AI-ming backwards: Vanishing archaeological landscapes in Mesopotamia and automatic detection of sites on CORONA imagery

Alessandro Pistola[1¶] ORCID, Valentina Orrù[2¶] ORCID, Nicolò Marchetti[2&] ORCID, Marco Roccetti[1*&] ORCID

[1] Department of Computer Science and Engineering, University of Bologna, Bologna, Emilia-Romagna, Italy

[2] Department of History and Cultures, University of Bologna, Emilia-Romagna, Italy

*Corresponding author

E-mail: marco.roccetti@unibo.it (MR)

¶ These authors contributed equally to this work.

&These authors also contributed equally to this work.

Abstract

This study discusses the results obtained by upgrading an existing deep learning model with the knowledge provided by one of the oldest sets of grayscale satellite imagery, known as CORONA. We improved the AI model's attitude towards the automatic identification of archaeological sites in an environment which has been completely transformed in the last five decades, including the complete destruction of many of those same sites. The initial Bing-based convolutional network model was re-trained using CORONA satellite imagery for the district of Abu Ghraib, west of Baghdad, central

Mesopotamian floodplain. The results were twofold and surprising. First, the detection precision obtained on the area of interest increased sensibly: in particular, the *Intersection-over-Union* (IoU) values, at the image segmentation level, surpassed 85%, while the general accuracy in detecting archeological sites reached 90%. Second, our re-trained model allowed the identification of four new sites of archaeological interest (confirmed through field verification), previously not identified by archaeologists with traditional techniques. This has confirmed the efficacy of using AI techniques and the CORONA imagery from the 1960s to discover archaeological sites currently no longer visible, a concrete breakthrough with significant consequences for the study of landscapes with vanishing archaeological evidence induced by anthropization.

Introduction

For the study of Near Eastern archaeological landscapes and the reconstruction of settlement patterns, the first objective to achieve through techniques like *remote sensing*, and subsequent ground truthing, is the identification and the localization of ancient settlements, which are known mostly as *Tells*, in this specific geographical area. A tell is an artificial mound created by the accumulated debris from centuries of human habitation. These mounds typically form in regions such as the Mesopotamian floodplain, where communities repeatedly built and rebuilt their settlements at the same site with perishable material. While remote sensing has been a key resource for archaeologists for many years, being a non-invasive method that contributes to the detection and preservation of cultural heritage, it requires a large amount of experts' human work and consequently requires a significant amount of time [1-6]. It goes without saying that using deep learning techniques as a support to human efforts could open new perspectives. For example, deep learning abilities can be put to good use for automatically analyzing satellite imagery, especially using a technique called *semantic segmentation* which, in simple words, consists in assigning a class label to each pixel in an image, up to a point where the entire image is recognized as easily interpretable by archaeologists. Several works in this field have confirmed the efficacy of this approach [7-10]. To conclude this part of the Introduction, it is worth mentioning that since the time when we began our research activity in the area of Deep Learning applied to archaeology many other similar initiatives have taken shape and are have been brought to the forefront of the scientific discussion. We are well aware that a wealth of studies have made relevant advancements to this field [11-14]. Our attention to these results is intense and witnessed by our citation to those studies,

over time. Nonetheless, this point needs a specific treatment to avoid unintentional misinterpretations of our thought. There are (at least) two axes along which experts can look at similarities or differences in all these kinds of researches. The first one is concerned with the remote sensing methodologies employed to gather the data on which various Artificial Intelligence techniques can work. Using passive or active approaches, satellite or aircrafts, radar or LiDAR or any other typology of sensors emitting their own signals, or a combination of some of them, constitutes a relevant difference that extends till the point of a distinction (or divergence) in the meaning of the results which can be obtained. Not to consider the issue of the type of data, as working with new rather than old (e.g., CORONA, in our specific case) high resolution images represents another important source of differentiation. The second axis is that of the distinction between Machine Learning (ML) and Deep Learning (DL). ML are algorithms that learn from *structured* data to predict outputs and discover patterns in that data. DL, instead, is always based on highly complex neural networks that mimic the way a human brain works to detect patterns in large *unstructured* data (like images). A traditional ML algorithm can be something as simple as linear regression or a search in a decisional tree, the driving force behind being often that of ordinary statistics. DL algorithms, instead, should be regarded as a sophisticated and mathematically complex evolution of ML. To achieve this result, DL mechanisms use a layered structure of algorithms, called artificial neural networks, with a specific design based on a cascade of several different computational blocks, inspired by the biological neural network of the human brain, and leading to a process of learning that is far more capable than that of standard ML models. It should be also considered that with DL one can fall often into an excess of inference to which it is difficult (even not possible) to give a formal explanation. An extension of this discussion, tailored to the archaeological field, is reported in [15]. Consequently, applying either ML or DL makes an important difference, being often unfair (or nonsensical), in the light of the explanation above, a comparison between studies that adopt different approaches. Given these premises, we look at the wealth of researches that have investigated how well AI techniques can work in the archaeological field, not with the aim of conducting one-to-one comparisons with specific papers that could have followed different approaches, rather with the responsibility of witnessing the level of productivity of the entire community in this specific field.

Our previous work

Building on a long-term scientific collaboration between AI-ers and archaeologists at the University of Bologna, Italy [16-19], a deep learning model has been recently proposed, enhanced with segmentation

and self-attention mechanisms, which was able to detect mounded archaeological sites in the Mesopotamian floodplain in southern Iraq. A set of modern (Bing-based) georeferenced vector shapes were used as data source, corresponding to the outlines of the previously mentioned *tells* and surrounding areas, totaling 4934 shapefiles in the southern Mesopotamian floodplain [Floodplains Project, 15]. Each image in that dataset was subjected to a variety of image manipulation techniques (including, for example, random rotation, mirroring, brightness and contrast correction) and it was then given as an input to train a convolutional neural network, augmented with segmentation and self-attention mechanisms. In the end, the result of this activity was a deep learning model able to detect archaeological sites in the area of interest, which achieved during a test on a set of already known sites an *Intersection Over Union* (IoU) score of 0.81, with a general accuracy in the neighborhood of 80%. This first work had, nonetheless, two important limitations. First, an initial attempt to exploit CORONA satellite imagery was unsuccessful, probably due to our inability to integrate this panchromatic imagery with full color pictures. Second, the entire testing activity was conducted on hundreds of already known archaeological sites, without the possibility of challenging the machine predictions on sites not already groundtruthed. Computer scientists' wish, instead, would have been to make those automatic predictions on sites not already certified as tells and, upon confirmation by the archaeologists, subjecting them to a process of ground-truthing, to understand in reality whether those predictions were accurate or not.

