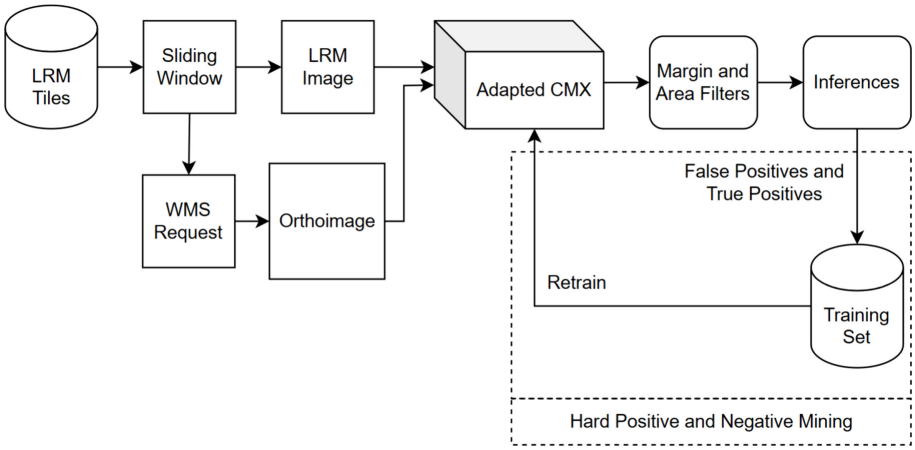

FIGURE 1 | Proposed computer vision system. [Colour figure can be viewed at wileyonlinelibrary.com]

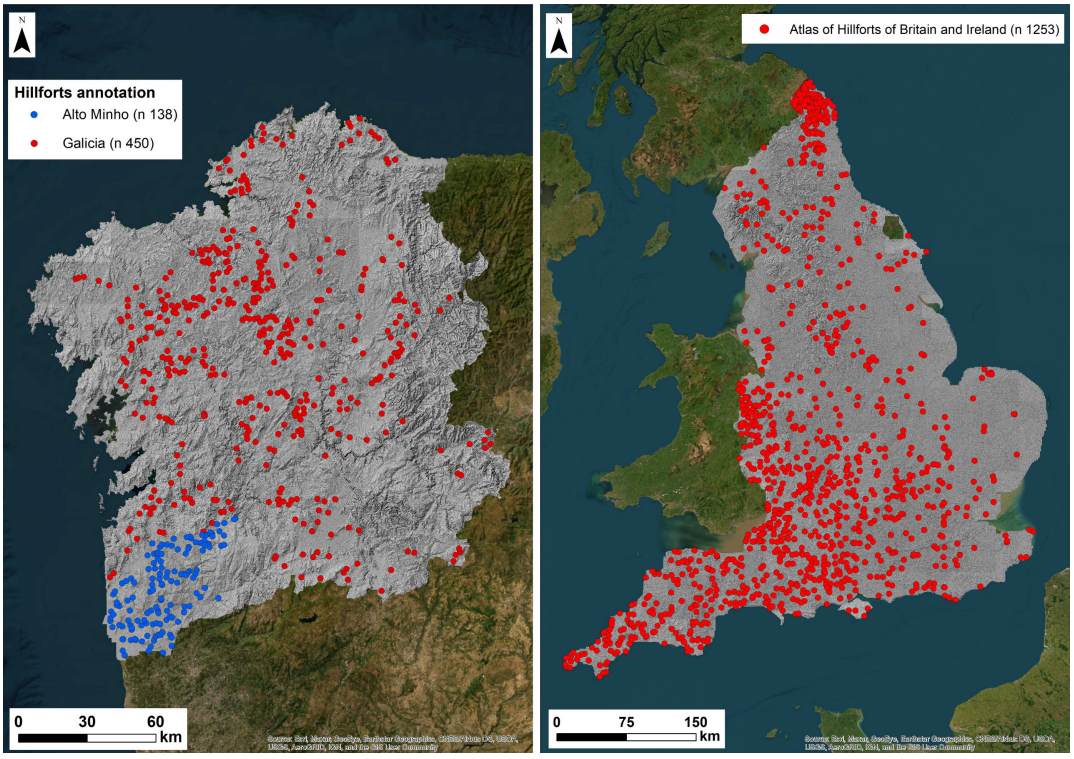

FIGURE 2 | LRMs and respective annotations. Alto Minho and Galicia on the left, England on the right. [Colour figure can be viewed at wileyonlinelibrary.com]

inevitably puts them near the decision boundary. Therefore, the main challenge of this particular task is to reduce as many false positives as possible while keeping the detected true positives.

With only a few exceptions, the works reported in the literature typically use a single data type to detect archaeological sites (Berganzo-Besga et al. 2021). This data type commonly results from the application of visualization techniques to LiDAR-derived digital terrain models (DTMs) (Bennett et al. 2012; Guyot, Lennon, and Hubert-Moy 2021; Kokalj and Somrak 2019; Štular et al. 2012). LiDAR technology offers a significant advantage over other aerial data collection methods, namely, its ability to penetrate vegetation cover and, as such, to create highly accurate DTM of the bare earth below, potentially revealing archaeological sites hidden beneath the canopy (Štular, Lozić, and Eichert 2021; Vinci et al. 2024). Although this represents a significant advantage, DTMs only provide visualization of microtopography, lacking other types of information and features, crucial to identifying archaeological sites and distinguishing them from other similar shapes in the landscape. Therefore, it is hypothesized that, similarly to archaeologists mapping archaeological sites in LiDAR data, machine learning models would also benefit from having access to multiple data types when faced with the same problem. In this work, the early fusion technique (Tziafas and Kasaei 2023) is proposed, which involves two types of data: LRMs generated from LiDAR-derived DTMs (Hesse 2010) and aerial orthoimages. By performing a fusion between these two data types, the rich information and features provided by the orthoimages can be expected to complement and contextualize the microtopographies depicted in the LRMs.

This work used airborne LiDAR data covering the Alto Minho region (≈2 points per square meter, ≈2220 km^2) in northern Portugal provided by the Comunidade Intermunicipal do Alto Minho, Galicia (≈0.5 points per square meter, ≈30 000 km^2) in Spain provided by Instituto Geográfico Nacional (IGN)-Plan Nacional de Ortofotografía Aérea (PNOA), as well as England (≈2 points per square meter, ≈130 000 km^2) provided by Environment Agency National LIDAR Programme. From the classified LiDAR point clouds, 1-m DTMs were generated, to which LRMs were applied to create greater contrast in the visualization of the positive

Algorithm 1 Pseudocode for the dataset generation algorithm.

```
processed_hillforts = []
for annotations do
    if annotation not in processed_hillforts then
        generate cropping window coordinates
        for cropping window coordinates do
            crop LRM image
            if hillforts visibility == 100% then
                WMS request to obtain the corresponding orthoimage
                store LRM image, orthoimage, and corresponding annotations
                append stored annotations to processed_hillforts
            end if
        end for
    end if
end for
```

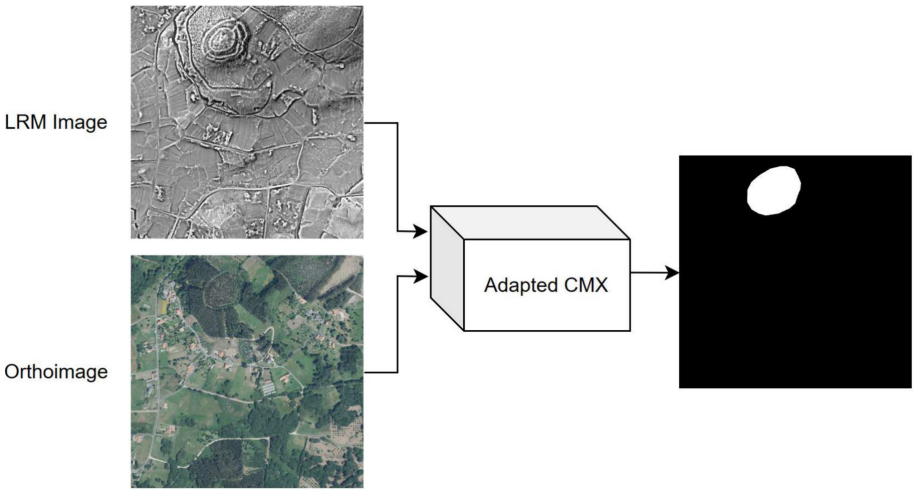

FIGURE 3 | Input/Output of the adapted CMX model. [Colour figure can be viewed at wileyonlinelibrary.com]

