

Article

Deep Learning Enables the Automatic Mapping of *Tell* Sites on Satellite Synthetic Aperture Radar Products

Elena Chiricallo ^{1,2} , Giulio Poggi ¹ , Sara Ferro ¹ , Sebastiano Vascon ^{2,†}  and Arianna Traviglia ^{1,*,†} 

¹ Center for Cultural Heritage Technology, Istituto Italiano di Tecnologia, 31056 Treviso, Italy; elena.chiricallo@iit.it (E.C.); giulio.poggi@iit.it (G.P.); sara.ferro@iit.it (S.F.)

² Department of Environmental Sciences, Informatics and Statistics, Università Ca' Foscari, 30171 Venezia, Italy; sebastiano.vascon@unive.it

* Correspondence: arianna.traviglia@iit.it

† These authors contributed equally to this work.

Highlights

What are the main findings?

- A Deep Learning pipeline using Sentinel-1 and COP-30 DEM data can map archaeological mounds, i.e., *tell* sites, in Iraq;
- The pipeline demonstrates promising transferability to a new region in Iran via minimal fine-tuning, while an explainability analysis provides archaeologists with insights into the model's decision logic.

What are the implications of the main findings?

- The pipeline offers a scalable approach for large-scale *tell* site mapping across Near and Middle East landscapes;
- It highlights the potential of SAR products as a high-value data source for automated archaeological prospection.

Abstract

Satellite Synthetic Aperture Radar (SAR) is an established technology for studying and monitoring archaeological landscapes, providing insights into surface morphology and the presence of near subsurface features. However, its application in large-scale archaeological prospection is limited by the lack of robust, automated methods for SAR data analysis. This study introduces a novel Deep Learning pipeline to automatically detect and segment archaeological settlement mounds, known as *tells*, in central Iraq on satellite SAR data. The pipeline leverages a state-of-the-art supervised method for instance segmentation, YOLOv8-Seg, and medium-resolution satellite SAR products, specifically the Copernicus Sentinel-1 Interferometric Wide Swath Mode Ground Range Detected and Copernicus Global 30-m Digital Elevation Model products. The model identifies *tell* sites with an Average Precision of 0.495 ± 0.010 and a pixel-wise Intersection over Union of 0.361 ± 0.048 over the test areas. Archaeological interpretation of the model's inferences confirms its reliability in locating and segmenting archaeological sites, leading also to the identification of previously unmapped potential sites. After a main test in central Iraq, the proposed workflow demonstrates promising transferability to a nearby test area in Iran, although with a need for regional fine-tuning to account for inherent variations in feature morphology and environmental context. This research establishes a baseline for future Deep Learning applications in Synthetic Aperture Radar-based archaeological prospection.



Academic Editor: John Trinder

Received: 30 April 2026

Revised: 30 June 2026

Accepted: 1 July 2026

Published: 7 July 2026

Copyright: © 2026 by the authors.

Licensee MDPI, Basel, Switzerland.

This article is an open access article

distributed under the terms and

conditions of the [Creative Commons](#)

[Attribution \(CC BY\)](#) license.

Keywords: deep learning; synthetic aperture radar; instance segmentation; *tell* sites; digital elevation model; Sentinel-1; COP-30 DEM; Iraq

1. Introduction

This study presents a novel Deep Learning pipeline for the detection and segmentation of archaeological settlement mounds (*tells*) in satellite Synthetic Aperture Radar products. Satellite Synthetic Aperture Radar (SAR) imaging systems have long been used for studying and monitoring archaeological landscapes and cultural heritage sites since the earliest civilian space radar missions in the 1980s [1–17]. They operate through active microwave remote sensing in a side-looking geometry and can provide high-resolution, day-and-night, and almost weather-independent images of the Earth’s surface on a global scale, complementing optical and LiDAR remote sensing data [18–23]. SAR transmits microwave pulses and listens for echoes, called backscatters, providing information about the geometric properties (shape, orientation, roughness) and the moisture content of the target. The chosen radio band and polarisation further impact what is observed from the scene, enabling the SAR signal to achieve a certain level of penetration through dry soil or vegetation cover and the sensitivity to differentiated scattering mechanisms. Moreover, SAR can be used for deriving the surface topography (InSAR Digital Elevation Model (DEM) products) and measuring the Earth’s surface displacements with millimetric precision (DInSAR products) through interferometric processing [18–21,24]. Among the various SAR products, anomalies in backscatter amplitude and topographic information derived from InSAR DEM data have been successfully used to identify archaeological features: these include shallow subsurface archaeological structures (e.g., earthworks, buildings, ancient roads, etc.), palaeochannels, archaeological micro-reliefs and archaeological mounds [1–13].

Although SAR data have long been recognised as a valuable source of information for archaeological mapping, their use in automated analyses remains overlooked. Most existing studies rely on manual visual interpretation or on pixel-based machine learning approaches, in which SAR is often treated as an ancillary variable [25]. Specifically, pixel-based approaches, which depend on the radiometric or elevation values of individual pixels, tend to struggle with inherent signal noise and geometric distortions when applied to amplitude images or InSAR DEMs, and overlook the spatial (textural and geometrical) structure and contextual information of archaeological features [26,27]. Recent works in Deep Learning for SAR have shown that automatic methods based on it outperform pixel-based or feature-engineered approaches across several applications such as ship detection, terrain surface classification, ice concentration estimation, and flood segmentation by learning hierarchical representations and capturing object-level patterns directly from the data [26–28]. Similar evidence has emerged in the LiDAR domain in archaeology, where several studies demonstrated that Deep Learning substantially improves the detection and segmentation of microtopographic structures [29–31]. These results indicate that Deep Learning offers a viable pathway to use SAR data as a primary source for large-scale archaeological detection. At the same time, the increasing number of missions and the expanding volume of SAR acquisitions make manual inspection unfeasible, leaving much of this information underused across extensive regions and longtime spans. By enabling large-scale inference from massive satellite acquisition, Deep Learning can address the dual challenge of SAR data complexity and volume.

This study presents a novel Deep Learning pipeline for the automatic detection and segmentation of *tell* sites in satellite SAR data. The proposed method is developed and tested in central Iraq, where thousands of *tells* have already been documented. In this

region, *tell* sites have been identified through visual inspection of satellite SAR amplitude imagery and interferometric DEMs, exploiting their characteristic soil appearance and coverage, morphology, and topographic elevation [10,11]. These approaches have proven particularly effective in remote and hazardous regions, where field surveys are impractical due to political instability and ongoing conflicts, and where LiDAR data from airborne or drone platforms are unavailable or restricted. However, manual interpretation proved to be time-consuming when analysing large areas containing a large number of scattered features throughout the landscape. At the same time, recent studies have demonstrated the potential of Deep Learning techniques for *tell* site detection using high-resolution satellite optical imagery [32–34]. However, the limited availability of free of charge archive images, the high costs of extensive new acquisitions, and security-related restrictions on sensitive areas are obstacles to their large-scale applicability.

In contrast, this study distinguishes itself by focusing exclusively on European Space Agency Copernicus SAR products, which are openly available and acquired with near-global coverage, enabling a reproducible and scalable framework for *tell* detection across different geographic locations. Building on this approach, the specific objectives of this work are (i) to provide a new Deep Learning tool for rapid and accurate large-scale *tell* site mapping using global coverage, open-access, medium-resolution SAR data; (ii) to develop a transferable workflow applicable to other study areas in the Near and Middle East with similar archaeological landscapes; and (iii) to establish a baseline for future Deep Learning research utilising SAR data in archaeological contexts. Finally, this work also marks one of the first applications of Explainable AI to archaeological remote sensing, addressing the “black box” nature of Deep Learning and its implications for the field.

2. Materials and Methods

2.1. Study Area

Tells are artificial hills peculiar to Near and Middle Eastern landscapes, formed from the accumulated and stratified remains of ancient settlements from the Neolithic (6th millennium BCE) through the Iron Age (10th to 6th century BCE in Iraq), with some sites showing later reuse [35–37]. Their systematic mapping can provide information about the emergence, development, and organisation of early complex human societies. Moreover, accurately identifying their locations and making this information available to relevant authorities represents a fundamental step toward their preservation, helping to protect them from looting and from large-scale landscape transformation driven by destructive anthropogenic activities, such as agriculture, industrial development, and transportation infrastructure [11,37–39].

Tells can be identified in SAR data by their elliptical shapes and distinct land-use, which create backscatter amplitude patterns that stand out against the agricultural background. In addition, their topographic prominence above the mean plain level creates signatures that can be detected in InSAR DEMs [10,11,17].

Building on these detectable morphological and geophysical characteristics, the analysis was conducted over a study area covering approximately 57,700 km² within the southern Mesopotamian floodplain, in Iraq, where the archaeological evidence is marked by the presence of thousands of tell sites, documented in extensive archaeological surveys conducted since the 1950s [37]. Here, Holocene deposits preserve an intricate network of fluvial ridges and relict distributary channels, indicating repeated episodes of avulsion and progradation [40]. Human occupation and irrigation practices since the mid-Holocene have reshaped channel courses and supported long-term settlement on elevated alluvial grounds [41]. Today, the region is characterised by a hot arid climate, with a mean annual precipitation ranging from 100 to 180 mm, occurring primarily during the humid season

between November and April [42]. The key landscape descriptors, including the median slope and the land cover, are presented in Table 1a and were derived from remote sensing products detailed in Section 2.2.

Table 1. Landscape (a) and tell sites (b) descriptors for the training, validation, and test areas. For specific descriptors, values in parentheses represent the median and interquartile range (IQR) to illustrate data distribution.

(a)				
Landscape Descriptor	Iraq			Iran
	Training	Validation	Test	Test
Extension (km ²)	37,009	15,850	4846	1470
Elev. Range (m AMSL) *	−14, +342	−7, +54	−2, +26	+32, +197
Slope (degrees) *, median IQR	0.41 (0.26–0.75)	0.34 (0.22–0.56)	0.36 (0.25–0.52)	0.52 (0.30–1.09)
Land Cover **				
Bare/Sparse Vegetation	35.55%	43.73%	32.11%	5.76%
Cropland	45.38%	30.58%	59.78%	70.94%
Shrubland and Grassland	7.91%	4.48%	5.08%	9.61%
Tree cover	5.70%	3.19%	0.63%	8.44%
Built-up	3.78%	2.82%	0.65%	3.96%
(b)				
Tell Site Descriptor	Iraq			Iran
	Training	Validation	Test	Test
Number of Sites	2634	825	290	110
Size (ha), median IQR	2.00 (1.01–4.34)	2.68 (1.35–5.52)	2.93 (1.81–5.60)	1.12 (0.59–2.10)
Length (m), median IQR	198 (134–301)	221 (157–341)	248 (178–366)	140 (104–211)
Elevation (m) †, median IQR	2.17 (1.36–3.34)	2.00 (1.35–3.22)	2.80 (1.99–3.97)	4.27 (2.50–7.44)

* Derived from the COP-30 DEM. Elevations are expressed in meters above mean sea level (AMSL). ** Derived from the ESA WorldCover 2021 dataset [43]; most representative classes. † Relative elevation calculated as max elevation within the site minus 150 m buffer average.

As part of the study, a geographically distinct test area was also included in Southwestern Iran, within Khuzestan Province. The region consists mostly of extensive alluvial plains that underwent phases of aggradation, presumably between the ninth and second millennia BCE, followed by a shift in the fluvial regime leading to the incision of meandering river valleys [44,45]. The present-day landscape is characterised by small agricultural plots occupying vast, slightly elevated surfaces, which possibly represent remnants of former terraces, standing approximately 20 m above the incised valley floors. Numerous archaeological surveys and excavations in the Khuzestan province have documented mounded settlements with occupation sequences spanning from the Neolithic to the Bronze Age, with the earliest phases dated to the late eighth millennium BCE [44]. The selection of a geographically distinct case study serves to validate the generalisation performance of the Deep Learning pipeline, specifically testing for robustness to domain-shift and the transferability of the results to a new prospection context.

2.2. Remote Sensing Data

The Deep Learning model presented here was trained using data from the Copernicus Global 30-m DEM and Sentinel-1 SAR imagery.

The Copernicus Global 30-m DEM (COP-30 DEM) [46] is derived from the interferometric TanDEM-X mission between 2011 and 2015. It provides a 30-m spatial resolution and an absolute vertical accuracy of approximately 2 m within the study area, outperforming previous global DEMs [47]. As an X-band radar-derived digital surface model, it includes both natural terrain and anthropogenic features such as buildings, infrastructure, and vegetation canopy, thereby providing essential topographic information for identifying *tell* sites based on their relative elevation.

Sentinel-1 (S-1), operational since 2014 with the S-1A and S-1C satellites (following the earlier S-1B mission), delivers systematic C-band SAR data with near-global coverage. This study uses dual-polarisation (VV + VH) Interferometric Wide (IW) Swath Mode Ground Range Detected (GRD) products, which consist of geocoded backscatter amplitude images with 10-m pixel spacing. These data are particularly sensitive to the surface roughness and moisture conditions, making them well-suited for detecting backscatter anomalies associated with *tell* sites. The C-band wavelength interacts with the landscape in a manner comparable to the X-band, with slightly deeper penetration in dry soils and sparse vegetation. The use of dual-polarisation further enhances interpretability by capturing different scattering mechanisms: Vertical–Vertical (VV) polarisation is primarily sensitive to surface and double-bounce scattering (e.g., bare soils or urban structures) while Vertical–Horizontal (VH) polarisation is more responsive to volume scattering (e.g., vegetation). This distinction is particularly useful for differentiating archaeological mounds from visually similar features such as modern urban clusters or densely vegetated areas. Moreover, the side-looking geometry of SAR acquisition can accentuate subtle topographic variations through effects such as shadowing and dihedral scattering (see Supplementary Materials Section S1).