New developments

In this new study, we decided to break the borders of the cultural shell of our previous research and try to learn from the CORONA satellite imagery (with its technical peculiarities: panchromatic and relatively low resolution), while then facing the challenge of validating the predictions of our model, not only on already confirmed archaeological sites, but also on areas where the presence of tells was never suggested before, based on traditional remote sensing techniques [20, 21]. The area of interest was located in the Abu Ghraib district in Iraqi (within the Greater Baghdad area) since this zone had not been included in previous archaeological surface surveys (only its easternmost tip was included in the pioneering researches conducted by Bob Adams [22]). In essence, our intent was that of integrating the CORONA satellite imagery into our already existing AI model, leveraging on the unique insights that this dataset can contribute. The CORONA satellite imagery, in fact, is by itself of extreme relevance in the archaeological field as it contains information dating back from 1960 to 1972, which for the greater

part is no longer visible on current basemaps due to agricultural operations and urbanization. In addition, while the archaeological significance of CORONA satellite imagery has been already empirically validated and widely acknowledged by the archaeologists, from the perspective of an AI model none has already demonstrated a tangible improvement in the accuracy detection, directly attributable to the inclusion of this imagery in the automated site detection process. This challenge is non-trivial, as previous researches have yielded inconsistent outcomes, so far. Therefore, a measurable improvement of the performances of our AI model, following the integration of the CORONA satellite imagery, would inform us about the model's ability to exploit the additional contextual information provided by these images, thus confirming what was already archaeologically recognized [24-28]. To conclude this Section, we anticipate here the scheme of the study we conducted. Based on the consideration that using geo-referenced CORONA imagery is now a standard practice in archaeology, because of its widely recognized value in providing information on heavily transformed landscapes, we first re-trained our convolutional neural model with the CORONA imagery, exploiting modern transfer learning techniques, and then we developed a two-stage fine-tuning procedure with the aim of obtaining the best possible deep learning model. In the end, the results were three new different configurations of our convolutional neural network: 1) one leveraging only on Bing basemaps, 2) another one leveraging only on CORONA basemaps and, finally, 3) one based on a combination of both the datasets. The final step of our current research consisted of two different phases. In the first one, the predictions generated by our three AI models were tested on already known sites (as well as on areas where the absence of sites was certain) using the traditional validation techniques of deep learning models. In the second, and more interesting, phase, instead, the attention of the archaeologists was focused on those areas predicted by the AI, comprising archaeological sites which had not been previously identified as such by the scientific community.

Materials and Methods

We first describe the data used in our study, and then we illustrate the methods employed to build our AI models (for accessing all the developed software and the data used in this study, see the Section: Data Availability Statement).

Data

We begin by noticing that all the remote sensing operations we carried out to identify archaeological evidence and to extract the usable data were conducted on the basis of various publicly available basemaps, specifically: current (Google, Bing, and Esri) and historical satellite imagery (CORONA), and topographic maps (US Army 1:100,000 from 1942). During this phase, 88 potential tells were identified, recorded as vector shapefiles and classified with the abbreviation *GHR* (i.e., the initials of the geographical district of interest: Abu Ghraib) and a sequential number. Starting from that information, the image creation process was based on the following five steps: **i)** all the shapefiles of the area of interest were imported from Bing and CORONA basemaps into an open-source GIS software [QGIS; 29], **ii)** sample squares each 2000 meters long, centered on the centroid of any given shapefile, were extracted from those images using a Python script developed by us (this was done in the same way both for Bing, through the QuickMapService plugin, and for the CORONA imagery, via the free services provided by the University of Arkansas' Center for Advanced Spatial Technologies). At that point, **iii)** we generated the truth masks, that is the masks that put in evidence, at a pixel level, the points either included in a tell or not. After that phase, we were in the obvious situation of having an unbalanced dataset, with a prevalence of non-empty truth masks. To fill this gap, additional images, with an empty truth mask, were generated, to balance the dataset. This was done, **iv)** by choosing 120 random points in the area of interest (with relative surrounding images, not containing any tell). The final dataset consisted of 88 images (around 41%) each including a *tell*, and 120 images (almost 59%) not portraying any tell or its parts, totaling a final amount of 208 pictures, on which a training activity could be conducted. Nonetheless, given the relatively small size of this dataset, **v)** an *aggressive* data augmentation procedure was exploited that has prevented the overfitting phenomenon. In particular, using the Albumentations library [30], three subsequent transformations (geometric, color space and kernel filters) were applied to all the images (Bing, CORONA and truth masks), where each transformation was chosen, in turn, from one of the three separate groups shown in Table 1, with a given probability. Fig 1 provides 3 examples of such a pipelined image augmentation process, where the transformations (RandomCrop, Flip, RandomRotate90, GaussNoise, Sharpen, Resize), (RandomCrop, Flip, RandomRotate90, CLAHE, GaussNoise, Sharpen, Resize) and (RandomCrop, Flip, RandomRotate90, RandomBrightnessContrast, MotionBlur, Sharpen, Resize) were applied in the reported cases following that exact sequence.

Table 1. Data augmentation pipeline.

Group	Technique	Probability of use
	RandomCrop	1
	Flip	0.5
	RandomRotate90	0.5
Geometric		0.2
	GridDistorsion	0.4
	RandomGridShuffle	0.6
Color space		0.5
	CLAHE	0.4
	RandomBrightnessContrast	0.8
	ChannelShuffle	0.1
	ColorJitter	0.2
	HueSaturationValue	0.2
Kernel filters		0.5
	Blur	0.4
	GaussNoise	0.4
	MotionBlur	0.2
	Sharpen	0.1
	Resize	1

Fig 1. Image augmentation pipeline: an example.

