Contents lists available at ScienceDirect

Journal of Cultural Heritage

journal homepage: www.elsevier.com/locate/culher

Review

From fragments to digital wholeness: An AI generative approach to reconstructing archaeological vessels

Lorenzo Cardarelli

Dipartimento di Scienze dell'Antichità, University of Rome La Sapienza, Piazzale Aldo Moro 5, 00185, Rome, Italy
CNR - Istituto di Scienze del Patrimonio Culturale, Area della Ricerca Roma 1, Via Salaria Km 29, 00015, Monterotondo Scalo (Rome), Italy

ARTICLE INFO

Article history:
Received 7 March 2024
Accepted 27 September 2024
Available online 15 October 2024

Keywords:
Pottery analysis
Deep learning
Quantitative archaeology
Generative AI

ABSTRACT

Reconstructing archaeological vessels from their fragments is a complex task that requires a long investment of time as well as in-depth knowledge of specific archaeological material. This paper proposes a framework based on generative artificial intelligence to reconstruct the entire vessel from a fragment. The proposed framework is based on a fragment simulation mechanism and the combination of three different deep learning models that position, reconstruct, and post-process the fragment to obtain a ready-to-use reconstruction. The method is applied as a case-study to a dataset of six Italian Bronze and Early Iron Age burial contexts, including about 4000 complete vessels and over 400 actual fragments. The results are evaluated using statistical metrics and expert human evaluation, showing promising results. The proposed method is a positive application of generative artificial intelligence in archaeology and provides a solution to the use of fragments in the digital and computational analysis of ceramics. The dataset, as well as the code used and the analytical pipeline, are fully available in the supplementary materials.

Introduction

Ceramics play an important role in archaeology and in reconstructing the past. The analysis of pottery provides information on social and productive organisation, relationships between groups, as well as establishing relative chronologies [1–5]. Over the years, the typological and traditional study that has dominated the analysis of pottery has been complemented by new computer-based approaches that allow the quantitative analysis of archaeological ceramics. These approaches cover different aspects, such as classification, unsupervised analysis, and reconstruction of fragments [6]. Although ceramics are an important and extremely valuable sample of archaeological reality, they are also extremely fragile and are often found in a fragmentary state. While it is traditionally possible to deal with ceramic fragments based in-depth knowledge of the archaeological material, the hurdle is more difficult to overcome using computational and quantitative methods. Faced with this problem, various methods and approaches have been developed to deal with fragments [7,8]. Among these approaches, the reconstruction of the entire vessel is an elegant and powerful solution, as it allows reconstruction and creates the foundations for the application of other computerised methods, which have been formulated and designed primarily for complete records [9–13]. To this end, several methods have been employed in the reconstruction of the entire shape of vessels, ranging from iterative to machine learning (ML) based methods [14]. Recently, the latter approaches seem to deliver the best results [15]. Indeed, ML is a valuable support to the archaeologist's work, and the number of application examples tends to increase [16]. Generally, ML is mainly used for classification and, to a lesser extent, for unsupervised aspects such as feature extraction or clustering [13]. More recently, generative AI has emerged, i.e., ML models capable of generating data such as images, text, or audio [17–21]. Early generative models include Variational Autoencoders (VAEs) [17,22] or Generative Adversarial Networks (GANs) [17,23], while the most recent and current include Large Language Models (LLMs) [24–26] or Diffusion Models (DMs) [27–30]. These latter models are extremely powerful and complex, but also extremely expensive in terms of computational resources [31]. As far as archaeology is concerned, the use of generative deep learning models is of great help, especially in the generalisation of tasks such as the reconstruction of ceramic fragments, which require a long investment of time, energy, and in-depth knowledge of the specific archaeological material.

E-mail address: lorenzo.cardarelli@uniroma1.it

https://doi.org/10.1016/j.culher.2024.09.012

Research aims

This paper introduces a generative AI-based framework for reconstructing vessels from fragments, removing assumptions about morphology or size. Reconstructing the entire shape of a pot is crucial for advancing computerised methods in ceramics analysis, extending applications to diverse contexts like settlement or survey scenarios where fragments are prevalent. In this sense, the complete reconstruction of the vase is considered the best solution for the application of computerised techniques for the analysis and management of ceramic fragments, firstly because it allows the result to be displayed graphically, and secondly, because the creation of a homogeneous database of raw data (images) allows the application of a wide range of quantitative analysis methods that have been extensively developed over time [7,10,32]. This approach complements traditional methods, supporting typological classification [33–35]. While embracing traditional approaches, the method establishes a foundation for analytically and reproducibly reconstructing vessel morphology, incorporating the untapped potential of generative AI [17] which is emerging as one of the most promising areas of AI and which has had little application in archaeology. The goal is to offer a precise and robust, applicable method for real fragments, incorporating contextual information like chronology and decoration. The framework's adaptability to diverse contexts will be tested, ensuring broader applicability beyond the training dataset.

Materials and methods

Previous works

As mentioned in the introduction, fragment reconstruction is a complex task and various methods have been proposed to tackle it: Eslami and collegues [14] propose a systematic review and analysis of these methods, showing a wide variety of approaches and solutions that have been developed [14]. These methods use different approaches, starting from the definition of the data itself (3D scanning, 2D representation), while agreeing on the need to reconstruct the profile using other existing fragments (matching). In this sense, the method proposed here has strong affinities with the one proposed by Navarro and collegues [15]. In this work, the reconstruction is proposed without the specific use of other fragments, but rather using a reference dataset of complete vessels. Using a dataset of two-dimensional profile representations, the authors propose and analyse a comparison between different customised models based on the GAN architecture. This type of architecture consists of two neural networks: a $G(x)$ generator and a $D(x)$ discriminator. In simple terms, the generator's task is to create the image, while the discriminator's task is to try to work out whether the image is real or generated and the joint training of these two networks leads to the creation of realistic generated images that are indistinguishable from the original ones. The model takes as input a fragmented half of the vessel [rim or foot, 15] and returns the complete vessel. The validity and quality of the reconstructions from the test set are confirmed by metrics [15, Table 1] and a human evaluation test [15]. The authors have also recently published a version of the model that uses three-dimensional models as training example [36]. Regarding the method proposed in this paper, if the methodological afferent sphere is common, differences can be highlighted, starting from the general architecture of the model, moving to the management of the training dataset and finally the possibility of using fragments not only related to the rim or foot.

Dataset creation, preprocessing and fragment-use issues

The dataset used to train the model is obtained from archaeological pottery drawings from a selection of Italian Bronze and Early Iron Age burial contexts [37–40]. The use of 2D archaeological drawings as a basis for this type of analysis is based on several considerations. They are primarily technical, standardised, and schematic representations of three-dimensional objects. Their use is widespread in archaeology, with no geographical or chronological differences. In fact, archaeological drawings can be defined as the standard for the graphic representation of archaeological ceramics. Although they are usually made by hand, new digitalisation techniques make it possible to obtain ceramic drawings that are analogous to those made by hand [8], underline the renewed need for this type of documentation. In addition, they can be found in large quantities, allowing for a vast and easily generated dataset. It should also be remembered that the drawings are made by specialists who have in-depth knowledge of the archaeological material, thus allowing for a high-quality dataset [41]. Furthermore, ceramic drawings have been practised for a long time and therefore allows for the creation of datasets and the retrieval of data from old publications (alias *legacy data*) [42,43]. Although the use of three-dimensional models [36] provide a better understanding of morphological characteristics and bring innumerable improvements, they are still difficult to apply to large and diverse quantities of material, unlike archaeological drawing. The training dataset only contains complete vessels and is created from the scans of the publications. The profiles are processed using a photo editing software, removing the prospectus, eliminating the handling elements and camping the profile in black. A Python script then pre-processed the images by creating images of a standardised $256 \cdot 256$ pixels size, without distorting the aspect-ratio of the vessel. This image dataset is complemented by tabular datasets containing and *ids*, some archaeological information and the various bibliographical references which is fully available in the Supplementary materials (SM).

