

Semi-automatic detection of building stones and wall segments of archaeological ruins

Erel Uziel^{a,*}, Motti Zohar^b, Yitzchak Jaffe^a

^a School of Archaeology and Maritime Cultures, University of Haifa, Israel

^b School of Environmental Sciences, University of Haifa, Israel

ABSTRACT

This study presents a semi-automatic methodology for detecting building stones and wall segments in archaeological research, using drone imagery and deep learning algorithms. The immediate outputs of the methodology are a georeferenced stones layer, with each stone detected as a separate instance, and a site plan layer, composed of stones considered part of detected wall segments. We developed this model via nine sites of varying size with different vegetation coverage, ground color, and material composition, exemplifying the model's ability to perform successfully even in challenging conditions. The digital layers, along with additional attributes associated with each shape, provide a foundation for further analysis, such as identifying multiple construction styles and site organization patterns, with significant potential for large-scale multi-site studies. Evaluation results demonstrate good model performance under varied conditions. We also provide trained models, trained on data from multiple sites, for immediate use and further refinement.

1. Introduction

1.1. Background

Documenting site features is a critical component of archaeological field reconnaissance, yet mapping standing buildings, architectural features, and wall segments is often time-consuming and methodologically challenging. While high-precision data collection is ideal, its labor-intensive nature means that in large-scale surveys, teams often resort to minimal documentation—basic descriptions, rough drawings, and provisional designations for future, more intensive investigations when time and resources permit. Approaches to mapping vary widely, from rough field sketches and simple outlines to GPS-defined perimeters and detailed 3D models generated with mobile Real-Time Kinematic (RTK) units (Attema et al., 2020; Banning, 2002; Lindsey, 2020; Knodell et al., 2023). Even so, loose scatters of stones or building materials that are not clearly in situ, or that lack distinctive characteristics (e.g., diagnostic stylistic features or datable elements), are often left undocumented. Peacock (1998), nicely captures the commonly limited attention given to building stones, “Firstly, in stone studies there is a need to proselytize, to attempt to ‘sell’ the subject by drawing attention to the wealth of information in stone finds. Secondly, there are many facets to stone study that are often seen as discrete entities [...] All too often stone had an ‘unjustifiably low’ priority in the unit’s work [...]”.

Unmanned Aerial Vehicles (UAVs or drones) have become a common

tool for archaeologists, providing high-resolution imagery, 3D models, Digital Elevation Models (DEMs), Digital Surface Models (DSMs), and more. Researchers are experimenting and evaluating their usefulness in cutting edge applications such as: large-area surveying (Homsher et al., 2017; Hammer, 2022), automated detection of small artifacts such as potsherds using machine learning algorithms (Orengo and Garcia-Molsosa, 2019; Orengo et al., 2021), automated mapping of cultural heritage (Trier et al., 2021), excavation documentation, 3D modelling of buildings (Sauerbier and Eisenbeiss, 2010), and the study of human activity in areas between conventional sites (Casana et al., 2023).

New and improved drones are released every year, pushing the boundaries of aerial photography. Longer flight times, higher-quality cameras, and enhanced flight-planning software now enable archaeologists to capture higher-resolution imagery over wider areas. Sub-centimeter Ground Sampling Distances (GSDs)—the real-world resolution per pixel—can now be achieved (Gilliot et al., 2022), compared to the multi-meter GSDs of satellite imagery (Lin et al., 2024) and the multi-centimeter GSDs of traditional manned photogrammetry (Pepe et al., 2018). UAVs equipped with beyond-visible-spectrum cameras assist in locating buried structures concealed by vegetation (Carmona et al., 2020; Vinci et al., 2025) and detecting underground anomalies (Waagen and van Hilst, 2025). LiDAR-equipped drones enable penetration of dense vegetation and detection of buried structures (Frachetti et al., 2024), identification of agricultural features and terraces (Casana

* Corresponding author.

E-mail address: euziel@humanities.haifa.ac.il (E. Uziel).

et al., 2023; Wang et al., 2022).

To the best of our knowledge, the automatic detection of stones using drone imagery and machine learning algorithms is a novel approach. This solution has only recently become feasible due to technological advances, including access to very high-resolution imagery, modern

computer vision algorithms, and increased computational power. Previous works, such as Trier et al. (2021), Altaweel et al. (2022), and Monna et al. (2020), combined machine learning with drone imagery to detect architectural features such as burial mounds. However, these methods operate at much lower resolutions and are not designed to

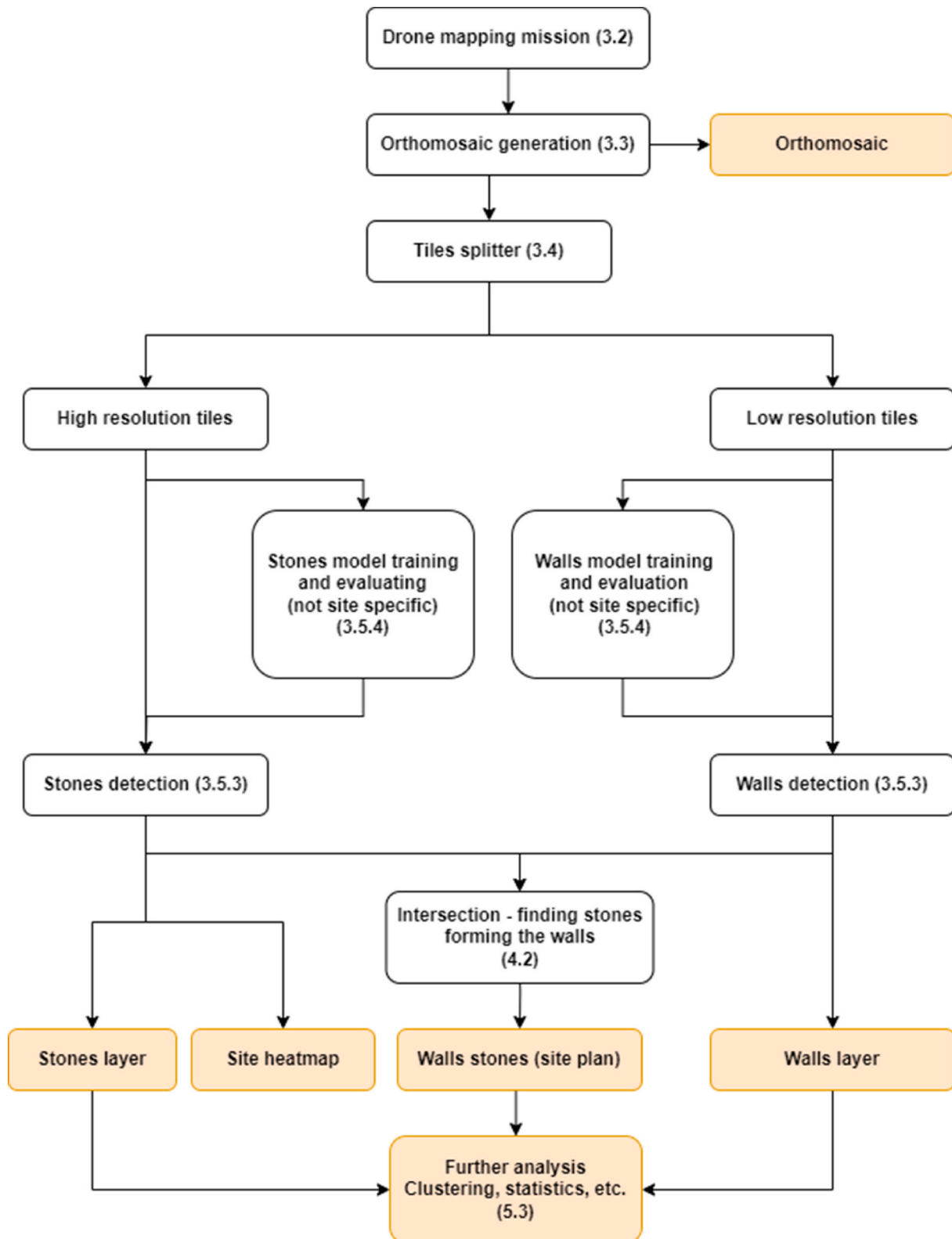


Fig. 1. Methodology steps and outputs. Outputs marked with orange fill. Relevant sections are indicated. (For interpretation of the references to color in this figure legend, the reader is referred to the Web version of this article.)

detect individual stones along with their shapes and characteristics, which, together with wall segment detection, form the basis for site plan generation and stone-level analysis.

This study presents a semi-automatic methodology for detecting building stones and wall segments in archaeological research, using drone imagery and deep learning algorithms. Georeferenced, digitized data can be imported into Geographic Information Systems (GIS), which are widely used by archaeologists for data management and spatial analysis (Menéndez-Marsh et al., 2023), enabling spatial analysis such as visibility analysis between multiple sites (Shmidov et al., 2023) and interregional research (Wernke et al., 2020). A digitized, georeferenced layer of stones, particularly when combined with a layer of wall segment, can unlock important insights. These include stone-level site plan, pattern detection, improved site documentation, intra-site and regional stones and walls comparisons, and more. Thus, by simplifying accurate stone mapping, important archaeological information can be extracted during surveys and excavations, information that might otherwise be missed without intensive efforts.

