ENHANCING AUTOMATION OF HERITAGE PROCESSES: GENERATION OF ARTIFICIAL TRAINING DATASETS FROM PHOTOGRAMMETRIC 3D MODELS

The International Archives of the Photogrammetry, Remote Sensing and Spatial Information Sciences, Volume XLVIII-M-2-2023
29th CIPA Symposium "Documenting, Understanding, Preserving Cultural Heritage:
Humanities and Digital Technologies for Shaping the Future", 25–30 June 2023, Florence, Italy

ENHANCING AUTOMATION OF HERITAGE PROCESSES: GENERATION OF ARTIFICIAL TRAINING DATASETS FROM PHOTOGRAMMETRIC 3D MODELS

G. Patrucco [1] *, F. Setragno [2]

[1] Laboratory of Geomatics for Cultural Heritage (G4CH Lab), Department of Architecture and Design (DAD), Politecnico di Torino, Viale Pier Andrea Mattioli 39, 10125 Torino, Italy - giacomo.patrucco@polito.it
[2] Volta®A.I. – Via Roberto Lepetit 34, 21040 Gerenzano, Italy - francesco@volta.ai

KEY WORDS: Movable heritage, Deep Learning, Artificial training datasets generation, Digital Photogrammetry, Automatic Classification, Semantic Segmentation.

ABSTRACT:

Nowadays, many efficient technologies have been developed with the aim of collecting digital images and other metric data, greatly optimising the acquisition procedures and techniques. However, processing this data can be onerous and time-consuming, and increasingly often, there is a need to develop new strategies to enhance the level of automation of these processes. Using artificial intelligence, and particularly Convolutional Neural Networks, it is possible to automate processing tasks such as classification and segmentation. However, a significant challenge is represented by the necessity of obtaining sufficient training data to properly train a deep learning model. These datasets are composed of a significant amount of data and need to be annotated, which sometimes represents an onerous and challenging task. Synthetic data can represent an effective solution to this problem, significantly reducing the time and effort required to manually create annotated datasets and can be particularly useful when studying objects characterised by specific features and high complexity, requiring tailored solutions and ad hoc training. The presented research explores the opportunity of using synthetic datasets – generated from photogrammetric 3D models – for deep-learning-based heritage digitisation applications. The use of synthetic data generated from textured 3D models derived from SfM photogrammetric processes is proposed, with the aim of enhancing automatic procedures in the framework of heritage processes.

1. INTRODUCTION

In the last decades, we have witnessed impressive technological signs of progress and improvements, which led to the development of very effective and efficient solutions in the framework of heritage digitisation. Nowadays, numerous effective and efficient solutions are available for collecting digital images, dense points clouds and other metric data, allowing rapid and sustainable acquisitions.

However, despite the development of these new technologies, different open issues still limit the efficiency of these digitisation processes. If, on the one hand, the acquisition phase has become increasingly faster, thanks to solutions that greatly enhance the rapidity of the primary data collection, on the other side, it should be underlined that these new technologies are extremely onerous from both a computational point of view and, above all, in terms of resources required from the operator during the processing of the data.

As a matter of fact, these procedures are often repetitive and time-spending, requiring manual operations that are often unsustainable when applied in the framework of massive digitisation projects.

Artificial Intelligence (AI) has established itself as a powerful and effective solution to improve the automation levels in the framework of the processing involving heritage datasets, especially concerning classification or semantic segmentation tasks (Grilli & Remondino 2019; Fan et al. 2018, Zia et al. 2022; Adamopoulos 2021).

In particular, in the framework of image processing, Convolutional neural networks (CNN) have been demonstrated to be one of the most effective – and efficient – techniques for these kinds of tasks, outperforming traditional methods in several application fields (semantic segmentation, classification, object detection, etc.) (Gu et al. 2018; Garcia-Garcia et al. 2018).

However, a significant limitation in efficiently utilising deep learning-based approaches – for optimising and enhancing the processing of heritage datasets – is represented by the necessity of a significant amount of input data to train the neural networks adequately (Ridnik et al. 2021). This represents a crucial aspect, considering that the heritage field is traditionally underfunded, and most researchers working in the conservation and valorisation field often have limited resources and consequently may not dispose of similar datasets. In this case, the complexity of the operation is not only related to the acquisition of the primary data – the digital image – but also to the need to perform annotation procedures (almost always involving manual procedures), exponentially increasing the time required to generate similar datasets.