While the design of the training dataset is straightforward and well established in relation to similar work [11–13,15], the creation and conceptualisation of the actual fragment dataset is more difficult. The formalisation of this dataset and its construction is closely related to the problems associated with the use of fragments. ML algorithms require images of the same size to be applied. However, pots are not the same size: trivially, there are tall, narrow pots (commonly called *closed shapes*) and short, wide pots (the *open*

Table 1
This table recap the dataset used, for the spatial distribution of the contexts please refers to Supplementary materials, Fig. 1. As far the chronology is concerned, here we have the absolute chronology for each phase: MBA (Middle Bronze Age): 1700–1350 BCE; LBA (Late Bronze Age): 1350–1200/1150 BCE; FBA (Final Bronze Age): 1200/1150–1000/950 BCE; EIA1 (Early Iron Age 1): 1000/950–820/800 BCE; EIA2 (Early Iron Age 2): 820/800–725 BCE [56].

Context	Abbreviation	Sample	Type	Chronology	Actual fragments	Bibliography
Casinalbo	CSLN	274	Training dataset	MBA/LBA	80	[44]
Pianello di Genga	PDG	250	Training dataset	FBA	79	[45]
Osteria dell'Osa	OSTS	1709	Training dataset	EIA1	86	[46]
Torre Galli	TRRGLL	490	Training dataset	EIA1/EIA2	54	[47]
Quattro Fontanili	QF	1140	Training dataset	EIA1/EIA2	98	[48–54]
Narde	NRD	306	Post training evaluation	FBA	20	[55]

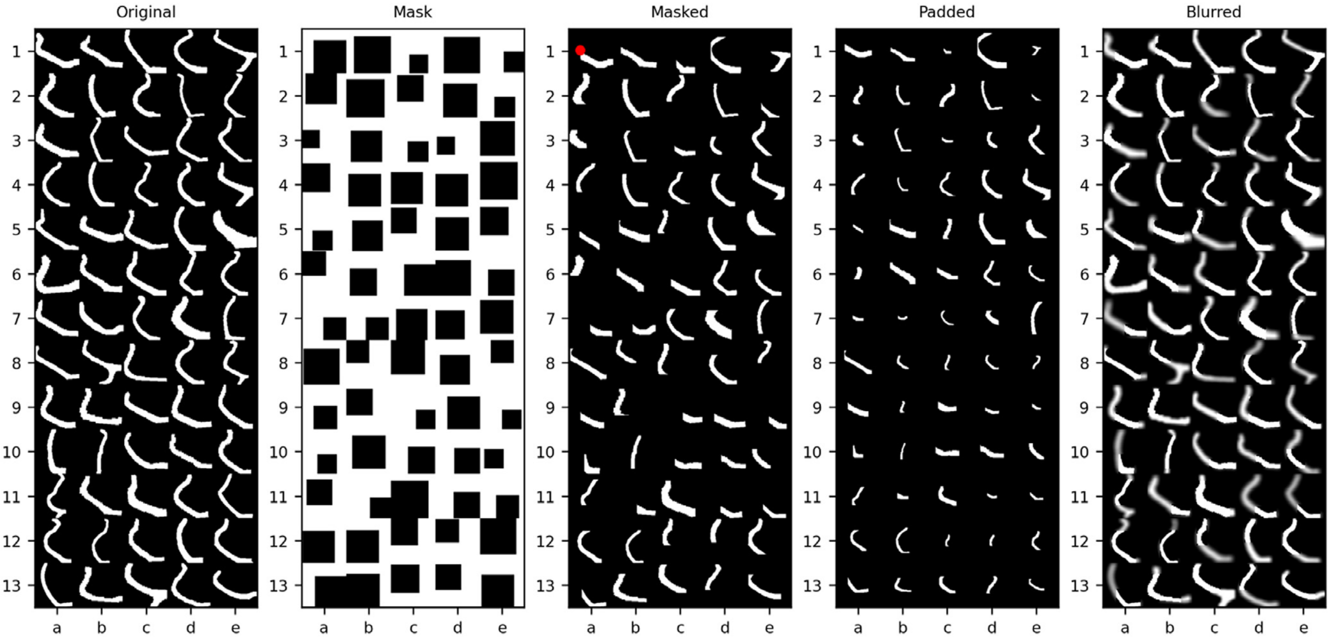

Fig. 1. A batch of images showing the fragment simulation process. Fragments are created according to different size and position (Mask, Masked). The fragments can therefore be very small (e.g. 3D, 10A), or very large (e.g. 1D, 6D). The Padded images simulate the input of the actual fragment dataset, while the Blurred images simulate specific outputs of the Reconstructor. The red dot in Masked 1A represent the coordinate to be predicted (see the Supplementary material for more information).

Table 2
A summary table of the architecture and tasks of the proposed deep learning pipeline. Full explanation is available in the Supplementary material.

Network	References	Architecture and task
Regressor	SM §1	A modified *ResNet101* network. Through an iterative training process, it positions the fragment within the standardised space of the image.
Reconstructor	SM §2	A customised convolutional Variational Autoencoder. Reconstructs the intact vessel from the positioned fragment.
Denoiser	SM §3	A customised convolutional Autoencoder. Post-processing of the reconstructed vessel to eliminate some imperfections

shapes). As far as entire vessels are concerned, the problem has been solved by using standardised size files into which the vessel has been inserted without distorting the aspect ratio. This solution cannot be applied to the fragments because we do not have the necessary information to scale them and place them correctly in the standardised image. In other words, without knowing the full dimensions, we cannot use the bounding box of the vessel as a criterion for maintaining the aspect ratio and positioning the object within the standardised images. As part of the proposed work, the first neural network attempts to get to the bottom of this problem by correctly placing the fragment in a standardised file based on morphological comparison with complete, similar vessels. With this in mind, the actual fragment dataset is created by placing the fragment in the centre of a standardised 256 · 256 pixels image.

The work analyses a dataset consisting of several contexts distributed throughout the Italian Bronze Age and Early Iron Age (Supplementary materials, Fig. 1). The training dataset of the model consists of 5 contexts with a total of about 4000 entire vessels. An additional context, not included in the training dataset, is used to evaluate the application of the method on contexts and materials other than those on which the model is specifically trained. For the practical application of the model, for each context, actual vessel fragments were also collected (Table 1; Supplementary materials, Table 1).

Deep learning pipeline

The deep learning proposed in this paper includes a processing/simulation step, where the data set is prepared for the training of the model, and the application of three different neural networks that perform different tasks. A table (Table 2) as well as the flowchart (Supplementary materials, Fig. 2), summarises the three models and their tasks. For details, please refer to SM.