1.2. Research goals

The main goals of our research are to: (1) provide an end-to-end methodology for generating georeferenced layers of site stones and walls; (2) train models on multiple sites and conditions that can be used as-is or further extended; and (3) demonstrate potential applications of the digitized layers.

Stone detection presents challenges due to the wide variability in size, shape, color, contrast, and surrounding environmental conditions. The detection of wall segments requires additional considerations to allow for flexible interpretation, as discussed in the evaluation section. Analyzing full-site high-resolution imagery further necessitates the use of optimized, task-specific algorithms and implementations.

New opportunities come with new challenges, including the need to understand the strengths and limitations of each method (Field et al., 2017; Casana, 2021). We regard this methodology as a tool to assist archaeologists in gaining insights more easily and in opening opportunities for further research. While stone detection results are impressive it is important to remember that wall detection, in particular, involves interpretation that is inherently archaeological and depends on broader contextual understanding.

2. Materials and methods

2.1. Overview

The various steps and outputs of our proposed methodology are highlighted in Fig. 1. The main steps are: (1) in-field images acquisition with drones; (2) orthomosaic generation; (3) splitting of the orthomosaic into georeferenced tiles; (4) stone model training and wall model training (not site-specific); (5) stone and wall detection; (6) intersection between the stone and wall layers; and (7) further analysis of the various outputs.

2.2. Images acquisition

The process begins with mission planning of the drone flight, followed by on-site execution. The drone operates based on predefined parameters, including flight altitude, speed, and image overlap (Mouget & Lucet, 2014; Rouse & Krumnow, 2020).

The GSD is determined by the drone camera specifications and flight altitude. As a rule of thumb, a minimum ratio of approximately 1:10 should be maintained between the target artifact size and the GSD (Field et al., 2017), particularly for high-contrast artifacts. In our case, as we aim to detect stones—many of which lack strong contrast, are partially covered, and with minimum dimensions of approximately 15 cm × 15 cm—our target GSD is around 0.7 cm/pixel. With commonly available

off-the-shelf drones such as the DJI Mavic 2 Pro and DJI Mavic 3 Enterprise, this resolution is typically achieved at altitudes between 25 and 30 m. For extensive areas with trees and other obstacles, and to ensure reliable connectivity, this range represents the practical base limit for safe flight operations. Special consideration and planning are required for non-flat terrain, as drone altitude is measured relative to the take-off point rather than the actual ground elevation beneath the drone (Orengo et al., 2021). Thus, the waypoint height of each point should be adjusted based on the relative terrain altitude.

2.3. Orthomosaic generation

The hundreds of acquired images are processed using photogrammetry software that identifies matching points between overlapping images. We performed this step using Pix4Dmapper software. Based on the stored image coordinates, camera specifications, and manually placed ground control points—georeferenced ground points that improve the accuracy, the software generates a georeferenced orthomosaic. In addition, it produces a 3D point cloud, DSM, and DEM. All these outputs are valuable for further site analysis (e.g., Homsher et al., 2017).

2.4. Tiles splitter

This step, along with the subsequent ones, was implemented in Python. The open-source project is available on Zenodo. The generated orthomosaic is stored as a single large image with embedded metadata. Image dimensions can exceed 50,000 × 50,000 pixels, with file sizes reaching several gigabytes. This large image along with the coordinates of its top-left corner, coordinate reference system, target scale, and margin settings, serves as the input to the “tiles splitter” tool. The tool rescales the image according to the source and target scales and divides it into tiles with predefined overlap. By setting the target scale to match the scale used during training, the trained model can be reused. The output consists of two georeferenced tile sets: one for stone detection and the other for wall detection. These tiles are used for model training, evaluation, and inference. Each set of tiles is stored in a dedicated folder, accompanied by a JSON file containing spatial metadata, enabling the geolocation of any detected shape within it. In contrast to stone detection, which requires high-resolution imagery, wall detection benefits more from broader spatial context than from maximum resolution. Therefore, the tiles for wall detection are rescaled to a lower resolution of approximately 3 cm/pixel. Each tile measures 800 × 800 pixels, corresponding to an area of approximately 25 × 25 square meters—suitable for detecting wall segments several meters in length.

2.5. Object detection model

2.5.1. Mask R-CNN overview

Mask R-CNN (Mask Region-based Convolutional Neural Network) is a deep learning algorithm that extends Faster R-CNN (Ren et al., 2015). Introduced by He et al. (2017), it provides a framework for object instance segmentation. The model combines *object detection*—the task of classifying individual objects using bounding boxes—with *semantic segmentation*, which involves classifying each pixel without distinguishing between object instances. It works in two stages: the Region Proposal Network (RPN) proposes object locations and these regions are used for classification, bounding box fine tuning and mask prediction. The output is a set of masks corresponding to the detected and classified objects.

Mask R-CNN employs a dedicated convolutional neural network (CNN) backbone to extract features from input images. In our implementation, we utilized ResNet-50 (He et al., 2016) as the backbone. Both ResNet-50 and the Mask R-CNN model incorporating the ResNet-50 backbone were initialized with pretrained weights. The weights of the final classifier layers were discarded, and only these layers were trained

on our dataset. This strategy facilitates effective fine-tuning of the model using a relatively small dataset, while still achieving strong performance, a concept referred to as transfer learning (Panda et al., 2024).

Mask R-CNN is a widely used algorithm in archaeological research. Examples include its use as a generic tool for feature detection (Altaweel et al., 2022), for detecting gold foil shedding (Hou et al., 2024), for identifying archaeological structures in LiDAR data (Guyot et al., 2021), and in various other applications.

In our case, we use two separate, fine-tuned models: one for stones detection and the other for wall segments detection. Although the tasks are related, there are subtle differences between them, which will be explained in the next sections.

2.5.2. Features thresholds

The model applies thresholds to exclude stones or wall segments that may be considered less relevant to certain types of archaeological research. Thresholds are especially important for stone detection, as they help exclude non-building stones such as surface bedrock. The first threshold concerns the minimum and maximum shape area in which stones that are smaller or larger than the defined range are discarded. A similar threshold is set for wall segments. The second threshold is the ratio between the stone area and the area of its enclosing Minimum Rotated Rectangle (MRR) (Fig. 2A)—the smallest rectangle that fully contains the polygon. This value ranges from 0 to 1, with values closer to 1 indicating that the stone's shape is more rectangular. The optimal values for these thresholds depend on the specific characteristics of the site.

2.5.3. From predicted masks to georeferenced shapes

Mask R-CNN, in our implementation, takes as an input a 3D matrix of size [3, 800, 800], containing the three-channel values for each pixel within the 800x800 tile. It returns a 3D matrix of size [n, 800, 800], where n is the number of inferred masks, with each pixel holding a confidence value. Several additional steps are required to convert these

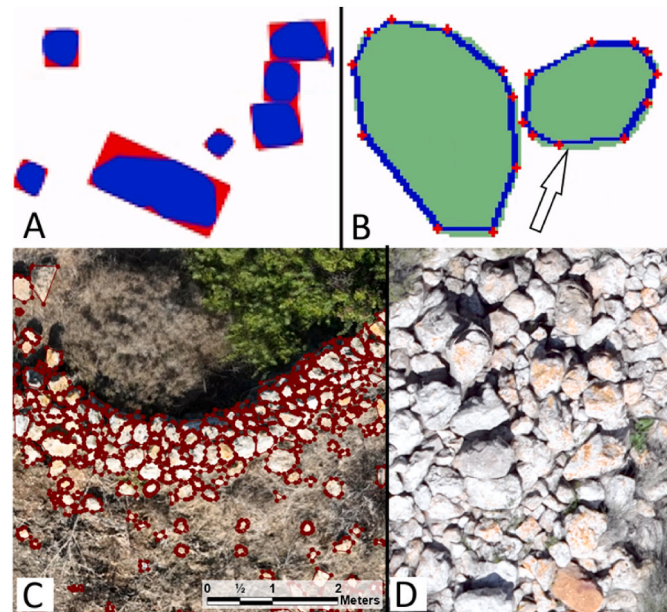


Fig. 2. (A) Stone polygons (blue) enclosed by MRRs (red). (B) Mask R-CNN predicts masks (green), which are approximated as polygons (blue lines with red points) using the Douglas-Peucker algorithm. The arrow highlights the difference between the predicted mask and the approximated polygon. (C) An annotated tile used for the stones model training. A single tile can include hundreds of stones. (D) An example of a stones pile from Khirbet Dorban, resulting from a collapsed wall. (For interpretation of the references to color in this figure legend, the reader is referred to the Web version of this article.)

results into georeferenced shapes. The first optional step involves merging overlapping shapes and subsequently removing shapes based on predefined area thresholds. The second step is converting the masks to polygons, which is done using the OpenCV library (Bradski, 2000). The contours of the detected masks are extracted, and then approximated with polygons using the Douglas-Peucker algorithm (Douglas & Peucker, 1973) (Fig. 2B). This is followed by an optional step of removing shapes close to the boundaries based on the polygons' centroids. The tiles splitter produces overlapping tiles to facilitate the removal of partially detected stones. The polygons are georeferenced using the tile's metadata. The polygons are then added to a GeoPandas GeoDataFrame (Jordahl et al., 2020), which stores the results from all inferred tiles. The GeoDataFrame can be used directly for further analysis in Python or exported to other GIS applications.