Disclaimer: All the displayed data fall under the condition of fair use utilization of geographical data for academic purposes. The list of all relevant data/software provider(s) is as follows: (i) original maps creation under the Section 5 of Microsoft Bing Maps Platform APIs' terms of use (https://www.microsoft.com/en-us/maps/product/print-rights); (ii) maps display achieved with an open source software, under the GNU licenses of QGIS (https://qgis.org/en/site/) and QuickMapsServices (https://github.com/nextgis/quickmapservices); (iii) final maps elaboration achieved with a software developed by the authors and available at (https://github.com/alepistola/AI_floodplains).

Methods

The type of convolutional neural network and the method used to train it were similar to those adopted in our previous works, to which we refer for a detailed description [18, 19]. Nonetheless, while there are precise limits on how much we can repeat from our previous cited studies, for the benefit of the readers it is important to remind what follows. Our work starts by using the PyTorch Segmentation Models library and by defining a Deep Convolutional Neural Network (DCNN) model, called MANet. Its complex

architecture is summarized in Fig 2. The following explanations on the schema portrayed in Fig 2 are in order. Our MANet (Multi-scale Attention Net) is a deep-learning neural network tailored to learn robust and discriminative features from high resolution remote sensing images. It stacks various multi-scale attention blocks, aiming at reducing spatial and channel redundancy to accelerate convolution. As shown in Fig 2, it is comprised of three main blocks: an encoder, a decoder and a segmentation head. The encoder, represented by the leftmost block in Fig 2, constitutes a proper convolutional architecture, based on the Efficientnet model (precisely, Efficientnet-b3) [31, 32]. Input to this encoder are high resolution images with a given number of channels (as satellite sensors can collect images at various regions of the electromagnetic spectrum) and a corresponding spatial resolution, expressed as a matrix of pixels (n x n). Those images pass through a cascade of multiple blocks (white-blue stacks of Fig 2) which implement a convolution procedure, followed by subsequent operations of batch normalization and swish activation. In essence, in this context, a convolution is an orderly procedure for image processing that converts many pixels in its receptive field (that is, the size of the input regions that produce the features of interest) into single values, aiming at reducing the number of free parameters, while allowing the network to be deeper. In fact, the final layer of our encoder returns feature maps at a reduced resolution of (16 x16) over 384 channels (as shown in Fig 2). To conclude the analysis of the encoder represented in Fig 2, one should put attention on the skip connections between the stacks of the encoder and the decoder (black arrows in Fig 2) that symbolize the passage of information from the input to the output stacks of the MANet. The role of the decoder of our MANet (rightmost stacks of Fig 2), instead, is that of performing a weighted recombination of the features extracted by the encoder, with the final aim to return segmentation maps (i.e., segmentation shapes) like those portrayed in the rightmost part of Fig 2. Those segmentation maps are the most-informative output of our MANet, as they can be used to put in evidence the sub-regions of archaeological interest within the images from which all this process has started. Our decoder incorporates both a Position-wise Attention Block (or PAB) and Multi-scale Fusion Attention Blocks (or MFABs). As to the use of these attention mechanisms, it should be reminded that they allow our models to weight different latent features in the images, specifying where attention is needed in this latent space to better learn [32]. Specifically, the PAB (orange block in Fig 2) implements a positional encoding mechanism that returns an attention map which indicates the greater importance of some pixels over the others, helping our architecture to identify those regions that deserve segmentation. In addition, by cascading several MFABs (red stacks

of the decoder of Fig 2), we have implemented a multi-scale strategy that aggregates features with the aim of capturing the relationships among different channels, thus making the segmentation more robust. The segmentation head represents the final layer responsible for computing the ultimate segmentation maps. This is the last step with which the initial remote sensing images are partitioned into different regions, with their pixel homogeneously classified and the final intent to identify sharper region boundaries. Finally, up-sampling blocks (purple stacks in Fig 2) are simply used to return output maps with the same resolution of the input images. It is worth noticing, in the end, that the entire procedure is able to process in the order of ten high resolution images in less than a second.

Fig 2. The architectural model of the Deep Convolutional Neural Network (DCNN).

Disclaimer: This Figure has been obtained starting from a royalty-free image for academic use and then customized by the authors with graphics and information coming from their research. It falls under the condition of fair use utilization of open source images for academic purposes.

Starting from the DCNN architecture we have discussed, we have re-trained our models, initially pre-trained on both Imagenet and on the images exploited in our previous work, with the new 208 images introduced in the previous Section, resorting to traditional transfer learning techniques [19, 33, 34]. In essence, we aimed at obtaining three new deep learning models, that added to the three ones built during our previous study [19], where the re-training activity was based on the use of the new imagery, respectively, provided by Bing, CORONA and a combination of both. As already anticipated, after these training activities, we concluded with a further incremental step of fine-tuning, applied to all the AI models of interest, using a particular technique called two-stage fine-tuning [35]. Technically speaking, this additional two-stage fine tuning activity, included, in turn, a first phase where, keeping the learning rate unchanged, the weights of the deep layers were frozen, subjecting to training only the *segmentation head*. In the second phase of this procedure, instead, the weights were unfrozen, reducing the learning rate by a factor of ten, and carrying out a re-training of the entire model, thus reducing the risk of both overfitting and catastrophic forgetting. We conducted this final procedure with the number of training epochs not fixed and proceeded until we detected a stagnation of the loss on the validation set, indicating a possible overfitting to avoid. For the sake of clarity, we summarize this (only apparently)

complex situation providing, in the list below, all the six different models, differentiated based on the combination of the training activities to which they were subjected and the type of image dataset used:

- **Bing**: the deep learning model trained on Bing basemaps during our previous study [19].
- **Bing_Bing**: the deep learning model previously trained on Bing basemaps and now re-trained on new Bing basemaps and finally fine-tuned as explained below.
- **CORONA**: the deep learning model trained on CORONA basemaps during our previous study [19].
- **CORONA_CORONA**: the deep learning model previously trained on CORONA basemaps and now re-trained on new CORONA basemaps and finally fine-tuned as explained below.
- **BingCORONA**: the deep learning model trained on a combination of Bing and CORONA basemaps during our previous study [19].
- **BingCORONA_BingCORONA**: the deep learning model previously trained on a combination of Bing and CORONA basemaps and now re-trained on new Bing and CORONA basemaps and finally fine-tuned as explained below.