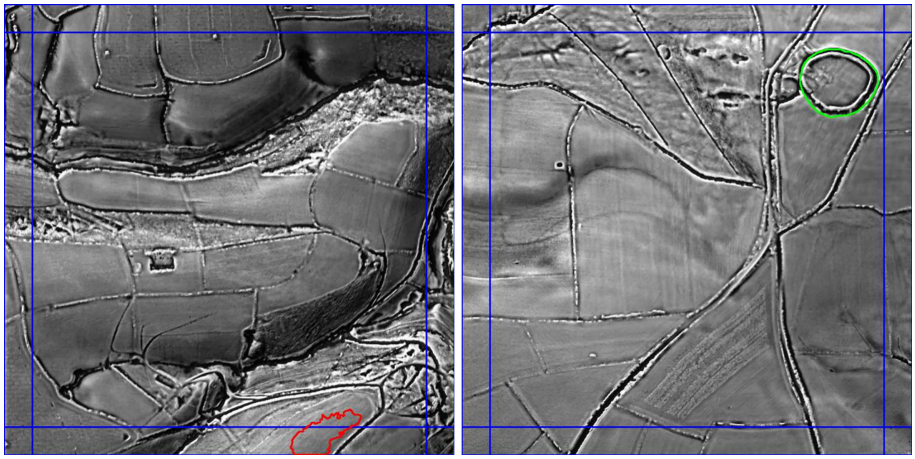

FIGURE 4 | The margin filtering process removes inferences of partially visible objects. On the left, the margin filter discarded a false positive inference (shown in red) because its IoU with the margin region exceeds 0.5. On the right, a true positive inference (shown in green) was retained as its IoU with the margin region is 0. [Colour figure can be viewed at wileyonlinelibrary.com]

(walls and ramparts) and negative (ditches) archaeological microtopographies of the hillforts on a local scale. Two expert archaeologists (Fonte J. and Menendez Marsh F.) manually annotated the full extent of some of the known hillforts as polygons in Alto Minho ($N = 138$) and Galicia ($N = 450$). The dataset Atlas of Hillforts of Britain and Ireland (Lock and Ralston 2022), which comprises 4147 entries, was considered for this work. Figure 2 illustrates the LRMs and annotations explored in this work.

Because LRMs are essentially large images, an image processing algorithm was developed to divide them into smaller images with a resolution of 800×800 pixels, creating a conventional dataset for machine learning. This algorithm iterates over the annotated hillforts and creates a list of top left coordinates of a cropping window with a resolution of 800×800 pixels around them, respecting the following equations:

$$\begin{aligned} X_{points} &= \left[X_{max} - \frac{width}{2}, X_{min} + \frac{width}{2}\right] \\ Y_{points} &= \left[Y_{max} - \frac{height}{2}, Y_{min} + \frac{height}{2}\right] \end{aligned} \tag{1}$$

X_{min}, X_{max}, Y_{min} and Y_{max} are the extremities of the annotation, and the width and height are set to 800×800 pixels. Next, the algorithm shuffles this list and iterates through it to crop LRM images around the current hillfort being processed. The hillforts depicted in the cropped image must be fully visible, and the algorithm aims to avoid cropping images that contain previously stored hillforts. Then, a Web Map Service (WMS) request is made to obtain the orthoimage. The LRM image, orthoimage and the corresponding annotations are stored, and the algorithm proceeds to the next hillfort. The pseudocode of the dataset generation algorithm is provided in Algorithm 1.

The semantic segmentation model used in this work was an adaptation of the cross-modal fusion for RGB-X Semantic Segmentation with Transformers (CMX) (Liu; Zhang et al. 2023) proposed by Zhang et al. This model uses transformer technology (Dosovitskiy et al. 2020), which was originally developed for natural language processing. This technology offers several advantages over conventional CNNs. It requires substantially fewer computational resources to train and can capture global dependencies in the image by splitting it into patches and using the self-attention mechanism. This allows for the consideration of relationships between distant pixels or regions, which is imperative in tasks where the global context of the image matters. As the name implies, CMX is a semantic segmentation model developed to perform an early fusion between two modalities: red, green and blue (RGB) and other arbitrary ones, referred to as X-modality. To ensure effective generalization across diverse modalities, the authors proposed a cross-modal feature rectification module to calibrate bimodal features by using features from one modality to rectify those of the other. Rectified feature pairs are then processed by a feature fusion module to facilitate extensive exchange of long-range contexts before integration. Because CMX was initially developed for multiclass segmentation, certain modifications were made to enable binary segmentation. Figure 3 illustrates the data types used in this work and the potential segmentation output.

TABLE 1 | Dataset generated by Algorithm 1.

Set	Images	Hillforts
Training	530	533
Validation	59	60

Although data fusion typically improves performance, remote sensing imagery remains challenging and requires additional considerations to address the associated false positive problem. Therefore, hard positive and negative mining (Davis et al. 2021; Schneider et al. 2015; Tang et al. 2017) are proposed. The hypothesis is that by augmenting the original training data, which was used to train the segmentation model, with false positives and true positives, and then retraining the model, it will gradually improve its ability to distinguish hillforts from other similar shapes in the landscape. This approach was inspired by a work conducted by Olivier and Verschoof-van der Vaart (2021), where they added negative examples (background images and no objects of interest) to their dataset, with the final goal of reducing false positives.

Furthermore, the training data were deliberately biased towards negative examples. This approach was inspired by a work conducted by Landauer and Verschoof-van der Vaart (2021). Although this might appear counter-intuitive, it serves a specific purpose: in deployment, the segmentation model is far more likely to encounter negative examples than hillforts. This is because during deployment, the segmentation model needs to sweep through tiles covering vast regions to search for hillforts. As a result, the vast majority of the data it processes does not contain these archaeological sites.

For deployment, several considerations were made to enhance the overall performance of the proposed solution. The adapted

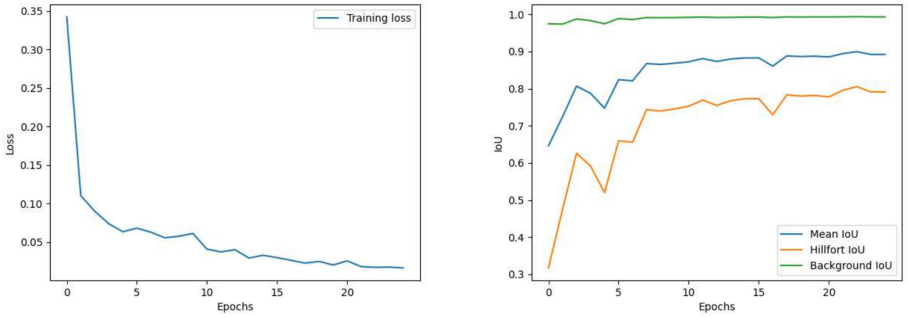

FIGURE 5 | Training results of the adapted CMX model on the original dataset. Training loss and validation IoU, respectively. [Colour figure can be viewed at wileyonlinelibrary.com]

CMX must process pairs of LRMs and orthoimages covering as many regions as required by the user. This translates to image processing where a sliding window scans through the LRM tiles. Using the coordinates of the sliding window, a WMS request is made to obtain the corresponding orthoimage. The proposed window step is set to 1/2, ensuring that a hypothetical site is fully visible to the algorithm at least once. The union operation is then applied to handle overlapping inferences. In addition, the area of these inferences is calculated, and a minimum area threshold is applied to remove smaller, noisy inferences that are not coherent with typical hillfort dimensions. Another crucial improvement to remove potential false positives was achieved by establishing a region along the image's margins with a width of 50 pixels. Any inference with an intersection over union (IoU) greater than 0.5 with this region is removed. This process ensures the elimination of inferences of partially visible objects situated on the image margins, which are often indicative of false positives. Figure 4 illustrates this process.