Fragment simulation and basic learning mechanism

ML algorithms learn by example. It follows that an active link between the fragment and the full-preserved vessel is required to reconstruct the entire vessel. The simulation pipeline described here applies to all three neural networks used in this work with some specific modification. Starting from a complete pot (Fig. 1, Original) a fragment is simulated by creating a mask (Fig. 1, Mask) and removing the part of the vessel outside the mask (Fig. 1, Masked). This squared mask varied in size and is created from a uniform distribution (30 to 120 pixels) to simulate fragments of different sizes and thus different preservation. The mask is also applied along the entire profile of the vessel, again using random values from a uniform distribution, so that both base and rim fragments and other morphological features within the vessel could be simulated. Obviously, during the training process, these values constantly change to obtain different fragments of the same vessel, thus increasing the training capacity of the model and its resistence to overfitting and allowing the model to trace back to the entire vessel from different fragments of the same shape.

Image basic preprocessing includes the transformation of the image into a tensor [57] and the resizing into a dimension of 128 · 128 pixels. Basic data augmentation [58] is also applied to the images, including random rotation (5°), translation and the application of a Gaussian blur. During the training, the dataset has been

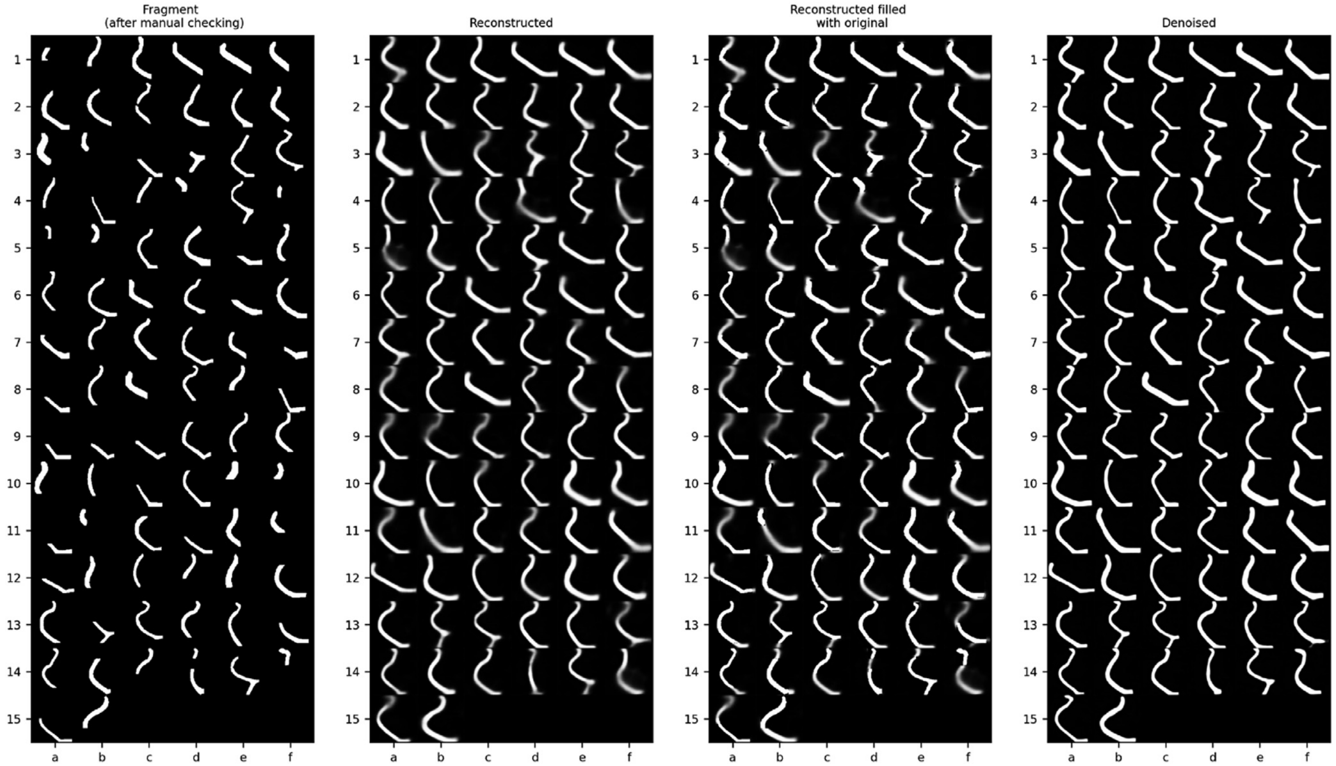

Fig. 2. Fragments and reconstructions (Reconstructed, Reconstructed filled with original and Denoised) from the context of Osteria dell'Osa.

divided into a train and a test [59] set with a ratio of 80 % (3090 records) and 20 % (773) respectively.

Motivating the proposed architecture

The proposed pipeline is based on different architectures with specific tasks. The division of the models helps to divide the workload, which can be carried out individually and independently, a useful approach for hardware resources that are not as powerful as those usually available to archaeologists, who often use their own PC or laptop (SM § 4). The positioning of the fragment is also an extremely important task from an archaeological point of view: in fact, it would still be possible to use only one model to carry out the reconstruction (using Padded as input image and the original as output image, Fig. 1), but this would not allow control over the positioning of the fragment within the image. Furthermore, the reconstruction process would inevitably change some morphological features of the fragment, which must be preserved. Additionally, separating the *Regressor* from the *Reconstructor* will, as we shall see, add an intermediate level of analysis and verification of the fragment's positioning within the image and manual correction. In conclusion, in an ever-growing landscape of generative AI [28], the VAE architecture (SM § 2) at the heart of the reconstruction mechanism is not as powerful as other methods [60], but in its probabilistic structure and ease of training, it proved to be an immediate solution, not wasteful in terms of computational power, and extremely effective for the problem at hand.

Application to real fragments

Once the model has been trained and tested, it will be applied to the actual fragments. In this case, the application will be divided into the different archaeological contexts (Table 1). Fine-tuning of the weights of each model will also be carried out using the individual contexts to obtain better results. Such fine-tuning will be done with a lower learning rate to adjust the parameters more precisely and to keep the learned information from the full training dataset.

Evaluation metrics

Evaluating the quality of generative models' outputs is a complex problem [61]. In this paper, the evaluation is carried out on two levels. First, the results of the test set are analysed with the same metrics used and proposed in [15], in order to also have a comparison with their work. The metrics used are SMSE,[1] DICE coefficient [62], Geometry Score (GS) [63] and Frechet Inception Distance (FID) [64]. The first two metrics measure the similarity/distance between the reconstructed image and the original image at the pixel level. The comparison is made for each image pair and then an average is taken to characterise the whole sample. For SMSE, which is a distance, a lower value indicates greater similarity, while the DICE coefficient varies between 0 (complete difference between the images) and 1 (complete match between the images). GS and FID, on the other hand, are complex metrics that operate at the level of the entire distribution and their use is specific to evaluating the performance of a generative model. The evaluation metrics are calculated on the test set and the outcomes are shown in the Results section. Such measures are mainly used to evaluate and characterise the output in a general way, to get some reference values and to test some post-processing procedures to get a result as close to the original as possible.

