Journal of Cultural Heritage 69 (2024) 18–26

Contents lists available at ScienceDirect

Journal of Cultural Heritage

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

Original article

Bridging the past and present: AI-driven 3D restoration of degraded artefacts for museum digital display

Ruxandra Stoean[a,b], Nebojsa Bacanin[c], Catalin Stoean[a,b,*], Leonard Ionescu[b,d]

[a] Department of Computer Science, University of Craiova, A. I. Cuza 13, Craiova, 200585, Romania
[b] Artificial Intelligence and Machine Learning, Romanian Institute of Science and Technology, Saturn 24-26, Cluj-Napoca, 400504, Romania
[c] Faculty of Informatics and Computing, Singidunum University, Danijelova 32, Belgrade, 11000, Serbia
[d] Restoration and Conservation Laboratory, Oltenia Museum, Madona Dudu 14, Craiova, 200410, Romania

ARTICLE INFO

Article history:
Received 2 December 2023
Revised 19 July 2024
Accepted 24 July 2024
Available online 9 August 2024

Keywords:
Restoration
3D replica
Deep learning
Semantic inpainting
Neural radiance fields

ABSTRACT

Artificial intelligence can lend a helpful digital "hand" in the restoration process of deteriorated cultural heritage items as well as towards an increased visitor interest in the museum exhibits. To this purpose, the present paper proposes a deep learning approach to repair the missing content and to recreate a visual counterpart of a degraded artefact by a 3D rendering of the semantic inpainted version. The new approach is constructed by means of some of the most recent and successful deep learning models for image inpainting and 3D reconstruction, namely stable diffusion and neural radiance fields. The method is tested in the scenario of ceramic artefacts, where the end visual result has a bigger impact. The ability of the novel technique to creatively reproduce a realistic and plausible 3D surrogate of broken archaeological objects shows the potential that AI has in supporting specialists with preserving the cultural heritage and bringing the museums into the public spotlight.

1. Introduction

One important recent step towards enhancing the preservation of our cultural heritage assets has been the possibility of creation of 3D digital twins for these items [1]. The digital surrogate can serve for restoration and conservation analysis, multi-faceted historical examination, as well as interactive, virtual online and on-site display. Using the 3D replica of the piece, the restoration and conservation specialists can get higher depths views on the object, with no further invasive manipulation. Then, after cleaning and treating the artefact, the experts can grasp the decorative details and historical specificity again from different angles and perspectives of a digital inspection. Subsequently, the possible restoration of the object can be inferred with the support of digital tools in the artificial environment prior to the actual human rehabilitation. Finally, the exhibit can be also watched in its digital lookalike, both on museum digital displays and in the online site. In the era of virtual experiences, museum visitors are thus also engaged in more interactively observing and understanding the gifts of the past.

Great museums worldwide have already made use of virtual twins [2]. The creation of a digital twin was made possible by 3D scanning technology and photogrammetry. However, highly performing 3D scanners are not always easily affordable, while the software operation needs well trained and experienced users. On the other hand, for possibly completing missing parts of the item, the available technology is currently endowed specifically with traditional image inpainting and 3D processing approaches.

In this context, the current paper puts forward a novel artificial intelligence (AI) alternative of a generative experience for creating 3D models of cultural heritage objects, starting from a possible semantic filling of the missing areas according to the existing context of the artefact. The recent neural radiance fields (NeRF) model for 3D rendering is combined with the Stable Diffusion (SD) architecture for image inpainting, and tailored for the particularities of the archaeological objects. This may result in a new promising direction of AI support in cultural heritage restoration, preservation and publicity. This promising direction in the AI role in cultural heritage preservation and restoration aligns with the ongoing discourse among international bodies and councils. UNESCO emphasizes the necessity of open and FAIR data for accurate models [3], while ICOM advocates for the convergence of museums towards digital transformation with AI [4]. Additionally, the American Alliance of Museums presents the possibilities of generative AI to recreate lost paintings from existing images [5] and the European Parliament conducted a briefing regarding the current opportuni-

* Corresponding author.
E-mail address: catalin.stoean@inf.ucv.ro (C. Stoean).

https://doi.org/10.1016/j.culher.2024.07.008
1296-2074/

ties and challenges of the interplay between AI and cultural heritage, and the effective means to support models trained on cultural heritage collections for information management and visitor engagement [6].

The paper is structured as follows. The research aim is formulated in Section 2. The material and the proposed methodology against the state of the art are provided in Section 3. The performance metrics are outlined in Section 4. The results are given and discussed in Section 5. The paper ends with the conclusions in Section 6.

2. Research aim

The scientific target of present research is to develop an AI generative framework to digitally repair deteriorated cultural heritage items and to recreate them in 3D. Additionally, the framework aims to facilitate the creation of video files showcasing the reconstructed models. State-of-the-art deep learning models for semantic inpainting (e.g. SD) and 3D reconstruction (e.g. NeRF) are tailored for artefact digital generation from 2D images.

The methodology is demonstrated on a test case of a ceramic artefact for reasons of better spatial visualization, but it can be applied to any type of archaeological object, irrespective of material composition, since the processing is digital.

The practical aim of the current work is to offer a straightforward and modern AI support tool for efficient examination, restoration and popular display through a 3D digital replica of cultural heritage degraded assets.

3. Material and methods

The proposed task and approach are presented against the state of the art in the direction of semantic completion and 3D reconstruction with deep learning for cultural heritage.

3.1. State of the art

AI and cultural heritage have begun sharing a beautiful story of bringing the remains of the past to digital life with reconstruction possibilities based on present deep learning models. A general survey on their interaction is conducted in [7].

Considering specifically the 3D AI recreation of heritage environments, the paper [8] performs an overview on machine learning and deep learning techniques for cultural heritage buildings and sites. A depth estimation network with an encoder based on DenseNet and a layer for edge guidance in the decoder is proposed in [9] for the 3D reconstruction of reliefs of a temple. A novel combination between a pretrained Minkowski-based deep neural network for semantic segmentation and a support vector machine was constructed for style classification of 3D heritage buildings in [10]. In the paper [11], the state-of-the-art NeRF were tailored for the 3D reconstruction of heritage data sets with available ground truth, consisting of architectural monuments. As observed, recent research in AI-based 3D reconstruction methods for heritage environments has predominantly focused on architectural structures.