In addition to SAR and DEM data, high-resolution optical imagery from Esri World Imagery was used for visual verification of annotations, while medium-resolution optical imagery from Sentinel-2 (S-2) supported control experiments. Finally, the ESA WorldCover 2021 dataset [43] was employed to provide environmental context for the archaeological sites (see Table 1).

2.3. Dataset Labels

Tell sites were manually delineated as polygons within a GIS environment. The annotation process was carried out using a composite of the S-1 and the COP-30 DEM data, supplemented by high-resolution optical RGB imagery from ArcGIS Pro (version 3.0.3) [48]. The annotations were validated using two established archaeological reference datasets covering the study area: the Ancient Near East (ANE) dataset [49,50], containing more than 3000 placemarks of potential archaeological sites, which was imported directly into the GIS environment; and the FloodPlains WebGIS [51], a public archaeological platform encompassing sites across the Mesopotamian alluvial plain. This latter dataset comprises almost 8000 entries, “about 4900 of which are archaeological sites documented by different survey projects, while the other 2900 are potential archaeological sites identified through remote sensing”. The FloodPlains WebGIS was consulted externally through its online platform to support cross-validation of the annotated dataset.

Each annotated polygon was enriched with specific attributes, including visibility on SAR data and optical RGB imagery, site statistics (e.g., size and elevation), visibility of archaeological traces (e.g., structures, evidence of looting), and the data validation origin. To ensure consistency and completeness, the labelling process involved cross-checking and revision by more than one operator. Potential *tell* sites identified during the process

but not present in the reference datasets were integrated into the dataset based on expert consensus, considering morphology, size, and landscape context. This reflects a necessary compromise when working at regional scales, where exhaustive ground truthing is not feasible or possible. The resulting preliminary dataset was then filtered for Deep Learning application, yielding a final, refined dataset of 3866 annotated features across the study area (see Section 2.4.3).

2.4. Description of the Deep Learning Pipeline

2.4.1. General Overview of the Pipeline

The proposed Deep Learning pipeline (Figure 1) performs instance segmentation of archaeological *tell* sites using false-colour RGB composites generated from S-1 and the COP-30 DEM. For this purpose, a state-of-the-art supervised instance segmentation model, YOLOv8-Seg [52], was employed. This model is designed to predict bounding boxes, segmentation masks, and class probabilities for objects within an image. In its original implementation, YOLOv8-Seg operates on RGB images and follows a modular architecture composed of a backbone, a neck, and three decoupled multi-scale heads. The backbone is responsible for extracting hierarchical features from the input data, while the neck aggregates and refines these features across multiple spatial scales. The network then branches into the prediction heads, which, at each scale, separate the instance segmentation process into three parallel tasks: localisation, classification, and mask generation (the latter inspired by the YOLACT approach [53]).

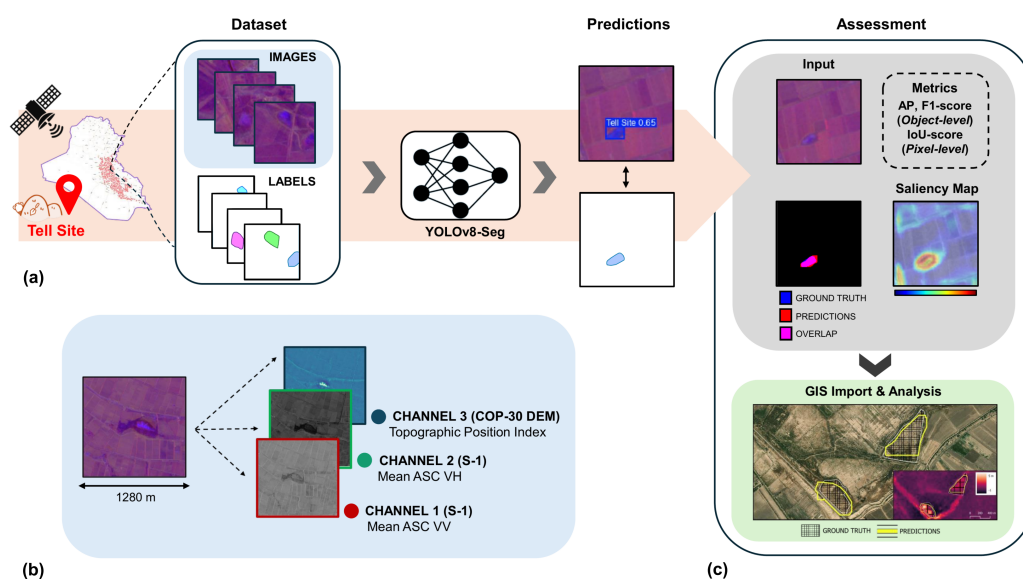


Figure 1. The proposed pipeline for the automatic detection and segmentation of *tell* sites in Central Iraq. (a) During the training phase, the YOLOv8-Seg model is trained using archaeological ground-truth labels to predict bounding boxes, segmentation masks, and class probabilities within false-colour RGB images. (b) Each input image has a size of 128 × 128 pixels (corresponding to 1280 × 1280 m on the ground) and is generated by combining Sentinel-1 (S-1) backscatter amplitude data (VV and VH bands) and COP-30 DEM topographic data. (c) Model performance is evaluated quantitatively using object- and pixel-level metrics, and qualitatively through GIS-based spatial analysis and saliency map interpretation. Saliency maps provide visual explanations of the model's predictions by highlighting specific pixels or regions in the input image that contribute most significantly to the identification of *tell* sites.

The model performance was assessed using standard Deep Learning evaluation metrics that capture both detection and segmentation capabilities. Average Precision (AP) was computed for bounding box detections at an Intersection over Union threshold of 0.5,

providing an overall measure of the model's ability to detect and localise *tell* sites across a range of confidence thresholds, independent of any single operating point.

To evaluate performance under optimal conditions, the F1-score was calculated for bounding boxes at the confidence threshold that maximises the balance between precision and recall on the validation dataset. This metric reflects the model's effectiveness in *tell* detection and localisation at its best-performing configuration.

Segmentation performance was quantified using the pixel-wise IoU-score for the "Tell Site" class, computed over the entire dataset at the same optimal confidence threshold. This measure captures the accuracy of the predicted masks, indicating how well the model delineates the spatial extent of the *tell* site.

The model outputs, including segmentation masks and the associated YOLO confidence reflecting the reliability of each bounding box detection, were subsequently converted into polygon features and imported into the GIS environment for subsequent archaeological interpretation and analysis.

2.4.2. Preprocessing of Remote Sensing Data

The preprocessing pipeline utilised the Google Earth Engine (GEE) data catalogue [54] to access the S-1 IW GRD and the COP-30 DEM image collections. These collections were processed using the GEE Python API (version 1.0.0) within a Python environment and then downloaded locally.

Specifically, a multitemporal series of dual-polarised S-1 IW GRD images, spanning the period from January 2021 to December 2023, was processed. The data were organised by orbit (ascending/descending) and polarisation (VV/VH). To mitigate speckle noise and enhance the signal from persistent features such as *tells*, the average amplitude image of each collection was computed, resulting in a 4-band S-1 data cube. This approach relies on the assumption that the *tell* sites remained relatively undisturbed within this temporal window. A Topographic Position Index (TPI) raster was generated from the COP-30 DEM using a neighbourhood radius of 8 pixels, corresponding to 240 m at the DEM's 30 m resolution. This radius was empirically chosen based on the average dimensions of the archaeological features in the dataset, as shown in Table 1. The resulting TPI raster was then downsampled to the S-1 resolution (10 m) using bilinear resampling.

Finally, a false-colour RGB composite was created for model input by combining the processed TPI raster and the mean ascending orbit VV (ASC VV) and VH (ASC VH) bands. To ensure compatibility with the model, the TPI band, fitted to values $[-6, 6]$, and the ASC VV and ASC VH bands, fitted to values $[-30, 5]$ dB, were each normalised to the $[0, 1]$ range and converted to an unsigned 8-bit format, with null values set to 0. The ASC VV and ASC VH bands were then assigned to the red and green channels, respectively, while the TPI band was assigned to the blue channel of the RGB composite. This specific combination was chosen to highlight both the backscatter characteristics and the topographic signature of the *tell* sites. A comparative analysis of this composite with two other SAR-based false RGB combinations, as well as an S-2 true-colour composite, is provided in Section 3. Details on their creation can be found in the Supplementary Materials Section S1.

2.4.3. Preprocessing of Archaeological Labels

The archaeological label dataset underwent extensive preprocessing to ensure its suitability for training a Deep Learning model on SAR data. Data filtering involved the removal of sites lacking clear visibility in both optical and SAR products (see Figure 2), as well as sites that were too small or too large for effective model training. Sites smaller than 0.3 hectares—approximately 30 pixels in the 10 m resolution false-colour composite—were

excluded. A summary of the median size and elevation of the remaining 3866 archaeological instances is provided in Table 1b.

Further filtering was applied during the creation of image tiles. Archaeological instances exceeding a predefined threshold of 80% of the tile area were excluded. This threshold was chosen a priori to ensure sufficient surrounding context for the model. Moreover, to mitigate border effects, instances truncated at tile borders were filtered out if the remnant polygon was smaller than 0.2 hectares. Finally, the labels were then converted into YOLO annotation format for network ingestion, and binary masks were generated for calculating pixel-wise IoU-scores during model evaluation.

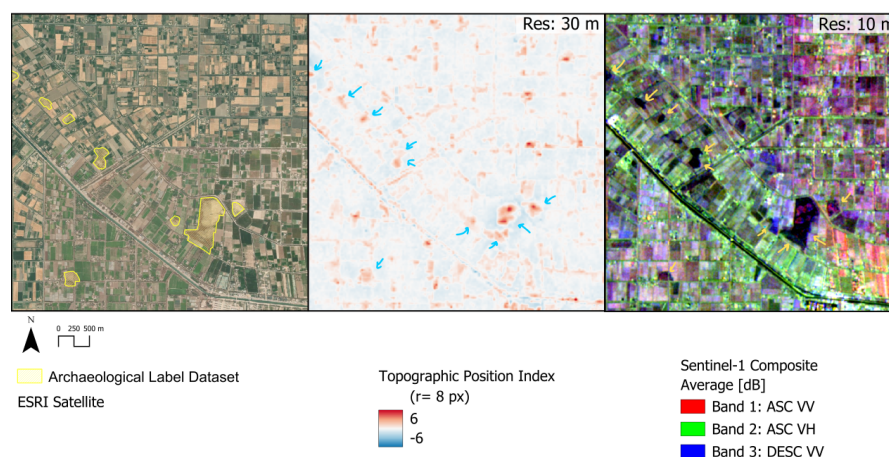


Figure 2. Identification of *tell* sites across different EO data products. From left to right: Esri high-resolution optical imagery, with ground truth polygons outlined in yellow; Topographic Position Index derived from the COP-30 Digital Elevation Model (30 m resolution); and a false-color Sentinel-1 composite (10 m resolution). Blue and yellow arrows highlight the physical boundaries of *tell* sites across the Copernicus dataset. Sources: Esri World Imagery; Copernicus WorldDEM-30; Copernicus Sentinel-1 data.

2.4.4. Deep Learning Experimental Setting

The experimental design utilised geographically distinct areas for training, validation, and testing, as shown in Figure 3. The training and validation areas covered approximately 53,000 km². The validation area was selected to be representative of the training area, matching both the *tell* site features and land cover characteristics. The model's generalisation was assessed on three distinct test areas in Iraq, each spanning approximately 1600 km². To evaluate transferability, an additional test area in Iran, in the Khuzestan province, spanning approximately 1400 km² and with different landscape and site characteristics, was considered. Detailed descriptors for all areas are shown in Table 1.

The YOLOv8-Seg model series comprises five architectures, which vary in size and number of parameters. Given the limited number of annotations on the training area ($N = 2634$), the small model, “yolov8s-seg”, was selected as an optimal balance between model size and performance. The model was trained on the IIT High-Performing Computing Cluster “Franklin” using NVIDIA Tesla V100 GPUs. Training was initialised using weights pretrained on the COCO dataset [55] to leverage existing feature recognition capabilities.

The primary experimental challenges were the severe class imbalance and the presence of label noise in the archaeological ground-truth across the vast study area, largely due to spatial sampling bias. While ground-truth annotations are typically more curated in regions with a high density of pre-existing ground-truth data, regions far from these areas may exhibit a higher proportion of missing (unlabelled) sites, resulting in a spatially non-

uniform annotation quality. To mitigate both challenges, we oversampled the target class by modifying the proportion of empty tiles (i.e., tiles without any archaeological instances) in the training and validation dataset. Specifically, we tested three percentages of empty tiles, which were randomly selected from the total pool: 20%, 10%, and 1%. This process artificially increased the overall representation of the “Tell Site” class and created a cleaner, higher-confidence training domain by reducing the influence of potentially mislabelled background areas.