2.5.4. Training

Training the two models requires the creation of two annotated datasets. Tiles from multiple sites were copied into separate folders, one for each model. We selected tiles with diverse features and characteristics, including stones of varying sizes, colors, and densities; different types of ground coverage; different types of walls; and both standing and collapsed walls (Fig. 7). The tiles were annotated using a tool called Labelme (Wada, 2021), an open-source software that enables the addition of annotated shape layers to images. These annotation layers, saved as JSON files, are read by our code and used for training and evaluating the models. As shown in Fig. 2C, a single tile can contain hundreds of stones that require annotation. To assist with the labor-intensive task of marking hundreds of polygons, we employed a “human-in-the-loop” approach (Wu et al., 2022), in which the model infers the tile and outputs a JSON file containing the detected polygons. This file can be read by the Labelme tool for further correction, automating the majority of the annotation work.

Annotating the wall tiles presents an archaeological challenge. Determining whether a series of stones constitutes a wall segment is ultimately an archaeological judgment, which depends on the archaeologist's perception and the specific research question. This is particularly difficult in poorly preserved sites, where only a single course of stones may be visible. The methodology offers flexibility by allowing archaeologists to annotate wall tiles based on their own interpretations. In our training, the model was exposed primarily to walls with clear contours. While most walls are relatively straight, curved walls were also included in the training and evaluation data. Fig. 4 (C, D) illustrates the difficulties posed by collapsed walls without clear contours. In our training data we employed a conservative approach that identifies walls with high confidence – based on well preserved segments.

The annotated dataset is randomly split, with 80 % of the tiles used for training and 20 % for testing, a commonly used and effective ratio (Gholamy et al., 2018). The training code is based on the PyTorch torchvision source code (Paszke et al., 2019). To help avoid overfitting, we applied data augmentation methods to the training dataset. The images were randomly rotated, scaled and flipped. Table 1 summarizes the data used for training and evaluation.

The archaeological sites used for training and evaluation are shown in Fig. 3 and described in Appendix B. In general, they are all located in the lower Galilee and Shephelah regions, which are typified by Mediterranean vegetation including low oak, pines, and carobs, as well as shrubs of maquis and mastic (*Pistacia lentiscus*), on terra rossa and rendzina soils, over limestones and dolomites. The sites were mapped over the course of two years, covering different seasons and

Table 1

Models training data. 80 % used for training and 20 % for evaluation.

Model	Sites	Tiles	Shapes
Stones	9	44	5,563
Walls	9	56	673

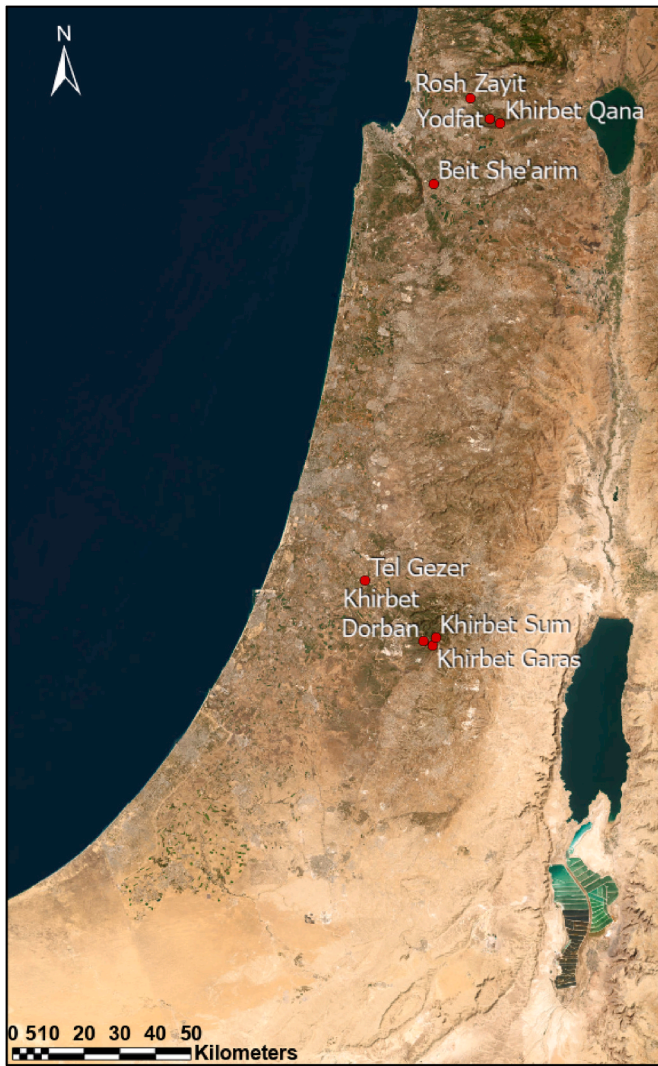


Fig. 3. Archaeological sites used for training and evaluation. Image created with ArcGIS Pro.

environmental settings (Fig. 7). All sites contained visible architectural features, and most were either unexcavated or only partially excavated.

2.5.5. Evaluation

2.5.5.1. Metrics. The following metrics were used to evaluate the two models:

$$\text{IoU} = \frac{P \cap G}{P \cup G} \quad (1)$$

Intersection ($\cap$) over Union ($\cup$) is a common metric for comparing the annotated (ground truth) mask and the predicted mask. **P** is the predicted mask area and **G** is the ground truth mask area. IoU is calculated for each pair of masks. The value ranges from 0 to 1, where values closer to 1 indicate better alignment between the two shapes. If the value exceeds a predefined threshold, the masks are considered a match.

$$\text{Precision} = \frac{TP}{TP + FP} \quad (2)$$

Where **TP** (True Positives) is the number of matched predictions, and **FP** (False Positives) is the number of predictions not matched. The value ranges from 0 to 1, where values closer to 1 indicate that most predictions were correct.

$$\text{Recall} = \frac{TP}{TP + FN} \quad (3)$$

Where **TP** as defined above and **FN** (False Negatives) is the number of missed ground truth objects. The value ranges from 0 to 1, where values closer to 1 indicate that most ground truth objects were detected.

$$\text{F1 Score} = 2 \times \frac{\text{Precision} \times \text{Recall}}{\text{Precision} + \text{Recall}} \quad (4)$$

The *F1 Score* is the harmonic mean of the precision and recall scores, providing a single measure that offers a balanced view of the model's performance (Molina et al., 2025). The value ranges from 0 to 1, where values closer to 1 indicate high precision and recall.

For the wall segments detection, we will also use the following metrics:

$$\text{Covered Prediction} = \frac{P \cap G}{P} > \text{coverage threshold} \quad (5)$$

$$\text{Covered Ground Truth} = \frac{P \cap G}{G} > \text{coverage threshold} \quad (6)$$

Where **P** and **G** represent the masks areas as defined above. Metrics (5) and (6) are similar to (1) but offer greater flexibility for evaluating wall segments that can be segmented in multiple valid ways. We calculate the ratio of covered predictions and ground truths to the total number of predictions and ground truths.

2.5.5.2. Stones model evaluation. The stones model is evaluated using these metrics, with an IoU threshold of 0.5 and minimum stone sizes ranging from 15cm × 15 cm–25cm × 25 cm. While building stone sizes vary, a size of 15cm × 15 cm, roughly the size of a hand stone, is expected to be smaller than most building stones. Examples of Roman-period stone sizes can be found in the study of ancient quarries by Gaied et al. (2010).

As seen in Fig. 2D, in many cases the stones are piled, with some partially obscured by others, making the contours less distinct. Due to the desire to detect even stones of small size used for filling wall segments (e.g. 15cm × 15 cm), this can lead to mismatches between the annotated stones and model predictions, the IoU threshold was set to 0.5, requiring at least 50 % overlap between the ground truth and the prediction. This issue is less significant when the stones are scattered on the ground or not densely piled, or with larger stones (for example with stone size of 25cm × 25 cm, and an IoU threshold of 0.7, the F1 score = 0.8).

2.5.5.3. Walls model evaluation. Wall segments often lack clear starting and ending edges. For instance, it is not always evident whether a gap between two stones should result in two separate wall segments or a single continuous segment (Fig. 4A). This ambiguity can lead to poor results during the evaluation process. However, in practice, this is not a significant concern, as the predictions of the walls model are later unified and intersected with the stones layer, producing a layer containing only the stones considered to be part of walls. To more correctly evaluate the wall segments, we chose to assess the results of the walls model using a combination of semantic segmentation evaluation and instance segmentation evaluation.

The Precision (2), Recall (3), and F1 Score (4) for the walls model are evaluated as a semantic segmentation task, meaning that each pixel is evaluated separately (Fig. 4B). Note that background pixels are ignored. Additionally, each predicted, and ground truth segment is tested for coverage, and the percentage of masks covered is reported, with the coverage threshold ranging from 0.5 to 0.7.