At the other end, concerning the digital semantic completion of deteriorated heritage pieces, the very recent literature points to the specific consideration of decorative art. A U-Net is used in [12] for restoring damaged ceramic tiles. In this sense, the study [13] employs the state-of-the-art generative models *CoModGAN* (generative adversarial networks with co-modulated stochastic and conditional styles), *LaMa* (large mask inpainting having the ResNet architecture with fast Fourier convolution) and *GLIDE* diffusion model for very degraded paintings. A generative adversarial network is also used in [14] with a structure reconstruction network with gated and fast Fourier convolution and a content restoration architecture for mural inpainting.

Therefore, to the best of our knowledge, this is the first research work on the combination between semantically spatially repairing deteriorated artefacts and obtaining the 3D digital reconstruction of the AI-restored object. This would be the third step of the framework for complete AI-support for the restoration of heritage objects, as begun with regression and semantic segmentation for the chemical composition of archaeological items outlined in [15]. The current work will be based on an earlier attempt for semantic inpainting of heritage items [16], where *LaMa* and SD, seconded by an autoencoder, were used for completing the appearance of textile artefacts. Once the digital images of the 3D real artefact are completed by the generative AI models, a NeRF architecture will recreate the 3D digitally repaired object.

3.2. Data

A video of the broken vase artefact is recorded in a laboratory room. Placed on a small table in the center of the room, the vase is encircled twice by a person holding the camera-first at the level of the vase and then from an overhead perspective. The resulting 58-second video exclusively captures views from the side and above the object; no views from below are recorded. Importantly, the object remains stationary on the table throughout the recording, with only the person and the camera rotating around it. Subsequently, the Nerfstudio preprocessing tool [17] is employed to convert the video into images, downscale them, and calculate camera poses for each image using COLMAP [18,19]. This tool selects 334 frames that are further used for experimentation throughout this research study. The data set is available at https://doi.org/10.6084/m9.figshare.24637428.v1 and includes the image files and the masks utilized in our research study.

3.3. Methodology

In a bird's-eye view, the procedure employed in the current study begins with recording a video file featuring the incomplete artefact. A human annotator then steps in to annotate the missing parts of the objects in the images extracted from the video, creating masks for instances where the object deficiencies are visible.

Following annotation, an inpainting procedure is applied to the images with associated masks, resulting in virtually restored images of the object. In cases where multiple outputs from the inpainting method are available for the same input image, a human expert selects the most appropriate one. The semantic inpainting approach tailored to the images of deteriorated artefacts is the state-of-the-art SD architecture [20]. This cutting-edge model for generative AI is based on noise learning and inference with a U-Net conditioned by text input. Contrastive learning is employed in analysing the prompt for signaling the model to look at the existing background of the artefact. Cross-attention is used for connecting text and image languages. The training takes place in the latent space of an autoencoder for image compression, for efficient generation speed. The *LaMa* mentioned in the literature survey above was also tried alternatively, but the results came less successful.

Next, a NeRF model is applied to the collection of virtually restored images, generating a 3D model of the object. NeRF is a fully connected deep network model that takes the 5D coordinates denoting the spatial location and the viewing direction for the object from a sparse set of 2D images and renders the volume density and radiance of the item [21]. This model allows for the export of a point cloud, mesh, and even the creation of a new video file featuring the virtually repaired object. For the video file, users can establish key frames, and the camera moves between them, creating

all intermediary views. It is important to note that the video output does not consist of the precise initial frames captured in the video file of the actual artefact; instead, new frames are derived by the NeRF model based on the inputs it receives.

4. Calculation

In order to numerically evaluate the quality of an inpainting method, the difference between the reconstructed version of an image and its original version needs to be assessed. Since the main goal of the current work is to virtually complete the missing parts of an object (a fragmented artefact), the metrics cannot be directly applied to the completion of the missing areas, as there is no ground truth for these cases. Therefore, in our experiments, images from which the missing parts are not visible are considered and masks are drawn on some parts that are already complete. In such cases, the virtually restored images can be compared against the original images.

We use six key measures for assessing the quality of image inpainting in the current study: Mean Squared Error (MSE), Peak Signal-to-Noise Ratio (PSNR), Structural Similarity Index (SSIM), Multi-Scale Structural Similarity Index Measure (MS-SSIM), Universal Image Quality Index (UIQI), and Visual Information Fidelity (VIF) [22].

4.1. Mean squared error

MSE is a widely used metric for quantifying the dissimilarity between an inpainted image and its original counterpart. It measures the average squared difference between corresponding pixel values in both images, indicating the overall pixel-wise error. The MSE is computed using Eq. (1), where M and N represent the height and width of the images, $I(i, j)$ is the pixel value at position (i, j) in the original image and $\hat{I}(i, j)$ is the pixel value at position (i, j) in the inpainted image.

$$\mathrm{MSE} = \frac{1}{MN}\sum_{i=1}^{M}\sum_{j=1}^{N}[I(i,j) - \hat{I}(i,j)]^2 \tag{1}$$

A lower MSE value indicates a closer similarity between the inpainted and original images.

4.2. Peak signal-to-noise ratio

PSNR is another important metric that quantifies image quality. It measures the ratio of the peak signal strength to the mean squared error and it is expressed in decibels (dB). The PSNR is computed as in Eq. (2), where MAX is the maximum possible pixel value (e.g., 255 for 8-bit images).

$$\mathrm{PSNR} = 10 \cdot \log_{10}\left(\frac{\mathrm{MAX}^2}{\mathrm{MSE}}\right) \tag{2}$$

Higher PSNR values signify a closer resemblance between the inpainted and original images in terms of pixel values.

4.3. Structural similarity index

SSIM is a metric that evaluates the structural and textural similarity between the inpainted image and the original image. It considers luminance, contrast, and structure, providing a more perceptually relevant assessment. As opposed to MSE which may not reflect well the human perceived similarity, SSIM addresses this limitation by considering texture [23].