The model’s performance on the validation dataset was used for the optimisation of training hyperparameters, including image size (512, 256, **128**), batch size (128, 64, **32**, 16), and the percentage of empty tiles (20%, **10%**, 1%). The validation dataset was also used for tuning the best confidence threshold for the final configuration. The optimisation details, along with a comparison with a model trained from scratch and a “yolov11s-seg” model, are detailed in the Supplementary Materials Section S2. The final hyperparameters and data augmentation settings are summarised in Tables 2 and 3, respectively. The model was trained using non-overlapping tiles.

A summary of the training, validation, and test datasets’ characteristics in their optimally configured state is shown in Table 4: training and validation datasets represent a highly unbalanced segmentation task, with less than 4% of all labelled pixels belonging to the “Tell Site” class. Unlike the training and validation datasets, the test datasets maintained the real-world tile distribution, resulting in a lower proportion of “Tell Site” pixels (0.47% and 0.35%). This domain shift may negatively impact F1-score and IoU-score metrics compared to validation results, potentially leading to a model that prioritises recall over precision. However, we intentionally maintained this workflow mirroring an operational scenario where the objective is to generate the most comprehensive list of potential sites for follow-up investigation.

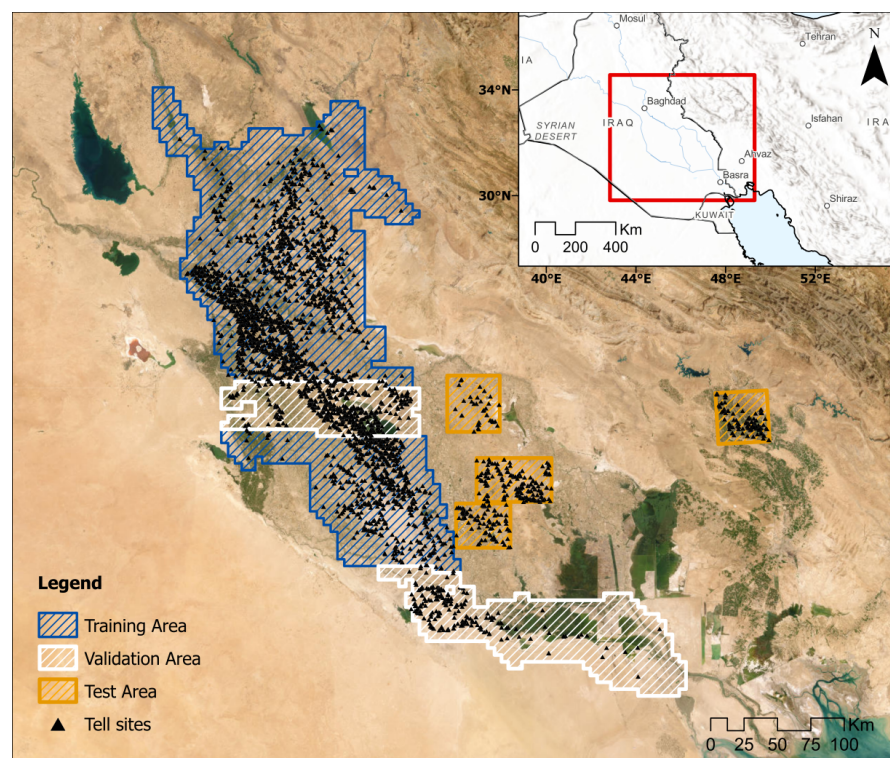


Figure 3. Spatial distribution of the study area in the Southern Mesopotamian floodplain. Training (blue), validation (white), and test areas (orange) are delineated across the landscape. Black triangles represent georeferenced archaeological annotations (*tell* sites) within the UTM 38N projection.

Table 2. Best model’s hyperparameters.

Hyperparameter	Value
Image Size	128 × 128
Percentage of Empty Tiles	10%
Batch Size	32
Optimizer	AdamW (auto)
Learning rate (lr0)	7.14×10^{-4} (auto)
Momentum	0.9 (auto)
Pretrained	True, COCO dataset

Table 3. Data augmentation settings.

Setting	Value
Saturation Adjustment (hsv_s)	0.1
Brightness Adjustment (hsv_v)	0.1
Hue Adjustment (hsv_h)	0.015 (default)
Mosaic (mosaic)	0
Scale (scale)	0.1
Translation (translate)	0.1 (default)
Flip Left-Right (fliplr)	0.5 (default)
Blur (blur)	$p = 0.01$, blur_limit = (3, 7) (default)
Median Blur (MedianBlur)	$p = 0.01$, blur_limit = (3, 7) (default)
Grayscale (ToGray)	$p = 0.01$, num_output_channels = 3, method = ‘weighted_average’ (default)
CLAHE (CLAHE)	$p = 0.01$, clip_limit = (1.0, 4.0), tile_grid_size = (8, 8) (default)

Table 4. Training, validation, and test datasets characteristics.

Characteristic	Iraq			Iran
	Training	Validation	Test	Test
Total Number of Tiles	2610	854	3150	930
Number of Filled Tiles	2349	769	313	97
Number of Archaeological Instances	3389	1114	429	125
Mean Instances per Filled Tiles	1.44	1.45	1.37	1.30
Percentage of Class “Tell Site” Pixels	3.46%	3.78%	0.47%	0.35%

2.4.5. Post-Processing of Model’s Inferences

The post-processing module prepares the output for a direct archaeological assessment in a GIS environment. Instance segmentation masks, generated by the best model over the area of interest, were converted into georeferenced polygons and saved along with the model’s confidence scores.

This step enables archaeologists to visually inspect predictions within their spatial context, comparing them against remote sensing and ground-truth data, to correct and refine the polygonal shapes, and to facilitate interpretation and analysis, thereby aiding in model debugging and the creation of final, validated site inventories. While the model was trained using non-overlapping tiles, the tiling strategy for prediction is adjusted depending on the desired output. Two outputs are generated:

- **Filtered Inferences (Non-Overlapping Tiles):** A collection of all instance inferences generated from non-overlapping tiles and converted into polygons, where the model’s confidence score is greater than a 5% threshold. This output is primarily useful during the model’s debugging and refinement phase, as it includes low-confidence predictions for further scrutiny of potential false positives or missed sites.

- **Merged Inventory (Overlapping Tiles):** A definitive collection of site inferences, derived from processing overlapping tiles (specifically, 50% overlap). This overlap-and-merge strategy ensures archaeological features appear near the centre of at least one tile, where the model can benefit from a richer local context. Only polygons with confidence scores greater than the optimal validation threshold are retained and then unified using spatial union (OR logic). The maximum confidence score among the overlapping polygons is assigned to the final, unified polygon. This procedure reduces the visual artifacts in predictions (e.g., seams) and fragmentation that result from predicting on non-overlapping tiles, thereby easing the visual archaeological interpretation and mitigating the adverse effects of tile borders on the performance metrics by correcting for potential sites that fall on the tile boundaries and could be missed by the model.

An example of the interpretation of these outputs is reported in Section 3.3. Additionally, two application-driven performance metrics, namely F1-score[†] and IoU-score[†], were computed over the test areas in Iraq using the merged inventory and will be discussed in the same section. These metrics provide a more realistic and comprehensive measure of model performance, as they evaluate the alignment between the final, merged site inferences and the ground-truth *tell* sites reported in Table 1b.

2.4.6. Explainability with Eigen-CAM

To align with current research directions on Explainability and Interpretability in Remote Sensing for Earth Observation [56–58], an explainability module was added to the pipeline. Deep Neural Networks are often referred to as “black boxes” because input data undergoes numerous layers of weighted multiplications and non-linear transformations, hindering the internal logic behind a specific output [59,60]. Saliency maps are the most common and intuitive method for explaining Convolutional Neural Networks. They provide visual explanations by highlighting specific pixels or regions in the input image that contribute most significantly to the model’s final decision.

For our pipeline, we adopted the Eigen-CAM method [61] as provided in the implementation by [62] due to its compatibility with the YOLO Ultralytics library. Eigen-CAM operates directly on the feature maps of a specific Convolutional Neural Network layer, typically the final layer, which contains the highest-level semantic representations of the image. The method applies Principal Component Analysis on the feature maps of the selected layer to identify the first principal component, i.e., the direction of maximum variance, which captures the most dominant patterns learnt by the model. These patterns are then projected back onto the input image as a saliency map.

In this study, explainability served to assist in model debugging, enhance the trustworthiness of model’s inferences and investigate human-AI alignment. It addressed questions such as which spatial regions the model prioritised during the detection and segmentation of the *tell* site and why a detection failed—specifically, whether the model focused on incorrect features, or its confidence in the prediction fell below the defined threshold. Furthermore, it investigated whether the model identified the same morphological characteristics that an archaeologist would use to identify a *tell* site. A qualitative analysis of these explanations is provided in Section 3.4.

3. Results

3.1. Inferences over the Iraqi Test Areas

Table 5 presents the model’s performance metrics over the validation area and three independent test areas in Iraq. It also includes the results of control experiments designed to evaluate the impact of different input bands and modalities, specifically COP-30 DEM,

S-1 SAR, and S-2 RGB data. These experiments were conducted by training, validating, and testing separate models using the fixed set of hyperparameters shown in Table 2 on different band composites. Metrics are presented together with 95% confidence intervals, calculated from 10 independent model runs with different random seeds.

The model trained with the proposed false-colour composite—the first configuration—achieved the highest performance over the test areas with an Average Precision of 0.495 ± 0.010 , an F1-score of 0.490 ± 0.021 , and a pixel-wise IoU-score of 0.361 ± 0.048 . The second configuration, based solely on COP-30 DEM data, followed with a decrease of up to 5 percentage points (0.442 ± 0.013 Average Precision, 0.468 ± 0.017 F1-score, and 0.313 ± 0.024 IoU-score). The third configuration, which includes only the S-1 backscatter information, performed worse with an Average Precision of 0.116 ± 0.013 and an IoU-score of 0.091 ± 0.010 . The fourth configuration, consisting of S-2 RGB data, performed similarly to the third configuration. The results suggest that the topographic information contained in the COP-30 DEM and its derivatives is the primary factor driving the model's performance. The S-1 SAR bands (VV, VH) offer valuable complementary information, although they are not sufficient alone for robust detection, like the S-2 RGB data. Notably, both Composites 1 and 2 overcome the S-2 RGB baseline.

Table 5. Model's performance over the validation and test area in *Iraq*. Best model in **bold** and the second best is underlined. 95% confidence intervals are shown and are calculated from 10 independent model runs with different random seeds.

(a) Validation Area			
3-Channel Composite	Object-Level Metric		Pixel-Level Metric
	Average Precision	F1-Score	IoU-Score
Our Composite (S-1 + COP-30)	0.474 ± 0.004	0.513 ± 0.004	0.389 ± 0.014
COP-30 DEM	<u>0.388 ± 0.009</u>	<u>0.447 ± 0.009</u>	<u>0.360 ± 0.017</u>
S-1 (VV-VH-VV)	0.231 ± 0.006	0.322 ± 0.005	0.207 ± 0.011
S-2 (R-G-B)	0.268 ± 0.005	0.345 ± 0.005	0.086 ± 0.012
(b) Test Area			
3-Channel Composite	Object-Level Metric		Pixel-Level Metric
	Average Precision	F1-Score	IoU-Score
Our Composite (S-1 + COP-30)	0.495 ± 0.010	0.490 ± 0.021	0.361 ± 0.048
COP-30 DEM	<u>0.442 ± 0.013</u>	<u>0.468 ± 0.017</u>	<u>0.313 ± 0.024</u>
S-1 (VV-VH-VV)	0.116 ± 0.013	0.149 ± 0.011	0.091 ± 0.010
S-2 (R-G-B)	0.120 ± 0.012	0.154 ± 0.014	0.086 ± 0.012

A visual evaluation of the predicted masks on the Iraq test dataset helps interpret these quantitative results. Figures 4 and 5 show examples of the model's segmentation inferences overlaid on the S-1 + COP-30 DEM false-colour composites. In the best-performing tiles (Figure 4a–d,h), the predicted masks closely follow the geometry of the ground-truth labels, proving high spatial reliability of the detections. This property is particularly relevant for any subsequent analyses requiring the precise delineation of site boundaries, such as morphometric measurements or spatial statistics on archaeological *tells*. In some instances, however, the visual inspection highlights inconsistencies with the ground-truth annotations. In Figure 4e, for example, the predicted masks encompass a large area outside the archaeological *tell*, overpredicting to nearby areas. In some examples, labels appear slightly over- or undersized (Figure 4f,g) or are potentially missing (Figure 5a,b). Such discrepancies represent inherent uncertainties in this type of modelling, which arise from the model's inaccuracy and also from the uncertainty typical of archaeological datasets,

particularly when ground truth information is derived from the manual interpretation of remote sensing data. In addition, the model seems to be prone to producing inaccurate predictions at the edges of image tiles, where the lack of contextual information prevents the model from accurately recognising features (Figures 4e and 5b–d). This behaviour was further investigated in Supplementary Materials Section S4.