It is important to consider the possibility of using existing datasets, if available, and modifying or integrating them according to the required tasks. This can significantly reduce the time and effort required to manually create annotated datasets. The increasing availability of benchmark datasets can represent an effective solution, considering that many of these datasets are often composed of a very high number of images (Kirillov et al. 2023; Matrone et al. 2020), and they are often shared and made available to different scientific communities for research and experimentation purposes.

However, it should be underlined that in many cases and application fields, such as heritage, the studied objects are characterised by high specificity, and the tasks required from the training of neural networks to support experts working in the framework of restoration and conservation, such as providing automatic methods for recognition and classification of decay,

* Corresponding author

are extremely complex (Hatır et al. 2021). For these reasons, these benchmark datasets may not always be entirely sufficient for the efficient training of deep learning models that accurately perform these tasks.
There is a crucial need to develop new techniques and strategies to deal with this problem focusing on the significant features of the objects and the tasks connected to the heritage studies, which always require tailored approaches and solutions. Specifically, the automatic generation of artificial training datasets can represent a viable solution (de Melo et al. 2021; Tremblay et al. 2018). It is worth mentioning the emergence of GAN (Generative Adversarial Network), which has enabled the production of realistic synthetic data and thereby made it possible to artificially augment training data (Bowles et al. 2018).

1.1 Motivation of the research

Despite the possible solutions mentioned in the previous section, the problem of generating training datasets and, above all, adequate labels represents a significant obstacle to the efficient use of deep learning in the field of cultural heritage, especially when the analysed objects are characterised by very specific features are when very specific tasks are required.
In particular, the main aim of this contribution is to propose a replicable methodology for generating a high amount of annotated synthetic images exploiting textured 3D models derived from SfM-based (Structure-from-Motion) photogrammetric processes, generating rendered images of textured photogrammetric 3D models. The goal is the generation of synthetic training datasets with the purpose of facilitating deep-learning-based applications without the onerous manual work required from labelling operations.

2. METHODS

The primary data have been collected in the framework of the B.A.C.K. TO T.H.E. F.U.T.U.RE. project (Lo Turco et al. 2018a). During the course of this project – carried out by a multidisciplinary research group from the Politecnico di Torino – a collection of wooden maquettes representing ancient Nubian temples ('Expedition models of Egyptian architectures') has been digitised using an image-based approach. The dataset comprises 2908 digital images (resolution: 8688 x 5792 pixels) of the different maquettes belonging to the collection (composed of 26 pieces, characterised by similar features in terms of size, morphology and material consistency), preserved in the Museo Egizio of Torino (Lo Turco et al. 2018b). The images have been acquired using a Canon EOS 5DS R DLSR camera (sensor: CMOS 50.3 Mpx; sensor size: 36 x 24 mm^2; focal length: 50 mm).
The data acquisition has been carried out as described in Patrucco & Setragno 2021, following photogrammetric criteria (overlap between adjacent images >80-90 % and high convergence of the cameras) with the final goal of performing a photogrammetric SfM-based 3D reconstruction of the wooden models.
Starting from these images, two different strategies have been developed in the framework of this research with the aim of training a model able to automatically identify and perform a pixel-based classification of the maquettes within a digital image.
For this reason, two neural networks have been trained following different strategies as regards the generation of the training datasets. The adopted strategies are as follows:

- In the first case, the first deep learning model has been trained using the digital 2908 images acquired as described in the previous paragraphs. The images have been manually annotated in order to generate adequate labels for the training. This dataset will be referred to as Dataset A. The results of the training performed using this dataset have been assumed as a ground truth to evaluate those obtained using the strategy proposed in the following paragraph.
- In the second case, the deep learning model has been trained from a dataset (which will be referred to as Dataset B) composed of rendered images of a photogrammetric 3D model (derived from the SfM-based processing of Dataset A).

The number of images and the resolution of each image are the same for both the considered datasets. The flowchart of the two proposed workflows can be observed in Figure 1.