The SSIM is calculated as in Eq. (3), where μ_I and $\mu_{\hat{I}}$ are the means of the original and inpainted images, σ_I and $\sigma_{\hat{I}}$ represent their standard deviations, $\sigma_{I\hat{I}}$ is the covariance between the original and inpainted images and C_1 and C_2 are constants to stabilize the division. The formula is applied on the gray-scale versions of the images in the current study.

$$\mathrm{SSIM}(I, \hat{I}) = \frac{(2\mu_I\mu_{\hat{I}} + C_1)(2\sigma_{I\hat{I}} + C_2)}{(\mu_I^2 + \mu_{\hat{I}}^2 + C_1)(\sigma_I^2 + \sigma_{\hat{I}}^2 + C_2)} \tag{3}$$

The SSIM ranges from -1 to 1, where 1 indicates a perfect match in terms of structural and textural content, 0 corresponds to no correlation and -1 means anti-correlation between the two images.

4.4. Multi-scale structural similarity index measure

MS-SSIM extends the SSIM metric by evaluating the similarity at multiple scales, capturing variations in image details across different resolutions [24]. This approach provides a more comprehensive assessment of image quality, particularly for high-resolution images. The MS-SSIM is calculated by combining the results of SSIM at multiple scales, to yield a single similarity score. Higher MS-SSIM values indicate better structural and textural similarity between the inpainted and original images.

4.5. Universal image quality index

UIQI quantifies image distortion by assessing three components: correlation loss, luminance distortion, and contrast distortion [25]. The UIQI is calculated as in Eq. (4), where $\sigma_{I\hat{I}}$ is the covariance between the original and inpainted images, and μ_I, $\mu_{\hat{I}}$, σ_I, and $\sigma_{\hat{I}}$ are the means and standard deviations of the original and inpainted images, respectively.

$$\mathrm{UIQI}(I, \hat{I}) = \frac{4\sigma_{I\hat{I}}\mu_I\mu_{\hat{I}}}{(\sigma_I^2 + \sigma_{\hat{I}}^2)(\mu_I^2 + \mu_{\hat{I}}^2)} \tag{4}$$

A UIQI value closer to 1 indicates higher image quality, showing that the inpainted image is more similar to the original in terms of correlation, luminance, and contrast.

4.6. Visual information fidelity

VIF measures image quality by assessing how much visual information is preserved in the inpainted image compared to the original. It is computed using a model of natural scene statistics and the human visual system, which evaluates the fidelity of visual information between the original and inpainted images [26]. Higher VIF values indicate better preservation of natural visual information, implying higher image quality.

These metrics, including MSE, PSNR, SSIM, MS-SSIM, UIQI, and VIF, collectively provide a quantitative assessment of the inpainting quality, allowing for a thorough evaluation of the virtually restored images concerning their original counterparts.

5. Results and discussion

The following subsections provide the results and discussion pertaining to both image inpainting and the subsequent construction of the 3D artefact. Various outcomes and considerations are explored within this section. The source code behind the implementation is provided here: https://github.com/catalinstoean/JCH-3D/.

5.1. Image inpainting

Our evaluation related to inpainting comprises two steps. Firstly, we focus on a scenario where a known reference image is available, allowing us to quantitatively assess the accuracy of the inpainting methods. This section provides a clear measure of how