A second level, related to human evaluation [65], will not concern the results obtained on the test set, but rather the application of the method to actual fragments. This level of evaluation will be carried out by experts with knowledge of ceramics and archaeological contexts and will be based on two tests: in the first test, the

[1] $SMSE = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(\widehat{x}_i - x_i)^2}$, where $\widehat{x}_i$ is the reconstructed image and x_i is the original image.

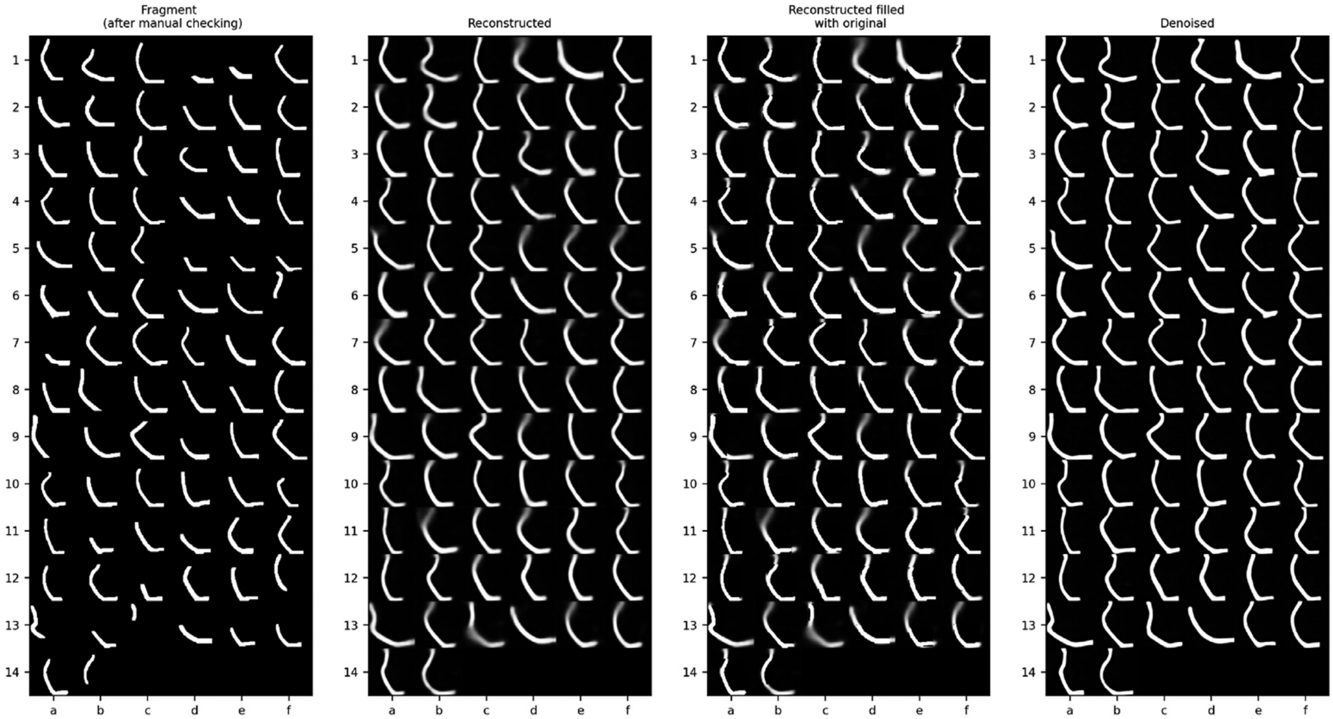

Fig. 3. Fragments and reconstructions (Reconstructed, Reconstructed filled with original and Denoised) from the context of Casinalbo.

evaluator (acting as a discriminator) will attempt to distinguish the reconstructed - therefore, generated - fragments from the original and fully preserved vessels within the training dataset. In the second test, the evaluator will be asked to assess the quality of the reconstruction using only the fragments and their reconstructed version. The rating will be made using a 5-point Likert scale [66], broken down as follows

- 1: The reconstruction presents serious problems, making it impossible to identify the original vessel / The reconstruction is not a vessel.
- 2: The reconstructed vessel can be identified, but its overall morphology has been significantly altered.
- 3: The reconstruction is correct, but the primary morphological features have been significantly altered.
- 4: The primary features have been preserved, but some secondary features have been altered.
- 5: Reconstruction is excellent, with all morphological features, including secondary features, preserved.

It should be noted that the evaluators are not aware of the method used for reconstruction and that they have no information about the data, including the number of generated fragments within their dataset. Not all evaluators will evaluate all datasets, but only those for which they have specific knowledge. Both evaluations were carried out using a web application. In the first evaluation, a sample of the same size of real and generated vessels was provided and the evaluation was carried out by analysing the profile and a three-dimensional reconstruction of it in order to better assess the volumetry of the pot (Supplementary materials, Fig. 3). The second evaluation was carried out by comparing the positioned fragment with the reconstruction proposed by the model (Supplementary materials, Fig. 3).

Results

The results will be presented in two sections: the first will cover the results obtained on the training dataset (§ 4.1), and the second will cover the results obtained by applying the model to real fragments (§ 4.2).

Table 3
Table showing the $\boldsymbol{R}^2$ values obtained by the Regressor on the test set.

	Height	Width	Scale
Using archaeological tabular data	0.93	0.86	0.83
Only using images	0.86	0.81	0.79

Evaluating the model on the training dataset

The *Regressor* is applied both to the image-only dataset and the image plus archaeological information dataset (SM § 1). In this case, the archaeological information concerns the arrangement into open and closed shapes and their classification according to functional elements.[2] The additional archaeological information is intended to add further layers of context to the image. The results are presented in the table (Table 3) and in the multiple scatter plots that relate the predicted and actual values (Supplementary materials, Fig. 4).

An identity line is also shown within each scatterplot showing the best theoretical result. The results obtained on the test set show a rather high R^2 value for all the variables to be predicted (ranging from 0 to 1). In particular, the use of additional archaeological information seems to improve the results, although not significantly. Analysis of the loss curve shows that the model has reached a plateau and therefore there is no need to increase the number of epochs and no evident overfitting phenomena appear to be present (Supplementary materials, Fig. 5).

[2] The classification is the same used in Cardarelli 2022 [13]: **Class 1**: open vessel with horizontal handle and non-articulate profile (*bowls, dishes*). **Class 2**: open vessel with articulated profile or one or two vertical handles (*cups, mugs, goblets, tankards*). **Class 3**: closed vessel with horizontal handle (*jars* and *necked jars*). **Class 4**: closed vessel with one or two vertical handles (*jugs* and *amphoras*).