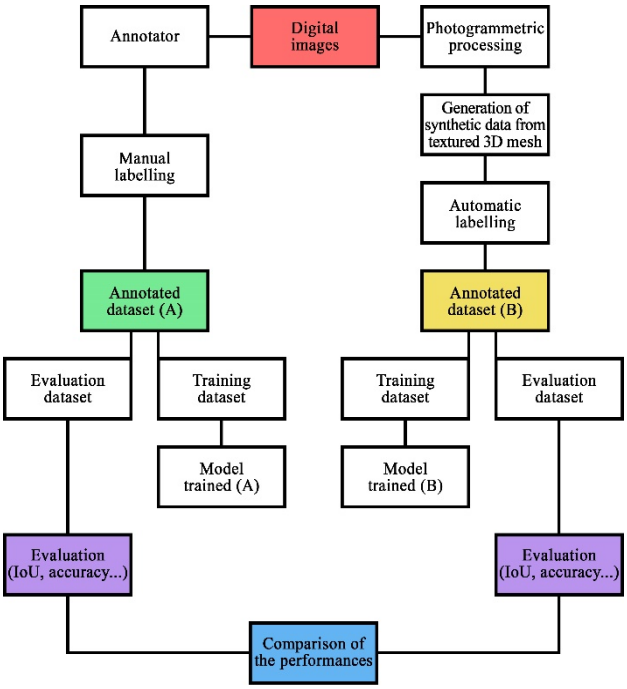

Figure 1. Flowchart of the two workflows presented in the current research.

While Dataset A has been generated following a standard approach, the generation of Dataset B required some less established operations. In the following section, it will be explained in detail how these synthetic data have been generated and annotated to make them suitable for use as input data for neural network training.

2.1 Generation of the synthetic dataset

The second dataset has been generated starting from the photogrammetric 3D model of each maquette. Specifically, a textured mesh is required in order to generate the rendered images that compose the synthetic dataset. The proposed workflow is as follows:

- Step 1. For each maquette dataset, a standard SfM-based workflow has been followed. The photogrammetric software used for the 3D reconstruction is the Agisoft Metashape platform (build 1.8). Specifically, the following methodology was adopted: interior orientation (determination of principal distance, principal point, lens distortions and affinity parameters) by self-calibration approach during the bundle adjustment (Granshaw 2020) and tie points generation by means of relative orientation.

The obtained model has been then scaled using scale bars placed on the acquisition stage during the collection of the digital images;

- Step 2. Depth maps generation and subsequent generation of a dense point cloud;
- Step 3. Triangulation of a high-resolution 3D mesh and generation of a UV map (Figure 2) from the projection of the oriented images.

Figure 2. (a) Texture 3D model ('Dendur portal' wooden model); (b) UV map.

- Step 4. Automatic generation of the synthetic rendered training dataset using open-source modelling software Blender and a Phyton script. By exporting the model in COLLADA format (.dae), the 3D mesh preserves the information related to the position and orientation of the cameras (Figure 3). Blender is able to read this information and use it to generate a virtual camera for the generation of the rendered images (Figure 4);

Figure 3. Oriented photogrammetric block of aligned images imported into the Blender platform as virtual cameras for generating the rendered dataset.

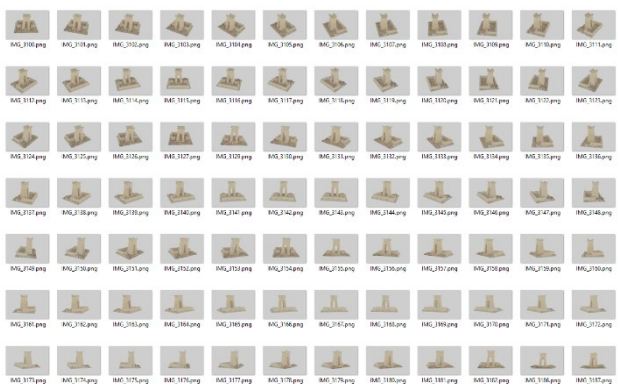

Figure 4. Example of rendered images generated from the textured photogrammetric 3D model.

- Step 5. Automatic generation of the rendered multi-class (class 1: wooden maquette; class 2: background) labels (by subtracting the pixels belonging to the background, automatically identified – using a script – according to an RGB-based selection);
- Step 6. Reclassification procedure to convert the rendered multi-class labels in a properly formatted ground truth for the training. Starting from the rendered images, it was possible to generate binary labels and automatically achieve the labels necessary for the training as ground truth (Figure 5).

Figure 5. Example of automatically generated labels. Each label corresponds to a rendered image from the training dataset.