Moving to the issue of the type of metrics used to evaluate the efficacy of our system, it is worth mentioning that the accuracy returned by our models was evaluated on the basis of the consideration of two different assessment perspectives: that is, a) through the lens of the semantic segmentation to which each image was subjected (i.e. trying to evaluate the achieved accuracy only at a pixel level), and b) at a more general level, analyzing the confusion matrix, to obtain an assessment of the accuracy and recall values. In particular, to evaluate the results produced by segmentation, we have used three different metrics: i) the Intersection over Union (IoU), ii) the binary Intersection Over Union, (bIoU), and finally iii) the Matthews Correlation Coefficient (MCC). The mathematical definitions of these metrics are beyond the scope of this paper and can be easily retrieved from the specialized literature, more interesting are the motivations for this choice which are as follows. The IoU metric is largely used in similar situations, but it may present various defects as it is recognized that it can return high values that could be not directly related to a better general object recognition, but just to a more precise identification of its contours. Indeed, we used it in this study, because it allowed a comparison with previous results. Its variant, the bIoU, is instead a better candidate to measure the accuracy of detection

in terms of the pixels being recognized as a part of a tell. Finally, the measurement of MCC values was added as it should represent the most appropriate metric for the problem under consideration based on findings described in recent literature [36, 37]. Coming to the final phase of our study, upon assessment of the performances of our AI models we chose the one with better accuracy: its results were plotted under the form of a corresponding heatmap, and then passed to the archaeological team on the field. Based on these heatmaps, the latter made the final decision on which were the more promising sites deserving a visit during the field survey campaigns.

Results

We present the results we have obtained during the testing phase of our deep learning models. The first fact to mention is that, of the 208 initial images, only 10% (i.e., 20 images) were used as subject of this testing phase. In fact, 156 images (75%) were used for re-training our models, with 32 of them (15%) used during the validation phase. Before passing to the results achieved on the 20 images which were never shown to our models before this testing phase, we consider it can be of some interest to readers being also informed on the results obtained with the 32 images of the validation phase. Nonetheless, it should be clear that this phase (i.e., the validation phase) constitutes an integral part of the re-training activities discussed in the previous Section. As such, the corresponding results do not represent the ultimate measurement of how well our models recognize tells when new images are proposed. Rather, they have given a preliminary assurance that our models have learnt effectively during re-training, with a generic propensity to generalize well to new images. Obviously, monitoring this tendency during training helped us to fine-tune our models for better results. With this in mind, Table 2 provides the results obtained with the 32 images of the validation phase. The following factors should be taken into consideration. First, being validation a part of the re-training activity, these results are provided only for the three re-trained models, namely: Bing_Bing, Bing_CORONA_BingCORONA, and CORONA_CORONA. Second, the numerical values of Table 2 are not given under the form of average and standard deviation, as they represent, instead, the better values the models achieved (during a specific epoch) before overfitting occurred. Third, these results only focus on the a pixel-wise accuracy.

Table 2. Validation: pixel-wise accuracy for the three re-trained models.

Model	IoU	MCC	bIoU	Epoch
Bing_Bing	83.07	55.25	37.00	15
Bing_CORONA_BingCORONA	84.42	69.99	56.86	9
CORONA_CORONA	82.25	59.30	43.70	27

Table 3 reports, instead, the average results (plus standard deviation) we achieved with our testing activity conducted with the 20 images our models have never seen before. These results are based on the metrics that we have already indicated being the most appropriate to recognize a tell at a pixel level (that is: IoU, bIoU and MCC). Each test was repeated ten times.

Table 3. Testing: pixel-wise accuracy for all models (average values with standard deviation).

Model	IoU	St.d.	MCC	St.d.	bIoU	St.d.
Bing (previous)	82.24	2.88	35.24	6.36	22.70	5.55
Bing_Bing	**86.12**	2.77	**34.03**	9.95	**21.53**	8.67
Bing_CORONA (previous)	84.30	1.56	45.76	7.93	28.80	6.45
Bing_CORONA_BingCORONA	**85.77**	2.03	**55.63**	5.88	**39.23**	6.37
CORONA (previous)	83.54	2.02	31.98	8.57	18.80	5.91
CORONA_CORONA	**85.09**	3.32	**47.27**	9.84	**33.19**	8.84

The results of Table 3 show that the re-training activity we conducted in the present study, combined with the effects of the two-stage fine tuning procedure, has had very positive effects on both the so called **BingCORONA_BingCORONA** and **CORONA_CORONA** models, also when compared with the results of our previous study, yielding a notable increase in terms of all the considered metrics. As to the **BING_BING** model, instead, it only improves on the IoU parameter**.** The following fact is of great interest: as the most notable increase in accuracy has been achieved in both models based on CORONA satellite imagery, while the simple Bing model presented no significant variation, this adds experimental evidence to the intuition that the combination of the two stage fine tuning procedure with the activities of transfer learning becomes really effective (with more accurate results) only when the corresponding model was built on top of the **CORONA** imagery. While it is true that in other researches, including the one we conducted previously, the integration of CORONA satellite imagery into generic AI models had produced inconclusive results (with motivations ranging from low resolution up to environmental factors, like cloud cover for example), the present study supports the hypothesis that the inclusion of CORONA imagery has the potential to enhance an AI model's performance that has the task of recognizing *tells* from satellite imagery. In other words, the improvement we have measured, at a pixel level, substantiates the thesis that transfer learning and complex fine-tuning activities may

benefit from the additional contextual information provided by these kinds of imagery, thus corroborating long-established archaeological insights of the same sign. Nonetheless, the results of Table 3 have (simply) informed us about the ability of our models to recognize if a given pixel is either comprised within a tell or not. We are, obviously, also interested in elevating our comprehension about the ability of our AI models to recognize a tell as a whole. Table 4 gives an answer to this question by providing the results we achieved in terms of the general accuracy in detecting tells, as emerging from the testing activities conducted with the same 20 images mentioned before. The used metrics, here, were those of *accuracy* and *recall* (the mathematical definitions of which can be easily retrieved from the specialized literature), while TP stands for true positives, TN: true negatives, FP: false positives and FN: false negatives. Table 3, again, highlights the increased ability of the BingCORONA_BingCORONA model in detecting tells, reaching a detection accuracy in the neighborhood of 90% (while our previous results hardly surpassed 80% [19]), with a very low percentage of both false positives and negatives.