TABLE 2 | Analysis of the images and corresponding inferences generated by the original model on unused data.

Images	Negative examples	Positive examples	Uncertain
4462	4042	127	293

TABLE 3 | Refined dataset obtained by hard positive and negative mining.

Set	Images	Hillforts	Background Images
Training	4281	648	3637
Validation	477	73	405

3 | Results

Several experiments were conducted to validate the proposed computer vision system to detect hillforts in remote sensing imagery. The following subsections provide detailed descriptions of these experiments. Any digital validation conducted by expert archaeologists was supported by multiple data types which helped them categorize each inference as true positive, false positive, or uncertain. Additionally, the training of the adapted CMX model was performed with the following parameters:

- Binary classification: hillfort or background
- Backbone: SegFormer (b2-sized) (Xie et al. 2021)
- Initial weights: pre-trained weights from ImageNet-1k (Deng et al. 2009)
- Optimizer: AdamW (Loshchilov and Hutter 2017)
- Learning rate: 6e-5
- Weight decay: 0.01
- Momentum: 0.9
- Batch size: 1
- Epochs: 25 or 50

The best weights are saved based on the mean IoU. Training was performed on the following hardware:

- GPU: Nvidia GeForce RTX 3080 10 GB GDDR6X
- CPU: AMD Ryzen 5 5600X 6-Core 3.7GHz

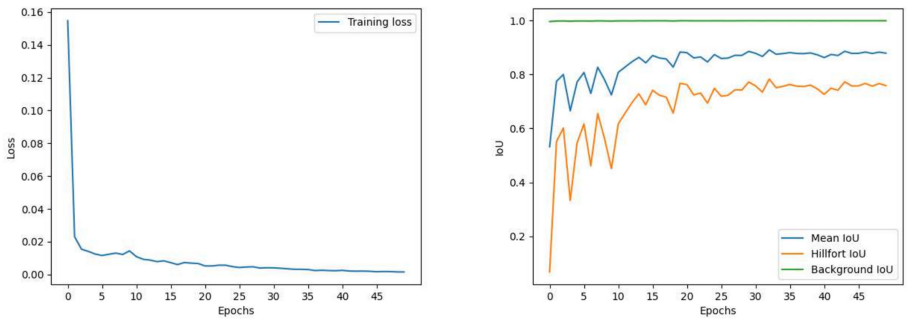

FIGURE 6 | Training results of the adapted CMX model on the refined dataset. Training loss and validation IoU, respectively. [Colour figure can be viewed at wileyonlinelibrary.com]

TABLE 4 | Inferences on the testing region.

Model	Inferences	True positives	False positives	False negatives
Original model	5687	9	5678	0
Refined model	49	9	40	0

TABLE 5 | Digital validation of inferences generated by the refined model.

True positives	Potential true positives	False positives	False negatives
15	9	25	0

3.1 | Northwest Iberia

The first experiment was conducted using the available LiDAR data for the Alto Minho region, in the north of Portugal, and for the Galicia region, in Spain. As described in Section 2, the LRMs were generated from the LiDAR-derived DTMs (divided into four tiles for Alto Minho and 87 for Galicia), both with 1-m spatial resolution. Direção Geral do Território 25 cm orthophotos from 2021 (DONP 2021) and IGN-PNOA 25 cm orthophotos from 2020 (PNOA 2020) were used for the RGB modality in Alto Minho and Galicia, respectively. It was decided to use the Parque Nacional da Peneda-Gerês (PNPG) area for testing, corresponding to the eastern tile of the Alto Minho region. Algorithm 1 used the remaining data to generate a conventional dataset. Table 1 summarizes the original dataset from this work.

TABLE 6 | Precision, recall and F_1 score of the digital validation in PNPG.

Precision	Recall	F_1 score
0.49	1.00	0.66

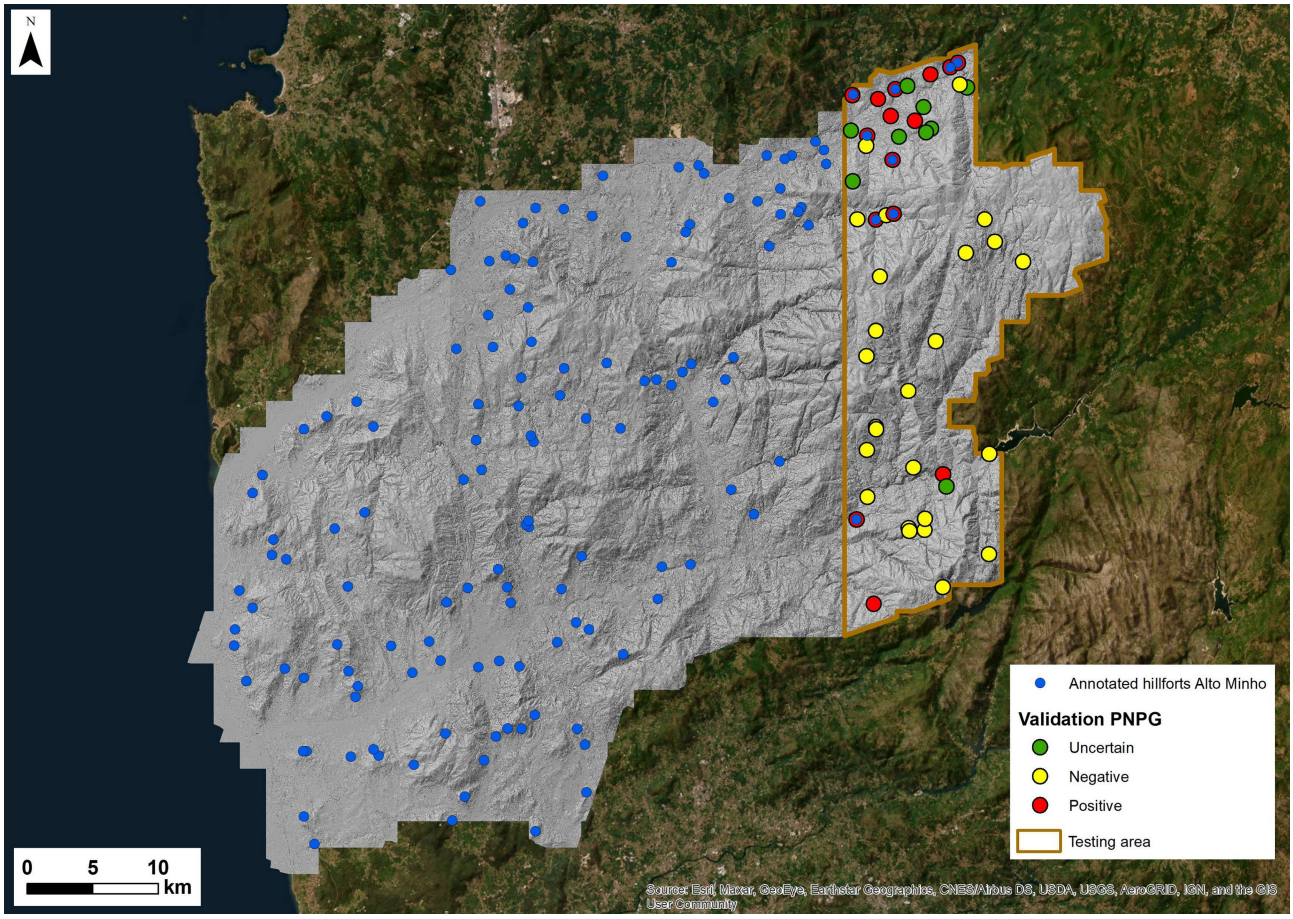

FIGURE 7 | Alto Minho and the respective testing area: PNPG. The blue points represent the annotations, and the remaining points represent classifications done by the archaeologists. [Colour figure can be viewed at wileyonlinelibrary.com]