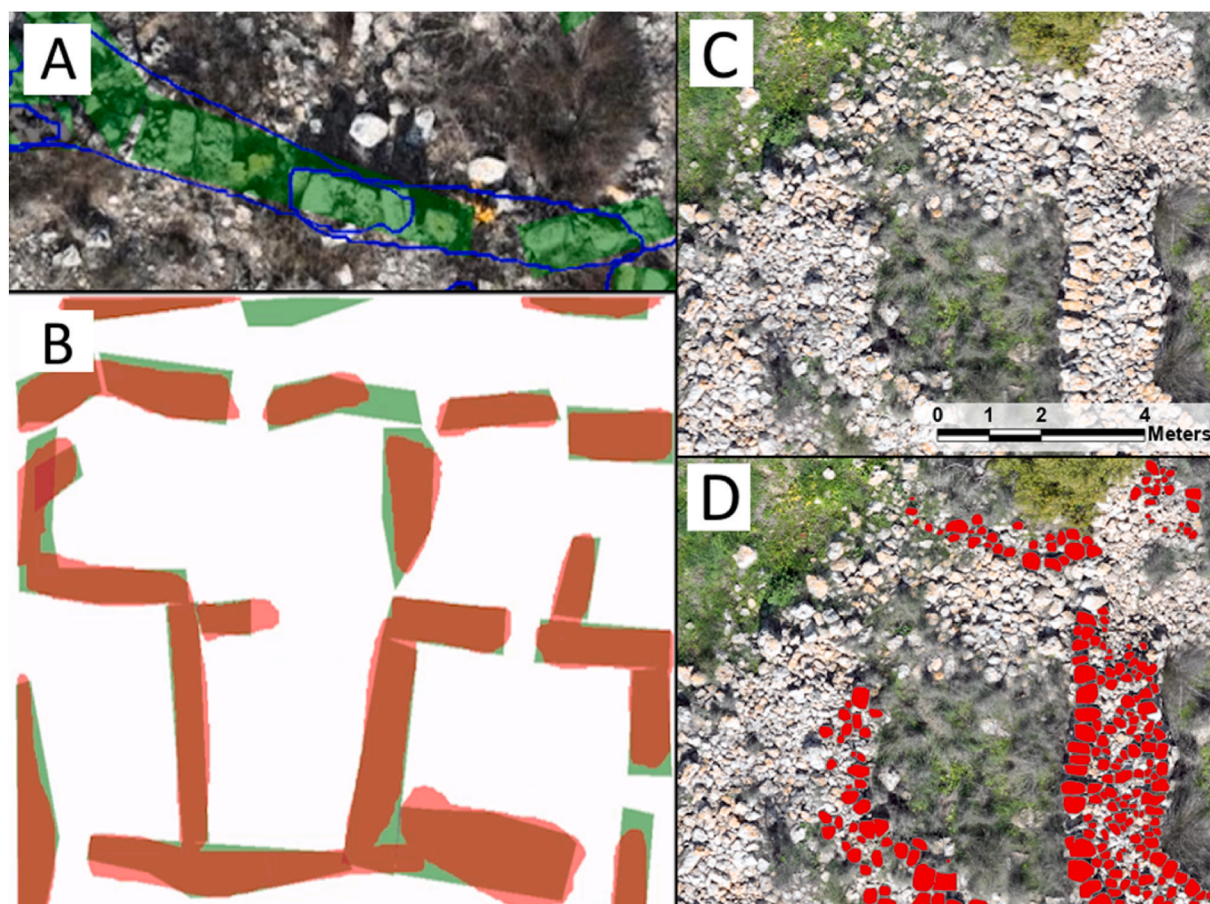


Fig. 4. (A) Annotated wall segments marked in green, and predicted wall segments marked with blue. This illustration highlights the ambiguity related to the segment's start and end edges. (B) Segmentation evaluation of the wall segments detection. Each ground truth (green) or predicted pixel (red) is tested separately. (C) Imagery showing partially collapsed walls. (D) The same imagery with stone walls highlighted in red, illustrating the importance of wall contours for wall detection. (For interpretation of the references to color in this figure legend, the reader is referred to the Web version of this article.)

3. Results

3.1. Evaluation results

The evaluation results of the stones model on the 9-site dataset, are presented in Table 2:

The evaluation results of the walls model are given in Table 3:

3.2. Outputs

The georeferenced outputs are: (1) the orthomosaic (Fig. 5B) and DTM generated by the photogrammetric software; (2) the stone layer (Fig. 5D); (3) the stone density heatmap (Fig. 6B); (4) the wall layer (Fig. 5C); and (5) the wall stones layer, which provides a stone-level site plan (Fig. 5A).

The stones model generates a layer of polygons, each representing an individual stone. Additional attributes are assigned to each stone, including its area and the MRR area ratio. The walls model similarly

Table 2

Stones model evaluation results. Precision (eq. (2)) – ratio of correct predictions. Recall (eq. (3)) – ratio of detected ground truths. F1 Score (eq. (4)) – harmonic mean of precision and recall scores.

Min size (cm ²)	Precision	Recall	F1 Score
15x15	0.87	0.79	0.83
20x20	0.86	0.85	0.86
25x25	0.86	0.87	0.86

Table 3

Walls model evaluation results. Each detected pixel is tested separately for the Precision, Recall and F1 Score. Covered Ground Truths (eq. (5)) – ratio of ground truth segments covered by the predictions. Covered Predictions (eq. (6)) – ratio of predicted segments covered by the ground truth. Coverage scores are reported for multiple thresholds.

Coverage Threshold	Precision (pixel based)	Recall (pixel based)	F1 Score (pixel based)	Covered Ground Truths	Covered Predictions
0.7	0.8	0.78	0.79	0.82	0.83
0.6				0.87	0.87
0.5				0.91	0.89

In summary, approximately 90 % of the ground truth wall segments were identified by the model, with a coverage of at least 50 %, and 85 % of the stones were correctly detected.

produces a polygon layer, where each polygon corresponds to a wall segment. Another layer is generated based on the intersection of the two layers, consisting of stones whose centroids fall within a wall segment. This is referred to as the “stone-level site plan” layer. Stones in this layer are grouped according to the containing wall segments, and additional features are computed, including the mean stone area, wall segment area, total stone area, and mean MRR value. A heatmap is produced using Kernel Density Estimation (KDE), with stones treated as weighted data points based on their area. KDE provides a simple and effective method for visually representing spatial data (Baxter et al., 1997), helping to highlight denser built areas within the site. All layers are

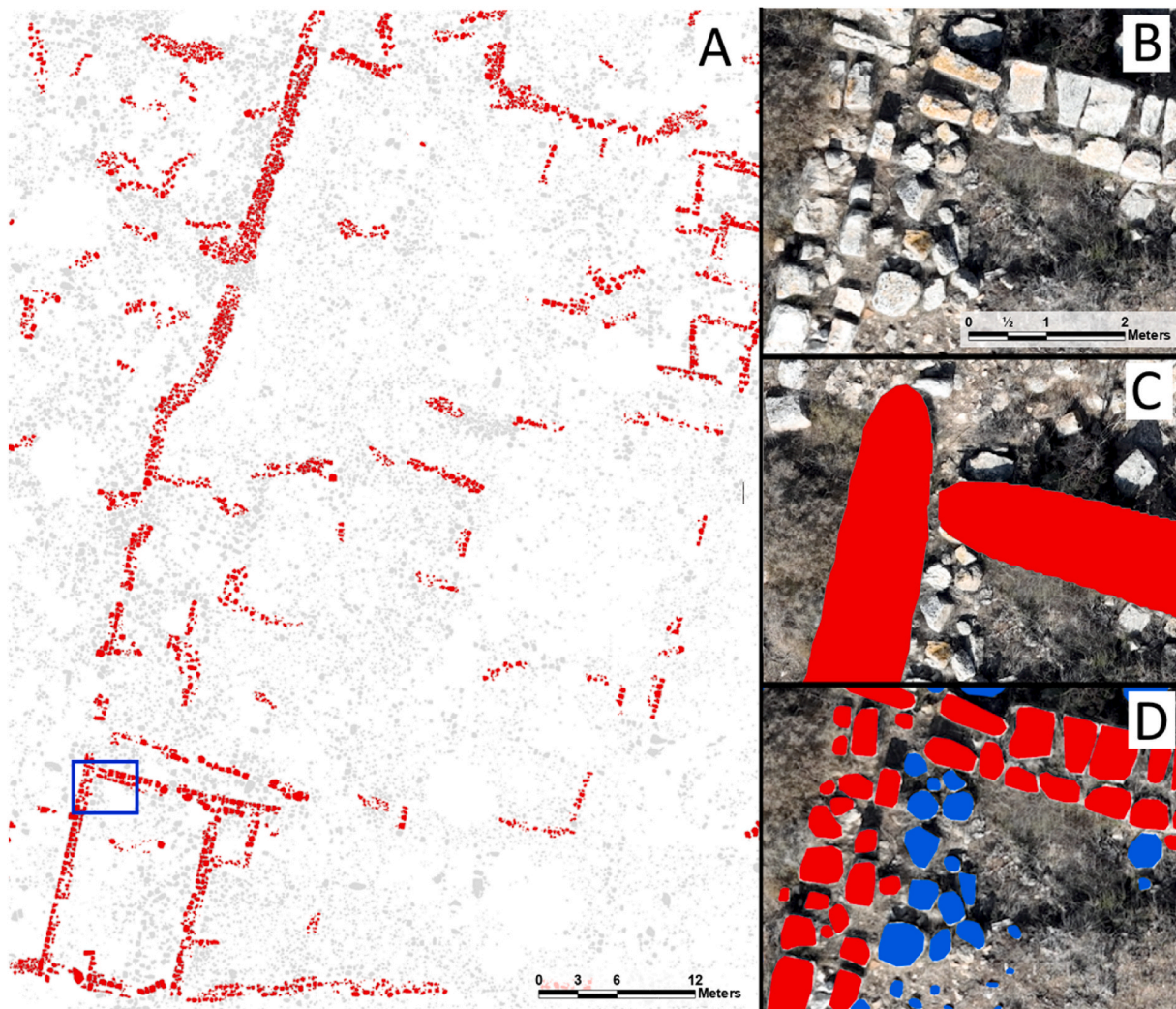


Fig. 5. Main outputs of the methodology, illustrated with Khirbet Qana (upper). (A) Wall stones layer – a stone-level site plan; the enlarged area is marked with a blue rectangle. (B) Orthomosaic. (C) Wall segments layer. (D) Stones layer, with stones that are part of wall segments marked in red and other stones marked in blue. (For interpretation of the references to color in this figure legend, the reader is referred to the Web version of this article.)

saved as GIS shapefiles (SHP) and can be opened using GIS software. Fig. 6 shows the orthomosaic, site plan, and heatmap of Khirbet Qana (upper). Site plans for the other mapped sites are provided in Appendix B.