Table 4. Testing: tell detection (accuracy and recall).

Model	Accuracy	Recall	TP	TN	FP	FN
Bing_Bing	0.75	0.50	4	11	1	4
BingCORONA_BingCORONA	**0.90**	**0.88**	7	11	1	1
CORONA_CORONA	0.70	0.67	4	10	4	2

New discoveries

Beyond the positive results reported in Tables 2, 3, and 4, the novelty of our work lies in the idea to use machine predictions (upon approval of the domain experts) to decide to extend the set of archaeological sites to be inspected during field survey campaigns, which we did. As already anticipated, our best AI model (i.e., BingCORONA_BingCORONA**)** produced prediction heatmaps, like that shown in Fig 3.

Fig 3. Example of an AI-generated heatmap used to predict the presence of archaeological sites.

Disclaimer: All the displayed data fall under the condition of fair use utilization of geographical data for academic purposes. The list of all relevant data/software provider(s) is as follows: (i) original maps creation under the Section 5 of Microsoft Bing Maps Platform APIs' terms of use (https://www.microsoft.com/en-us/maps/product/print-rights); (ii) maps display achieved with an open source software, under the GNU licenses of QGIS (https://qgis.org/en/site/) and QuickMapsServices

(https://github.com/nextgis/quickmapservices); (iii) final maps elaboration achieved with a software developed by the authors and available at (https://github.com/alepistola/AI_floodplains).

These heatmaps were analyzed by the archaeologists who compared them with the list of sites of potential interest identified through standard remote sensing operations. In our case, of which the heatmap of Fig 3 is an example, our attention was mainly attracted by machine predictions for a number of specific sites that were not previously recognized as potential tells before through traditional methods. Of these sites, eight were accompanied with very high values of probability of being positive cases as returned by the AI model, which led to one of the key results presented in this paper. Subsequently, two field reconnaissance campaigns were conducted in January 2023 and January 2024, covering the Iraqi district of Abu Ghraib at the northwestern apex of the Mesopotamian floodplain. Field activities were directed at verifying sites identified using the CORONA imagery (both those identified with standard remote sensing procedures and those suggested by the AI model described above). During these two campaigns, a total of 96 potential sites were investigated (including the eight suggested by the AI). Of these 96, only 15 showed no signs of ancient anthropogenic activity and were thus false positives. The field survey results revealed, in fact, that 81 turned out to be positively confirmed sites. Of the eight sites suggested by the AI, four were among the 81 confirmed sites of archaeological relevance. To be noticed, again, is the fact that all the 81 sites were discovered by virtue of the analyses conducted on the CORONA satellite imagery, being based either on remote sensing or through AI. The validation of these sites was achieved through the collection and subsequent study of superficial ancient ceramics, which also enabled their dating. Fig 4 summarizes these field-survey results, showing the entire survey area inside which both remote sensing- and AI- based predicted sites are shown using dots of different colors (also based on the fact they were confirmed as either positive or negative cases).

Fig 4. Discovered sites during the Abu Ghraib archaeological survey campaigns. Red: positive cases discovered by AI. Blue: positive cases discovered with remote sensing.

Disclaimer: All the displayed data fall under the condition of fair use utilization of geographical data for academic purposes. The list of all relevant data/software provider(s) is as follows: (i) original maps creation under the Section 5 of Microsoft Bing Maps Platform APIs' terms of use

(https://www.microsoft.com/en-us/maps/product/print-rights); (ii) maps display achieved with an open source software, under the GNU licenses of QGIS (https://qgis.org/en/site/) and QuickMapsServices (https://github.com/nextgis/quickmapservices); (iii) final maps elaboration achieved with a software developed by the authors and available at (https://github.com/alepistola/AI_floodplains).

As to the four archaeological sites discovered based on the suggestion of our BingCORONA_BingCORONA deep learning model, Figs 5-8 show, from the left, the heatmap produced based on the Bing imagery, the heatmap produced with the CORONA imagery and, finally an actual ground photo of the site (all the geographical coordinates of these four confirmed sites are listed in the S1 Appendix below).

Fig 5: GHR.036: heatmaps produced by our BingCORONA_BingCORONA model (Bing, left; CORONA, right) and an on-site photo overview (rightmost).

Disclaimer: All the displayed data fall under the condition of fair use utilization of geographical data for academic purposes. The list of all relevant data/software provider(s) is as follows: (i) original maps creation under the Section 5 of Microsoft Bing Maps Platform APIs' terms of use (https://www.microsoft.com/en-us/maps/product/print-rights); (ii) maps display achieved with an open source software, under the GNU licenses of QGIS (https://qgis.org/en/site/) and QuickMapsServices (https://github.com/nextgis/quickmapservices); (iii) final maps elaboration achieved with a software developed by the authors and available at (https://github.com/alepistola/AI_floodplains).

Fig 6: GHR.077: heatmaps produced by our BingCORONA_BingCORONA model (Bing, left; CORONA, right) and an on-site photo overview (rightmost).

Disclaimer: All the displayed data fall under the condition of fair use utilization of geographical data for academic purposes. The list of all relevant data/software provider(s) is as follows: (i) original maps creation under the Section 5 of Microsoft Bing Maps Platform APIs' terms of use (https://www.microsoft.com/en-us/maps/product/print-rights); (ii) maps display achieved with an open source software, under the GNU licenses of QGIS (https://qgis.org/en/site/) and QuickMapsServices

(https://github.com/nextgis/quickmapservices); (iii) final maps elaboration achieved with a software developed by the authors and available at (https://github.com/alepistola/AI_floodplains).