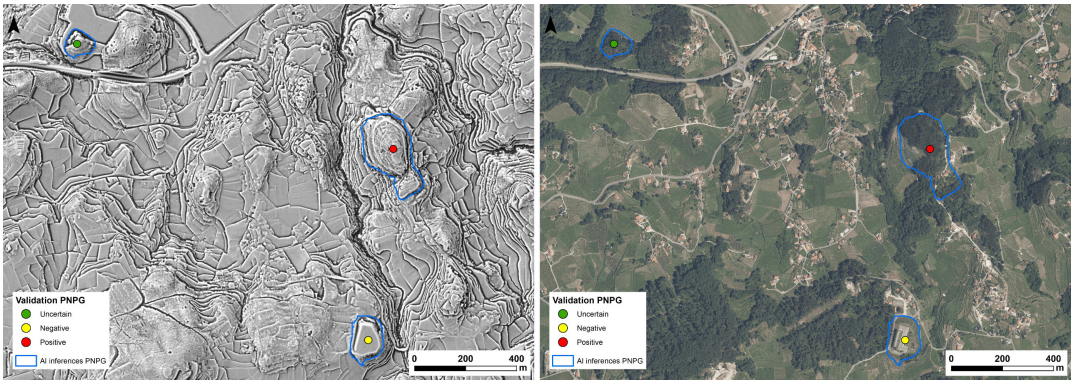

FIGURE 8 | Digital validation in PNPG. The points represent the classifications done by the archaeologists, and the blue polygons represent the inferences. [Colour figure can be viewed at wileyonlinelibrary.com]

Following the methodology described in Section 2, it is hypothesized that by biasing the training data towards negative examples, the model will enhance its ability to distinguish hillforts from similar landscape shapes. However, in order to acquire high-quality positive and negative examples, the aim is not simply to select them randomly. Instead, the approach involves training a model on the original dataset, applying it to unused data, and incorporating images containing true positive and false positive inferences into the dataset. This process is essentially the hard positive and negative mining depicted in Figure 1. Figure 5 illustrates the training results of the adapted CMX model on the original dataset.

Let us refer to this model as the *original model*. As discussed in previous sections, this model is expected to produce a significant number of false positives. The model was deployed in randomly cropped images from unused data from Alto Minho and Galicia, except for the PNPG area (≈534 km^2), which is reserved for testing. An expert archaeologist then analysed the resulting images and corresponding inferences before populating the original dataset with new samples. The analysis is shown in Table 2.

TABLE 7 | Analysis of the images and corresponding inferences generated by the refined model on four LRM tiles from England.

Images	Negative examples	Positive examples	Uncertain
256	188	45	23

TABLE 8 | Dataset obtained by hard positive and negative mining on four LRM tiles from England.

Set	Images	Hillforts	Background images
Training	209	40	169
Validation	24	5	19

Negative examples refer to images containing false positive inferences, *positive examples* refer to images containing true positive inferences, and *uncertain* refers to images where the inferences were not clear enough to be classified with certainty. Following this, the negative and positive examples were used to populate the original dataset. Table 3 summarizes the refined dataset.

Figure 6 illustrates the training results of the adapted CMX model on the refined dataset.

Let us refer to this model as the *refined model*. Both models, original and refined, were deployed in the testing region with nine annotated hillforts, as shown in Figure 1. Table 4 illustrates the number of inferences per model in the testing region.

As observed, the refined model reduced the number of false positives in 99.3% while retaining the true positives. Two expert archaeologists (Fonte J., Hipólito J.) digitally validated the inferences of the refined model, as it was believed that some of the false positives could actually be hillforts that were not originally annotated. Table 5 illustrates the results of the digital validation, according to which, of the 40 false positives that the refined model identified, only 25 were actually false positives.

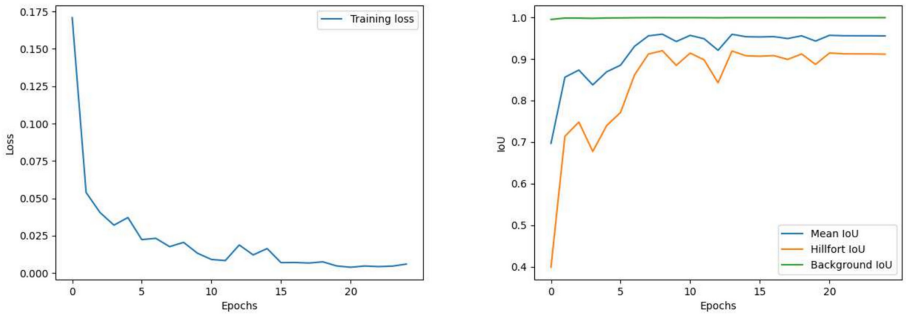

FIGURE 9 | Fine-tuning results of the adapted CMX model on the English data: training loss and validation IoU, respectively. [Colour figure can be viewed at wileyonlinelibrary.com]

FIGURE 10 | Examples of Atlas points not representing the exact locations of hillforts: with offsets of 103 and 212 m, respectively. [Colour figure can be viewed at wileyonlinelibrary.com]

Furthermore, it appears that the model identified six hillforts that had not been initially annotated and nine potential hillforts.

Table 6 presents the precision, recall and F_1 score obtained when grouping the potential true positives and the true positives classifications.

Figure 7 illustrates Alto Minho and the respective testing area, and Figure 8 illustrates some inferences generated by the refined model in PNPG.

3.2 | England

The second experiment aims to demonstrate the adaptability of the refined model discussed in Section 3.1 by applying it in England, which features a completely different landscape and even hillfort typology. Geospatial Commission Aerial Photography Great Britain 25 cm orthophotos (Bluesky International Limited and Getmapping Plc 2023) were used for the RGB modality. Four LRM tiles out of 232 covering England were selected to generate a dataset for fine-tuning the refined model to the new landscape and hillfort typology. The same hard positive and negative mining process was conducted by deploying the refined model on the selected tiles and having an archaeologist analyse the generated inferences. The results are shown in Table 7.

Negative examples refer to images containing false positive inferences, *positive examples* refer to images containing true positive inferences, and *uncertain* refers to images where the inferences were not clear enough to be classified with certainty. Following this, the negative and positive examples were used to create the dataset. Table 8 shows the created dataset.

Figure 9 illustrates the training results of the adapted CMX model on this dataset.

The computer vision system was tested on the 232 LRM tiles and orthoimages, following the pipeline shown in Figure 1. Given the substantial amount of LiDAR-derived data, totaling over 600 GB, a high-performance computing (HPC) system was utilized for processing. The HPC system consisted of two Nvidia GeForce RTX 3090 GPUs (10 GB GDDR6X each) and an AMD Ryzen 7 5800X processor (8 cores, 3.7 GHz). The processing took approximately 30 days due to the slow WMS service and the sliding window step of 1/2. In a production environment, implementing a caching mechanism could reduce the processing time to a week or less. The generated inferences were compared with the Atlas of Hillforts of Britain and Ireland dataset. As the Atlas represents the locations of known hillforts as points, inferences were classified as true positives, false positives, or false negatives based on whether the points overlapped with the inferences. However, some points in the Atlas are offset by a considerable distance from the actual location of the hillforts, as illustrated in Figure 10. In addition, some inferences only partially detected the full extent of the hillforts.