3.3. Test case sites

The results for the mapped sites are summarized in Table 4. Approximately 350,000 stones were detected, of which 18 % were classified as wall stones.

4. Discussion

4.1. Model robustness

The models were trained on data from multiple sites under varied conditions, aiming to provide robust performance across a wide range of archaeological contexts out of the box. Fig. 7 presents representative imagery from each site. Site conditions and special characteristics are listed in Table 5 (appendix A). The models evaluated on our dataset achieved good results. While the thresholds help distinguish between bedrock stones and building stones, bedrock stones still pose a challenge for the neural network model. Additional weaknesses include irregular

ground-cover shapes (such as cast shadows), partially covered stones and walls, low contrast between stones and the ground or vegetation, and unclear stone contours—for example, in walls with mortar. As a general rule, imagery with clearly visible stone will yield good performance.

4.2. Understanding the site structure

Fig. 6 illustrates the value of the methodology's direct outputs. As shown, Khirbet Qana (upper) is highly sloped and densely vegetated, making it difficult to interpret spatial patterns from ground level. While the orthomosaic is useful, it demands a labor-intensive, step-by-step examination to "connect the dots." In contrast, the automatically generated site plan offers immediate insights and a clearer understanding of the site's structure. The heatmap further enhances the analysis by highlighting areas of higher land-use intensity. These outputs contain additional valuable information that can be further extracted and analyzed to deepen our understanding.

4.3. Insights from detected features

4.3.1. Uncovering styles and periods through clustering and site statistics

The attributes of the grouped stones within each wall segment can be

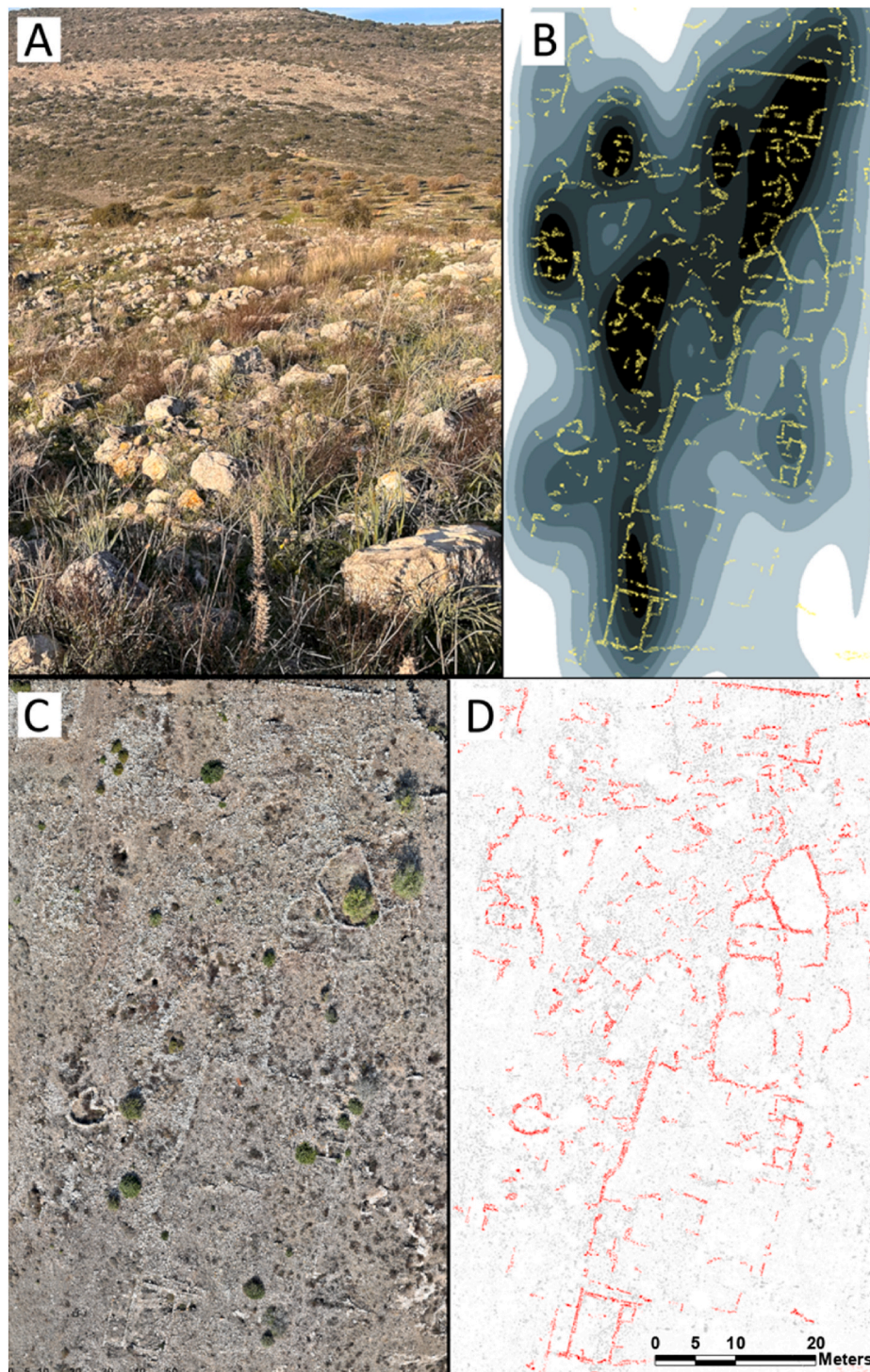


Fig. 6. Khirbet Qana (upper) – (A) ground-level photograph illustrating the difficulty of interpreting the scattered stone patterns, (B) orthomosaic generated from the drone images, (C) detected stones in light gray and detected stones that are part of detected walls in red, heatmap, representing the stones density (darker colors represent denser regions). The site plan layer, marked in yellow, is overlaid on it. (For interpretation of the references to color in this figure legend, the reader is referred to the Web version of this article.)

used to portray the wall segments and support the identification of wall periods and typologies. Clustering can be applied to a single site (local clustering) to decipher the typology of its architectural features, or to multiple sites (regional clustering) to compare architectural characteristics across sites. We applied the K-Means multivariate clustering algorithm (Baxter, 2015), using ArcGIS Pro, with the wall segments' mean stone area and mean MRR value as features, to the data from Khirbet Qana. We began with the top hill village, a multiperiod site (Edwards

and Sion, 2002) that contains visible remnants from different periods. The remains of most of the older walls, dating to the Roman and Byzantine periods, are not well preserved with one or two courses. Later aged walls, from the Mamluk and Ottoman periods, were built over the existing village and used for livestock pen. The scattered state of the building stones, leads to noisy data, but at least some of the different clusters provide important insights about the walls. Such an example is demonstrated in Fig. 8, where a 3-m-long wall made of large hewn

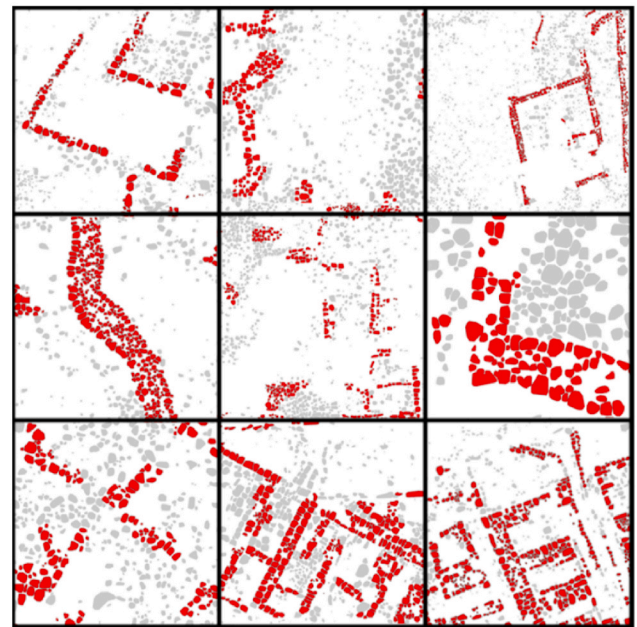


Fig. 7. Left - Representative imagery from each of the sites: (top) Khirbet Garas, Khirbet Sum, Rosh Zayit; (middle) Khirbet Dorban, Yodfat, Khirbet Qana lower; (bottom) Khirbet Qana upper, Beit She'arim, Tel Gezer. Right - Representative site plan layer from each of the sites.

Table 4

Sites statistics extracted from the model outputs. Tel Gezer and Beit She'arim were excluded as they were only partially mapped.