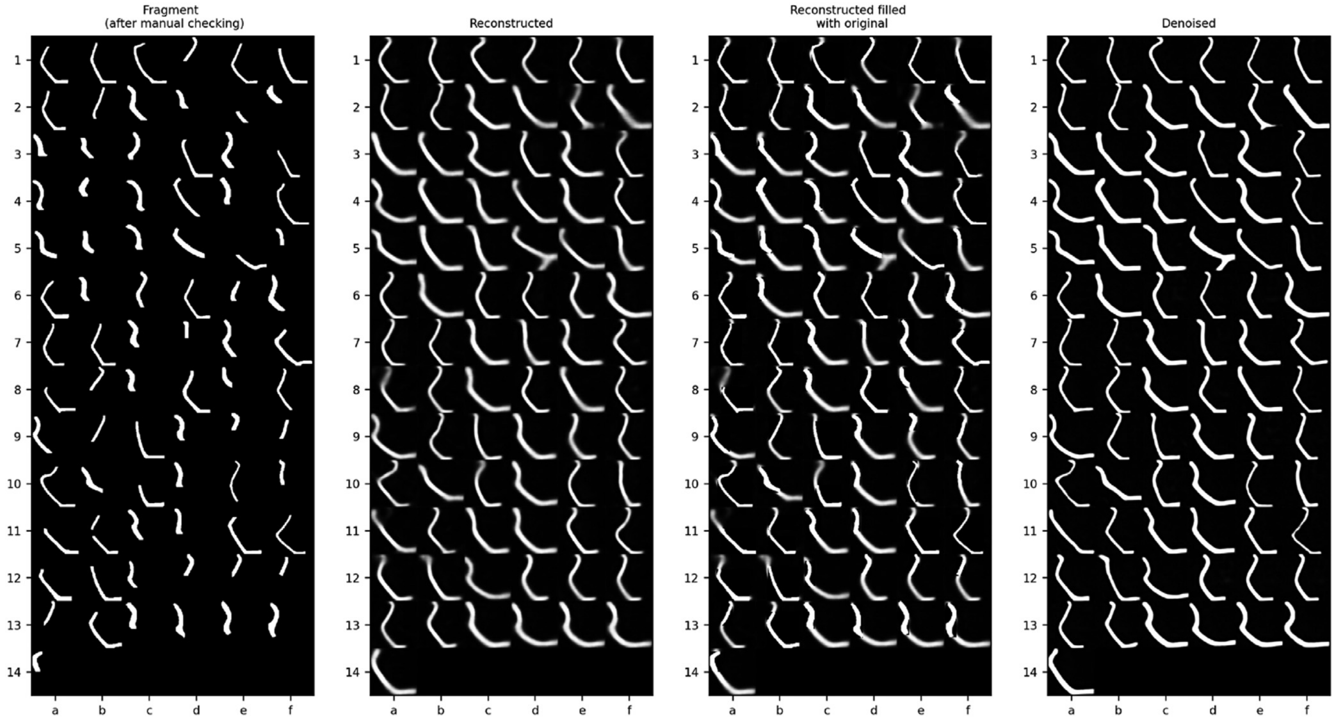

Fig. 4. Fragments and reconstructions (Reconstructed, Reconstructed filled with original and Denoised) from the context of Pianello di Genga.

Table 4
Table showing the results obtained on the test set using the metrics proposed by Navarro and colleagues [15].

Metrics	Reconstructed vanilla	Reconstructed filled with original	Denoised
RMSE	0.2127	0.1772	0.1871
Dice coefficient	0.8441	0.8851	0.8721
GS score	0.0012	0.0008	0.0013
FID score	0.0174	0.0094	0.0018

After the training, the results obtained on the test set are shown both for the *Reconstructor* as well as for the *Denoiser* (Supplementary materials, Fig. 6). The proposed images show the results of the model with input and output images and the loss curve, which in both cases appears to have converged and shows no overfitting phenomena (Supplementary materials, Fig. 7).

For the evaluation of the results, evaluation metrics are tested on 3 different outputs, namely the reconstruction of the fragment as output by the *Reconstructor* (Reconstructed vanilla), the reconstruction with the original fragment superimposed (Reconstructed filled with original), and finally the output of the *Denoiser* (Denoised, where the input is the Reconstructed filled with original) (Fig. 2). At a general level, the results appear to be very positive [also compared to those obtained by 15] and show small differences at the level of specific metrics: while Reconstructed filled with original shows the best results at the level of RSME, Dice coefficient and GS, Decoded shows the best FID and very similar related to the other metrics. In view of the applications on real fragments, *Denoised* outputs are chosen as the best result (Table 4).

Practical application of the model: the actual fragments

To implement the model in practice, the networks trained on the training dataset is specialised for each archaeological context. This process, known as fine-tuning, is implemented to adjust the reconstruction of vessels in each specific context, with the aim of achieving optimal results and avoiding, for example, the reconstruction of a vessel in a way that is inconsistent with the archaeological context. Fine-tuning involves using a flexible number of epochs, determined by the quantity of records. Also, the mini-batch size varies based on the number of samples within the context, and the ADAM optimizer is employed with a learning rate set at 0.0005. This allow for more precise parameter tuning and to avoid excessive changes to the model weights, which were modelled extensively on the training dataset during training process. A detailed visualisation of the results of the training process can be found in the Supplementary materials, where each context is examined in a dedicated Jupyter notebook [67,68] including all the diagnostic plots and metrics. Here, we focus instead on applying the method to real fragments and analysing the results through human evaluation. The results are displayed for each context. Starting with Osteria dell'Osa context, which contains 86 fragments. The figure shows the input fragments positioned with the *Regressor*, together with their reconstructed versions (Fig. 2). Of the total number of fragments, 12 have been identified as requiring correction with respect to the predicted position, because the initial reconstructed version (without manual correction) results an artefact or an implausible vessel. (3C, 3F, 6E, 8A, 9B, 9C, 9D, 10C, 11A, 11D, 14F). Fig. 3 shows the results of the reconstruction of the Casinalbo context. In this case, fragments 3D, 5D, 5E, 5F, 6D, 11B, 11D, 13B, 13E required manual positioning correction. Turning instead to the Pianello di Genga context (Fig. 4), the position of fragments 1F, 2F, 5B, 8A, 8D, 9C, 10B, 10C, 10E had to be corrected. In the context of Quattro Fontanili (Supplementary materials, Fig. 8), the fragments that had to be corrected were 2D, 4C, 4F, 5B, 5D, 7B, 9B, 9D, 9F, 11D, 12D, 15A, 16E. Finally, the results are shown for the Torre Galli context (Supplementary materials, Fig. 9). In this case, no fragments required manual correction. Moving to the Narde contest (Supplementary materials, Fig. 10), which was not present in the training dataset but only reconstructed by fine-tuning, fragments 1F, 2D, 3B required manual correction. It is important to note that

Table 5
Results of human evaluation phase 1. The table shows the accuracy values (percentage of correctly identified real or generated vessels) for the first phase of the human evaluation, divided by evaluator and context, the latter being identified by an abbreviation defined in Table 1.

Evaluator	CSLN	PDG	OSTS	TRRGLL	QF	NRD
Evaluator 1	0.58	0.49				0.5
Evaluator 2		0.51	0.48			
Evaluator 3				0.52		
Evaluator 4					0.52	

Table 6
Results of human evaluation phase 2. The table shows the mean score (according to § 3.5) for the second phase of the human evaluation, divided by evaluator and context.

Evaluator	CSLN	PDG	OSTS	TRRGLL	QF	NRD
Evaluator 5	4.1	4.22				4.15
Evaluator 6		4.3	4.53			
Evaluator 7				4.3		
Evaluator 8					4.34	

the fragments that required manual placement were mainly those belonging to the bases, which are characterised by extremely poor preservation and a lack of diagnostic features. Visual analysis of the combined output of the Reconstructor and Denoiser, pending expert evaluation, shows very positive results.

Human evaluation: results

The results of the first phase of the human evaluation are shown in the Table 5, while the results of the second phase are shown in the Table 6. Briefly, we recall that the first phase of human evaluation consists of asking the expert to distinguish between real and generated fragments, while the second phase consists of evaluating the quality of the reconstruction.