A buffer of 300 m was applied to each inference to mitigate this problem. Finally, inferences that overlapped with points from the Atlas of Hillforts dataset were considered true positives, inferences with no overlapping points were considered false positives, and isolated points were considered false negatives. Table 9 illustrates the results.

Similarly to Section 3.1, three expert archaeologists (Fonte J., Dias R. and Pereiro T.) digitally validated some of the inferences. Based on the Atlas of Hillforts dataset, areas in northern, western, southwestern and southern England were selected for their significant clusters of hillforts. The analysis identified potential false negatives, where no hillfort in the Atlas of Hillforts dataset was visually identified in the LRM tiles or orthoimages, and potential false positives, which could represent previously unknown hillforts or other types of enclosures. Because 232 LRM tiles cover England, validating all of them would be a daunting task. Therefore, it was decided to validate the inferences within 34 tiles (≈15% of the total) as a compromise between human resources and the margin of error. This resulted in 1461 validated inferences, which is approximately 18% of the total inferences. Table 10 illustrates the results of the digital validation in England.

Table 11 presents the values of the precision, recall and F_1 score obtained when grouping the potential true positives and the true

TABLE 9 | Inferences in England.

Inferences	True positives	False positives	False negatives
8113	675	7438	578

TABLE 11 | Precision, recall and F_1 score of the digital validation in England.

Precision	Recall	F_1 score
0.45	0.84	0.59

TABLE 10 | Digital validation of inferences generated in England.

Region	Validated inferences	True positives	Potential true positives	False positives	False negatives
Northern England	353	108	95	150	170
Western England	276	57	63	156	30
Southwestern England	564	88	170	306	53
Southern England	268	25	56	187	14
Total	1461	278	384	799	267

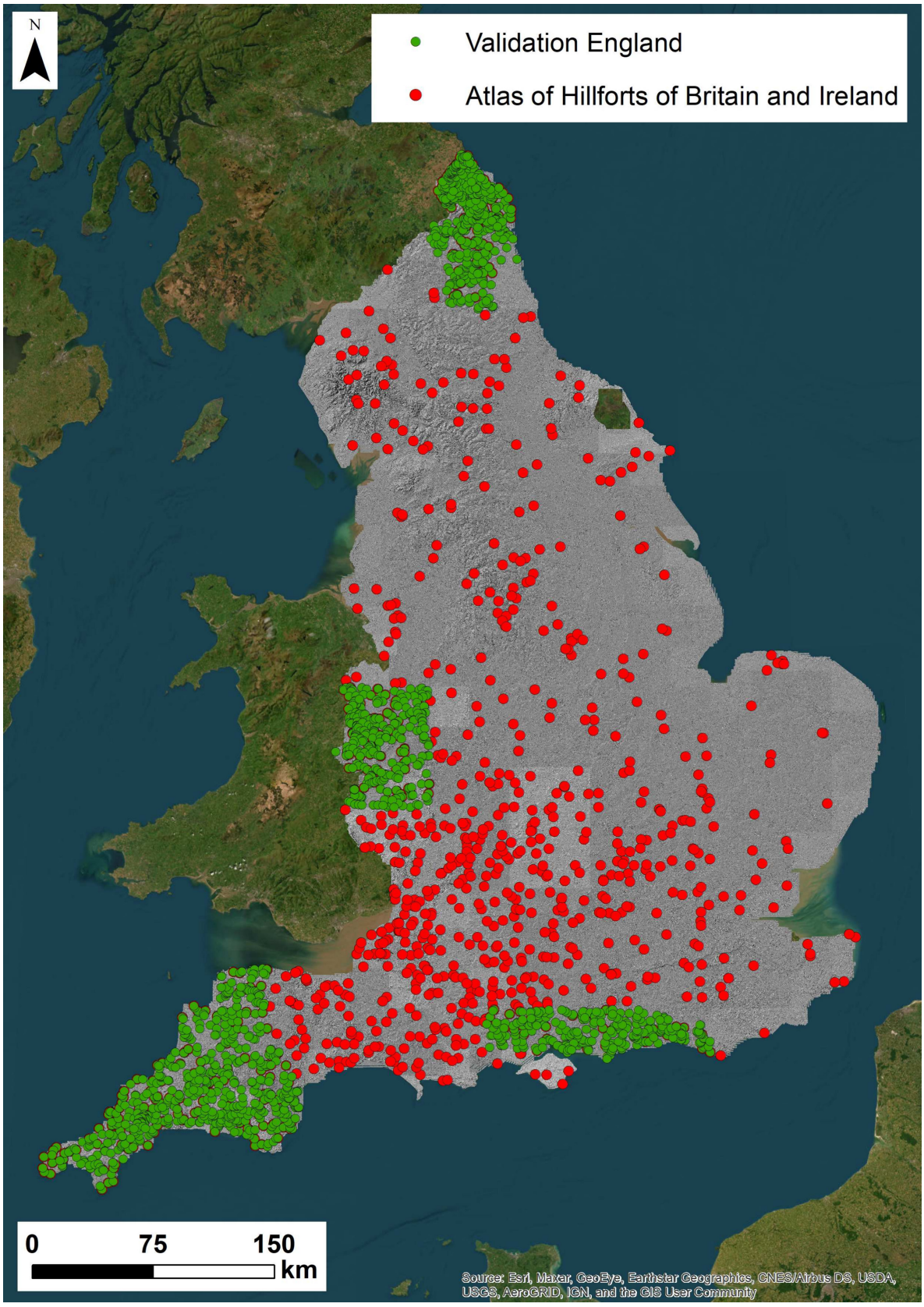

FIGURE 11 | Digital validation in England. [Colour figure can be viewed at wileyonlinelibrary.com]

positives classifications. It also accounts for the fact that only 126 out of the 267 false negatives were visible in both LRMs and aerial orthoimages. Figure 11 illustrates the inferences and regions in England that were validated.

To understand the factors contributing to the model's errors and identify potential strategies to improve its performance in the future, false positives were analysed and classified during the digital validation process. Table 12 shows this analysis.

Figure 12 illustrates some examples of the digital validation.

4 | Discussion

One of the main goals of this work was to reduce the frequent occurrence of false positives when applying machine learning models to remote sensing imagery to detect archaeological sites. Hillforts are characterized by their intricate and diverse shape, often mistaken for similar landscape features, either through human observation or artificial intelligence analysis. The initial hypothesis proposed that combining LiDAR-derived LRMs and aerial optical images through data fusion would improve the effectiveness of the machine learning model. LRMs delineate local microtopographical changes free of vegetation, thus enhancing the visual contrast of hillforts' positive features, such as walls and ramparts, as well as negative ones, such as ditches. However, aerial optical imagery captures what is perceptible to the human eye, providing context to these microtopographies within the real world. This fusion aims to emulate the approach of expert archaeologists during digital archaeological surveys, leveraging these complementary data types. Although this work revolved around these two data types, it is important to note that there are several other valid options, such as multispectral and hyperspectral aerial and satellite imagery (Altaweel et al. 2022; Bachagha et al. 2023; Casini et al. 2023; Chaidron and Lemernier 2023; Karamitrou et al. 2022; Orengo et al. 2020; Zimmer-Dauphinee, VanValkenburgh, and Wernke 2024).

TABLE 12 | Analysis of false positives during the digital validation.