2.1 Neural network training

Once both datasets – and relative labels – were obtained, it was possible to proceed with training aimed at achieving a deep learning model capable of automatically recognising the objects of interest (the wooden maquettes) within an image. Both neural network training procedures have been performed using the same architecture (DeepLab V3+) (Chen et al. 2017), the same parameters and the same data augmentation strategies (in particular, among others, changing artificial backgrounds to improve the capability to identify the object of interest). Concerning both datasets, 90% of the images have been used as training data, while the remaining 10% has been employed as validation datasets for the assessment and evaluation of the trained models' effectiveness.

2.2 Strategies for training enhancement

Different strategies have been followed in order to improve the effectiveness of the training. First of all, different data augmentation strategies were applied. As a matter of fact, very often, the number of available images to perform training is limited (as in the case presented in this research), and this may affect the effectiveness of the trained model and the generalisation capability. This represents a common criticality for many research groups that often do not have the resources to acquire and annotate an adequate dataset. In these cases, different data augmentation strategies have been developed in the last few years (Wang & Perez 2017; Mikołajczyk & Grochowski 2018) with the aim of increasing the volume of the dataset and diversifying its features in order to improve the generalisation capability of the training. Traditional methods have been used (as described in the previous experience, reported in Patrucco & Setragno 2021):

- Random rotation;
- Random cropping;

- Horizontal flipping;
- Vertical flipping;
- Random brightness changes;
- Random occlusions.
- Random background changes.

The latter strategy listed above was particularly important for Training B since the rendered images are characterised by the presence of a white background. In this case, adding an artificial background is almost mandatory since otherwise, there would be a risk that the model learns to recognise the object only when that feature is identified, and clearly, this cannot happen when the classification task is applied not to artificial images, but to real images.

For this reason, two different typologies of artificial background have been used to provide context for the wooden models during the training:

- Generic backgrounds randomly downloaded from the internet (Figure 6a);
- Realistic empty backgrounds acquired using a DSLR camera (Figure 6b).

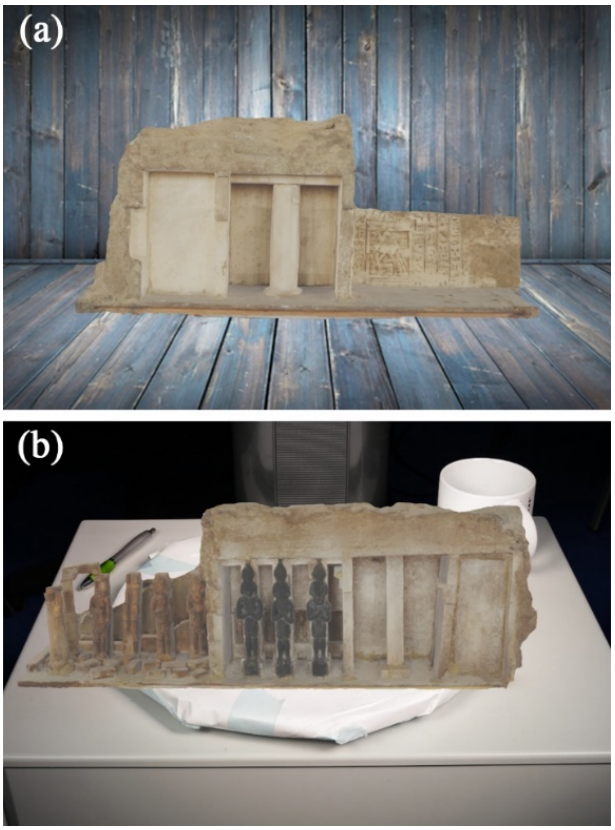

Figure 6. Example of artificial backgrounds used during the training procedures: (a) Artificial background randomly downloaded from the internet; (b) Empy background acquired with a DSLR camera.

Additionally, another strategy has been implemented with the aim of improving training performance and avoiding out-of-memory errors caused by the high resolution of the images acquired with a DSLR camera (8688 x 5792 pixels). A common issue during training is the risk of out-of-memory errors caused by high-resolution images. To avoid this, a common strategy is to downsample the original images. However, in this project, a different approach was tested: each digital image was divided into different tiles of 512 x 512 pixels rather than simply downsampling the images. This prevented memory errors and, at the same time, allowed to exploit the original high spatial resolution of the images maintaining the level of detail. Additionally, another advantage is represented by the fact that, by following this strategy, the number of images composing the training dataset increases significantly.