For the first evaluation phase, it is decided to measure the performance of the model in terms of the *accuracy* of the evaluator, defined as

$$\text{Accuracy} = \frac{\text{TP} + \text{TN}}{\text{TP} + \text{TN} + \text{FP} + \text{FN}}$$

where TP is the number of *True Positive* (i.e., the number of times the evaluator correctly identified a real vessel), TN is the number of *True Negative* (i.e., the number of times the evaluator correctly identified a generated vessel), FP is the number of *False Positive* (i.e., the number of times the evaluator incorrectly identified a real vessel) and FN is the number of *False Negative* (i.e., the number of times the evaluator incorrectly identified a generated vessel). A perfect score by the evaluator tends towards 100, i.e. he/she was able to correctly distinguish all real and generated pots, while a score tending towards 0 indicates that the evaluator made a mistake, confusing generated pots and identifying them as real, and vice versa. From a preliminary point of view, assuming that the model has produced high quality output (based on the metrics in Table 4), it is expected that the accuracy score will tend towards 0.5, i.e. the evaluator has basically guessed, producing a random result for a binary classification.

As for the second evaluation step, it was decided to measure the performance of the model by an average of the scores assigned by the evaluators, according to the Likert scale proposed in Section 3.5. Mean scores tending towards 5 indicate a high ability of the proposed model to produce results that preserve the morphological characteristics of the fragment, while scores tending towards 0 would indicate poor quality results that alter the morphological characteristics of the fragment.

Discussion

The evaluation metrics and the aesthetic quality of the results are extremely positive, but an evaluation based on these aspects can only be used as a preliminary check, as they work at the level of pixel distribution and may not pick up important differences from an archaeological point of view. In this sense, the most interesting results are those obtained by expert's human evaluation. The results of the first phase (Table 5) show a clear inability of the evaluator to distinguish the generated vessels from the original ones, with an accuracy tending towards 0.5 and in line with the expected score. This result is extremely interesting as it shows the model's ability to generate output that is indistinguishable from the training data, even in the eyes of an expert in the field. Good results are also obtained in the second phase of human evaluation (Table 6), where the average scores are greater than 4, indicating an advanced (but not perfect) ability of the model to preserve the morphological characteristics of the fragment during reconstruction. While the reconstruction can be considered excellent and extremely high performance, it should be noted that some records had to be placed manually to allow a proper reconstruction. This is undoubtedly a flaw in the process and in the robustness and applicability of the method, but several considerations need to be considered: 1) the misplaced or manually placed vessels are a minority (approximately 10–15 % of the total) compared to the rest of the correctly placed fragments; 2) the misplaced sherds are almost exclusively bases of extremely poorly preserved vessels. Starting from the latter point, there are many things to consider. First, we are using an extremely poorly preserved fragment whose reconstruction, while valid, is not robust. Therefore, simply retraining the model with a different *seed* [69] would most likely produce a different result. Secondly, the positioning of the bases is an extremely simple task, even for an archaeologist who does not have in-depth knowledge of archaeological material. In this sense, the author was quite surprised that the model did not learn how to perform such a simple task, therefore, the positioning of these fragments can be done by hand under the supervision of an operator. Based on these considerations, the issue of manual positioning should not be seen as a problem, but as an opportunity to reflect on the use of material that is clearly inadequately preserved. Indeed, the proposed method is extremely robust and provides high quality results even in the case of contexts not directly present in the training dataset, as we can see from the example of the Narde context. For the application of the method to other contexts, the only limitation and necessity is to have a sufficiently large and varied dataset for fine-tuning: the estimation is entirely preliminary and also depends on material characteristics such as morphological diversity (defined as the presence of different ceramic classes: bowls, plates, amphorae) and morphological variability (defined as morphological variability within the same ceramic class), but a dataset of at least 250–300 complete vessels is recommended. Based on the analyses carried out and the results obtained, it seems that the best solution is to present the context in its entirety, with all the data available, as this best defines the intrinsic variability of the ceramic assemblage. This means that data is important, this is true, but we do not need huge, material-rich contexts: as far as application aspects are concerned, the problem of defining this dataset must be archaeological in nature: for example, several small contexts with few materials can be used and combined to have a sufficiently large training dataset. In this sense, if the Narde context did not have enough entire vessels, the Pianello di Genga context, the most similar in terms of chronology and culture (Table 1; Supplementary materials, Fig. 1), would have been used. The use of this approach can considerably extend the scope of computational methods, making it possible to analyse contexts characterised by a high degree of fragmentation. In the case of the Casinalbo or Pianello di Genga contexts, for example, the amount of data that can be analysed has increased by about 30 %. I would like to emphasise that this increase in data can be analysed using any analysis method like an autoencoder [13,70], but also methods not related to the world of deep learning, such as elliptic Fourier analysis [10]. The analyses that can be performed (classification, dimensionality reduction or clustering) depend on our research objective, but the key point is that the proposed method allows us to integrate and analyse more data, more quickly, with standardised and repeatable results. Additionally, the resulting images can also be used for traditional typological analysis or visualisation purpose.

Conclusion and future works

The results obtained in this work are very positive and promising. The suggested method can reconstruct the vessels from the fragments with high accuracy, robustness and without any size or preservation constraints. The results obtained from the human evaluation show that the model produces an output that is indistinguishable from the original dataset, even for an expert in the field, but also that the model produces a reconstruction that can modify some secondary morphological characteristics of the original fragment. This is an aspect that can certainly be improved in the future with approaches that better preserve the spatial information of the image, such as U-Net or Pix2Pix. These models could also be used in a seamless solution (one complex neural network instead of 3), this approach could also reduce manual positioning of images, but it would also be more computationally expensive. Nevertheless, the proposed method is also extremely flexible and can be applied to different contexts, even those not present in the training dataset, with the only requirement being a sufficiently large dataset for fine-tuning. However, future work is needed to test the application of the method to non-protohistoric contexts (such as historical contexts). Fragment reconstruction is a step forward in the computational analysis of archaeological ceramics and the application of generative artificial intelligence to archaeology: it lays the foundations for reproducible and rapid reconstruction of a set of vessels and, by producing an image, it also proposes a standardised result, allowing integration with other methods of digital analysis and beyond.

Acknowledgements

I would like to thank all those who participated in the evaluation of the results on the basis of their profound knowledge of the archaeological material: A.C., M.B., A.D.R., F.Q., A.L., F.E., C.P., L.P. I would also like to thank R.R. for his advice during the revision of this work.

Fundings

This research did not receive any specific grant from funding agencies in the public, commercial, or not-for-profit sectors.

Supplementary materials

Supplementary material associated with this article can be found, in the online version, at doi:10.1016/j.culher.2024.09.012.

References

[1] A.O. Shepard, Ceramics For the archaeologist. Repr, Carnegie Inst, Washington, DC, 1985.

[2] C.M. Sinopoli, Approaches to Archaeological Ceramics, Springer US, Boston, MA, 1991, doi:10.1007/978-1-4757-9274-4.

[3] C. Orton, M. Hughes, Pottery in Archaeology, 2nd ed., Cambridge University Press, 2013, doi:10.1017/CBO9780511920066.

[4] P.M. Rice, Pottery analysis: A sourcebook, University of Chicago Press, Chicago; London, 2015 Second edition.