Other examples of misclassification are presented in Figure 5f–h as, for example, challenges in correctly predicting large *tells* which heavily affect pixel-based performance metrics. Extremely large *tells*, in fact, occupy a high number of pixels, spanning multiple tiles. In addition, they are often composed of aggregated or combined mounds, which may not have been internally discerned in the ground truth (Figure 5e). These complex formations combine multiple archaeological and geomorphological features, producing heterogeneous topographic signatures that make it difficult to delineate which parts correspond to actual human occupation and which derive solely from fluvial or sedimentary processes. Such ambiguous cases should therefore not necessarily be seen as model errors but rather reflect the complex nature of the archaeological landscape and segmentation tasks under such ambiguous conditions, which is rich in potential candidates that show the same features of archaeological abandoned settlements.

Overall, the results show that the performance on the separate test dataset remains consistent with that obtained on the validation dataset. While a drop in performance is observed for Composites 3 and 4, Composites 1 and 2 show solid generalisation ability, confirming the model's capacity to perform robustly on spatially distinct but morphologically similar areas. The evaluation of the trained YOLOv8-Seg model on the three independent test areas demonstrated its ability to automatically detect and segment *tell* sites in the satellite SAR products. The observed performance metrics, Average Precision and F1-score, suggest a promising level of automation for large-scale prospection, significantly reducing the time and effort required for visual inspection. A comparison with a pixel-based Random Forest classification baseline and an evaluation of the pipeline's robustness against seasonal effects are provided in Supplementary Materials Sections S5 and S6, respectively. Moreover, a discussion of how this pipeline can reduce the visual inspection effort by operating as a pre-screening tool is reported in Supplementary Materials Section S7. Finally, Table 6 shows the performance metrics for the final merged inventory, following the post-processing steps described in Section 2.4.5.

3.2. Transferability and Regional Fine-Tuning to Iran

To evaluate the broader applicability of the proposed workflow, we conducted a transferability test on an independent archaeological landscape in Iran, in the Khuzestan province.

The results of this direct model transfer to the Iranian test area are reported in Table 7. Applying the model trained on the Iraqi dataset directly to this new domain yielded a significant performance degradation in configurations that include DEM-derived information. Specifically, the DEM-based model (Configuration 2) experienced the most severe performance drop, with its Average Precision falling to 0.136 ± 0.021 (versus 0.442 ± 0.013 in the Iraqi test area). Similarly, the multimodal composite (Configuration 1) achieved an Average Precision of 0.214 ± 0.021 , an F1-score of 0.234 ± 0.018 , and a pixel-wise IoU-score of 0.069 ± 0.011 , to be compared to 0.495 ± 0.010 , 0.490 ± 0.021 , and 0.361 ± 0.048 , respectively, on the Iraqi test area.

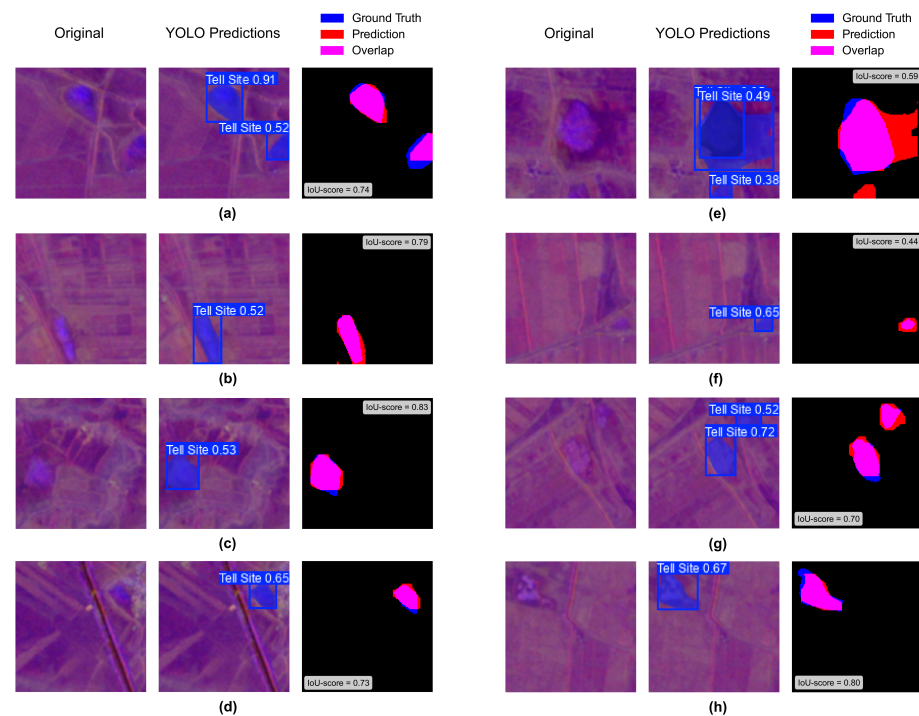


Figure 4. Examples (a–h) of YOLO inferences on the Iraqi test dataset for the best-performing seed. YOLO returns both object detection boxes with a confidence level indicating the model’s prediction and segmentation masks. The quality of the segmentation masks is also assessed through the per-tile pixel-wise IoU-score between the ground truth binary masks and the binarized YOLO predictions. In the visualization, purple regions indicate the intersection (overlap) between ground truth (blue) and predicted masks (red).

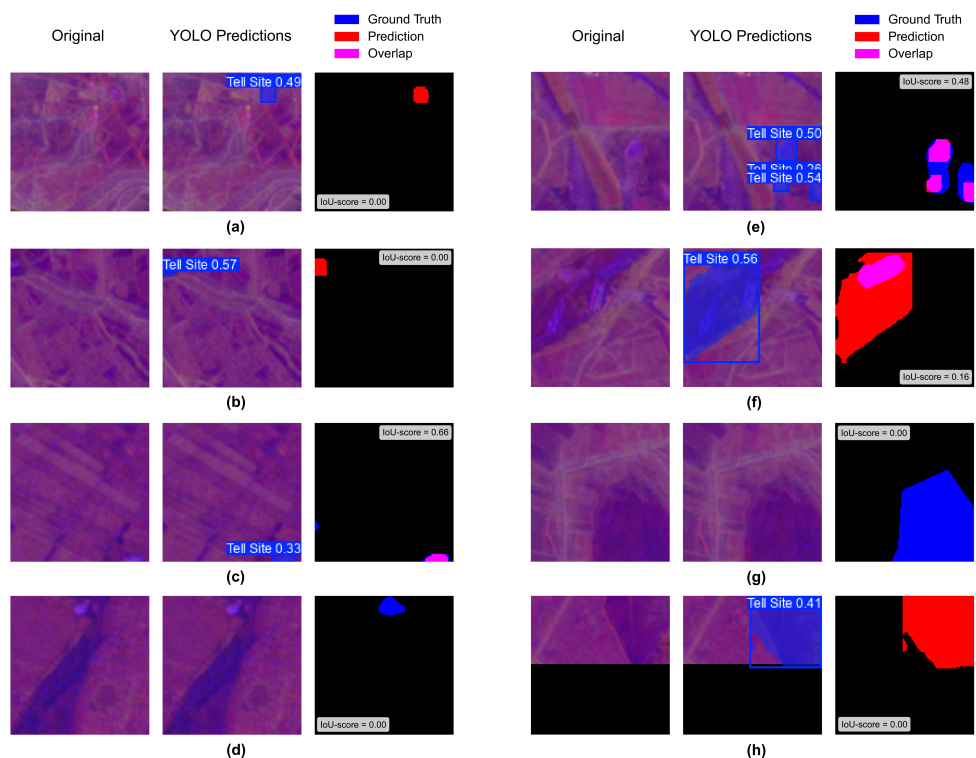


Figure 5. Examples (a–h) of incorrect YOLO inferences on the Iraqi test dataset. The quality of the segmentation masks is also assessed through the per-tile pixel-wise IoU-score between the ground truth binary masks and the binarized YOLO predictions. In the visualization, purple regions indicate the intersection (overlap) between ground truth (blue) and predicted masks (red).

Table 6. F1-score[∧] and IoU-score[∧] metrics for the best performing model over the Iraqi test area. These metrics were computed using the merged inventory of site inferences, generated by unifying predictions over the test area from tiles with 50% overlap. The Recall value (in parentheses) indicates the proportion of successfully detected sites over the actual ground truth (290 sites, see Table 1b).

	Object-Level Metric	Pixel-Level Metric
	F1-Score [∧] (Recall)	IoU-Score [∧]
Test Area Iraq	0.52 (0.66)	0.39

Table 7. Transferability test over *Iran*. Best model in **bold** and the second best is underlined. 95% confidence intervals are shown and are calculated from 10 independent model runs with different random seeds.

3-Channel Composite	Object-Level Metric		Pixel-Level Metric
	Average Precision	F1-Score	IoU-Score
Our Composite (S-1 + COP-30)	0.214 ± 0.021	0.234 ± 0.018	0.069 ± 0.011
COP-30 DEM	0.136 ± 0.021	0.193 ± 0.021	0.043 ± 0.012
S-1 (VV-VH-VV)	0.220 ± 0.018	0.198 ± 0.020	0.070 ± 0.010
S-2 (R-G-B)	0.185 ± 0.026	<u>0.206 ± 0.021</u>	0.070 ± 0.014

Notably, the S-1-based configuration (Configuration 3) demonstrated the highest stability in zero-shot transfer, with an Average Precision of 0.220 ± 0.018 . This performance pattern suggests a topographic domain shift between the two landscapes. While the proposed multimodal configuration appears less context-dependent than the DEM-only mode, likely because the consistency of the S-1 features counterbalances the topographic shift, confirming that complementary SAR and elevation datasets improve geographic transferability, a noticeable performance gap remains compared to the Iraqi test area. This suggests that data multimodality alone cannot fully bypass regional geomorphological shift, indicating the necessity of weight adaptation through regional fine-tuning.

Visual inspection of the model's inferences in the Iranian context (Figure 6) revealed a higher number of false positives and false negatives (Figure 6e,f). Label-related issues similar to those observed in the Iraqi test data also occur in the Iran dataset, where large false positives or false negatives are typically associated with extensive mounds (Figure 6g,h). The lower performance observed in the Iranian test area can likely be attributed to distinct geomorphological and land use conditions compared to the Iraqi landscape (see Section 2.1 for an overview of the contexts).

Whereas the Iraqi plains are predominantly flat and characterised by extensive fluvial ridges, the Iranian terrain is more irregular, with incised valleys, localised aggradation, and slope terracing. Land-use patterns are also different, with Iraq characterised by large, continuous agricultural fields, while Iran has smaller, fragmented plots with denser and higher vegetation that further modifies the surface topography. Differences are also evident in *tell* morphology, with Iranian *tells* exhibiting a more pronounced relative elevation to the mean plain level over smaller areas than those in Iraq (on average twice as high for half the areal extent, see Table 1).

The initial transferability analysis suggested that regional fine-tuning would be necessary to achieve comparable performance to that observed in Iraq. To test this, we evaluated several fine-tuning strategies over a subset of the Iranian test area, detailed in Supplementary Materials Section S3. The results, presented in Table 8, confirm the effectiveness of this approach. The fine-tuning strategies resulted in a substantial increase in performance across all metrics. Specifically, the best fine-tuned model (*Fine-tuning 200 tiles, with 20 new sites, frozen backbone*) achieved an Average Precision of 0.497 ± 0.044 , restoring it to levels comparable to

those achieved in the Iraqi test area. Compared to the initial *No Fine-tuning* baseline (Average Precision: 0.230 ± 0.026), fine-tuning led to an approximate doubling of the Average Precision. Even compared to the *Histogram Matching* baseline (Average Precision: 0.269 ± 0.027), the fine-tuned model showed a substantial improvement, confirming that simply mitigating the TPI distribution shift via histogram matching is insufficient compared to adapting the model weights to the new target distribution, highlighting the need to learn complex, region-specific morphological patterns. Notably, the IoU-score did not improve in line with the Average Precision and F1-score, exhibiting a saturation-like behaviour. We attribute this to the presence of a very large *tell* site (an annotation outlier) that comprises almost half of the total “Tell Site” pixels in the subset test area. This specific site was not detected, which tends to distort pixel-based metrics such as IoU score. In such cases, object-based metrics become a more reliable indicator of overall model performance. The effectiveness of the small dataset size (200 tiles, with only 10% containing at least one archaeological instance) suggests that the learned weights are highly transferable and only require minor local adaptation.

From a machine learning perspective, the success of the frozen backbone strategy demonstrates that the deep spatial features learned by the YOLOv8-Seg backbone in Iraq are fundamentally robust and transferable. The network does not suffer from representational collapse, but rather, the underlying representation of a *tell* remains valid, requiring only an adaptation of the final prediction heads and confidence threshold to suppress the natural background topography of the Iranian landscape. Consequently, while zero-shot transfer is constrained by regional geomorphology, targeted fine-tuning, with the incorporation of a small number of Iran-specific tell site data and characteristic landscape features, offers an effective and necessary step that lowers the operational barrier for reusing pre-trained models across highly diverse geographical contexts.