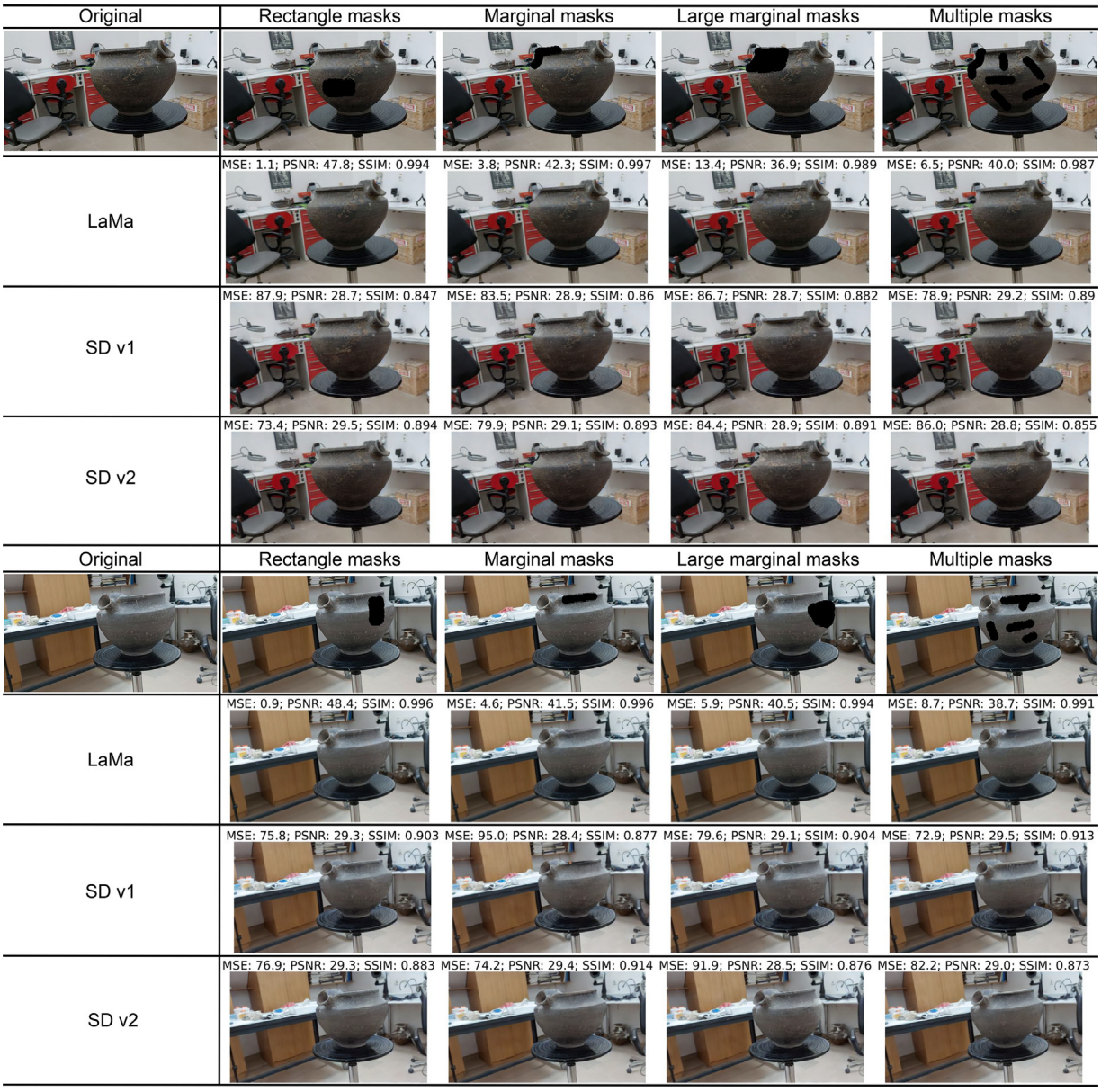

Fig. 1. Four types of masks are created for each image where the defects are not visible for checking the efficiency of the inpainting techniques. The figure presents the four types of masks (black spots) for two distinct images. UQI and MS-S are further abbreviated to represent UIQI and MS-SSIM.

well the techniques perform when the ground truth is present. Next, we describe the steps taken for reconstructing those missing regions, that is dealing with the regions where no definitive reference exists.

5.1.1. Evaluation using ground truth

Two of the most successful recent approaches to image inpainting are SD and LaMa. To assess the effectiveness of the two inpainting techniques in restoring the missing sections of the vase in the images, we employ ground truth images as reference points for evaluating the authenticity of the recreated areas. To accomplish this, we deliberately select vase images where any imperfections or damage are not visible. We then use masking to conceal various segments of the vase, allowing the inpainting procedures to virtually reconstruct, and compare these concealed portions with the original, undamaged appearance.

For each image, we consider four types of masking, as illustrated in the first and fifth rows from Fig. 1. All the masks are made with a free-hand tool: while the first one (second column, rows 1 and 5, in the figure), resembling a rectangle, is made in the interior of the vase, the marginal masks are smaller ones but made at the edge of the vase, the ones from the fourth column are again taken at the margins of the vase but they cover a larger area and finally, more masks of distinct sizes and on various positions of the vase are considered in the cases of the last column in Fig. 1. The masks are exclusively applied to the vase. While they may occasionally extend just slightly beyond the vase edge, their primary purpose is to cover the vase itself. Each image requires custom masks since the vase position varies slightly from one image to another.

We employed the default settings for both inpainting methods. It is important to note that LaMa consistently produces identical results when applied to the same file and its associated mask in multiple runs. In contrast, SD generates varying result images for identical inputs in different runs. To address this variability, we conducted 10 separate runs of the SD method and subsequently calculated the mean result. This mean result was used for a direct comparison with the LaMa method. Fig. 1 also shows results for

Table 1
Comparison between SD and LaMa for the cases when ground truth is known.

Model	Mask type	MSE	PSNR	SSIM	VIF	UIQI	MS-SSIM
SD	Large-marginal	95.56	28.36	0.869	0.869	0.308	0.969
	Marginal	88.16	28.7	0.873	0.875	0.310	0.971
	Multiple	88.08	28.7	0.867	0.878	0.306	0.970
	Rectangle	84.89	28.86	0.87	0.877	0.310	0.971
LaMa	Large-marginal	13	37.27	0.992	0.989	0.538	0.996
	Marginal	3.56	43.2	0.998	0.995	0.545	0.999
	Multiple	5.33	41.52	0.992	0.994	0.532	0.997
	Rectangle	1.29	47.88	0.995	0.998	0.539	0.998

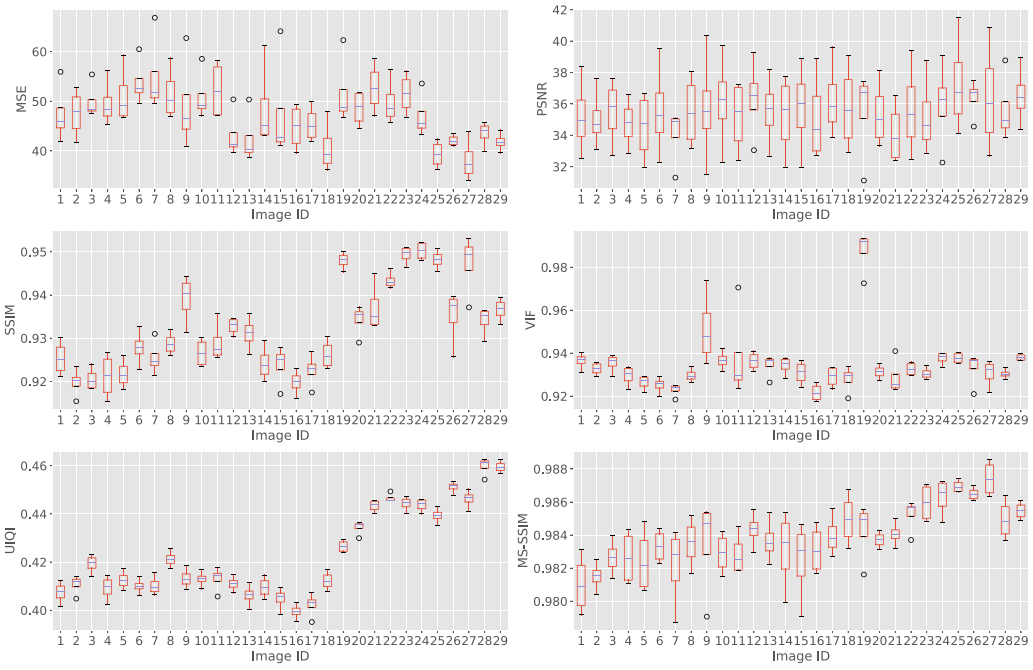

Fig. 2. Differences illustrated as box plots from the images obtained using SD for generating content as compared to the original images when using MSE, PSNR, SSIM, VIF, UIQI and MS-SSIM. For each of the 29 images, 4 different sections are covered and generated respectively and, for each case, there are 10 solutions generated.

the 2 images and each type of masks when applying LaMa and the ones obtained from 2 distinct runs of the SD model. For each case, the evaluation metrics are computed as compared to the complete original images.