However, to ensure the network could learn features from the contextualisation of each tile, a subsampled version of the original image was added to the cropped tiles, creating a multiband raster (Figure 7). As a result, each image in the training dataset has a resolution of 512 x 512 pixels and is composed of the following six bands:

- First band: Red band (original resolution);
- Second band: Green band (original resolution);
- Third band: Blue band (original resolution);
- Fourth band: Red band (downsampled image);
- Fifth band: Green band (downsampled image);
- Sixth band: Blue band (downsampled image).

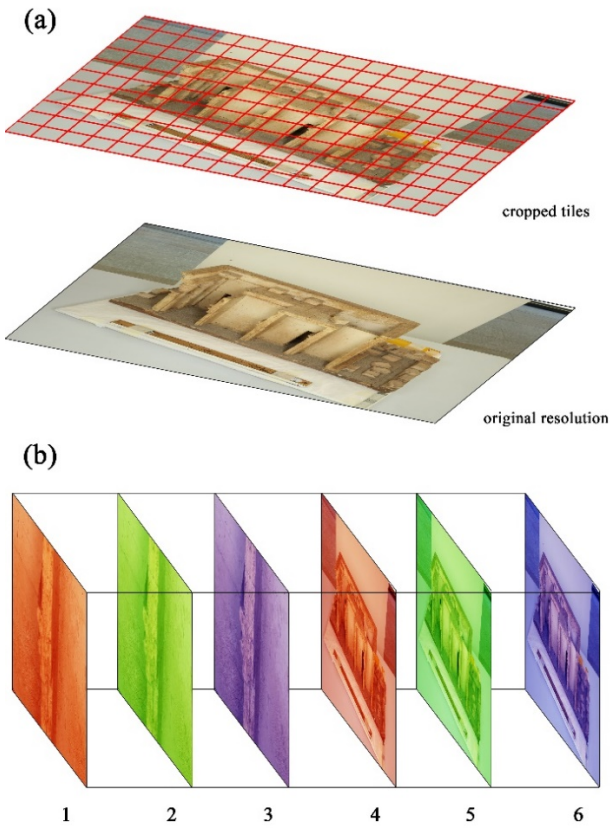

Figure 7. (a) Tiles subdivision of the original image raster; (b) Example of multiband raster used as input training dataset: (1) Red band, original resolution (cropped tile); (2) Green band, original resolution (cropped tile); (3) Blue band, original resolution (cropped tile); (4) Red band, downsampled image; (5) Green band, downsampled image; (6) Blue band, downsampled image.

3. RESULTS

3.1 Validation of the trained models

The following performance evaluation metrics have been considered for evaluating the quality and the effectiveness of the neural networks training: Accuracy, IoU (Intersection over Union), Precision, Recall and F1-score. These metrics – which are the ones traditionally used to define and evaluate the overall

quality and effectiveness of training in the deep learning field – are defined as:

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

$$\text{IoU} = \frac{\text{Area of Overlap}}{\text{Area of Union}}$$

$$\text{Precision} = \frac{\text{TP}}{\text{TP} + \text{FP}}$$

$$\text{Recall} = \frac{\text{TP}}{\text{TP} + \text{FN}}$$

$$\text{F1} - \text{score} = \frac{2 \text{ x Precision x Recall}}{\text{Precision} + \text{Recall}}$$

where TP = True Positive (pixels correctly predicted as belonging to a specific class)
TN = True Negative (pixels correctly predicted as not belonging to a specific class)
FP = False Positive (pixels erroneously predicted as belonging to a specific class)
FN = False Negative (pixels erroneously predicted as not belonging to a specific class)

After both trainings, the performances of the two neural networks have then been compared and evaluated in order to assess the effectiveness of this strategy. At first, validation was performed by considering, for each predictive model, the reference evaluation dataset, consisting of 10% of the overall image dataset. In both cases, the results highlight high performances, indicating that in both cases, the neural network training produced a predictive model capable of adequately performing the given task. The performance evaluation metrics are reported in Table 1. As it is possible to observe (Figure 8), the results obtained from the training developed from the synthetic dataset are completely consistent with those obtained from the traditional manually annotated dataset.