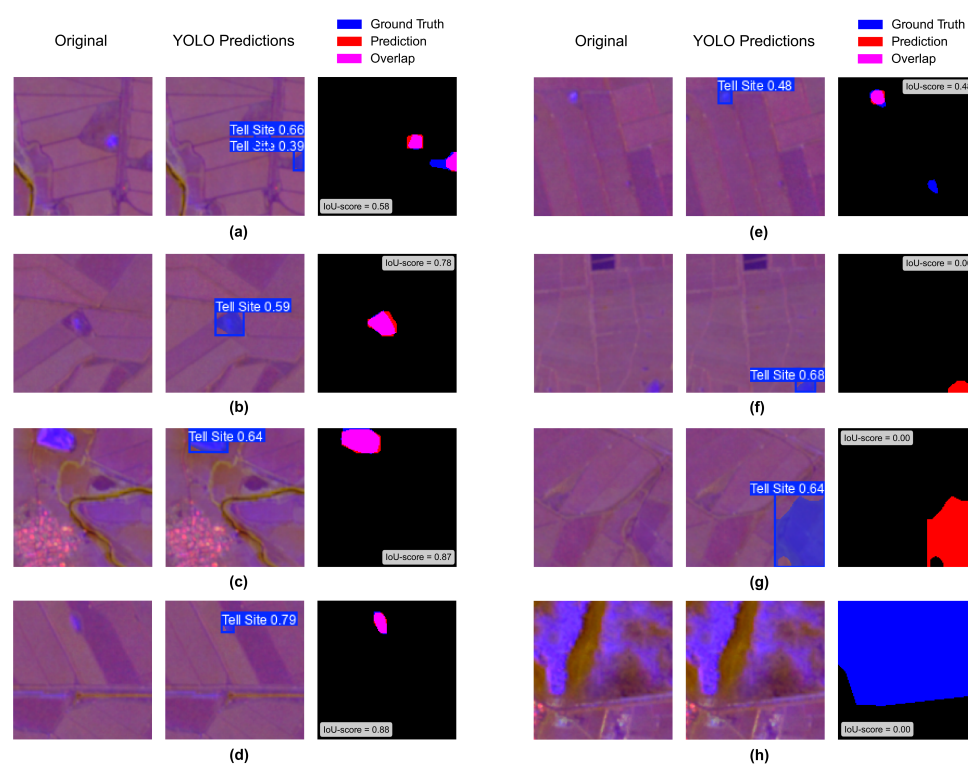


Figure 6. Examples (a–h) of both successful and unsuccessful YOLO inferences on the Iranian test dataset for the best-performing seed. The quality of the segmentation masks is also assessed through the per-tile pixel-wise IoU-score between the ground truth binary masks and the binarized YOLO predictions. In the visualization, purple regions indicate the intersection (overlap) between ground truth (blue) and predicted masks (red).

Table 8. Performance of fine-tuned models over a subset of the Iranian test area. Best model in **bold**. 95% confidence intervals are shown and are calculated from 10 independent model runs with different random seeds. Results are compared against two baselines: (1) Zero-shot, where the original models are applied directly to the test data; and (2) Zero-shot, histogram matching, where the original models are applied after preprocessing the test data using a histogram matching algorithm [63] to mitigate domain shift. This technique was applied selectively to the TPI band.

Model	Object-Level Metric		Pixel-Level Metric
	Average Precision	F1-Score	IoU-Score
Fine-Tuned Models			
Fine-tuning 100 tiles	0.288 ± 0.032	0.294 ± 0.032	0.086 ± 0.014
Fine-tuning 200 tiles	0.488 ± 0.050	0.487 ± 0.055	0.123 ± 0.022
Fine-tuning 200 tiles, frozen backbone	0.497 ± 0.044	0.515 ± 0.035	0.123 ± 0.032
Baseline Models			
Zero-shot	0.230 ± 0.026	0.252 ± 0.021	0.048 ± 0.012
Zero-shot, Histogram Matching	0.269 ± 0.027	0.277 ± 0.014	0.109 ± 0.025

3.3. Archaeological Interpretation

This section analyses the model's predictions in relation to the contextual information available in the GIS environment (e.g., orthophotos, thematic cartography, and topographic data) and assesses whether the observed reliabilities and uncertainties can be explained by the characteristics of landscape and remote sensing data (see Figure 7). Even if it is not possible to establish a direct comparison between the SAR products and the high-resolution RGB imagery due to date mismatches, it can be reasonably assumed that, in most situations, land divisions and, therefore, agricultural land use remain stable over multiple years within the same land parcels.

The highest spatial overlap between predicted masks and ground-truth labels is observed in relatively clear situations, where *tell* morphologies are small and approximately oval, and land divisions closely follow the mound outline (Figure 7a). In some cases, detection is further supported by a coherent association between topographic change and land-use patterns, such as persistent agricultural parcel boundaries that follow the outline of the mound. From a modelling perspective, this association is relevant because it reinforces the spatial coherence between surface morphology and landscape texture visible in the GIS data. The observed patterns may equally reflect geomorphological constraints, soil moisture conditions, or modern land-management decisions, none of which can be independently verified within the scope of the available data. In several cases, however, more challenging conditions such as non-uniform land cover, e.g., uncultivated areas left as bare soil and characterised by sparse and patchy vegetation, do not appear to significantly affect the model's predictions (Figure 7b). From a topographic perspective, additional challenges include the spatial continuity between many *tells* and fluvial ridge systems, where elevation remains nearly constant over long stretches of the plain. In these contexts, *tells* are not isolated high-relief features but rather embedded within the ridge morphology (Figure 7c). Despite these geomorphological conditions, the model is frequently able to delineate *tell* features with a high degree of spatial accuracy.

The most frequent mismatches occur when the ground-truth annotations extend beyond the actual topographic mound and include portions of adjacent flat land (Figure 7e,f). These patches are clearly evident in the RGB imagery due to their irregular or oval shapes and their frequently bare or uncultivated surfaces, which contrast with the regular agricultural landscape. Annotations derived from legacy studies, therefore, sometimes enclose the entire patch and are not limited to the topographic prominence of the *tells*, especially in large archaeological areas where materials are scattered over wide extents and multiple

tells are present. This ambiguity, inherent in the dataset characteristics, leads to two complementary outcomes in the model's prediction. In some cases (Figure 7g), the model slightly overpredicts the ground truth, extending beyond the mound limits and into the entire land parcel. In other instances, the model underpredicts by fragmenting large composite areas into smaller units that closely follow the topographic prominence of individual *tells* (see Figure 5e).

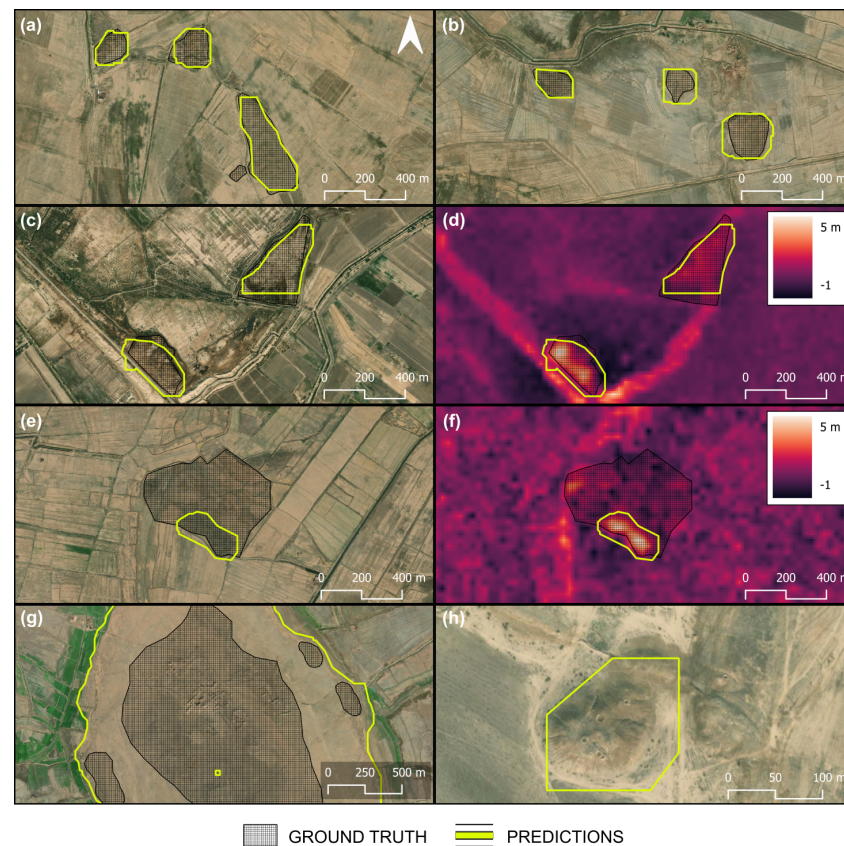


Figure 7. Archaeological assessment of the significance of the predicted masks compared to ground-truth labels and contextual landscape information in GIS, visualised over RGB and COP-30 DEM imagery: (a) high spatial overlap in clear cases; (b) high spatial overlap in challenging land cover conditions and surface disturbance; (c,d) high spatial overlap in challenging topographic conditions, in which *tell* sites are in continuity with the relief of a fluvial ridge; (e,f) underprediction over a large ground-truth label corresponding to bare soil parcel without any topographic prominence; (g) overprediction over a large archaeological site whose ground-truth is represented by multiple mounded features; (h) False positive prediction archaeologically validated by the presence of potential looting pits. Background imagery in (a–c,e,g,h): Esri World Imagery. Background imagery in (d,f): Copernicus COP-30 DEM.

A visual inspection of the 253 false positive detections in the Iraq test area reveals that 57 instances (22.5%) may correspond to previously unrecorded *tell* sites (Figure 7h). These features are consistent with known sites in terms of size and morphology, with a median area of 1.66 hectares and typical dimensions of less than 200 m, both lower than median values computed for the test area (see Table 1b). Those *tells* are often a bit smaller, low in relief, or lack distinctive surface features such as agricultural boundaries or distinct land use from the surrounding area, making their identification challenging. The presence of these sites among the false positives is primarily due to limitations in the original visual survey. During large-scale visual inspections, less prominent or previously unknown *tells* can be occasionally overlooked. Factors such as vegetation cover, subtle morphological

boundaries, modern disturbances, and visually noisy contexts can make sites difficult to identify. Beyond these limitations, the archaeological dataset itself should be regarded as inherently incomplete, as existing inventories inevitably reflect uneven survey intensity and historical research biases. Therefore, these detections should not be considered model errors but rather highlight potential archaeological features that are missing from the original test dataset. The reliability of these predictions, as legitimate archaeological sites, is further demonstrated by the identification of looting activity in at least six of these newly identified sites. Since looting can be a direct indicator of buried archaeological deposits, these cases provide independent evidence of the model's ability to identify locations of high interest. An additional 68 false positive detections correspond to elongated depositional alluvial features (e.g., levees, crevasse splays) that share morphological similarities with *tell* sites. Those features are valuable for landscape characterisation and archaeological prospection, as they highlight candidate locations for future ground-truth survey.

A separate evaluation was also conducted on predictions that fall below the detection threshold (see Section 2.4.5). Visual inspection confirmed the archaeological plausibility of a subset of these predictions, some of which correspond to features already present in the ground truth, while others represent previously unrecorded candidates. A subset of these detections consists of small, oval *tells*, often less than 1 hectare in size, characterised by minimal elevation change but a clear visual contrast with the surrounding agricultural landscape, for example, through sparse vegetation or regular field patterns. At a broader territorial scale, these sub-threshold detections may still represent archaeological evidence. While their inclusion increases false positives and reduces precision, it also highlights the model's potential to improve recall by detecting subtle features that are easily overlooked during standard visual interpretation, particularly in extensive and low-contrast landscapes.

The model supports the archaeological prospection in two ways. On the one hand, by focusing on a reduced number of candidates, it enables a more efficient allocation of resources for field survey inspections, prioritising areas with the highest probability of archaeological presence and reducing the time and costs associated with traditional landscape surveys. Moreover, the model can identify potential sites that human operators might miss during manual documentation and traditional archaeological surveys, greatly improving the reliability of prospection in large-scale studies. On the other hand, the confidence score associated with each detection provides a measure of model certainty, allowing researchers to rank candidates prioritising higher-confidence detections over lower-confidence ones. While the integration with other remote sensing datasets (e.g., optical imagery and historical maps) and field ground truthing activities remains necessary to confirm the nature and significance of the detected features, the model offers a systematic and scalable method that overcomes some of the limitations of visual inspection and incomplete ground truth, supporting informed decision making during the exploratory phases of research.

3.4. Explainability

The inspection of EigenCAM visualisations [61] provides insights into how the model identifies and prioritises spatial patterns potentially associated with archaeological *tells*. This analysis was conducted within the GIS environment, comparing the model's explanations with SAR products and high-resolution RGB imagery. In several cases (e.g., Figure 8a–c), the model's attention maps highlight localised topographic reliefs suggesting that the model effectively exploits elevation-derived information. However, in some cases (e.g., Figure 9a), the model focuses on bare-soil uncultivated areas enclosed by curvilinear roads or ditches, whose shape resembles the oval morphologies typical of *tells*, but without

actual topographic relief. This indicates a sensitivity to textural and soil moisture patterns that are similar to those associated with the *tell* features in the S-1 data.