There are 29 images selected where the defect is not visible. As for each image, there are 4 different types of masks made, we end up having 126 individual cases for the inpainting techniques. The results for each type of mask in turn and over the 2 inpainting models can be seen in Table 1. The most difficult case for both models proves to be having large marginal masks.

The outputs in Table 1 clearly outline the LaMa model as the winning technique for virtually repairing the missing parts. This is partly attributed to the fact that LaMa exhibits a less imaginative approach in comparison to the SD model. While SD occasionally generates inventive completions that seamlessly blend into the overall image, it tends to produce images with greater contrast. In comparison, LaMa outputs often lean towards a smoother, albeit occasionally blurrier, restoration. Consequently, we considered all images generated by SD, i.e. results from 10 repeated runs obtained for each input image and mask pair, and went further in analyzing the results for each image in turn. The MSE results box plot for each image ID in Fig. 2 is calculated based on the MSE differences between 40 pairs of images, original and inpainted sample. The 40 inpainted items for a single image ID are derived from the 10 repeated runs of the SD model across the 4 different mask types prepared for that particular sample. Similar corresponding calculations are made for PSNR, SSIM, VIF, UIQI and MS-SSIM. The box plots indicate that indeed for some samples the standard deviation is very large, as the images may be very different.

The weaker results in the case of the SD model occur however due to the reason that the model does not only apply changes to the masked parts, but also to the rest of the image to blend the new spots better into the whole picture. Fig. 3 presents a small experiment that proves this statement. We consider the masks from the image in the first row and column and LaMa and SD are separately applied to generate the masked parts. Then, we compute the metrics in turn, for observing the differences between the parts of the images excluding the masks, for the masked parts alone, and for the complete images. The first row illustrates the outputs representing the differences between the original and the generated images, but without the parts that were generated. It can be observed that LaMa presents an image that is identical with the original image, based on the results from the image in the first row, second column. However, the outputs of the image in the first row, third column, show that SD indeed changes as well beyond the masked parts. The second row computes the differences only considering the newly generated parts (the masked regions), where the two approaches have similar results, and the third row shows the differences between the complete images. Naturally, in all cases, the same masks are used. The second column proves that LaMa does not affect any of the parts beyond the masked ones, while the third column demonstrates that SD has

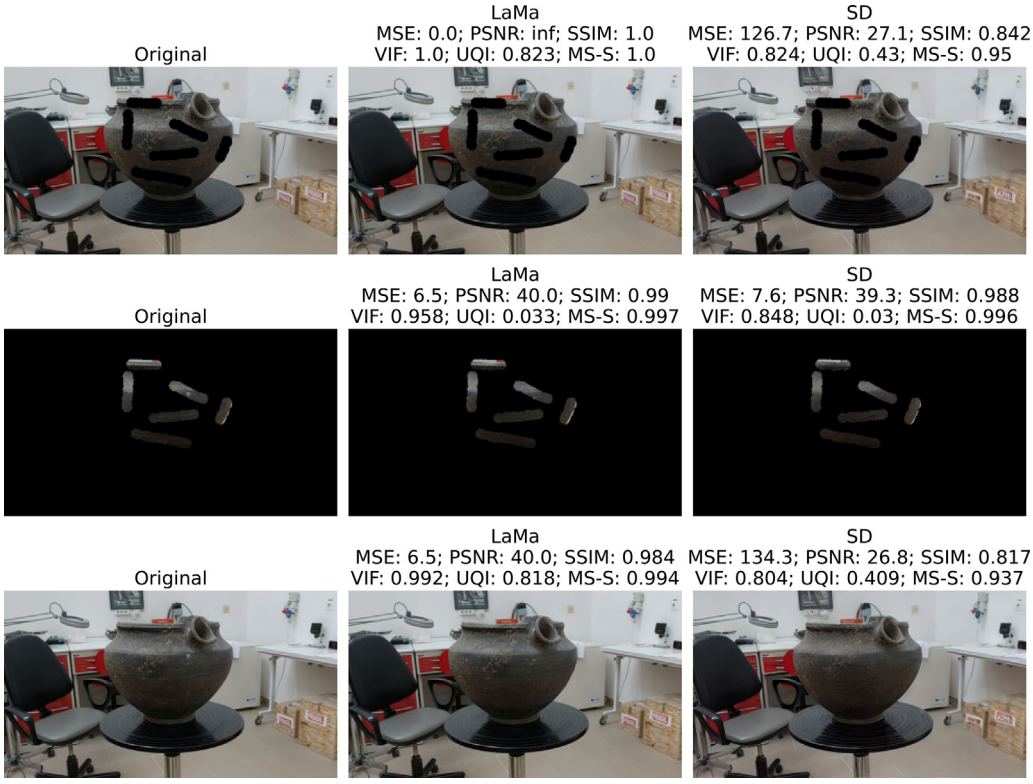

Fig. 3. Comparison between LaMa and SD models on the same image overall, as well as on the mask alone and excluding the mask, respectively.

Fig. 4. Delineation of the missing parts to create masks for inpainting.

significantly weaker results than LaMa especially because the entire image is affected by the model. For SD, we used the implementation from [20], using Python language, PyTorch library, and revision $fp16$.

5.1.2. Addressing missing parts

Masks are created to represent the missing sections of the objects, and the same two inpainting methods are employed to recreate these areas. Again, LaMa yields a single result per image, while SD generates 10 versions for the virtual restoration of the vase. The human expert then determines which of the 10 versions best represents the intended correction.

Fig. 4 illustrates, for two images (first column), how the human expert annotates the missing parts (second column) and generates the masks (third column) used by the inpainting techniques. The human annotator typically covers an area at least as large as the entire object. If the segmented area is too small, the inpainting procedures may create a completion that appears unnatural and lacks symmetry. Therefore, it is preferable to provide sufficient space for the models to accommodate the object, even if parts of the background need to be recreated. This approach helps avoid unnatural completions and minimizes noise added to the overall scene. Our focus on the object is on the large missing part and that is the only section we intend to cover.