Fig 7: GHR.078: heatmaps produced by our BingCORONA_BingCORONA model (Bing, left; CORONA, right) and an on-site photo overview (rightmost).

Disclaimer: All the displayed data fall under the condition of fair use utilization of geographical data for academic purposes. The list of all relevant data/software provider(s) is as follows: (i) original maps creation under the Section 5 of Microsoft Bing Maps Platform APIs' terms of use (https://www.microsoft.com/en-us/maps/product/print-rights); (ii) maps display achieved with an open source software, under the GNU licenses of QGIS (https://qgis.org/en/site/) and QuickMapsServices (https://github.com/nextgis/quickmapservices); (iii) final maps elaboration achieved with a software developed by the authors and available at (https://github.com/alepistola/AI_floodplains).

Fig 8: GHR.079: heatmaps produced by our BingCORONA_BingCORONA model (Bing, left; CORONA, right) and an on-site photo overview (rightmost).

Disclaimer: All the displayed data fall under the condition of fair use utilization of geographical data for academic purposes. The list of all relevant data/software provider(s) is as follows: (i) original maps creation under the Section 5 of Microsoft Bing Maps Platform APIs' terms of use (https://www.microsoft.com/en-us/maps/product/print-rights); (ii) maps display achieved with an open source software, under the GNU licenses of QGIS (https://qgis.org/en/site/) and QuickMapsServices (https://github.com/nextgis/quickmapservices); (iii) final maps elaboration achieved with a software developed by the authors and available at (https://github.com/alepistola/AI_floodplains).

We conclude this Section by returning to the importance of having updated our AI models both with new deep learning procedures (transfer learning and two stage fine-tuning) and with the CORONA satellite imagery: in Fig 9 the reader can realize the quantity of information revealed by the CORONA imagery (top rightmost picture, Fig 9) with respect to the corresponding Bing one (top leftmost picture, Fig 9) for two of the four discovered sites suggested by AI, namely GHR.078 and GHR.079. It is also worth noting

that, in the same Fig 9, the green contours represent the on-map predictions provided by our BingCORONA_BingCORONA AI model.

Fig 9: Predictions of the BingCORONA_BingCORONA model for sites GHR.078 and GHR.079 (Bing, left; CORONA, right).

Disclaimer: All the displayed data fall under the condition of fair use utilization of geographical data for academic purposes. The list of all relevant data/software provider(s) is as follows: (i) original maps creation under the Section 5 of Microsoft Bing Maps Platform APIs' terms of use (https://www.microsoft.com/en-us/maps/product/print-rights); (ii) maps display achieved with an open source software, under the GNU licenses of QGIS (https://qgis.org/en/site/) and QuickMapsServices (https://github.com/nextgis/quickmapservices); (iii) final maps elaboration achieved with a software developed by the authors and available at (https://github.com/alepistola/AI_floodplains).

Discussion

In this research, the use of AI techniques has been of great help in support to a process which remains, nonetheless, guided by the archaeologist's knowledge and expertise. Deep learning models have helped towards the aim to identify areas potentially containing archaeological sites, albeit neglected during normal remote sensing operations. It has remained an archaeologists' task to take the final decision about the precise locations of the sites to visit, based on their professional experience. In this sense, our proposed AI-based approach to archaeology has been conceived just to provide additional support to the archaeologists, rather than to replace them. Beyond the impact of the AI, in some sense already documented in a previous study of ours [19], it has emerged here also the fundamental role played by the CORONA imagery dataset, and its versatility, especially when used in combination with a deep learning model. This consideration derives not only from our direct experience on the four archaeological sites which were not detectable using the Bing imagery alone, thereby highlighting the substantive impact of incorporating the CORONA satellite imagery into the process, but also considering the state of preservation of the sites which were the subject of the on-field campaigns of 2023 and 2024. To better illustrate this point, Fig 10 shows that, of the 81 archaeological sites discovered during those campaigns, almost all had been destroyed over the past decades: either completely (31) or largely (19) or partially (19); where *completely* means a destruction of almost 100%,

largely means over the threshold of 50% and finally *partially* means below 50%. In this context, the inclusion of CORONA satellite imagery has been fundamental because many of the destroyed sites were no longer visible from modern basemaps (like Bing maps). The CORONA satellite imagery, from the 1960s and early 1970s, has the ability to document a world that has almost disappeared: in the specific case of Abu Ghraib, the loss of the possibility to identify sites with modern basemaps, in fact, would range from 40% to 55%, if totally destroyed sites alone, or totally plus largely destroyed ones, were considered. Thus, the development of an automatic process that is able to identify disappearing sites, by including historical imagery, allows everyone to start a fundamental reflection for the protection of the existing/remaining archaeological evidence.

Fig 10: State of preservation of the 81 archaeological sites discovered during the 2023-2024 on-field campaigns. Disclaimer: All the displayed data fall under the condition of fair use utilization of geographical data for academic purposes. The list of all relevant data/software provider(s) is as follows: (i) original maps creation under the Section 5 of Microsoft Bing Maps Platform APIs' terms of use (https://www.microsoft.com/en-us/maps/product/print-rights); (ii) maps display achieved with an open source software, under the GNU licenses of QGIS (https://qgis.org/en/site/) and QuickMapsServices (https://github.com/nextgis/quickmapservices); (iii) final maps elaboration achieved with a software developed by the authors and available at (https://github.com/alepistola/AI_floodplains).

To conclude this Section, we would like to add that, while we have documented that an AI-based identification process has the potential to make unexpected discoveries, nonetheless, what should not be forgotten is the awareness that we still do not know how this happens. Precisely, this should be the reason that pushes towards the integration of AI with human experts, through collaborative processes, aimed at mitigating classification errors and incorrect interpretations [38-42].