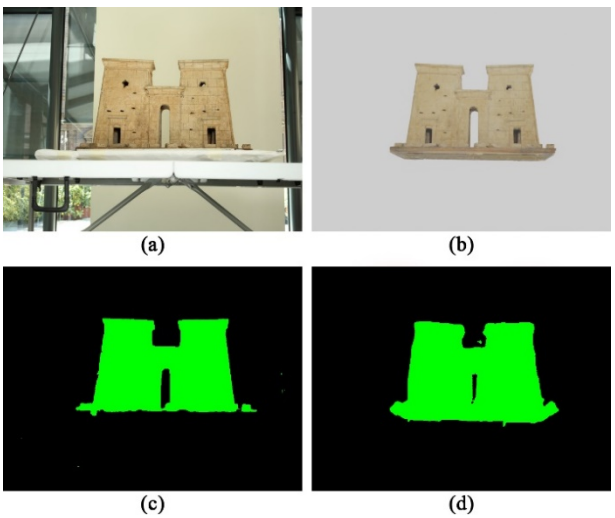

Figure 8. (a) Original digital image belonging to Dataset A; (b) Synthetic rendered image belonging to Dataset B; (c) Predicted classified image derived from the algorithm trained with Training A; (d) Predicted classified image derived from the algorithm trained with Training B.

	Training A (traditional dataset)	**Training B** (synthetic dataset)
Mean IoU	88%	88%
Accuracy	93%	97%
Precision	90%	94%
Recall	97%	91%
F1-score	93%	92%

Table 1. Comparison between the performance evaluation metrics of the two different training procedures.

4. DISCUSSION

In the previous section, it was observed how the performance of both trained models is adequate with respect to the expectations and requirements for the considered task (identification and segmentation of the object of interest, i.e., the wooden models).
From the comparison between the performances of the two trained neural networks (Table 1) and from a visual inspection of the generated predicted classified images (Figure 8), it is clear that the strategy of generating rendered synthetic images for the automatic production of training datasets can represent a successful alternative to onerous and time-spending manual labelling operations.
However, the goal of Training B (trained using synthetic images) is to recognise – with a good level of accuracy – not only the rendered images (that compose the validation dataset of Dataset B) but also real images. Therefore, the validation dataset of Dataset A was processed using the deep learning model derived from Training B. In Figure 9, it is possible to observe four images (belonging to the 'Temple of Abu Oda' wooden model dataset) classified using the trained model B, while the evaluation performance metrics achieved on the evaluation dataset A are reported in Table 2.

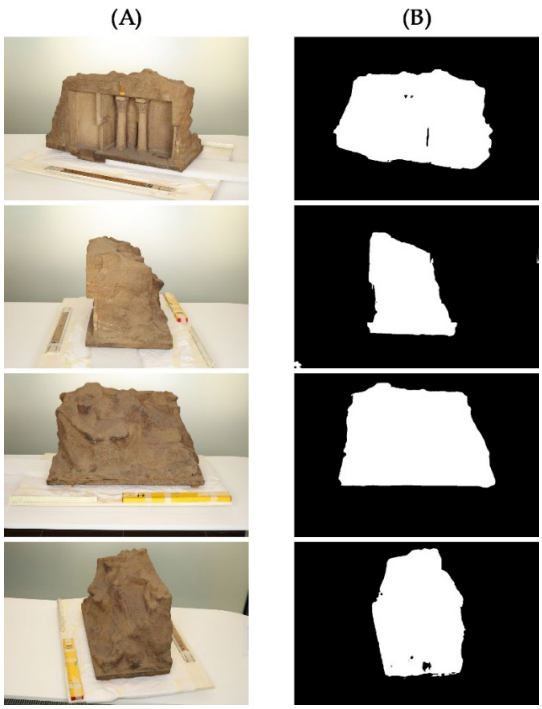

Figure 9. Examples of images belonging to the evaluation dataset of Dataset A classified using the model trained from the synthetic images of Dataset B. (A) Original digital image; (B) Classified image.

	Training B Evaluation dataset belonging to the traditional dataset (Dataset A)
Mean IoU	80%
Accuracy	89%
Precision	72%
Recall	98%
F1-score	83%

Table 2. Performance evaluation metrics observed on the classification of the validation dataset (belonging to the traditional dataset) using the deep learning model derived from Training B.