False positives	Number of instances
Parcel	285
Water reservoir	94
Urban	91
Extraction	88
Geomorphology	60
Park	53
Hydrology	33
Vegetation	29
Agriculture	28
Other	24
Factory	10
Noise	4

The second hypothesis proposed the integration of archaeological expertise in the hard positive and negative mining process (Figure 1). The third and final hypothesis involved biasing the training data towards negative examples, as the machine learning model is more likely to encounter negative instances during the inference process. These two hypotheses should enrich the dataset with carefully analysed samples, leading the machine learning model to significantly reduce the number of false positives, hence improving its performance. This improvement is observed in Table 4, which compares the results of the computer vision system before and after hard positive and negative mining. After retraining the segmentation model with the dataset obtained from hard positive and negative mining, shown in Table 3, the number of false positives in PNPG was reduced from 5678 to 40, representing a reduction of 99.3% while keeping the true positives. This is a drastic improvement with only one iteration of hard positive and negative mining. The performance of the proposed computer vision system is expected to keep improving with the following iterations, as the dataset used to retrain the segmentation model becomes increasingly refined with new samples analysed by expert archaeologists.

Although the proposed solution in this work shows promising results, it is not without its challenges. The first evident challenge is associated with computational resources, particularly storage. Geospatial data tends to be heavy, often encompassing expansive regions. Multimodal approaches involve exploring various data types from the same regions, which can present storage challenges. Additionally, training deep learning models with multiple inputs requires substantial computational resources, mainly GPUs. The second challenge is associated with the required human resources and time investment of the proposed hard positive and negative mining. The archaeological analysis of images and corresponding inferences generated by the computer vision system has to be done manually. Although software was developed to speed up this process, the visual analysis of each image and corresponding inferences still needs to be done by an expert archaeologist. The third challenge is associated with model generalization and adaptability, which can be observed in Section 3.2. This section discusses the second experiment of this work, where the computer vision system trained with Iberian data was applied in England. Given England's distinct hillfort typology and landscape from Northwest Iberia, the computer vision system encountered difficulties in performing as effectively as it did in PNPG. These difficulties can be observed in Table 7, which presents the archaeological analysis of images and corresponding inferences generated by the computer vision system trained with Iberian data on four LRM tiles from England. A small dataset resulted from this analysis, which was then utilized to fine-tune the segmentation model. Such fine-tuning is essential when applying machine learning models in different site typologies and landscapes from those present in the original dataset. Although retraining the segmentation model with a rich dataset composed solely of data from England would likely yield even better results than the ones presented in Tables 10 and 11, fine-tuning the model with a small dataset is a practical and efficient compromise. It efficiently

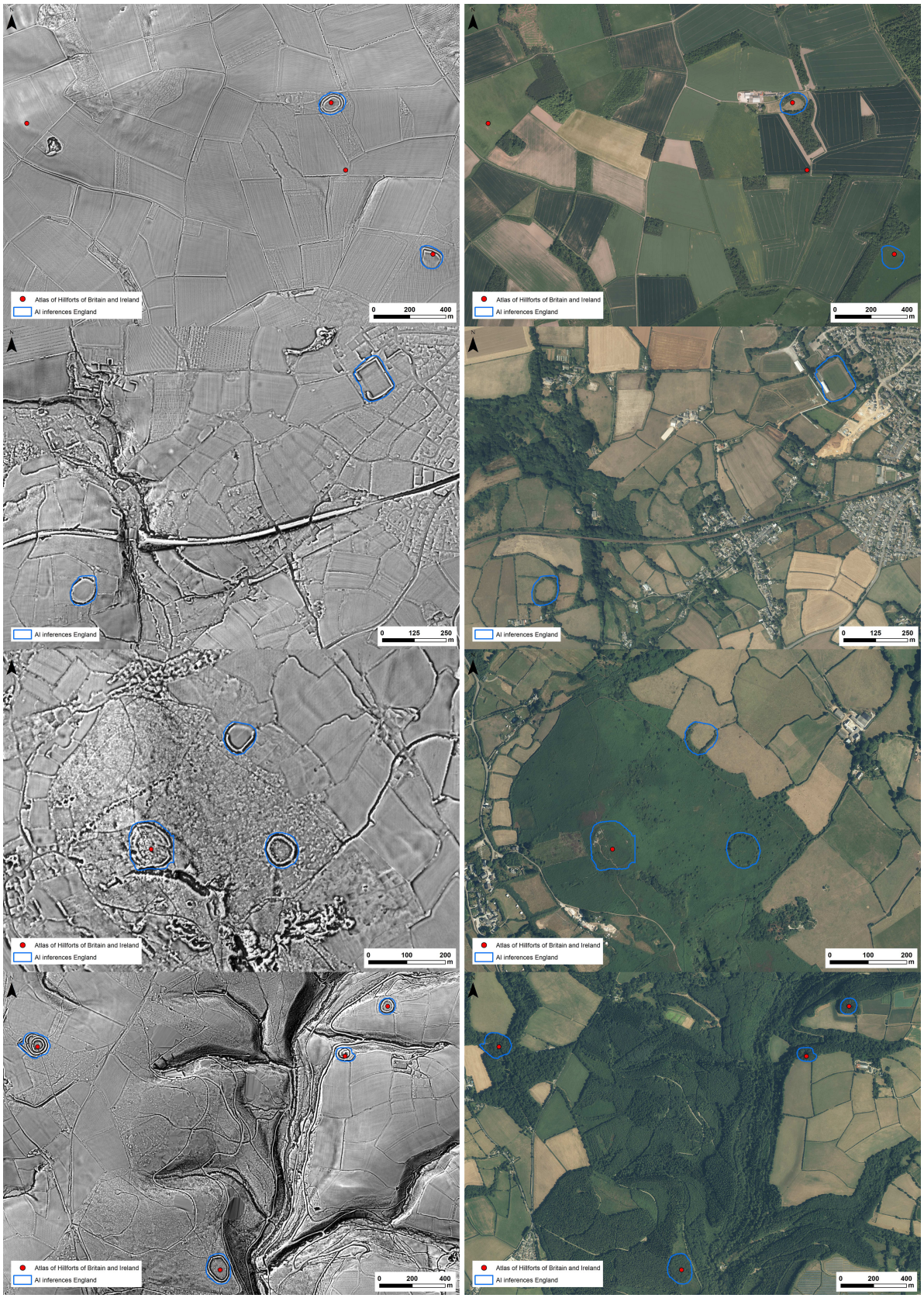

FIGURE 12 | Digital validation in England. The red points represent the Atlas of Hillforts of Britain and Ireland dataset, and the blue polygons represent the generated inferences. The first line illustrates two hillforts that were not detected out of four, as they were not visible in the LRM or the orthoimage. The second line illustrates a potential true positive on the left and a false positive on the right. The third line illustrates a true positive and two potential true positives. The fourth line illustrates four true positives. [Colour figure can be viewed at wileyonlinelibrary.com]

balances human resources and time investment. However, this still implies that the segmentation model requires fine-tuning every time the proposed computer vision system needs to be deployed in regions with different site typologies and landscapes. The fourth and final challenge is associated with processing time. As mentioned in Section 3.2, processing England with the proposed computer vision system took approximately 30 days. Sweeping through LRM tiles across England with a sliding window, using a step size of 1/2, and making a WMS request for the orthoimage for each window, is time-consuming. Implementing a caching mechanism and focusing the proposed computer vision system on regions with higher archaeological potential could significantly reduce processing time. Archaeological predictive modelling could be explored as a means to identify these regions, optimizing the deployment of the computer vision system (Podgórski et al. 2021). Additionally, Table 12 provides useful insights on what is inducing the computer vision system in error when searching for hillforts. Based on this information, the hard positive and negative mining should prioritize introducing negative examples in the dataset containing these types of false positives. This should gradually improve the computer vision system's ability to distinguish between hillforts and these false positives.

5 | Conclusion

This work introduces a computer vision system designed to effectively detect hillforts in remote sensing imagery. The system adapts the multimodal semantic segmentation model proposed by Zhang et al. (2023). This model requires two modalities, a LiDAR-derived LRM image and an aerial orthoimage, and performs feature fusion to generate a binary mask highlighting the detected hillforts. The binary mask undergoes post-processing, which includes margin and area filters. The margin filter removes inferences at the edges of the image, typically consisting of partially visible objects and probable false positives. The area filter removes inferences smaller than the average hillfort size, which are often false positives. Inferences that survive the filtering process are considered. These inferences can then be used for hard positive and negative mining, where an expert archaeologist digitally analyzes them and classifies them as true positives or false positives. Based on these classifications, the original dataset can be enriched with more positive and negative samples. Given that the computer vision system is more likely to encounter negative samples during the deployment phase, the training data are intentionally biased towards negative samples.