Conclusions

We have described a deep learning model designed to identify sites of potential archaeological interest in the Abu Ghraib district, West of Baghdad in the Mesopotamian floodplain. This AI model has been built incrementally over the past years, using transfer learning techniques and a final two stage fine tuning procedure that has elevated the level of detection accuracy up to 90% (while previous results did not surpass the threshold of 80%). The role played by the CORONA imagery dataset has been

fundamental in this context of vanishing archaeological evidence, as it has allowed the AI to *see* sites no longer visible due to the process of anthropization. Surprisingly, this process has also led to the identification of unexpected archaeological sites, which thus far had not been identified in standard remote sensing operations. In particular, our archaeological team visited the eight new sites suggested by the AI model, also because they had the appropriate morphological characteristics. During our field survey campaign, four of these eight sites have been confirmed as positive cases. In fact, even if they were totally destroyed and no longer visible on the ground, some ceramic sherds could still be collected, making it possible to confirm their existence and date them. It must be acknowledged that without the AI's suggestions, the areas where the sites were confirmed would not have been investigated during routinary field surveys. In the end, the development of AI models able to automatically identify potential sites, no more visible from current basemaps, represents a real breakthrough which could be further extended exploring the possibility of adding other technologies and methods like, for example, LIDAR and super-resolution ones [43-48].

Data Availability Statement

All results were obtained using open-source software and models, as well as publicly available data (images, annotations) and computational resources (Google Colab), making this type of work highly accessible and replicable even in resource-limited research environments. In addition to the specific information provided within the document, all the code, data, archeological annotations and various resources are available on GitHub (https://github.com/alepistola/AI_floodplains). Moreover, all the geographical data displayed in this paper falls within the conditions of correct use of geographical data for academic purposes. The creation of the maps took place respecting the terms of use of the Microsoft Bing Maps API (https://www.microsoft.com/en-us/maps/product/print-rights) while the related visualization of the maps was made possible via an open-source software regulated by the GNU licenses of QGIS (https://qgis.org/en/site/) and QuickMapsServices (https://github.com/nextgis/quickmapservices). The final processing of the maps was obtained instead through the software we developed and made available on Github at the address above.

References

1. Wilkinson TJ. Archaeological landscapes of the Near East. University of Arizona Press; 2003.

2. Bickler SH. Machine learning arrives in archaeology. Advances in Archaeological Practice. 2021;9(2):186–191.
3. Mantovan L, Nanni L. The computerization of archaeology: Survey on artificial intelligence techniques. SN Computer Science. 2020; 1:1–32.
4. Tenzer M, Pistilli G, Bransden A, Shenfield A. Debating AI in archaeology: applications, implications, and ethical considerations. Internet Archaeology. 2024;(67).
5. Lyons TR, Hitchcock RK. Aerial remote sensing techniques in archeology. 2. Chaco Center; 1977.
6. Kucukkaya AG. Photogrammetry and remote sensing in archeology. Journal of Quantitative Spectroscopy and Radiative Transfer. 2004;88(1-3):83–88.
7. Orengo HA, Conesa FC, Garcia-Molsosa A, Lobo A, Green AS, Madella M, et al. Automated detection of archaeological mounds using machine-learning classification of multisensor and multitemporal satellite data. Proceedings of the National Academy of Sciences. 2020;117(31):18240–18250.
8. Guyot A, Lennon M, Lorho T, Hubert-Moy L. Combined detection and segmentation of archeological structures from LiDAR data using a deep learning approach. Journal of Computer Applications in Archaeology. 2021;4(1):1.
9. Argyrou A, Agapiou A. A Review of Artificial Intelligence and Remote Sensing for Archaeological Research. Remote Sensing. 2022;14(23):6000.
10. Caspari G, Crespo P. Convolutional neural networks for archaeological site detection–Finding “princely” tombs. Journal of Archaeological Science. 2019; 110:104998.
11. Guyot A, Hubert-Moy L, Lorho T. Detecting Neolithic Burial Mounds from LiDAR-Derived Elevation Data Using a Multi-Scale Approach and Machine Learning Techniques. Remote Sensing, 2018; 10(2):225.
12. Soroush M, Mehrtash A, Khazraee E, Ur J. Deep Learning in Archaeological Remote Sensing: Automated Qanat Detection in the Kurdistan Region of Iraq. Remote Sensing, 2020; 12(3):500.
13. Ur J, Babakr N, Palermo R, Creamer P, Soroush M, Ramand S, et al. The Erbil Plain Archaeological Survey: Preliminary Results, 2012–2020. Iraq. 2021; 83:205-243.

14. Landauer J, Klassen S, Wijker AP, van der Kroon J, Jaszkowski A, Verschoof-van der Vaart WB. Beyond the Greater Angkor Region: Automatic large-scale mapping of Angkorian-period reservoirs in satellite imagery using deep learning, PLOS One. 2025; 20(3): e0320452.
15. Gattiglia G. Managing Artificial Intelligence in Archeology. An overview. Journal of Cultural Heritage. 2025; 71:225-233.
16. Roccetti M, Casini L, Delnevo G, Orrù V, Marchetti N. Potential and limitations of designing a deep learning model for discovering new archaeological sites: A case with the Mesopotamian floodplain. In: Proceedings of the 6th EAI International Conference on Smart Objects and Technologies for Social Good; 2020. p. 216–221.
17. Casini L, Roccetti M, Delnevo G, Marchetti N, Orrù V. The barrier of meaning in archaeological data science. arXiv preprint arXiv:210206022. 2021.
18. 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; 2022. p. 378–383.
19. Casini L, Marchetti N, Montanucci A, Orrù V, Roccetti M. A human–AI collaboration workflow for archaeological sites detection. Scientific Reports. 2023;13(1):8699.
20. Marchetti N, Bortolini E, Menghi Sartorio JC, Orrù V, Zaina F. Long-Term Urban and Population Trends in the Southern Mesopotamian Floodplains. Journal of Archaeological Research. 2024; p. 1–42.
21. Traviglia A, Cowley D, Lambers K, et al. Finding common ground: Human and computer vision in archaeological prospection. AARGnews. 2016; 53:11–24.
22. Adams RM. Settlement and Irrigation Patterns in Ancient Akkad. In: McG Gibson, The City and Area of Kish. Field Research Projects; 1972. p. 182-208.
23. Comer DC, Harrower MJ, Casana J, Cothren J. The CORONA atlas project: Orthorectification of CORONA satellite imagery and regional-scale archaeological exploration in the Near East. Mapping archaeological landscapes from space. 2013; p. 33–43.
24. Kennedy D. Declassified satellite photographs and archaeology in the Middle East: case studies from Turkey. Antiquity. 1998;72(277):553–561.