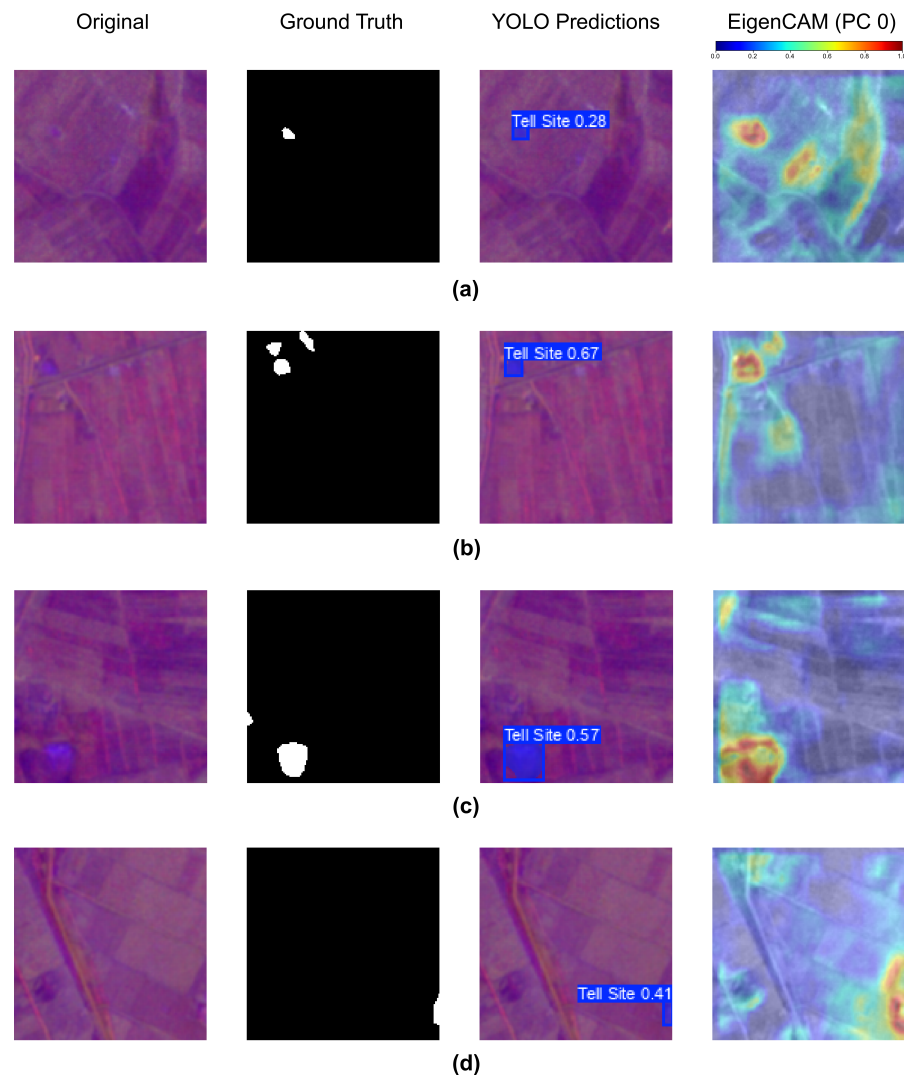


Figure 8. Examples (a–d) of EigenCAM explanations over the Iraqi test area, compared to the archaeological ground truth and model’s inferences. From left to right, the original false-color image, the binary ground-truth mask, the YOLOv8-Seg predictions with confidence score, and the corresponding EigenCAM map. Reddish pixels in the EigenCAM maps highlight the regions that contribute most significantly to the model’s final decision during *tell* sites identification.

In areas with incised channels and strong elevation contrasts (e.g., Figure 9b), the model’s saliency maps mark abrupt elevation changes. In addition, depositional features at river bends (e.g., Figure 8d) are highlighted, likely due to their morphological similarity to *tells*. Such activations reflect the influence of local relief and sedimentary patterns, suggesting that geomorphological features can either generate classification ambiguity or provide clues to support correct predictions. In the Iranian test area, the presence of high vegetation clustered in small patches (e.g., Figure 9c,d), which is largely absent from the Iraqi training dataset, creates topographic anomalies leading to detection errors: these failures are reflected in the high level of noise characterising the corresponding activation maps.

Overall, the EigenCAM analysis suggests that the model captures meaningful topographic and textural signals associated with potential archaeological objects, but it can also be confused by natural or agricultural structures that share similar morphometric characteristics. The interpretability analysis is therefore essential not only to validate the

model's reasoning but also to guide future refinements, which may include expanding the training dataset with hard negative and positive examples, such as agricultural terraces and vegetated *tells*.

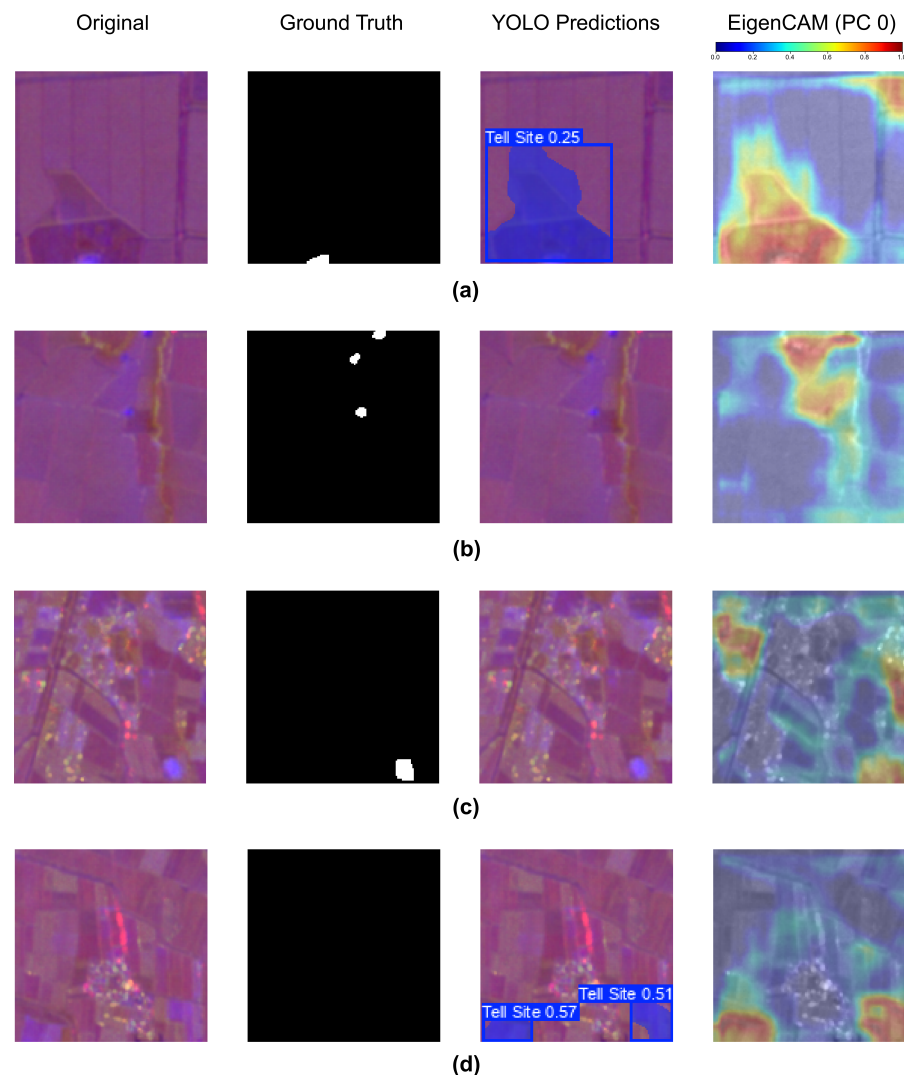


Figure 9. Examples (a–d) of EigenCAM explanations over the Iranian test area, showing the model's response under direct transfer to a new geographical context.

4. Discussion

This section summarises and discusses the main findings of this work:

- *Open-access, medium-resolution SAR products (i.e., Sentinel-1 amplitude images and the InSAR COP-30 DEM) can be used in a Deep Learning pipeline to map tell sites at large scales in Iraq (Section 3.1):*
 - At medium resolution (10–30 m), a multimodal composite combining SAR dual-polarimetric amplitude and COP-30 DEM topographic information consistently outperformed single-modality baselines (optical RGB-only, topographic-only, and SAR amplitude-only). Results confirmed that performance is robust to seasonality, as shown in the Supplementary Materials Section S6;
 - The higher performance of the COP-30 DEM over S-1-only and S-2 RGB-only configurations underscores that topographic prominence is the primary discriminative signal for the presence of a potential site. S-1 backscatter provides meaningful but insufficient information when used alone;

- Although transitioning this SAR-based baseline model into a fully automated pipeline without expert supervision on predictions remains an ongoing challenge (the best model achieved an object-level F1-score of 52% and a pixel-level IoU-score of 39%), the pipeline can operate as a high-accuracy binary tile-level classifier. In this configuration, it achieves an Area Under the Receiver Operating Characteristic Curve of up to 0.94, as discussed in the Supplementary Section S7. By operating as a pre-screening tool, the model can reduce the manual visual inspection effort required to detect the presence of potential archaeological features across large landscapes;
- Archaeological interpretation of the model's false predictions provides deeper context for these performance metrics, revealing that 22% actually correspond to highly plausible, unrecorded *tell* candidates, some with evidence of looting. This highlights a common challenge in large-scale remote sensing datasets, where the large volume of data can degrade the accuracy of initial manual visual analysis. Consequently, such inherent data limitations must be accounted for when evaluating model performance on complex archaeological features and landscapes.
- *Cross-regional transferability of the pipeline can be achieved, though minimal fine-tuning is required to mitigate performance drops (Section 3.2):*
 - Under direct cross-regional transfer from the main Iraqi study area to the Iranian test area, the proposed pipeline experienced a significant performance drop. This degradation is likely attributable to the differing geomorphological and land use conditions between the two regions, which introduced elements of confusion for the model and generated false positive predictions. This interpretation seems further supported by the explainability analysis, see Section 3.4;
 - The multimodal composite demonstrated greater robustness to regional variations than the COP-30 DEM baseline, confirming that the model effectively exploits the complementary information provided by the S-1 VV and VH bands rather than degenerating into a detector of topographic anomalies;
 - Regional fine-tuning on a small subset of Iranian data (amounting to less than 8% of the size of the original Iraqi training area and containing only 20 new local sites) substantially recovered the model's performance, restoring it to the same level achieved in the main study area. This suggests that the pre-trained weights encode transferable morphological representations requiring only minor local adaptation.

A fair comparison of this work within the current state-of-the-art research on Deep Learning for *tell* site mapping is not easily achievable due to differing rationales guiding the existing experimental designs, ranging from the choice of the study areas to data resolution and ground-truth curation. A tentative comparison with the work by [34], which shares the same main study area in Iraq but utilises high-resolution optical historical and modern satellite images, shows that the IoU-score for the “*Tell Site*” class falls within the same range: 0.36 ± 0.05 for our pipeline, and the equivalent bIoU metric of 0.39 ± 0.06 reported by [34]. Consequently, these results suggest SAR products as a robust and complementary alternative for archaeological mapping of *tell* sites. This approach can leverage the operational advantages inherent to Copernicus Sentinel-1 data, such as the availability of open-access, regularly updated global coverage, and all-weather imaging capabilities.

Nevertheless, we acknowledge that the current workflow does not exploit the full S-1 radar echo signal, as it relies on dual-polarimetric backscattering amplitude and omits phase-derived products such as InSAR coherence. Despite the complementary value of coherence products for landscape characterisation, integrating them into a large-scale

framework for archaeological site detection involves a trade-off among data availability, spatial resolution, and computational efficiency.

First, globally available S-1 coherence products [64] are currently limited to a 3 arcsec pixel spacing (around 90 m). This pixel spacing is suboptimal for identifying archaeological features such as *tell* sites, which have a median size of approximately 2 hectares. Secondly, to enhance reproducibility and scalability, remote sensing data preprocessing was conducted on the GEE platform, which can process only S-1 amplitude data. Generating custom coherence maps over extensive regional areas from S-1 IW Single Look Complex data would demand an ad hoc external preprocessing pipeline and be computationally intensive, especially since multitemporal median coherence products are needed for noise reduction. This would negate the advantages of a fully cloud-based preprocessing workflow, while the spatial multi-looking required for coherence estimation would inevitably degrade the spatial resolution compared to the original amplitude products [24]. These constraints prevented the integration of InSAR coherence into the current workflow, despite its value and its potential for automated site detection and contextual landscape characterisation, both of which remain not yet fully explored within archaeological remote sensing.

Finally, some of the remaining uncertainties may not be fully attributable to sensor characteristics or model design, but may reflect the nature of the *tell* mapping task itself and the extent to which *tell* sites can be represented within a binary segmentation framework. *Tell* sites are not discrete archaeological objects with clearly defined boundaries, but the result of long-term interactions between human occupation, geomorphological processes, and taphonomic dynamics, including erosion, accumulation, and modern agricultural practices. The transition between archaeological space and the surrounding landscape is often gradual and fuzzy by definition. This ambiguity is reflected in available datasets derived from legacy surveys and historical studies, where annotations may include topographic prominence, but also artifact scatters, associated uncultivated areas, or archaeological complexes containing multiple nearby *tells*. Remote sensing analyses themselves are often insufficient to delineate the broader archaeological spaces associated with *tells*, even when integrating multiple sources such as cartographic records, historical photographs, legacy imagery, and modern satellite acquisitions. From this perspective, the uncertainty observed in both the annotations and the model predictions may partly reflect ambiguity inherent to the archaeological landscape itself, rather than solely limitations of the Machine Learning algorithms.