After only one iteration of hard positive and negative mining, the results of the Northwest Iberia experiment were drastically improved, reducing the false positives encountered by the original model from 5678 to 40, representing a reduction of 99.3%. After the digital validation conducted by expert archaeologists, the achieved F_1 score for PNPG was 66%. For the next experiment, the computer vision system was tested in England, which has a different typology and landscape of sites. The segmentation model was fine-tuned with a small dataset built from four LRM tiles from England. Then, the computer vision system was applied to the entire country of England to search for hillforts. The achieved F_1 score, based on the Atlas of Hillforts of Britain and Ireland dataset (Lock and Ralston 2022) and the digital validation by expert archaeologists, was 59%.

The proposed computer vision system is believed to be effectively expanded to other archaeological sites, which has major implications for archaeological research and cultural heritage management. This assumption is supported by the demonstrated effectiveness in detecting hillforts and different types of enclosure, which often present challenging and complex shapes. However, certain shortcomings of the proposed computer vision system were discussed, mainly related to human resources, computational resources and processing time. For future work, exploring archaeological predictive modelling to identify regions with higher archaeological potential could enable the search to focus within those areas. This focus should alleviate the shortcomings mentioned above. Additionally, although fusion techniques generally lead to better results, it is important to conduct an ablation study to numerically verify the contribution of using aerial orthoimagery alongside LRMs. However, due to the extensive scope of the datasets and the project's time constraints, such a study was not possible in this work. Future research should also address additional challenges, including the effect of cropmark seasonality on detection accuracy, variations in the quality and colour distribution of aerial orthoimages from different regions, and the potential use of other data types.

Acknowledgements

The authors are grateful to César Parcero Oubiña and Jesús García Sánchez for their insightful comments on a draft version of this paper. We would like to thank the Comunidade Intermunicipal do Alto Minho (CIM Alto Minho) for providing the airborne LiDAR data for the Alto Minho region. We also acknowledge the Instituto Geográfico Nacional from Spain and the Environment Agency from England for making their geographic information openly available, which has facilitated our research and the generation of new knowledge. For the purpose of open access, the author has applied a Creative Commons Attribution (CC BY) licence to any Author Accepted Manuscript version arising from this submission.

Conflicts of Interest

The authors declare no conflicts of interest.

Data Availability Statement

Research data are not shared.

References

Altaweel, M., A. Khelifi, Z. Li, A. Squitieri, T. Basmaji, and M. Ghazal. 2022. "Automated Archaeological Feature Detection Using Deep Learning on Optical UAV Imagery: Preliminary Results." *Remote Sensing* 14, no. 3: 553. https://doi.org/10.3390/rs14030553.

Argyrou, A., and A. Agapiou. 2022. "A Review of Artificial Intelligence and Remote Sensing for Archaeological Research." *Remote Sensing* 14, no. 23: 6000. https://doi.org/10.3390/rs14236000.

Bachagha, N., A. Elnashar, M. Tababi, F. Souei, and W. Xu. 2023. "The Use of Machine Learning and Satellite Imagery to Detect Roman Fortified Sites: The Case Study of Blad Talh (Tunisia Section)." *Applied Sciences* 13, no. 4: 2613. https://doi.org/10.3390/app13042613.

Bennett, R., K. Welham, R. A. Hill, and A. Ford. 2012. "A Comparison of Visualization Techniques for Models Created From Airborne Laser Scanned Data." *Archaeological Prospection* 19, no. 1: 41–48. https://doi.org/10.1002/arp.1414.

Berganzo-Besga, I., H. A. Orengo, F. Lumbreras, M. Carrero-Pazos, J. Fonte, and B. Vilas-Estévez. 2021. "Hybrid MSRM-Based Deep Learning and Multitemporal Sentinel 2-Based Machine Learning Algorithm Detects Near 10k Archaeological Tumuli in North-Western Iberia." *Remote Sensing* 13, no. 20: 4181. https://doi.org/10.3390/rs13204181.

Bluesky International Limited and Getmapping Plc. 2023. WMS APGB Aerial Imagery 25cm resolution. Accessed April 25, 2024. https://www.arcgis.com/home/item.html?id=4d9d5e1b0e044c7b9a26f07d5f63bd48.

Câmara, A., A. de Almeida, D. Caçador, and J. Oliveira. 2023. "Automated Methods for Image Detection of Cultural Heritage: Overviews and Perspectives." *Archaeological Prospection* 30, no. 2: 153–169. https://doi.org/10.1002/arp.1883.

Canedo, D., J. Fonte, L. G. Seco, et al. 2023. "Uncovering Archaeological Sites in Airborne Lidar Data With Data-Centric Artificial Intelligence." *IEEE Access* 11: 65608–65619. https://doi.org/10.1109/ACCESS.2023.3290305.

Casini, L., N. Marchetti, A. Montanucci, V. Orrù, and M. Roccetti. 2023. "A Human–AI Collaboration Workflow for Archaeological Sites Detection." *Scientific Reports* 13, no. 1: 8699. https://doi.org/10.1038/s41598-023-36015-5.

Cerrillo-Cuenca, E. 2017. "An Approach to the Automatic Surveying of Prehistoric Barrows Through Lidar." *Quaternary International* 435: 135–145. https://doi.org/10.1016/j.quaint.2015.12.099.

Chaidron, C., and S. Lemernier. 2023. "Multispectral Imaging and Artificial Intelligence for Archaeology: First Results and First Projects." *SCIRES-IT-SCIentific RESearch and Information Technology* 12, no. 2: 87–98. https://doi.org/10.2423/i22394303v12n2p87.

DONP. 2021. "Digital Orthophotos of North Portugal." Accessed April 25, 2024. https://www.dgterritorio.gov.pt/Ortofotomapas-2021.

Davis, D. 2021. "Theoretical Repositioning of Automated Remote Sensing Archaeology: Shifting From Features to Ephemeral Landscapes." *Journal of Computer Applications in Archaeology* 4, no. 1: 94–109. https://doi.org/10.5334/jcaa.72.

Davis, D. S. 2020. "Defining What We Study: The Contribution of Machine Automation in Archaeological Research." *Digital Applications in Archaeology and Cultural Heritage* 18: e00152. https://doi.org/10.1016/j.daach.2020.e00152.

Davis, D. S., G. Caspari, C. P. Lipo, and M. C. Sanger. 2021. "Deep Learning Reveals Extent of Archaic Native American Shell-Ring Building Practices." *Journal of Archaeological Science* 132: 105433. https://doi.org/10.1016/j.jas.2021.105433.

Deng, J., W. Dong, R. Socher, L. -J. Li, K. Li, and L. Fei-Fei. 2009. "Imagenet: A Large-Scale Hierarchical Image Database." In *2009 IEEE Conference on Computer Vision and Pattern Recognition*, 248–255. https://doi.org/10.1109/CVPR.2009.5206848.

Dosovitskiy, A., L. Beyer, A. Kolesnikov, et al. 2020. "An Image is Worth 16x16 Words: Transformers for Image Recognition at Scale." arXiv preprint arXiv:2010.11929. https://doi.org/10.48550/arXiv.2010.11929.

Fiorucci, M., M. Khoroshiltseva, M. Pontil, A. Traviglia, A. Del Bue, and S. James. 2020. "Machine Learning for Cultural Heritage: A Survey." *Pattern Recognition Letters* 133: 102–108. https://doi.org/10.1016/j.patrec.2020.02.017.