Fig. 5 displays four sample images in the first row, where the inpainted results produced by LaMa (in the second row) are visually less successful than the SD ones (in the third row). Notably, SD has the advantage of generating distinct output samples for the same input. This variability enables the human expert to select the optimal or representative images from the multiple outputs.

A total of 334 images are extracted from the entire video file of the vase using the NeRF implementation. Out of these, 29 images are identified where the damaged section is not visible (the ones used in the previous subsection). Masks are created for the remaining 305 images, and the inpainting techniques are applied to address these cases. Consequently, we have two folders

Fig. 5. Samples where inpainting is applied and the results for LaMa are not successful. First row shows the original images, middle row illustrates the LaMa outputs and the last row contains images generated with SD.

containing 334 images each, depicting the fully restored vase - one folder created by LaMa and the other by SD. As stated before, in the case of the SD folder, a human restorer selected the most representative result for each sample.

5.2. 3D artefact generation via NeRF

The NeRF model selects a limited number of images from the LaMa inpainting procedure, resulting in a 3D model that is neither clear nor usable. This, corroborated with the major visual difference between LaMa (second row) and SD (third row) in Fig. 5, determined us to count on the images produced by the SD alone. Although this choice may appear counter-intuitive to the numerical advantages in scores for LaMa over SD, it was illustrated in the middle row from Fig. 3 that, when focusing on the inpainted areas alone, the differences in numerical results are not very significant in any of the 6 metrics. For clarity in further discussions, we will refer to the 3D model obtained from the SD inpainting.

5.2.1. SD-based 3D model

The NeRF-provided model allows for versatile manipulation of the artefact, facilitating changes in views through actions such as rotation or zoom. When exporting the object as a point cloud using the default settings, a few additional objects, such as a part of the table leg, appear in the distance. These distant objects are subsequently removed in a post-processing step. While the removal of well-separated objects is straightforward, the points of the vase are closely intertwined with those of the upper surface of the supporting table. Consequently, we have chosen to keep these points together for the subsequent evaluations, as any separation would involve a subjective intervention.

The broken part of the vase is effectively reconstructed in the Nerfstudio application, the one that also allows for the creation of video files from different camera positions. Fig. 6 illustrates six different views of the artefact from the Nerfstudio interface. Since the initial video file lacked any views from below the vase, the NeRF model could not recreate the area above the vase, resulting in a slightly blurry depiction. The primary focus is on the main object, and the background objects appear less crisp in the images. However, these are minor issues; capturing the ceiling on camera could ameliorate the first drawback, and ideally, filming the object against a blank background would facilitate inpainting and later 3D reconstruction, and could eventually address the second issue.

A more significant challenge arises from the unclear view from above the vase (see the last 2 images in the second row from Fig. 6). This is due to the vase being reconstructed from multiple inpainted images, and the opening of the object is not always consistent in size in the obtained SD outputs. One potential mitigation measure to explore in future work involves creating masks for subsequent frames by altering the masks drawn for previous images. This approach could help maintain better proportions of the reconstructed vase from one image to the next. The next subsection presents an experiment where the number of input images is reduced by sequentially eliminating less successful images.

5.2.2. Ablation experiment

Some inpainting cases present challenges where the SD results, much like LaMa, are not optimal. These suboptimal images obtained through inpainting are assumed to potentially diminish the quality of the final 3D model. Consequently, we opt to exclude such samples in an experiment where we systematically reduce the number of training input pictures for NeRF. This reduction is done in increments of 10%, starting from 334 images (which is the scenario presented in the previous subsection), then to 301, followed by 267, and finally to 234 images. This experiment serves a dual purpose: first, to observe how the NeRF model performs with fewer but better quality input images, and second, to assess how much we can reduce the number of images to alleviate the workload for the human annotator.

Fig. 7 displays, from left to right, the progression of point clouds model: first, the initial broken vase is depicted; then, the representation obtained from all 334 images, including the inpainted images from SD, is shown and finally, the point clouds obtained from a reduced number of input images by sequentially removing less realistically looking images at each stage. As the number of input images decreases, the missing part of the vase exhibits a reduction in represented points. This reduction is visible in the point clouds and can be also noticed in the model generated by the Nerfstudio application.

The differences between the point clouds are next visually and numerically compared. For this, the difference between the model obtained from the 334 images is directly compared to each other model, in turn. Accordingly, for each pair of point clouds, the same procedure is applied and it is subsequently described. Their scales are matched using principal component analysis (PCA). Scale matching using PCA for point clouds involves analyzing principal axes to ensure consistent scaling across different parts of the 3D model and it represents a usual procedure for this task [27]. Next, the two 3D models are registered via the Iterative Clos-

Fig. 6. Captions from Nerfstudio of the reconstructed artefact using SD and NeRF.

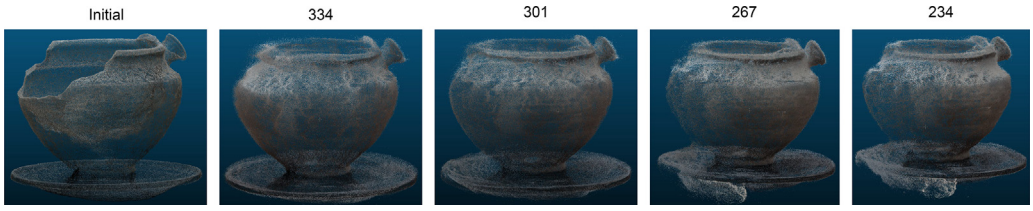

Fig. 7. Point clouds differences between the initial vase and the NeRF reconstructed one when using 334 images, 301, 267 and 234, respectively.