5. Conclusions

In this study, we introduced a novel Deep Learning pipeline designed to automatically detect and segment *tell* sites using satellite SAR data products, which are still overlooked in automated archaeological analyses.

The YOLOv8-Seg pipeline proposed in this study demonstrated a promising level of performance, achieving an average precision of 0.49 and a pixel-wise IoU-score of 0.36 on independent test areas comprising 290 ground-truth *tell* sites. Archaeological interpretation of the model's inferences confirmed its potential as a valuable tool for large-scale archaeological prospection, specifically for mapping *tell* sites bigger than 0.3 hectares (3000 m²) across vast landscapes. The pipeline also allows for the identification of previously unmapped potential sites, demonstrating its usefulness for the preservation and protection of *tell* sites in regions where ground access is limited.

A primary strength of this pipeline is its reliance on freely available satellite data, which enhances its accessibility and transferability. We showed that the model can achieve comparable performance in geo-morphologically distinct regions such as Iran requiring only minimal fine-tuning with a limited number of local annotations. Moreover, this re-

search establishes a baseline for future investigations into the application of Deep Learning techniques to SAR data for archaeological research. Such work is timely as advances in Deep Learning for SAR are currently addressing challenges commonly associated with SAR data interpretation, promising to improve the visibility and interpretability of archaeological proxies through Deep Learning-based denoisers and provide new strategies for data fusion with LiDAR, optical, and historical data [26,28]. This work is also among the first in archaeological prospection to integrate an explainability module in the Deep Learning pipeline, increasing the transparency and trustworthiness of the AI's output.

Despite these achievements, the current methodology leaves room for further methodological refinement and to inform future research. First, in terms of SAR data exploitation, the proposed pipeline intentionally focused on SAR amplitude and ready-to-use InSAR products (i.e., the COP-30 DEM), while leaving additional phase-based information content, such as interferometric coherence products, for future integration. Moreover, during the preprocessing phase, these input data were converted to an 8-bit false RGB format, a practical choice that facilitated the integration with YOLOv8, but may not fully preserve their original numerical richness. [65]. Second, as an initial baseline, this study directly transferred a Deep Learning architecture designed for natural computer vision images to this archaeological remote sensing task. While this demonstrated good cross-domain applicability, future work could explore domain-specific architectural modifications or tailored loss and evaluation functions. Finally, model performance is directly linked to the characteristics of the available remote sensing data (e.g., data resolution and temporal gaps): the medium-resolution satellite products (10–30 m) may be limited in detecting smaller sites; the different temporal range of the input data (S-1 from 2021–2023 and COP-30 DEM from 2011–2015) may introduce uncertainty, as it assumes *tell* sites remained undisturbed over this decade despite ongoing urbanisation and agricultural expansion.

Future research will focus on overcoming some of the identified limitations by concentrating on two key directions: multimodality and data-efficient learning. As our findings strongly motivate the adoption of a multimodal approach, we would investigate innovative, natively remote sensing Deep Learning methods for the integration of optical, SAR, and topographic information. Furthermore, we would address the challenge of data scarcity and noise inherent in archaeological studies. This involves exploring the potential of Geospatial Foundation Models as a novel paradigm for Earth Observation [66]. Geospatial Foundation Models offer a promising path to building robust multimodal Deep Learning models, even in data-scarce scenarios, as the requirement for large, high-quality datasets in traditional supervised Deep Learning often limits its applicability to archaeological features.

Supplementary Materials: The following supporting information can be downloaded at: <https://www.mdpi.com/article/10.3390/rs18132255/s1>, Figure S1: Visual identification of a *tell* site across different satellite EO data products. Figure S2: Hyperparameter tuning of image size and batch size on the validation dataset. Figure S3: Hyperparameter tuning of the percentage of empty tiles on the validation dataset for the optimal image size and batch size configuration. Figure S4: Performance comparison between a YOLOv8-Seg model trained from scratch and a model initialised with COCO weights. Figure S5: Performance comparison between YOLOv8s-Seg and YOLOv11s-Seg. Figure S6: Spatial distribution of confidence levels for the validation dataset. Figure S7: Confidence level distribution for truncated and non truncated inferences that exceed the model's operational threshold. Figure S8: Qualitative comparison between YOLOv8-Seg predictions and the outputs generated via the Random Forest pixel-based classification on two samples of the independent test set in Iraq. Figure S9: Average feature importance across the five Random Forest model runs, derived using the Gini impurity criterion. Orange, blue, and green bars represent topographic, SAR amplitude-based, and optical features, respectively. Figure S10: Seasonal effects on the visibility of *tell* sites over S-1 dual-polarisation composites. Figure S11: Model's performance over the Iraqi test area for the dry

and humid season composite and the full-year composite. Figure S12: ROC curve for the proposed pipeline operating as a binary tile-level classifier for the presence or absence of *tell* sites. Table S1: Random Forest configuration and hyperparameters used for the pixel-based classification of *tell* sites. Table S2: Performance comparison between the Random Forest baseline and the YOLOv8-Seg pipeline on the independent Iraqi test set. Table S3: Performance metrics over the Iraqi test area for the dry and humid season composites and the full-year composite, calculated at the object-level and pixel-level. Table S4: Tile-level classification performance computed over the Iraqi test area, which comprises a total of 3150 tiles. References [67–70] are cited in the supplementary materials.

Author Contributions: Conceptualization, E.C., G.P., S.V. and A.T.; methodology, E.C. and S.V.; software, E.C.; validation, E.C. and G.P.; formal analysis, E.C.; data curation, E.C. and G.P.; writing—original draft preparation, E.C. and G.P.; writing—review and editing, E.C., G.P., S.F., S.V., and A.T.; visualization, E.C. and G.P.; supervision, S.V. and A.T. All authors have read and agreed to the published version of the manuscript.

Funding: This research received no external funding.

Data Availability Statement: The data supporting this study’s findings are available from the corresponding author, A.T., upon reasonable request.

Acknowledgments: We acknowledge that the research activity herein was carried out using the IIT Franklin HPC infrastructure; we gratefully acknowledge the Data Science and Computation Facility and its Support Team for their support and assistance on the IIT High Performance Computing Infrastructure.

Conflicts of Interest: The authors declare no conflicts of interest.

Abbreviations

The following abbreviations are used in this manuscript:

SAR	Synthetic Aperture Radar
TPI	Topographic Position Index
InSAR	Interferometric Synthetic Aperture Radar
DEM	Digital Elevation Model
VV	Vertical–Vertical Polarisation
VH	Vertical–Horizontal Polarisation
S-1	Sentinel-1
S-2	Sentinel-2
COP-30 DEM	Copernicus Global 30-m DEM

References

1. Comer, D.C.; Harrower, M.J. *Mapping Archaeological Landscapes from Space*; Springer: New York, NY, USA, 2013; Volume 5.
2. Cigna, F.; Tapete, D.; Lasaponara, R.; Masini, N. Amplitude change detection with ENVISAT ASAR to image the cultural landscape of the Nasca region, Peru. *Archaeol. Prospect.* **2013**, *20*, 117–131. [\[CrossRef\]](#)
3. Linck, R.; Busche, T.; Buckreuss, S.; Fassbinder, J.; Seren, S. Possibilities of archaeological prospection by high-resolution X-band satellite radar—A case study from Syria. *Archaeol. Prospect.* **2013**, *20*, 97–108. [\[CrossRef\]](#)
4. Dore, N.; Patruno, J.; Pottier, E.; Crespi, M. New research in polarimetric SAR technique for archaeological purposes using ALOS PALSAR data. *Archaeol. Prospect.* **2013**, *20*, 79–87. [\[CrossRef\]](#)
5. Chen, F.; Masini, N.; Yang, R.; Milillo, P.; Feng, D.; Lasaponara, R. A space view of radar archaeological marks: First applications of COSMO-SkyMed X-band data. *Remote Sens.* **2014**, *7*, 24–50. [\[CrossRef\]](#)
6. Tapete, D.; Cigna, F. SAR for landscape archaeology. In *Sensing the Past: From Artifact to Historical Site*; Springer: Cham, Switzerland, 2017; pp. 101–116. [\[CrossRef\]](#)
7. Paillou, P. Mapping palaeohydrography in deserts: Contribution from space-borne imaging radar. *Water* **2017**, *9*, 194. [\[CrossRef\]](#)
8. Chen, F.; Lasaponara, R.; Masini, N. An overview of satellite synthetic aperture radar remote sensing in archaeology: From site detection to monitoring. *J. Cult. Herit.* **2017**, *23*, 5–11. [\[CrossRef\]](#)
9. Tapete, D.; Cigna, F. COSMO-SkyMed SAR for detection and monitoring of archaeological and cultural heritage sites. *Remote Sens.* **2019**, *11*, 1326. [\[CrossRef\]](#)

10. Tapete, D.; Traviglia, A.; Delpozzo, E.; Cigna, F. Regional-scale systematic mapping of archaeological mounds and detection of looting using COSMO-SkyMed high resolution DEM and satellite imagery. *Remote Sens.* **2021**, *13*, 3106. [\[CrossRef\]](#)
11. Tapete, D.; Cigna, F. Detection, morphometric analysis and digital surveying of archaeological mounds in southern Iraq with CartoSat-1 and COSMO-SkyMed DEMs. *Land* **2022**, *11*, 1406. [\[CrossRef\]](#)
12. Cigna, F.; Balz, T.; Tapete, D.; Caspari, G.; Fu, B.; Abballe, M.; Jiang, H. Exploiting satellite SAR for archaeological prospection and heritage site protection. *Geo-Spat. Inf. Sci.* **2024**, *27*, 526–551. [\[CrossRef\]](#)
13. Cigna, F.; Tapete, D. Archaeological Prospection and Site Monitoring with Medium to Very High Resolution SAR Imagery: Case Studies in Rome (Italy). In *Proceedings of the IGARSS 2024-2024 IEEE International Geoscience and Remote Sensing Symposium*; IEEE: Piscataway, NJ, USA, 2024; pp. 3384–3387. [\[CrossRef\]](#)
14. Cigna, F.; Lasaponara, R.; Masini, N.; Milillo, P.; Tapete, D. Persistent scatterer interferometry processing of COSMO-SkyMed StripMap HIMAGE time series to depict deformation of the historic centre of Rome, Italy. *Remote Sens.* **2014**, *6*, 12593–12618. [\[CrossRef\]](#)
15. Tapete, D.; Cigna, F.; Donoghue, D.N. ‘Looting marks’ in space-borne SAR imagery: Measuring rates of archaeological looting in Apamea (Syria) with TerraSAR-X Staring Spotlight. *Remote Sens. Environ.* **2016**, *178*, 42–58. [\[CrossRef\]](#)
16. Chen, F.; Guo, H.; Ma, P.; Lin, H.; Wang, C.; Ishwaran, N.; Hang, P. Radar interferometry offers new insights into threats to the Angkor site. *Sci. Adv.* **2017**, *3*, e1601284. [\[CrossRef\]](#) [\[PubMed\]](#)
17. Zaina, F.; Tapete, D. Satellite-based methodology for purposes of rescue archaeology of cultural heritage threatened by dam construction. *Remote Sens.* **2022**, *14*, 1009. [\[CrossRef\]](#)
18. Richards, J.A. *Remote Sensing with Imaging Radar*; Springer: Berlin/Heidelberg, Germany, 2009; Volume 1.
19. Long, D.; Ulaby, F. *Microwave Radar and Radiometric Remote Sensing*; University of Michigan Press: Ann Arbor, MI, USA, 2015.
20. Woodhouse, I.H. *Introduction to Microwave Remote Sensing*; CRC Press: Boca Raton, FL, USA, 2017.
21. Shimada, M. *Imaging from Spaceborne and Airborne SARs, Calibration, and Applications*; CRC Press: Boca Raton, FL, USA, 2018.
22. Flores-Anderson, A.I.; Herndon, K.E.; Thapa, R.B.; Cherrington, E. *The SAR Handbook: Comprehensive Methodologies for Forest Monitoring and Biomass Estimation*; NASA: Washington, DC, USA, 2019.
23. National Aeronautics and Space Administration; Indian Space Research Organisation. *NASA-ISRO SAR (NISAR) Mission Science Users’ Handbook*, 2nd ed.; NASA Jet Propulsion Laboratory: Pasadena, CA, USA, 2025.
24. Ferretti, A.; Monti-Guarnieri, A.; Prati, C.; Rocca, F.; Massonet, D. *InSAR Principles-Guidelines for SAR Interferometry Processing and Interpretation*; ESA Publications: Noordwijk, The Netherlands, 2007.
25. Orengo, H.A.; Conesa, F.C.; Garcia-Molsosa, A.; Lobo, A.; Green, A.S.; Madella, M.; Petrie, C.A. Automated detection of archaeological mounds using machine-learning classification of multisensor and multitemporal satellite data. *Proc. Natl. Acad. Sci. USA* **2020**, *117*, 18240–18250. [\[CrossRef\]](#)
26. Zhu, X.X.; Montazeri, S.; Ali, M.; Hua, Y.; Wang, Y.; Mou, L.; Shi, Y.; Xu, F.; Bamler, R. Deep learning meets SAR: Concepts, models, pitfalls, and perspectives. *IEEE Geosci. Remote Sens. Mag.* **2021**, *9*, 143–172. [\[CrossRef\]](#)
27. Schmitt, M.; Hänsch, R. *Deep Learning for Synthetic Aperture Radar Remote Sensing*; Elsevier: Amsterdam, The Netherlands, 2025.
28. Datcu, M.; Huang, Z.; Anghel, A.; Zhao, J.; Căcovăanu, R. Explainable, physics-aware, trustworthy artificial intelligence: A paradigm shift for synthetic aperture radar. *IEEE Geosci. Remote Sens. Mag.* **2023**, *11*, 8–25. [\[CrossRef\]](#)
29. van der Vaart, W.B.V.; Lambers, K. Learning to look at LiDAR: The use of R-CNN in the automated detection of archaeological objects in LiDAR data from the Netherlands. *J. Comput. Appl. Archaeol.* **2019**, *2*, 31–40. [\[CrossRef\]](#)
30. Fiorucci, M.; Verschoof-Van Der Vaart, W.B.; Soleni, P.; Le Saux, B.; Traviglia, A. Deep learning for archaeological object detection on LiDAR: New evaluation measures and insights. *Remote Sens.* **2022**, *14*, 1694. [\[CrossRef\]](#)
31. Kokalj, Ž.; Džeroski, S.; Šprajc, I.; Štajdohar, J.; Draksler, A.; Somrak, M. Machine learning-ready remote sensing data for Maya archaeology. *Sci. Data* **2023**, *10*, 558. [\[CrossRef\]](#) [\[PubMed\]](#)
32. Casini, L.; Orrù, V.; Roccetti, M.; Marchetti, N. When machines find sites for the archaeologists: A preliminary study with semantic segmentation applied on satellite imagery of the Mesopotamian floodplain. In *Proceedings of the 2022 ACM Conference on Information Technology for Social Good*; ACM: New York, NY, USA, 2022; pp. 378–383. [\[CrossRef\]](#)
33. Casini, L.; Marchetti, N.; Montanucci, A.; Orrù, V.; Roccetti, M. A human–AI collaboration workflow for archaeological sites detection. *Sci. Rep.* **2023**, *13*, 8699. [\[CrossRef\]](#) [\[PubMed\]](#)
34. Pistola, A.; Orrù, V.; Marchetti, N.; Roccetti, M. AI-ming backwards: Vanishing archaeological landscapes in Mesopotamia and automatic detection of sites on CORONA imagery. *PLoS ONE* **2025**, *20*, e0330419. [\[CrossRef\]](#) [\[PubMed\]](#)
35. Matthews, W. Tells in Archaeology. In *Encyclopedia of Global Archaeology*, 2nd ed.; Smith, C., Ed.; Springer: Cham, Switzerland, 2014; pp. 10553–10556.
36. Wilkinson, T.J. *Archaeological Landscapes of the Near East*; University of Arizona Press: Tucson, AZ, USA, 2003.
37. Marchetti, N.; Bortolini, E.; Menghi Sartorio, J.C.; Orrù, V.; Zaina, F. Long-term urban and population trends in the Southern Mesopotamian floodplains. *J. Archaeol. Res.* **2025**, *33*, 117–158. [\[CrossRef\]](#)