Site	Area (hectare)	Detected stones	Detected wall stones	<i>Walls area</i> <i>Site area</i>	Mean wall stones area (cm ²)
Rosh Zayit	0.85	17,591	2,170	136	116
Khirbet Dorban	1.14	29,693	9,141	302	344
Khirbet Sum	1.95	40,169	10,888	301	586
Khirbet Garas	2.06	31,405	4,104	176	363
Khirbet Qana (lower)	2.21	56,683	12,603	324	716
Yodfat (B)	2.9	46,073	6,147	158	457
Khirbet Qana (upper)	4.24	97,396	12,899	210	890

An analysis of these results is provided in section 5.3.2.

stones (MRR = 82 %, average stone area = 1,210 cm²) is located near a large wall segment built from many small stones (MRR = 78 %, average stone area = 574 cm²). Despite extensive manual examination of the orthomosaic, this wall became apparent only after applying the clustering algorithm. The clustering algorithm revealed a spatial outlier—a wall constructed differently from the surrounding walls.

Applying the clustering algorithm to multiple sites allows for a comparison of their features. In the case of the two sites located on Khirbet Qana hill, we observe that the lower village was built mostly with smaller stones and lower MRR values. These characteristics resemble the walls of the pens in the upper village (Fig. 9), which may suggest a relationship between the pens in the upper village and the buildings in the lower village.

Extracting site-wide statistics from the digitized layers is straightforward. Although this approach averages features from different periods or cultures in the same site, it provides useful outputs for large-scale comparisons. Table 4 lists the statistics of the mapped sites. Various insights can be gained from these statistics. For example, the

walls at Khirbet Garas are sparse and built with relatively large stones, whereas Khirbet Dorban exhibits a denser layout with smaller stones. It is important to note that the extracted data is two-dimensional, which makes volume measurements problematic unless wall heights are manually incorporated. Fig. 10 visualizes the data. The chart reveals three site clusters based on mean wall stone area and layout density ratio. We hypothesize that the clusters reflect differences in the periods of the sites' visible remains, excluding Rosh Zayit, which is primarily a storage fort with different characteristics (Gal and Alexandre, 2000).

These quantitative measures enable systematic comparison of construction techniques across archaeological sites. The resulting site clusters likely reflect distinct chronological or cultural building practices. This result demonstrates the significant potential of semi-automated wall and stone detection, even during the early reconnaissance phase of pedestrian field surveys. It provides valuable preliminary insights that can effectively inform future research directions and be further validated through subsequent excavation work.

5. Conclusion

The potential value of recording stones and architectural fragments, even in low-density or ambiguous contexts, is considerable. When documented systematically, such features can serve as critical anchors for site definition and contribute meaningfully to the resolution and comparability of both extensive and intensive survey frameworks (Bintliff et al., 2013; Meyer, 2022; Wernke et al., 2024; Wilkinson, 2003). Yet, time and labor remain significant constraints in stone mapping. Based on fieldwork at the Mountaineer site, Colorado (Andrews et al., 2021), Meltzer (personal communication, 2025) reports that georeferencing a ~25 cm rock in maximum length using an EDM/Total Station system with an average of 6 points (to account for rough shape, height, orientation and inclination) took an average of 6 min per rock. Thus, compiling the spatial data for the 3,834 rocks analyzed by Jayalath et al. (2015) required 23,004 min total, or 383 h in the field devoted to mapping alone. Using this standard, digitizing the 350,000 stones from 9 sites in this study would take 35,000 h, compared to roughly 45–60 min in the field and several hours of computer processing (mostly unattended) per site, totalling roughly 20 h.

The methodology presented in this paper provides a relatively simple

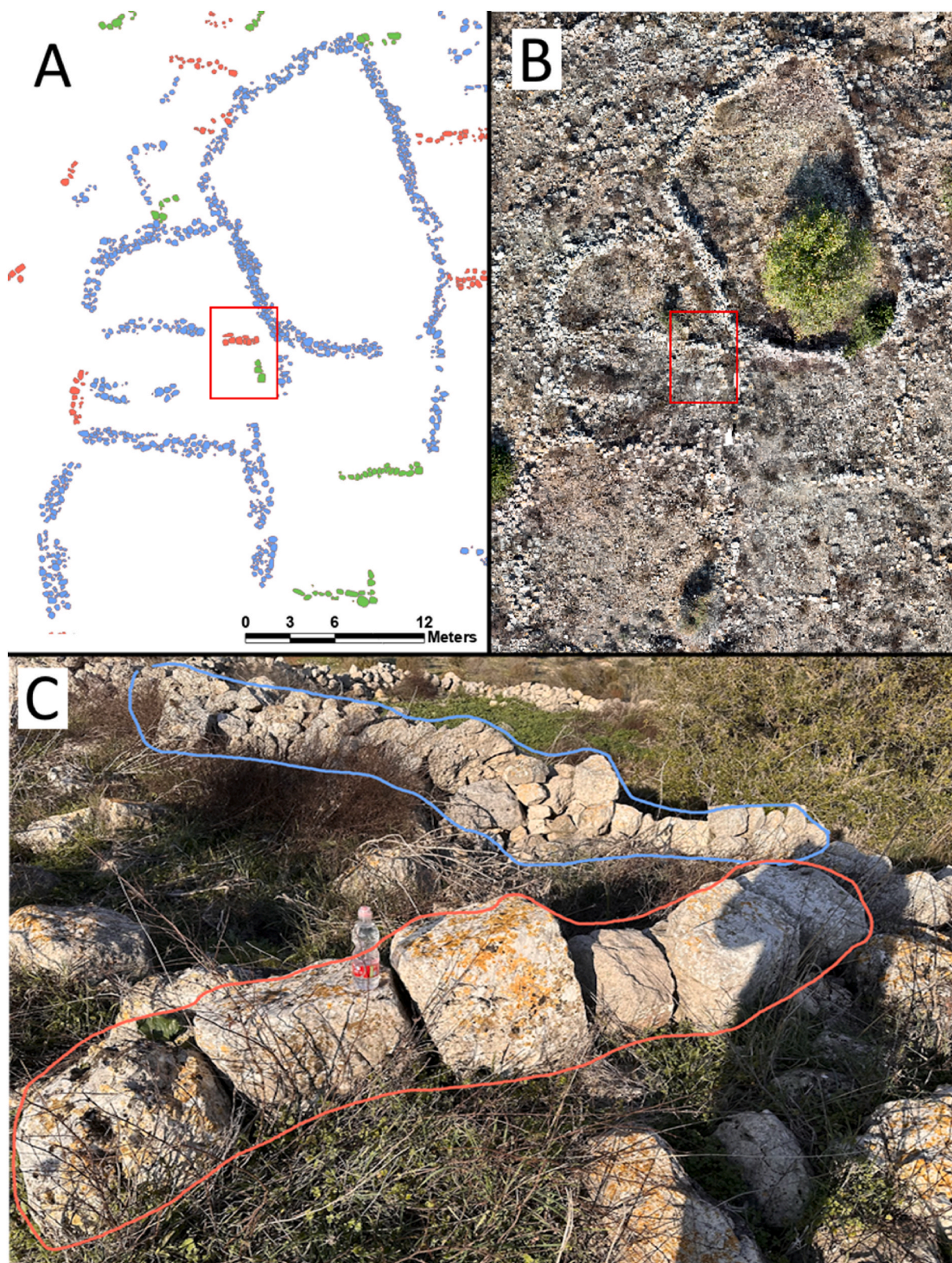


Fig. 8. Khirbet Qana (upper) (A) Clustered wall stones. (B) Orthomosaic. (C) Ground-level image of the area marked by the red rectangle, showing a 3-m-long wall with a clearly distinct typology. (For interpretation of the references to color in this figure legend, the reader is referred to the Web version of this article.)

process for mapping site stones and walls. Although beyond the scope of the present study, high-resolution imagery obtained from alternative sensors, such as multispectral systems or LiDAR, may allow development of the models in different areas with highly obstructive vegetation cover

and under other challenging conditions. Our methodology relies on high-resolution imagery, which has become accessible in recent years with advancements in drone technology. Stone is a common feature at many sites—one that cannot be overlooked. Mapping stones and walls