[5] V. Roux, Ceramics and society: A technological Approach to Archaeological Assemblages, Springer International Publishing, Cham, 2019, doi:10.1007/978-3-030-03973-8.

[6] S. Karl, P. Houska, S. Lengauer, J. Haring, E. Trinkl, R. Preiner, Advances in digital pottery analysis, It - Inform. Technol. 64 (2022) 195–216, doi:10.1515/itit-2022-0006.

[7] A. Karasik, U. Smilansky, Computerized morphological classification of ceramics, J. Archaeol. Sci. 38 (2011) 2644–2657, doi:10.1016/j.jas.2011.05.023.

[8] P. Demján, P. Pavúk, C.H. Roosevelt, Laser-aided profile measurement and cluster analysis of ceramic shapes, J. Field Archaeol. 48 (2023) 1–18, doi:10.1080/00934690.2022.2128549.

[9] L. Van Der Maaten, G. Lange, P. Boon, Visualization and automatic typology construction of pottery profiles, in: Making History Interactive Computer Applications and Quantitative Methods in Archaeology Proceedings of the 37th International Conference, 2009, pp. 1–12.

[10] J. Wilczek, F. Monna, P. Barral, L. Burlet, C. Chateau, N. Navarro, Morphometrics of second iron age ceramics – strengths, weaknesses, and comparison with traditional typology, J. Archaeol. Sci. 50 (2014) 39–50, doi:10.1016/j.jas.2014.05.033.

[11] C. Cintas, M. Lucena, J.M. Fuertes, C. Delrieux, P. Navarro, R. González-José, et al., Automatic feature extraction and classification of iberian ceramics based on deep convolutional networks, J. Cult. Herit. 41 (2020) 106–112, doi:10.1016/j.culher.2019.06.005.

[12] P. Navarro, C. Cintas, M. Lucena, J.M. Fuertes, C. Delrieux, M. Molinos, Learning feature representation of iberian ceramics with automatic classification models, J. Cult. Herit. 48 (2021) 65–73, doi:10.1016/j.culher.2021.01.003.

[13] L. Cardarelli, A deep variational convolutional autoencoder for unsupervised features extraction of ceramic profiles. A case study from central italy, J. Archaeol. Sci. 144 (2022) 105640, doi:10.1016/j.jas.2022.105640.

[14] D. Eslami, L. Di Angelo, P. Di Stefano, C. Pane, Review of computer-based methods for archaeological ceramic sherds reconstruction, Virt. Archaeol. Rev. 11 (2020) 34, doi:10.4995/var.2020.13134.

[15] P. Navarro, C. Cintas, M. Lucena, J.M. Fuertes, R. Segura, C. Delrieux, et al., Reconstruction of iberian ceramic potteries using generative adversarial networks, Sci. Rep. 12 (2022) 10644, doi:10.1038/s41598-022-14910-7.

[16] S.H. Bickler, Machine learning arrives in archaeology, Adv Archaeol Pract 9 (2021) 186–191, doi:10.1017/aap.2021.6.

[17] D. Foster, Generative Deep learning: Teaching machines to paint, write, compose, and Play, 1st edition, CA: O'Reilly Media, Inc, Sebastopol, 2019.

[18] S. Raschka, Y. Liu, V. Mirjalili, D. Dzhulgakov, Machine Learning With PyTorch and scikit-learn: Develop machine Learning and Deep Learning Models With Python, Packt Publishing, Birmingham, 2022.

[19] Epstein Z., Hertzmann A., the Investigators of Human Creativity, Akten M, Farid H, Fjeld J, et al. Art and the science of generative AI. Science 2023;380:1110–1. https://doi.org/10.1126/science.adh4451.

[20] C. Stokel-Walker, R. Van Noorden, What ChatGPT and generative AI mean for science, Nature 614 (2023) 214–216, doi:10.1038/d41586-023-00340-6.

[21] R. Gozalo-Brizuela, E.C. Garrido-Merchan, ChatGPT is not all you need, A State Art Rev. Larg. Generat. AI Model. (2023), doi:10.48550/arXiv.2301.04655.

[22] D.P. Kingma, M. Welling, Auto-Encoding Variational Bayes, 2013, doi:10.48550/arXiv.1312.6114.

[23] I. Goodfellow, J. Pouget-Abadie, M. Mirza, B. Xu, D. Warde-Farley, S. Ozair, et al., Generative adversarial nets, Advances in Neural Information Processing Systems, 27, Curran Associates, Inc., 2014.

[24] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A.N. Gomez, et al., Attention is all You Need, 2023, doi:10.48550/arXiv.1706.03762.

[25] Y. Chang, X. Wang, J. Wang, Y. Wu, L. Yang, K. Zhu, et al., A Survey on Evaluation of Large Language Models, 2023, doi:10.48550/arXiv.2307.03109.

[26] J. Kaddour, J. Harris, M. Mozes, H. Bradley, R. Raileanu, R. McHardy, Challenges and Applications of Large Language Models, 2023, doi:10.48550/arXiv.2307.10169.

[27] J. Ho, A. Jain, P. Abbeel, Denoising Diffusion Probabilistic Models, 2020, doi:10.48550/arXiv.2006.11239.

[28] P. Dhariwal, A. Nichol, Diffusion Models Beat GANs on Image Synthesis, 2021, doi:10.48550/arXiv.2105.05233.

[29] F.-A. Croitoru, V. Hondru, R.T. Ionescu, M. Shah, Diffusion models in vision: a survey, IEEE Trans. Pattern. Anal. Mach. Intell. 45 (2023) 10850–10869, doi:10.1109/TPAMI.2023.3261988.

[30] Z. Chang, G.A. Koulieris, H.P.H. Shum, On the Design Fundamentals of Diffusion Models: a Survey, 2023, doi:10.48550/arXiv.2306.04542.

[31] Z. Wang, Y. Jiang, H. Zheng, P. Wang, P. He, Z. Wang, et al., Patch Diffusion: Faster and More Data-Efficient Training of Diffusion Models, 2023, doi:10.48550/arXiv.2304.12526.

[32] L. Cardarelli, Traditional and digital typologies compared: the example of italian protohistory, Origini (2023) XLVIIpress.

[33] D.L. Clarke, Analytical Archaeology, Routledge, London, 1968.

[34] W.Y. Adams, E.W. Adams, Archaeological Typology and Practical reality: A dialectical Approach to Artifact Classification and Sorting, Cambridge University Press, Cambridge, 1991, doi:10.1017/CBO9780511558207.

[35] D.W. Read, Artifact classification: A conceptual and Methodological approach. 1. Paperback ed, Left Coast Press, Walnut Creek, Calif, 2009.

[36] P. Navarro, C. Cintas, M. Lucena, J.M. Fuertes, A. Rueda, R. Segura, et al., IberianVoxel: automatic completion of iberian ceramics for cultural heritage studies, in: Proceedings of the thirty-second international joint conference on artificial

intelligence, International Joint Conferences on Artificial Intelligence Organization, Macau, SAR China, 2023, pp. 5833–5841, doi:10.24963/ijcai.2023/647.

[37] A.M Bietti Sestieri, Italy in europe in the early iron age, in: Proceedings of the Prehistoric Society, 63, 1997, pp. 371–402, doi:10.1017/S0079497X00002498.

[38] A.M Bietti Sestieri, Peninsular Italy, Oxford University Press, 2013, doi:10.1093/oxfordhb/9780199572861.013.0035.