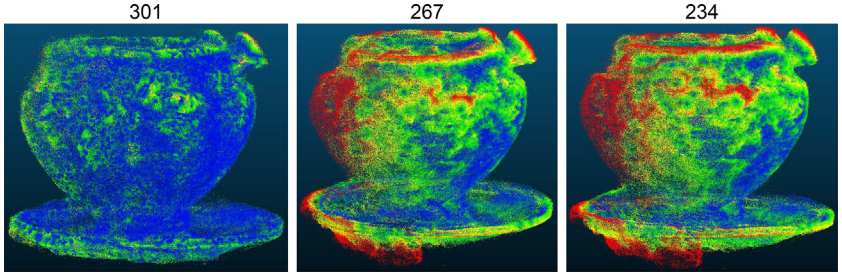

Fig. 8. Point clouds differences between the NeRF reconstructed vase from 334 images and the versions obtained from 301, 267 and 234 images, respectively. Blue points indicate similarities, green and yellow points gradually correspond to more distant ones, and the most distant ones are colored in red.

est Point (ICP) algorithm. ICP aligns multiple 3D point clouds by minimizing the distance between their corresponding points through iterative steps and it is widely employed in 3D scanning [28]. The stopping criterion utilizes the Root Mean Square (RMS) difference with a threshold value set at 1.0E-05. Following this, the Hausdorff distance is employed to measure the difference between registered point clouds, known as a conventional choice for assessing 3D models [29]. This metric calculates the maximum distance between points in one set (the reference point cloud derived from the 334 images) and their nearest counterparts in the other set, effectively representing the dissimilarity or nearest neighbor distance between the two point clouds.

The compared results can be visualized in Fig. 8. The color scale used follows blue-green-yellow-red, where blue corresponds to the closer points, and then gradually goes to pointing out the most distant ones in red. The obtained representations use an active scalar field histogram with a maximum saturation value set to 0.2 in all cases, for objectivity reasons. The actual mean distances indicate an increase in distance when comparing the differences of the point cloud obtained from 334 images with the ones resulting from 301, 267, and 234, respectively, i.e. 2.74E-03, 7.45E-03 and 8.04E-03. Similarly, the standard deviations are increasing, having 2.81E-03, 6.24E-03, and 6.66E-03. These findings suggest that, as fewer images are utilized, albeit removing the less successful ones, there is a noticeable decline in the quality of the resulting 3D model. The CloudCompare software is utilized for manipulating the cloud points.

The model derived from 334 images using the SD method proved to be the most successful among those generated in our current experimentation. These 334 images were obtained through the pre-processing procedure in Nerfstudio. Looking ahead, a potential future avenue involves augmenting this quantity by extracting additional frames from the initial video file to produce more inpainted images. This prospective expansion may contribute to refining the final model further.

6. Conclusions

The current study ventured into the domains of image semantic inpainting and 3D digital artefact generation utilizing advanced techniques, notably stable diffusion and neural radiance fields (NeRF). Moreover, our approach showcased the capability to translate this digitally reconstructed artefact model into a versatile video representation. This video's flexibility allows for the manipulation of various key frames (camera positions) as starting points, while NeRF seamlessly creates intermediary frames to compose a cohesive and complete visual representation. These findings emphasize the potential of employing image-based 3D reconstruction methodologies to transform broken artefact movies into informative 3D visualizations. Our experimental exploration revealed a correlation between the number of input images derived from a movie of the incomplete artefact and the resulting quality of reconstructed 3D models. Notably, our observations highlight the pivotal

role played by the input data quantity in shaping the fidelity and accuracy of these 3D models.

The proposed approach offers a fast, straightforward and affordable 3D semantic recreation for deteriorated artefacts. The digital counterpart could be displayed next to the historic object and even propose a possibility of its lookalike for those heavily degraded items that are beyond acceptable restoration. This is important for the museum specialists both from the preservation and historic perspectives, but also from the possibility for audience increase with a virtual engagement.

An important avenue for future work involves automating the detection of missing parts in objects, leading to the self-regulating creation of masks that cover artefact defects. This task, currently time-consuming and reliant on human expertise, could benefit from the application of image segmentation tools such as a U-Net. However, a crucial consideration in training such a model is the need for a sufficiently large data set. This data set should encompass a diverse range of items, paired with annotated masks. This pairing is essential for the model to learn effectively and generalize well when encountering new artefact items.

Another possible enhancement for future 3D virtual content restoration would be to allow the expert to give textual information to the generative model. This would be knowledge about the historical period or even more guided advice on adding certain decorative elements or colors.

Acknowledgements

This work was supported by a grant of the Romanian Ministry of Research and Innovation, CCCDI - UEFISCDI, project number 178PCE/2021, PN-III-P4-ID-PCE-2020-0788, *Object PErception and Reconstruction with deep neural Architectures* (*OPERA*), within PNCDI III.