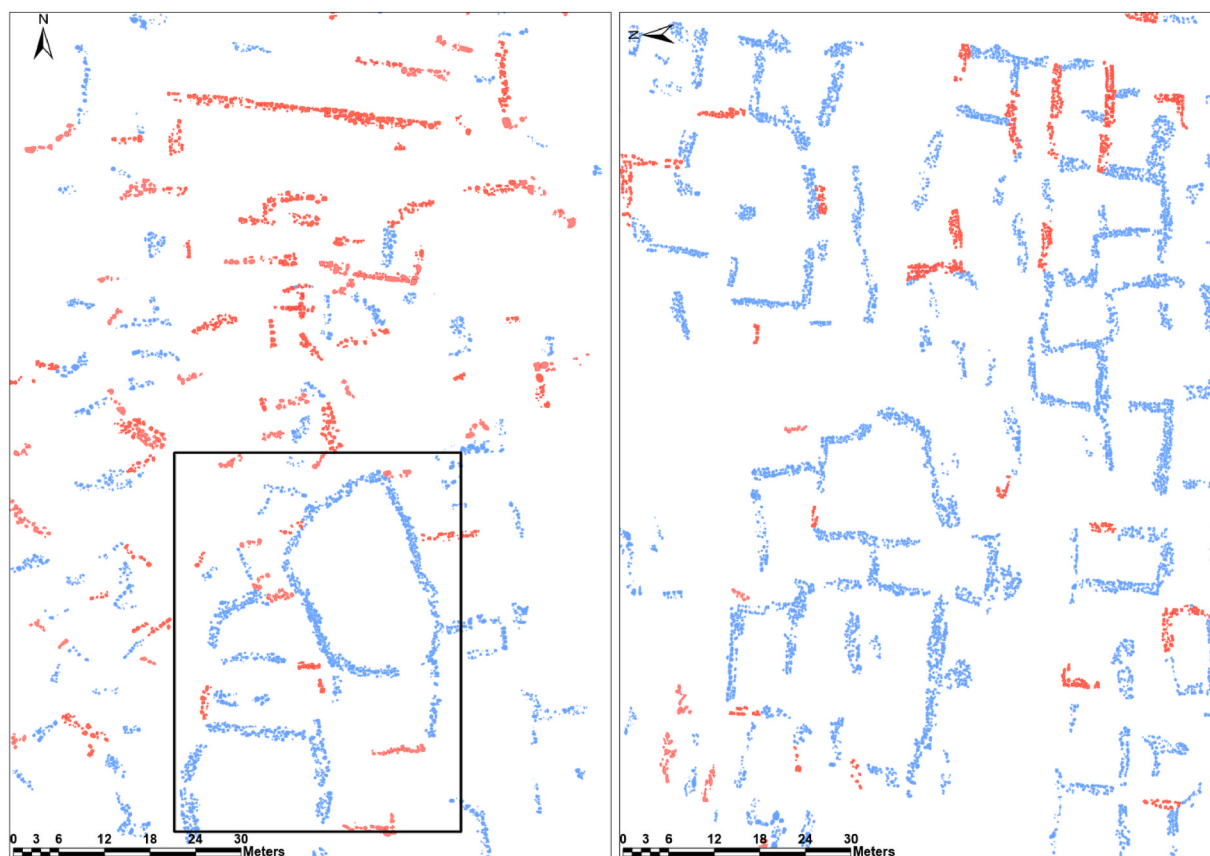


Fig. 9. Clustering based on average stone size and MRR to compare two nearby sites: Khirbet Qana – upper (left) and Khirbet Qana – lower (right). The walls at the lower site resemble the more recent sections of the upper site, highlighted by the black rectangle.

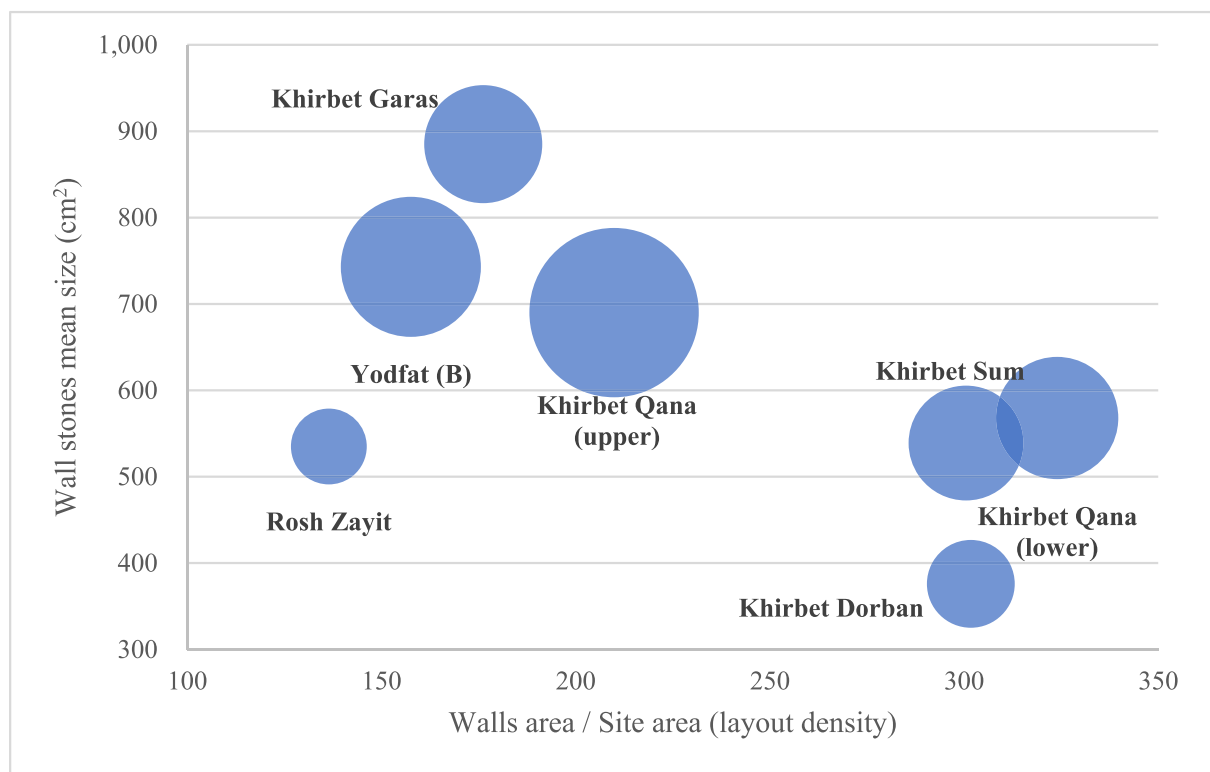


Fig. 10. Bubble chart visualization of multiple site statistics. Bubble size represents each site's area. The chart reveals three clusters of sites.

across multiple sites using this methodology opens the door to large-area research, enabling comparisons between sites, cultures, and periods. While comparison with excavation data is ultimately necessary, it represents a future stage of research that would require an additional one to two years of work, following what has already been a multi-year process. The tool is especially well suited to preliminary archaeological surveys: it provides a means of systematically analyzing and comparing complex or poorly understood ruins, generating hypotheses about site structure, and guiding decisions on where to excavate and what to expect. Its primary value lies in creating a comprehensive digital framework of all visible stones, enabling interpretive work to proceed in a more informed and structured manner.

CRedit authorship contribution statement

Erel Uziel: Writing – review & editing, Writing – original draft, Software, Methodology, Investigation, Data curation, Conceptualization. **Motti Zohar:** Writing – review & editing, Validation, Supervision, Data curation, Conceptualization. **Yitzchak Jaffe:** Writing – review & editing, Writing – original draft, Supervision, Project administration, Investigation, Data curation, Conceptualization.

Data availability statement

The data and code that support the findings of this study are openly available in Zenodo at <https://doi.org/10.5281/zenodo.15673668>.

Explanations on the code and other files, and step by step instructions are given in [appendix A](#).

Reproducible results: The Associate Editor for Reproducibility downloaded all materials and could reproduce the results presented by the authors.

Declaration of competing interest

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

Appendix A. Supplementary data

Supplementary data to this article can be found online at <https://doi.org/10.1016/j.jas.2025.106430>.