[39] F. Nicolis, Northern Italy, Oxford University Press, 2013, doi:10.1093/oxfordhb/9780199572861.013.0038.

[40] A.M Bietti Sestieri, L'italia Nell'età Del Bronzo e Del ferro: Dalla palafitte a Romolo (2200-700 a. c.), Roma: Carocci, 2010.

[41] C. Morgan, H. Wright, Pencils and pixels: drawing and digital media in archaeological field recording, J. Field Archaeol. 43 (2018) 136–151, doi:10.1080/00934690.2018.1428488.

[42] P. Allison, Dealing with legacy data - an introduction, IA (2008), doi:10.11141/ia.24.8.

[43] D.R. Snow, Making legacy literature and data accessible in archaeology, in: Making History Interactive Computer Applications and Quantitative Methods in Archaeology (CAA) Proceedings of the 37th International Conference, Williamsburg, Virginia, United States of America, 2010, pp. 350–355. March 22-26.

[44] A. Cardarelli, La Necropoli Della Terramara Di Casinalbo editor, All'insegna del giglio, Borgo San Lorenzo (Fi), 2014.

[45] V. Bianco Peroni, R. Peroni, A. Vanzetti, La Necropoli Del Bronzo Finale Di Pianello Di Genga, All'insegna del giglio, Borgo San Lorenzo (FI) [i.e. Florence, Italy], 2010.

[46] A.M. Bietti Sestieri, La Necropoli Laziale Di Osteria Dell'osa editor, Quasar, Roma, 1992.

[47] M Pacciarelli, in: Torre galli: La necropoli Della Prima Età Del ferro: Scavi paolo Orsi, Soveria Mannelli (Catanzaro), Rubbettino, 1999, pp. 1922–1923.

[48] M. Moretti, A. De Agostino, J.B. Ward-Perkins, R. Staccioli, A.P. Vianello, D. Ridgway, et al., Veio (isola farnese). – scavi in una necropoli villanoviana in località «quattro fontanili», Notizie Degli Scavi Di Antichità XVII (1963) 77–272.

[49] J.B. Ward-Perkins, R. Staccioli, J. Close-Brooks, A. Batchvarova, Veio (isola farnese). – continuazione degli scavi nella necropoli villanoviana in località «quattro fontanili», Notizie Degli Scavi Di Antichità XIX (1965) 49–236.

[50] J.B. Ward-Perkins, R. Staccioli, M. Torelli, A. Batchvarova, M.T Falconi Amorelli, Veio (isola farnese). – continuazione degli scavi nella necropoli villanoviana in località «quattro fontanili», Notizie Degli Scavi Di Antichità XXI (1967) 87–286.

[51] J.B. Ward-Perkins, M.T. Falconi Amorelli, A. Batchvarova, M. Wheeler, E. Fabbricotti, M. Meagher, et al., Veio (isola farnese). - continuazione degli scavi nella necropoli villanoviana in località «quattro fontanili», Notizie Degli Scavi Di Antichità XXIV (1970) 178–329.

[52] L. Cavagnaro Vanoni, M. Moretti, L. Berni Brizio, M. Meagher, M. Pandolfini, F. Healey, et al., Veio (isola farnese). – continuazione degli scavi nella necropoli villanoviana in località «quattro fontanili», Notizie Degli Scavi Di Antichità XXVI (1972) 195–384.

[53] M. Bedello, E. Fabbricotti, Veio (isola farnese). - continuazione degli scavi nella necropoli villanoviana in località «quattro fontanili», Notizie Degli Scavi Di Antichità XXIX (1975) 63–184.

[54] E. Fabbricotti, Veio (isola farnese). – continuazione degli scavi nella necropoli villanoviana in località «quattro fontanili», Notizie Degli Scavi Di Antichità XL (1976) 149–184.

[55] L. Salzani, C. Colonna, La fragilità dell'urna. I recenti scavi a narde necropoli di frattesina (XII-IX sec. A.c.), Catalogo Della Mostra, 2010.

[56] M. Pacciarelli, Dal villaggio alla città: la svolta protourbana del 1000 a.c. Nell'italia tirrenica, Firenze: all'insegna Del Giglio, 2000.

[57] Y. Panagakis, J. Kossaifi, G.G. Chrysos, J. Oldfield, M.A. Nicolaou, A. Anandkumar, et al., Tensor methods in computer vision and deep learning, Proceed. IEEE 109 (2021) 863–890, doi:10.1109/JPROC.2021.3074329.

[58] C. Shorten, T.M. Khoshgoftaar, A survey on image data augmentation for deep learning, J. Big Data 6 (2019) 60, doi:10.1186/s40537-019-0197-0.

[59] G. James, D. Witten, T. Hastie, R. Tibshirani, An Introduction to Statistical Learning, 103, NY: Springer New York, New York, 2013, doi:10.1007/978-1-4614-7138-7.

[60] S. Bond-Taylor, A. Leach, Y. Long, C.G. Willcocks, Deep generative modelling: a comparative review of VAEs, GANs, normalizing flows, energy-based and autoregressive models, IEEE Trans. Pattern Anal. Mach. Intell. 44 (2022) 7327–7347, doi:10.1109/TPAMI.2021.3116668.

[61] A. Borji, Pros and cons of GAN evaluation measures: new developments, Comput. Visi. Image Understand. 215 (2022) 103329, doi:10.1016/j.cviu.2021.103329.

[62] L.R. Dice, Measures of the amount of ecologic association between species, Ecology 26 (1945) 297–302, doi:10.2307/1932409.

[63] V. Khrulkov, I. Oseledets, Geometry Score: a Method for Comparing Generative Adversarial Networks, 2018.

[64] M. Heusel, H. Ramsauer, T. Unterthiner, B. Nessler, S. Hochreiter, GANs Trained by a two Time-Scale Update Rule Converge to a Local Nash Equilibrium, 2018, doi:10.48550/arXiv.1706.08500.

[65] E. Betzalel, C. Penso, A. Navon, E. Fetaya, A study on the evaluation of generative models, 2022, doi:10.48550/arXiv.2206.10935.

[66] A.T. Jebb, V. Ng, L. Tay, A review of key likert scale development advances: 1995–2019, Front. Psychol. 12 (2021) 637547, doi:10.3389/fpsyg.2021.637547.

[67] B.M. Randles, I.V. Pasquetto, M.S. Golshan, C.L. Borgman, Using the jupyter notebook as a tool for open science: an empirical study, in: 2017 ACM/IEEE joint conference on digital libraries (JCDL), 2017, pp. 1–2, doi:10.1109/JCDL.2017.7991618.

[68] M. Beg, J. Taka, T. Kluyver, A. Konovalov, M. Ragan-Kelley, N.M. Thiéry, et al., Using jupyter for reproducible scientific workflows, Comput. Sci. Eng. 23 (2021) 36–46, doi:10.1109/MCSE.2021.3052101.

[69] S. Bethard, We need to Talk About Random Seeds, 2022, doi:10.48550/arXiv.2210.13393.

[70] S. Parisotto, N. Leone, C.-B. Schönlieb, A. Launaro, Unsupervised clustering of roman potsherds via variational autoencoders, J. Archaeol. Sci. 142 (2022) 105598, doi:10.1016/j.jas.2022.105598.