25. Kouchoukos N. Satellite images and Near Eastern landscapes. Near Eastern Archaeology. 2001;64(1-2):80–91.
26. Philip G, Donoghue D, Beck A, Galiatsatos N. CORONA satellite photography: an archaeological application from the Middle East. Antiquity. 2002;76(291):109–118.
27. Ur JA. Settlement and landscape in northern Mesopotamia: the Tell Hamoukar survey 2000-2001. Akkadica. 2002;123(1):57–88.
28. Casana J, Wilkinson TJ. Settlement and landscapes in the Amuq region. The Amuq Valley Regional Projects. 2005; 1:1995–2002.
29. QUANTUM G. Development Team. Quantum GIS geographic information system. http://qgis osgeo org. 2011.
30. Buslaev A, Iglovikov VI, Khvedchenya E, Parinov A, Druzhinin M, Kalinin AA. Albumentations: Fast and Flexible Image Augmentations. Information. 2020;11(2). doi:10.3390/info11020125.
31. Iakubovskii P. Segmentation Models Pytorch. https://smpreadthedocsio/en/latest/. 2020.
32. Li R, Zheng S, Zhang C, Duan C, Su J, Wang L, et al. Multiattention network for semantic segmentation of fine-resolution remote sensing images. IEEE Transactions on Geoscience and Remote Sensing. 2021; 60:1–13.
33. Torrey L, Shavlik J. Transfer learning. In: Handbook of research on machine learning applications and trends: algorithms, methods, and techniques. IGI global; 2010. p. 242–264.
34. Deng J, Dong W, Socher R, Li LJ, Li K, Fei-Fei L. Imagenet: A large-scale hierarchical image database. In: 2009 IEEE conference on computer vision and pattern recognition. Ieee; 2009. p. 248–255.
35. Valizadeh Aslani T, Shi Y, Wang J, Ren P, Zhang Y, Hu M, et al. Two-stage fine-tuning: A novel strategy for learning class-imbalanced data. arXiv preprint arXiv:220710858. 2022.
36. Chicco D, Jurman G. The Matthews correlation coefficient (MCC) should replace the ROC AUC as the standard metric for assessing binary classification. BioData Mining. 2023;16(1):1–23.
37. Sech G, Soleni P, Verschoof-van der Vaart WB, Kokalj Z, Traviglia A, Fiorucci M. Tranfer Learning of Semantic Segmentation Methods for Identifying Buried Archaeological Structures on LiDAR Data. In: IGARSS 2023-2023 IEEE International Geoscience and Remote Sensing Symposium. IEEE; 2023. p. 6987–6990.

620 38. Baeza Yates R, Estevez Almenzar M. The relevance of non-human errors in machine learning.
621 In: Hernandez-Orallo J, Cheke L, Tenebaum J, Ullman T, Martınez-Plumed F, Rutar D, Burden
622 J, Burnell R, Schellaert W, editors. Proceedings of the Workshop on AI Evaluation Beyond
623 Metrics (EBeM 2022); 2022 Jul 25; Vienna, Austria. [Aachen]: CEUR-WS; 2022. CEUR
624 Workshop Proceedings; 2022. p. 6.
625 39. Yampolskiy RV. On monitorability of AI. AI and Ethics. 2024; p. 1–19.
626 40. 35. Roccetti M, Tenace M, Cappiello G. Prescient Perspectives on Football Tactics: A Case
627 with Liverpool FC, Corners and AI. 2024; doi:10.13140/RG.2.2.27842.59847/16.
628 41. Gao L, Guan L. Interpretability of Machine Learning: Recent Advances and Future Prospects.
629 IEEE MultiMedia. 2023;30(4):105–118. doi:10.1109/MMUL.2023.3272513.
630 42. Messeri L, Crockett M. Artificial intelligence and illusions of understanding in scientific research.
631 Nature. 2024;627(8002):49–58.
632 43. Ji J, Qiu T, Chen B, Zhang B, Lou H, Wang K, et al. Ai alignment: A comprehensive survey.
633 arXiv preprint arXiv:231019852. 2023.
634 44. Thabeng OL, Adam E, Merlo S. Evaluating the Performance of Geographic Object-Based
635 Image Analysis in Mapping Archaeological Landscapes Previously Occupied by Farming
636 Communities: A Case of Shashi–Limpopo Confluence Area. Remote Sensing. 2023;15(23).
637 doi:10.3390/rs15235491.
638 45. Canedo D, Hipolito J, Fonte J, Dias R, do Pereiro T, Georgieva P, et al. The Synergy between
639 Artificial Intelligence, Remote Sensing, and Archaeological Fieldwork Validation. Remote
640 Sensing. 2024;16(11):1933.
641 46. Lepcha DC, Goyal B, Dogra A, Goyal V. Image super-resolution: A comprehensive review,
642 recent trends, challenges and applications. Information Fusion. 2023; 91:230–260. doi:
643 https://doi.org/10.1016/j.inffus.2022.10.007.
644 47. Salvetti F, Mazzia V, Khaliq A, Chiaberge M. Multi-image super resolution of remotely sensed
645 images using residual attention deep neural networks. Remote Sensing. 2020;12(14):2207.
646 48. Wei Z, Zhang S. Small object detection in satellite remote sensing images based on super-
647 resolution enhanced DETR. In: International Conference on Remote Sensing, Mapping, and
648 Image Processing (RSMIP 2024). vol. 13167. SPIE; 2024. p. 33–42.

Supporting information captions

S1 Appendix. Geographical coordinates of GHR.036, GHR.077, GHR.078 and GHR.079 archaeological sites.