Fiorucci, M., W. B. Verschoof-Van Der Vaart, P. Soleni, B. Le Saux, and A. Traviglia. 2022. "Deep Learning for Archaeological Object Detection on Lidar: New Evaluation Measures and Insights." *Remote Sensing* 14, no. 7: 1694. https://doi.org/10.3390/rs14071694.

Fontana, G. 2022. "Italy's Hidden Hillforts: A Large-Scale Lidar-Based Mapping of Samnium." *Journal of Field Archaeology* 47, no. 4: 245–261. https://doi.org/10.1080/00934690.2022.2031465.

Guyot, A., M. Lennon, and L. Hubert-Moy. 2021. "Objective Comparison of Relief Visualization Techniques With Deep CNN for Archaeology." *Journal of Archaeological Science: Reports* 38: 103027. https://doi.org/10.1016/j.jasrep.2021.103027.

Guyot, A., M. Lennon, T. Lorho, and L. Hubert-Moy. 2021. "Combined Detection and Segmentation of Archeological Structures From Lidar Data Using a Deep Learning Approach." *Journal of Computer Applications in Archaeology* 4, no. 1: 1. https://doi.org/10.5334/jcaa.64.

Hesse, R. 2010. "Lidar-Derived Local Relief Models–A New Tool for Archaeological Prospection." *Archaeological prospection* 17, no. 2: 67–72. https://doi.org/10.1002/arp.374.

Karamitrou, A., F. Sturt, P. Bogiatzis, and D. Beresford-Jones. 2022. "Towards the Use of Artificial Intelligence Deep Learning Networks for Detection of Archaeological Sites." *Surface Topography: Metrology and Properties* 10, no. 4: 44001. https://doi.org/10.1088/2051-672X/ac9492.

Kazimi, B., and M. Sester. 2023. "Self-Supervised Learning for Semantic Segmentation of Archaeological Monuments in DTMs." *Journal of Computer Applications in Archaeology 6 (2023), Nr. 1* 6, no. 1: 155–173. https://doi.org/10.5334/jcaa.110.

Kokalj, Z., and M. Somrak. 2019. "Why Not a Single Image? Combining Visualizations to Facilitate Fieldwork and On-Screen Mapping." *Remote Sensing* 11, no. 7: 747. https://doi.org/10.3390/rs11070747.

Landauer, J. "Hillfort Finder App." Accessed May 10, 2024. https://juergenlandauer-hillfortfinderapp-hillfortfinder-s7m8nz.streamlit.app/.

Landauer, J., S. Maddison, G. Fontana, and A. Posluschny. 2023. "Site Detection Across Europe and the HillfortFinderApp–Latest Results From a Deep Learning Based Hillfort Search." https://doi.org/10.5281/zenodo.7997347.

Landauer, J., and W. Verschoof-van der Vaart. 2021. "Find 'em All: Large-Scale Automation to Detect Complex Archaeological Objects With Deep Learning - A Case Study on English Hillforts." In *Proceedings CAA2021, LIMASSOL 2021*.

Liu, H. "RGBX Semantic Segmentation." Accessed May 10, 2024. https://github.com/huaaaliu/RGBX_Semantic_Segmentation.

Lock, G., and I. Ralston. 2022. *Atlas of the Hillforts of Britain and Ireland*. Edinburgh: Edinburgh University Press. Accessed 10 May 2024. https://hillforts.arch.ox.ac.uk/.

Loshchilov, I., and F. Hutter. 2017. "Decoupled Weight Decay Regularization." arXiv preprint arXiv:1711.05101. https://doi.org/10.48550/arXiv.1711.05101.

Murrugarra-Llerena, J., L. N. Kirsten, and C. R. Jung. 2022. "Can We Trust Bounding Box Annotations for Object Detection?." In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 4813–4822. https://doi.org/10.1109/CVPRW56347.2022.00528.

Olivier, M., and W. Verschoof-van der Vaart. 2021. "Implementing State-Of-The-Art Deep Learning Approaches for Archaeological Object Detection in Remotely-Sensed Data: The Results of Cross-Domain Collaboration." *Journal of Computer Applications in Archaeology* 4, no. 1: 274–289. https://doi.org/10.5334/jcaa.78.

Orengo, H. A., F. C. Conesa, A. Garcia-Molsosa, et al. 2020. "Automated Detection of Archaeological Mounds Using Machine-Learning Classification of Multisensor and Multitemporal Satellite data." *Proceedings of the National Academy of Sciences* 117, no. 31: 18240–18250. https://doi.org/10.1073/pnas.2005583117.

PNOA. 2020. "Plan Nacional de Ortofotografía Aérea." Accessed April 25, 2024. https://pnoa.ign.es/.

Podgórski, Z., D. Szatten, M. Brzezińska, and M. Maerker. 2021. "Spatial Analysis of Hillfort Locations in the Chełmno Land (Poland) Using

Digital Terrain Analysis and Stochastic Data Exploration." *Journal of Archaeological Science: Reports* 39: 103170. https://doi.org/10.1016/j.jasrep.2021.103170.

Schneider, A., M. Takla, A. Nicolay, A. Raab, and T. Raab. 2015. "A Template-Matching Approach Combining Morphometric Variables for Automated Mapping of Charcoal Kiln Sites." *Archaeological Prospection* 22, no. 1: 45–62. https://doi.org/10.1002/arp.1497.

Štular, B., Ž. Kokalj, K. Oštir, and L. Nuninger. 2012. "Visualization of Lidar-Derived Relief Models for Detection of Archaeological Features." *Journal of Archaeological Science* 39, no. 11: 3354–3360. https://doi.org/10.1016/j.jas.2012.05.029.

Štular, B., E. Lozić, and S. Eichert. 2021. "Airborne Lidar-Derived Digital Elevation Model for Archaeology." *Remote Sensing* 13, no. 9: 1855. https://doi.org/10.3390/rs13091855.

Tang, T., S. Zhou, Z. Deng, H. Zou, and L. Lei. 2017. "Vehicle Detection in Aerial Images Based on Region Convolutional Neural Networks and Hard Negative Example Mining." *Sensors* 17, no. 2: 336. https://doi.org/10.3390/s17020336.

Tziafas, G., and H. Kasaei. 2023. "Early or Late Fusion Matters: Efficient RGB-D Fusion in Vision Transformers for 3D Object Recognition." In *2023 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, 9558–9565. https://doi.org/10.1109/IROS55552.2023.10341422.

Verschoof-van der Vaart, W. B., and K. Lambers. 2022. "Applying Automated Object Detection in Archaeological Practice: A Case Study From the Southern Netherlands." *Archaeological Prospection* 29, no. 1: 15–31. https://doi.org/10.1002/arp.1833.

Vinci, G., F. Vanzani, A. Fontana, and S. Campana. 2024. "Lidar Applications in Archaeology: A Systematic Review." *Archaeological Prospection* 1–21. https://doi.org/10.1002/arp.1931.

Xie, E., W. Wang, Z. Yu, A. Anandkumar, J. M. Alvarez, and P. Luo. 2021. "Segformer: Simple and Efficient Design for Semantic Segmentation With Transformers." *Advances in Neural Information Processing Systems* 34: 12077–12090.

Zhang, J., H. Liu, K. Yang, X. Hu, R. Liu, and R. Stiefelhagen. 2023. "CMX: Cross-Modal Fusion for RGB-X Semantic Segmentation With Transformers." *IEEE Transactions on Intelligent Transportation Systems* 24, no. 12: 14679–14694. https://doi.org/10.1109/TITS.2023.3300537.

Zimmer-Dauphinee, J., P. VanValkenburgh, and S. A. Wernke. 2024. "Eyes of the Machine: AI-Assisted Satellite Archaeological Survey in the ANDES." *Antiquity* 98, no. 397: 245–259. https://doi.org/10.15184/aqy.2023.175.