38. Zaina, F. A risk assessment for cultural heritage in southern Iraq: Framing drivers, threats and actions affecting archaeological sites. *Conserv. Manag. Archaeol. Sites* **2019**, *21*, 184–206. [\[CrossRef\]](#)
39. Cigna, F.; Rayne, L.; Makovics, J.L.; Irvine, H.K.; Jotheri, J.; Algabri, A.; Tapete, D. Environmental Challenges and Vanishing Archaeological Landscapes: Remotely Sensed Insights into the Climate–Water–Agriculture–Heritage Nexus in Southern Iraq. *Land* **2025**, *14*, 1013. [\[CrossRef\]](#)
40. Iacobucci, G.; Troiani, F.; Milli, S.; Nadali, D. Geomorphology of the lower Mesopotamian plain at Tell Zurghul archaeological site. *J. Maps* **2023**, *19*, 2112772. [\[CrossRef\]](#)
41. Nadali, D. Cities in the Water: Waterscape and Evolution of Urban Civilisation in Southern Mesopotamia as Seen from Tell Zurghul, Iraq. In *Southern Iraq's Marshes*; Jawad, L.A., Ed.; Series Title: Coastal Research Library; Springer International Publishing: Cham, Switzerland, 2021; Volume 36, pp. 15–31. [\[CrossRef\]](#)
42. The World Bank Group. World Bank Climate Change Knowledge Portal: Iraq. Available online: <https://climateknowledgeportal.worldbank.org/country/iraq> (accessed on 6 March 2026).
43. Zanaga, D.; Van De Kerchove, R.; Daems, D.; De Keersmaecker, W.; Brockmann, C.; Kirches, G.; Wevers, J.; Cartus, O.; Santoro, M.; Fritz, S.; et al. *ESA WorldCover 10 m 2021 v200*, 2022. Data Set. Available online: <https://zenodo.org/records/7254221> (accessed on 6 March 2026)
44. Alizadeh, A.; Kouchoukos, N.; Bauer, A.M.; Wilkinson, T.J.; Mashkour, M. Human-environment interactions on the Upper Khuzestan Plains, southwest Iran. Recent investigations. *Paléorient* **2004**, *30*, 69–88. [\[CrossRef\]](#)
45. Rashidian, E. The Dez-Karun-Confluence in Lower Susiana SW Iran) in the Last Millennia; a Case Study for the Human-Environment-Interaction. *Iran. J. Archaeol. Stud.* **2021**, *11*, 9–23. [\[CrossRef\]](#)
46. Copernicus Data Space Ecosystem. Copernicus DEM—Global and European Digital Elevation Model | Copernicus Data Space Ecosystem. Available online: <https://dataspace.copernicus.eu/explore-data/data-collections/copernicus-contributing-missions/collections-description/COP-DEM> (accessed on 6 March 2026).
47. Bielski, C.; López-Vázquez, C.; Grohmann, C.H.; Guth, P.L.; Hawker, L.; Gesch, D.; Trevisani, S.; Herrera-Cruz, V.; Riazanoff, S.; Corseaux, A.; et al. Novel approach for ranking DEMs: Copernicus DEM improves one arc second open global topography. *IEEE Trans. Geosci. Remote Sens.* **2024**, *62*, 4503922. [\[CrossRef\]](#)
48. *ArcGIS Pro*, Version 3.0.3; Environmental Systems Research Institute (Esri): Redlands, CA, USA, 2022.
49. Pedersén, O. Ancient Near East on Google Earth: Problems, Preliminary Results, and Prospects. In *Proceedings of the 7th International Congress on the Archaeology of the Ancient Near East*; Harrassowitz: Wiesbaden, Germany, 2012; Volume 3, pp. 385–393.
50. Pedersén, O. ANE Site Placemarks for Google Earth (Version 10), 2007. Data Set. Available online: <https://zenodo.org/records/6384045>. (accessed on 6 March 2026).
51. Marchetti, N.; Baldassarri, P.; Bertossa, S.; Orrù, V.; Valeri, M.; Zaina, F. FloodPlains. Developing a public archaeological WebGIS for the Southern Mesopotamian alluvium. In *Proceedings of the 12th International Congress on the Archaeology of the Ancient Near East*; Harrassowitz: Wiesbaden, Germany, 2023; Volume 1, pp. 921–939. [\[CrossRef\]](#)
52. Jocher, G.; Chaurasia, A.; Qiu, J. *Ultralytics Yolo*, Version 8.0.0; Ultralytics: Frederick, MD, USA, 2023.
53. Bolya, D.; Zhou, C.; Xiao, F.; Lee, Y.J. Yolact: Real-time instance segmentation. In *Proceedings of the 2019 IEEE/CVF International Conference on Computer Vision (ICCV)*; IEEE: Piscataway, NJ, USA, 2019; pp. 9157–9166. [\[CrossRef\]](#)
54. Gorelick, N.; Hancher, M.; Dixon, M.; Ilyushchenko, S.; Thau, D.; Moore, R. Google Earth Engine: Planetary-scale geospatial analysis for everyone. *Remote Sens. Environ.* **2017**, *202*, 18–27. [\[CrossRef\]](#)
55. Lin, T.Y.; Maire, M.; Belongie, S.; Hays, J.; Perona, P.; Ramanan, D.; Dollár, P.; Zitnick, C.L. Microsoft coco: Common objects in context. In *Proceedings of the European Conference on Computer Vision*; Springer: Cham, Switzerland, 2014; pp. 740–755. [\[CrossRef\]](#)
56. Tuia, D.; Roscher, R.; Wegner, J.D.; Jacobs, N.; Zhu, X.; Camps-Valls, G. Toward a collective agenda on AI for earth science data analysis. *IEEE Geosci. Remote Sens. Mag.* **2021**, *9*, 88–104. [\[CrossRef\]](#)
57. Tuia, D.; Schindler, K.; Demir, B.; Zhu, X.X.; Kochupillai, M.; Džeroski, S.; van Rijn, J.N.; Hoos, H.H.; Del Frate, F.; Datcu, M.; et al. Artificial Intelligence to Advance Earth Observation: A review of models, recent trends, and pathways forward. *IEEE Geosci. Remote Sens. Mag.* **2024**, *13*, 119–141. [\[CrossRef\]](#)
58. Höhl, A.; Obadic, I.; Fernandez-Torres, M.A.; Najjar, H.; Oliveira, D.A.B.; Akata, Z.; Dengel, A.; Zhu, X.X. Opening the Black Box: A systematic review on explainable artificial intelligence in remote sensing. *IEEE Geosci. Remote Sens. Mag.* **2024**, *12*, 261–304. [\[CrossRef\]](#)
59. Chollet, F. *Deep Learning with Python*; Manning Publications Co.: Shelter Island, NY, USA, 2018.
60. Molnar, C. Interpretable Machine Learning. Available online: <https://christophm.github.io/interpretable-ml-book/> (accessed on 6 March 2026).
61. Muhammad, M.B.; Yeasin, M. Eigen-cam: Class activation map using principal components. In *Proceedings of the 2020 International Joint Conference on Neural Networks (IJCNN)*; IEEE: Piscataway, NJ, USA, 2020; pp. 1–7. [\[CrossRef\]](#)
62. Rigved, S. YOLO-V12-CAM. Available online: <https://github.com/rigvedrs/YOLO-V12-CAM> (accessed on 6 March 2026).

63. Van der Walt, S.; Schönberger, J.L.; Nunez-Iglesias, J.; Boulogne, F.; Warner, J.D.; Yager, N.; Gouillart, E.; Yu, T. scikit-image: Image processing in Python. *PeerJ* **2014**, *2*, e453. [[CrossRef](#)] [[PubMed](#)]
64. Kellndorfer, J.; Cartus, O.; Laval, M.; Magnard, C.; Milillo, P.; Oveisgharan, S.; Osmanoglu, B.; Rosen, P.A.; Wegmüller, U. Global seasonal Sentinel-1 interferometric coherence and backscatter data set. *Sci. Data* **2022**, *9*, 73. [[CrossRef](#)] [[PubMed](#)]
65. Rolf, E.; Klemmer, K.; Robinson, C.; Kerner, H. Position: Mission critical—Satellite data is a distinct modality in machine learning. In *Proceedings of the Forty-First International Conference on Machine Learning*; JMLR.org; ACM Digital Library: New York, NY, USA, 2024; Volume 235, pp. 42691–42706. <https://dl.acm.org/doi/10.5555/3692070.3693807>. [[CrossRef](#)]
66. Zhu, X.X.; Xiong, Z.; Wang, Y.; Stewart, A.J.; Heidler, K.; Wang, Y.; Yuan, Z.; Dujardin, T.; Xu, Q.; Shi, Y. On the foundations of Earth foundation models. *Commun. Earth Environ.* **2026**, *7*, 103.. [[CrossRef](#)]
67. Leberl, F. *Radargrammetric Image Processing*; Artech House: Norwood, MA, USA, 1989.
68. Breiman, L. Random forests. *Mach. Learn.* **2001**, *45*, 5–32. [[CrossRef](#)]
69. Menze, B.H.; Ur, J.A. Mapping patterns of long-term settlement in Northern Mesopotamia at a large scale. *Proc. Natl. Acad. Sci. USA* **2012**, *109*, E778–E787. [[CrossRef](#)] [[PubMed](#)]
70. Food and Agriculture Organization. GIEWS—Global Information and Early Warning System Country Brief: Iraq. Available online: <https://www.fao.org/giews/countrybrief/country.jsp?lang=en&code=IRQ> (accessed on 6 March 2026).

Disclaimer/Publisher’s Note: The statements, opinions and data contained in all publications are solely those of the individual author(s) and contributor(s) and not of MDPI and/or the editor(s). MDPI and/or the editor(s) disclaim responsibility for any injury to people or property resulting from any ideas, methods, instructions or products referred to in the content.