References

[1] X. Dang, W. Liu, Q. Hong, Y. Wang, X. Chen, Digital twin applications on cultural world heritage sites in China: a state-of-the-art overview, J. Cult. Heritage 64 (2023) 228–243, doi:10.1016/j.culher.2023.10.005.
[2] Cultural heritage in 3D, 2023, https://patrimoni.gencat.cat/en/stories/cultural-heritage-3d.
[3] S. Ziesche, Open data for AI: what now?, 2023, https://www.unesco.org/en/articles/open-data-ai-what-now.
[4] F. Croizet, Smart museums to face the crisis, 2021, https://icom.museum/en/news/smart-museums-to-face-the-crisis/.
[5] D. Fonner, Empowering provenance research in the age of big data and (re)generative artificial intelligence, 2023, https://www.aam-us.org/2023/08/23/empowering-provenance-research-in-the-age-of-big-data-and-regenerative-artificial-intelligence/.
[6] M. Pasikowska-Schnass, Artificial intelligence in the context of cultural heritage and museums: complex challenges and new opportunities, 2023, https://www.europarl.europa.eu/thinktank/en/document/EPRS_BRI(2023)747120.
[7] M. Fiorucci, M. Khoroshiltseva, M. Pontil, A. Traviglia, A. Del Bue, S. James, Machine learning for cultural heritage: a survey, Pattern Recognit. Lett. 133 (2020) 102–108, doi:10.1016/j.patrec.2020.02.017.
[8] G. Pandi, K.P. Aggarwal, Deep learning-based 3-D model for the cultural heritage sites in the state of Gujarat, India, in: M. Pandit, M.K. Gaur, S. Kumar (Eds.), Artificial Intelligence and Sustainable Computing, Springer Nature Singapore, Singapore, 2023, pp. 737–750.
[9] J. Pan, L. Li, H. Yamaguchi, K. Hasegawa, F.I. Thufail, Brahmantara, S. Tanaka, 3D reconstruction of Borobudur reliefs from 2D monocular photographs based on soft-edge enhanced deep learning, ISPRS J. Photogramm. Remote Sens. 183 (2022) 439–450, doi:10.1016/j.isprsjprs.2021.11.007.
[10] G. Artopoulos, M.I. Maslioukova, C. Zavou, M. Loizou, M. Deligiorgi, M. Averkiou, An artificial neural network framework for classifying the style of cypriot hybrid examples of built heritage in 3D, J. Cult. Heritage 63 (2023) 135–147, doi:10.1016/j.culher.2023.07.016.
[11] G. Mazzacca, A. Karami, S. Rigon, E. Farella, P. Trybala, F. Remondino, NeRF for heritage 3D reconstruction, Int. Arch. Photogramm. Remote Sens. Spatial Inf. Sci. XLVIII-M-2-2023 (2023) 1051–1058, doi:10.5194/isprs-archives-XLVIII-M-2-2023-1051-2023.
[12] N. Farajzadeh, M. Hashemzadeh, A deep neural network based framework for restoring the damaged persian pottery via digital inpainting, J. Comput. Sci. 56 (2021) 101486, doi:10.1016/j.jocs.2021.101486.
[13] L. Cipolina-Kun, S. Caenazzo, G. Mazzei, Comparison of CoModGANs, LaMa and GLIDE for art inpainting completing M.C Eschers print gallery, in: 2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW), IEEE Computer Society, Los Alamitos, CA, USA, 2022, pp. 715–723, doi:10.1109/CVPRW56347.2022.00087.
[14] X. Deng, Y. Yu, Ancient mural inpainting via structure information guided two-branch model, Heritage Sci. 11 (2023), doi:10.1186/s40494-023-00972-x.
[15] R. Stoean, N. Bacanin, C. Stoean, L. Ionescu, M. Atencia, G. Joya, Computational framework for the evaluation of the composition and degradation state of metal heritage assets by deep learning, J. Cult. Heritage 64 (2023) 198–206, doi:10.1016/j.culher.2023.10.007.
[16] C. Stoean, N. Bacanin, Z. Volkovich, L. Ionescu, R. Stoean, Study on semantic inpainting deep learning models for artefacts with traditional motifs, in: I. Rojas, G. Joya, A. Catala (Eds.), Advances in Computational Intelligence, Springer Nature Switzerland, Cham, 2023, pp. 479–490.
[17] M. Tancik, E. Weber, E. Ng, R. Li, B. Yi, T. Wang, A. Kristoffersen, J. Austin, K. Salahi, A. Ahuja, D. Mcallister, J. Kerr, A. Kanazawa, Nerfstudio: a modular framework for neural radiance field development, in: SIGGRAPH '23: ACM SIGGRAPH 2023 Conference Proceedings, Association for Computing Machinery, New York, NY, USA, 2023, pp. 1–12, doi:10.1145/3588432.3591516.
[18] J.L. Schönberger, J.-M. Frahm, Structure-from-motion revisited, in: 2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2016, pp. 4104–4113, doi:10.1109/CVPR.2016.445.
[19] J.L. Schönberger, T. Price, T. Sattler, J.-M. Frahm, M. Pollefeys, A vote-and-verify strategy for fast spatial verification in image retrieval, in: S.-H. Lai, V. Lepetit, K. Nishino, Y. Sato (Eds.), Computer Vision – ACCV 2016, Springer International Publishing, Cham, 2017, pp. 321–337.
[20] R. Rombach, A. Blattmann, D. Lorenz, P. Esser, B. Ommer, High-resolution image synthesis with latent diffusion models, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2022, pp. 10684–10695.
[21] B. Mildenhall, P.P. Srinivasan, M. Tancik, J.T. Barron, R. Ramamoorthi, R. Ng, NeRF: representing scenes as neural radiance fields for view synthesis, Commun. ACM 65 (1) (2021) 99-106, doi:10.1145/3503250.
[22] W. Burger, M.J. Burge, Digital Image Processing: An Algorithmic Introduction, Springer Nature, 2022.
[23] Z. Wang, A.C. Bovik, Mean squared error: love it or leave it? A new look at signal fidelity measures, IEEE Signal Process. Mag. 26 (1) (2009) 98–117, doi:10.1109/MSP.2008.930649.
[24] Z. Wang, E. Simoncelli, A. Bovik, Multiscale structural similarity for image quality assessment, in: The Thrity-Seventh Asilomar Conference on Signals, Systems & Computers, 2003, 2, 2003, pp. 1398–1402Vol.2, doi:10.1109/ACSSC.2003.1292216.
[25] Z. Wang, A. Bovik, A universal image quality index, IEEE Signal Process. Lett. 9 (3) (2002) 81–84, doi:10.1109/97.995823.
[26] H. Sheikh, A. Bovik, Image information and visual quality, IEEE Trans. Image Process. 15 (2) (2006) 430–444, doi:10.1109/TIP.2005.859378.
[27] E. Straková, D. Lukáš, Z. Bobovský, T. Kot, M. Mihola, P. Novák, Matching point clouds with STL models by using the principle component analysis and a decomposition into geometric primitives, Appl. Sci. 11 (5) (2021), doi:10.3390/app11052268.
[28] H. Liu, T. Liu, Y. Li, M. Xi, T. Li, Y. Wang, Point cloud registration based on MCMC-SA ICP algorithm, IEEE Access 7 (2019) 73637–73648, doi:10.1109/ACCESS.2019.2919989.
[29] J. Ryu, S. ichiro Kamata, An efficient computational algorithm for Hausdorff distance based on points-ruling-out and systematic random sampling, Pattern Recognit. 114 (2021) 107857, doi:10.1016/j.patcog.2021.107857.