References

- Altaweel, M., Khelifi, A., Li, Z., Squitieri, A., Basmaji, T., Ghazal, M., 2022. Automated archaeological feature detection using deep learning on optical UAV imagery: preliminary results. *Remote Sens.* 14 (3), 553. <https://doi.org/10.3390/rs14030553>.
- Andrews, B.N., Meltzer, D.J., Stiger, M., 2021. *The Mountaineer Site: a Folsom Winter Camp in the Rockies*. University Press of Colorado.
- Attema, P., Bintliff, J., Leusen, M.V., Bes, P., Haas, T.D., Donev, D., et al., 2020. *A Guide to Good Practice in Mediterranean Surface Survey Projects*.
- Banning, E.B., 2002. *Archaeological Survey*. Kluwer Academic/Plenum Publishers, New York.
- Baxter, M.J., 2015. Spatial k-means Clustering in archaeology—variations on a Theme. *Academia*. <https://www.academia.edu/18142974>.
- Baxter, M.J., Beardah, C.C., Wright, R.V., 1997. Some archaeological applications of kernel density estimates. *J. Archaeol. Sci.* 24 (4), 347–354.
- Bintliff, John, et al., 2013. *The Leiden-Ljubljana Ancient cities of boeotia project, 2010–2012 seasons*. *Pharos* 19 (2), 1–34.
- Bradski, G., 2000. *The OpenCV Library*. Dr. Dobb's Journal of Software Tools.
- Carmona, J.A.S., Quirós, E., Mayoral, V., Charro, C., 2020. Assessing the potential of multispectral and thermal UAV imagery from archaeological sites. A case study from the Iron Age hillfort of Villavieja del Tamuja (Cáceres, Spain). *J. Archaeol. Sci.: Reports* 31, 102312.
- Casana, J., 2021. Rethinking the landscape: emerging approaches to archaeological remote sensing. *Annu. Rev. Anthropol.* 50 (1), 167–186.
- Casana, J., Fowles, S., Montgomery, L.M., Mermejo, R., Ferwerda, C., Hill, A.C., Adler, M., 2023. Multi-sensor drone survey of ancestral agricultural landscapes at Picuris Pueblo, New Mexico. *J. Archaeol. Sci.* 157, 105837.
- Douglas, D.H., Peucker, T.K., 1973. Algorithms for the reduction of the number of points required to represent a digitized line or its caricature. *Cartographica: Int. J. Geogr. Inf. Geovisualization* 10 (2), 112–122.
- Edwards, D., Sion, D., 2002. *Khirbet Qana: from Jewish village to Christian pilgrim site*. *J. Rom. Archaeol. Supplementary Series*. No 49 (Volume 3), 101–132.
- Field, S., Waite, M., Wandsnider, L., 2017. The utility of UAVs for archaeological surface survey: a comparative study. *J. Archaeol. Sci.: Reports* 13, 577–582.
- Frachetti, M.D., Berner, J., Liu, X., Henry, E.R., Maksudov, F., Ju, T., 2024. Large-scale medieval urbanism traced by UAV–lidar in highland Central Asia. *Nature* 1–7.
- Gaied, M.E., Younes, A., Gallala, W., 2010. A geoarchaeological study of the ancient quarries of Sidi Ghedamsy Island (monastir, Tunisia). *Archaeometry* 52 (4), 531–549.
- Gal, Z., Alexandre, Y., 2000. *Horbat Rosh Zayit: an Iron Age Storage Fort and Village* (IAA Reports—Monograph Series of the Israel Antiquities Authority, vol. 68).
- Gholamy, A., Kreinovich, V., Kosheleva, O., 2018. Why 70/30 or 80/20 relation between training and testing sets: a pedagogical explanation. *Int. J. Intell. Technol. Appl. Stat.* 11 (2), 105–111.
- Gilliot, J.M., Hadjar, D., Michelin, J., 2022. Potential of ultra-high-resolution uav images with centimeter gnss positioning for plant scale crop monitoring. *Remote Sens.* 14 (10), 2391.
- Guyot, A., Lennon, M., Lorho, T., Hubert-Moy, L., 2021. Combined detection and segmentation of archeological structures from LiDAR data using a deep learning approach. *Journal of Computer Applications in Archaeology* 4 (1), 1–19. <https://doi.org/10.5334/jcaa.64>.
- Hammer, E., 2022. Multi-centric, Marsh-based Urbanism at the early Mesopotamian city of Lagash (Tell al-Hiba, Iraq). *J. Anthropol. Archaeol.* 68, 101458.
- He, K., Zhang, X., Ren, S., Sun, J., 2016. Deep residual learning for image recognition. In: *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 770–778.
- He, K., Gkioxari, G., Dollár, P., Girshick, R., 2017. Mask R-CNN. In: *Proceedings of the IEEE International Conference on Computer Vision*.
- Homsher, R.S., Adams, M.J., Prins, A.B., Gardner-Cook, R., Tepper, Y., 2017. New directions with digital archaeology and spatial analysis in the Jezreel Valley. *Journal of Landscape Ecology* 10 (3), 154–164.
- Hou, M., Huo, D., Yang, Y., et al., 2024. Using mask R-CNN to rapidly detect the gold foil shedding of stone cultural heritage in images. *Heritage Science* 12, 46. <https://doi.org/10.1186/s40494-024-01158-9>.
- Jayalath, K.P., Gunst, R.F., Meltzer, D.J., 2015. Spatial point pattern identification of an apparent Ice-Age house structure. *Spatial Statistics* 14, 563–580.
- Jordahl, K., Van den Bossche, J., Fleischmann, M., et al., 2020. *Geopandas/geopandas: v1.0.1 (Version v1.0.1)*. Zenodo [Software]. <https://zenodo.org/records/12625316>.
- Knodell, A.R., Wilkinson, T.C., Leppard, T.P., Orengo, H.A., 2023. Survey archaeology in the Mediterranean world: regional traditions and contributions to long-term history. *J. Archaeol. Res.* 31 (2), 263–329.
- Lin, G.G., Semple, A., Tan, B., Zhang, P., 2024. A review on sensor spatial resolution and a discussion on geometric specifications. In: *JACIE 2024 Workshop*.
- Lindsay, I., Kong, N.N., 2020. Using the ArcGIS collector mobile app for settlement survey data collection in Armenia. *Advances in Archaeological Practice* 8 (4), 322–336.
- Menéndez-Marsh, F., Al-Rawi, M., Fonte, J., Dias, R., Gonçalves, L.J., Seco, L.G., et al., 2023. Geographic information systems in archaeology: a systematic review. *Journal of Computer Applications in Archaeology* 6 (1).
- Meyer, N., 2022. Finding sites in Mediterranean survey. *J. Mediterr. Archaeol.* 35 (2), 143–168.
- Molina, J.M., Llerena, J.P., Usero, L., Patricio, M.A., 2025. Advances in instance segmentation: technologies, metrics and applications in computer vision. *Neurocomputing*, 129584.
- Monna, F., Magail, J., Rolland, T., Navarro, N., Wilczek, J., Gantulga, J.O., et al., 2020. Machine learning for rapid mapping of archaeological structures made of dry stones—Example of burial monuments from the Khirgisuur culture, Mongolia. *J. Cult. Herit.* 43, 118–128.
- Mouget, A., Lucet, G., 2014. Photogrammetric Archaeological Survey with UAV. *ISPRS Annals of Photogrammetry, Remote Sensing and Spatial Information Sciences II-5*, 251–258. <https://doi.org/10.5194/isprsnals-II-5-251-2014>.
- Orengo, H.A., Garcia-Molsosa, A., 2019. A brave new world for archaeological survey: automated machine learning-based potsherd detection using high-resolution drone imagery. *J. Archaeol. Sci.* 112, 105013.
- Orengo, H.A., Garcia-Molsosa, A., Berganzo-Besga, I., Landauer, J., Aliende, P., Tres-Martínez, S., 2021. New developments in drone-based automated surface survey: towards a functional and effective survey system. *Archaeol. Prospect.* 28 (4), 519–526.
- Panda, A., Panigrahi, D., Mitra, S., Mittal, S., Rahimi, S., 2024. Transfer learning applied to computer vision problems: survey on current progress, limitations, and opportunities. *arXiv preprint arXiv:2409.07736*.
- Paszke, A., Gross, S., Massa, F., et al., 2019. *PyTorch: an imperative style, high-performance deep learning library*. *arXiv*. <https://arxiv.org/abs/1912.01703>.
- Peacock, D.P.S., 1998. *The Archaeology of Stone: a Report for English Heritage*. English Heritage.
- Pepe, M., Fregonese, L., Scaioni, M., 2018. Planning airborne photogrammetry and remote-sensing missions with modern platforms and sensors. *European Journal of Remote Sensing* 51 (1), 412–436.
- Ren, S., He, K., Girshick, R.B., Sun, J., 2015. Faster R-CNN: towards real-time object detection with region proposal networks. *Adv. Neural Inf. Process. Syst.* 28.
- Rouse, L., Krumnow, J., 2020. On the fly: strategies for UAV-based archaeological survey in mountainous areas of Central Asia and their implications for landscape research.

- J. Archaeol. Sci.: Reports 30, 102275. <https://doi.org/10.1016/j.jasrep.2020.102275>.
- Saubier, M., Eisenbeiss, H., 2010. UAVs for the documentation of archaeological excavations. *Int. Arch. Photogram. Rem. Sens. Spatial Inf. Sci.* 38 (5), 526–531.
- Shmidov, Y., Wiegmann, A., Sion, O., Uziel, J., Ganor, A., Klein, E., 2023. Photogrammetric documentation and visibility analysis of the Roman siege works at Masada. *New Studies in the Archaeology of the Judean Desert: Collected Papers*, pp. 419–449.
- Trier, Ø.D., Reksten, J.H., Løseth, K., 2021. Automated mapping of cultural heritage in Norway from airborne lidar data using faster R-CNN. *Int. J. Appl. Earth Obs. Geoinf.* 95, 102241.
- Vinci, G., Vanzani, F., Fontana, A., Campana, S., 2025. LiDAR applications in archaeology: a systematic review. *Archaeol. Prospect.* 32 (1), 81–101.
- Waagen, J., van Hilst, A., 2025. Drones over excavations—Enhancing the archaeologist's eyes. An empirical case study into the deployment of UAS remote sensing as a research tool for archaeological excavation, in Veldhoven, the Netherlands. *J. Archaeol. Sci.: Reports* 62, 105043.
- Wada, K., 2021. Labelme: image polygonal annotation with Python. Zenodo. <https://doi.org/10.5281/zenodo.5711226> [Software], Version 1.2.0.
- Wang, Y., Liu, C., Tiwari, A., Silver, M., Karnieli, A., Zhu, X.X., Albrecht, C.M., 2022. Deep semantic model fusion for ancient agricultural terrace detection. In: 2022 IEEE International Conference on Big Data (Big Data). IEEE, pp. 4888–4892.
- Wernke, S., VanValkenburgh, P., Saito, A., 2020. Interregional archaeology in the age of big data: building online collaborative platforms for virtual survey in the Andes. *J. Field Archaeol.* 45 (Suppl. 1), S61–S74.
- Wernke, Steven A., et al., 2024. Large-scale, collaborative imagery survey in archaeology: the Geospatial platform for andean culture, history and archaeology (GeoPACHA). *Antiquity* 98 (397), 155–171.
- Wilkinson, T.J., 2003. *Archaeological Landscapes of the near East*. University of Arizona Press.
- Wu, X., Xiao, L., Sun, Y., et al., 2022. A survey of human-in-the-loop for machine learning. *Future Gener. Comput. Syst.* 135, 364–381.