The first observation that emerges from the analysis of these pieces of evidence is that most of the results show a similar order of magnitude, evidencing that the model is able to fulfil the task of recognising and segmenting the wooden maquettes. Based on the performance evaluation metrics observed in Table 1 (concerning Training A) and in Table 2, it appears that both deep learning models achieved good results in terms of mean IoU, Accuracy, Recall, and F1-score. However, there is some discrepancy between the two models in terms of Precision.

The deep learning model A achieved a Precision score of 90%, which means that out of all the predicted positive instances, 90% were actually true positives. This can be considered a good result, indicating that the model made relatively few false positive predictions.

In contrast, the deep learning model B – when processing real images – achieved a Precision score of 72%, which means that a higher number of false positive predictions were made. This indicates that the model may not have performed as well in terms of identifying only relevant instances.

Overall, the second model performed slightly worse than the first one in terms of all the metrics except for Recall, where it achieved a higher score. This means that the second model was better at correctly identifying all instances of the object of interest in the image but at the cost of a higher number of false positives.

In a deep learning model, having lower Precision but high Recall means that the model is able to identify a high number of relevant instances (i.e., true positives) but also a high number of irrelevant instances (i.e., false positives).

It's important to underline that the synthetic images used to train the second model may not have perfectly represented the real images that the model was applied to, which could have affected its performance. Therefore, there is still room for improvement to enhance the synthetic dataset and, consequently, increase the efficiency of the trained model in terms of automatic recognition.

However, generally, from the analysis of these metrics, it can be deduced that the obtained results are adequate compared to the expectations, and the model is able to perform the desired task (automatic detection and segmentation of the object of interest). Additionally, the approach based on the generation of synthetic datasets for the training of CNNs is considerably advantageous concerning the time required for annotating the datasets. In order to properly annotate Dataset A with manual procedures, considering approximately 1-2 minutes for the generation of each label, 50-100 hours were necessary for the annotation procedures. In the case of Dataset B, the labels were generated in a few minutes (with an inference time <1 second for each label), requiring minimal operator involvement. This represents a significant advantage both in terms of increasing efficiency and optimisation in the processes related to the processing and analysis of heritage-related datasets and for the contribution required from the operator who, following the presented strategy, avoids a time-consuming and repetitive procedure.

5. CONCLUSIONS AND FUTURE PERSPECTIVES

In the presented paper, a methodology for the automatic identification and segmentation of heritage datasets (specifically, in the case of the current research, wooden maquettes belonging to a museum collection) from synthetic images was developed and tested. Classification and semantic segmentation represent critical tasks in the field of cultural heritage nowadays. The obtained results showed that both models achieved good performances in terms of the evaluation metrics used. However, the second model – trained on synthetic images – showed slightly lower metrics (in particular, in terms of Precision), compared to the first model, with the exception of the Recall, which is higher. This may suggest that the synthetic dataset may not have perfectly represented the real images, indicating a potential area for improvement.

However, the proposed strategy – and, in general, the new increasingly developed strategies involving the use of automatically generated artificial data – may potentially represent a step ahead towards an automated approach, helping in optimising – from a time and cost savings perspective – many processes involving heritage datasets. The proposed methodology can be extended and applied to other heritage domains, facilitating the identification and classification of objects of interest in images.

Additionally, an opportunity that should be considered is represented by the abundance of 3D models of heritage assets derived from photogrammetric surveys within the geomatics scientific community. If properly shared – through appropriate platforms or online 3D viewers – these models could represent a valid starting point for generating synthetic datasets to be used for the training of deep learning models, with the aim to perform tasks supporting the study, the investigation and the analysis of built heritage assets.

Additionally, future perspectives include the exploration of new strategies for generating synthetic datasets that better represent the real object of interest, investigating the generalisation capabilities of the models on different datasets, and developing a more comprehensive methodology that integrates different techniques for image analysis and deep learning. Overall, the proposed methodology represents a promising approach to support the preservation and analysis of cultural heritage.

REFERENCES

Adamopoulos, E., 2021. Learning-based classification of multispectral images for deterioration of historic structures. *Journal of Building Pathology and Rehabilitation*, 6(41).

Bowles, C., Chen, L., Guerrero, R., Bentley, P., Gunn, R., Hammers, A., Dickie, D. A., Hernández, M. V., Wardlaw, J., Rueckert, D., 2018. GAN Augmentation: Augmenting Training Data using Generative Adversarial Networks. *arXiv preprint*, arXiv:1810.10863.

Chen, L. C., Papandreou, G., Kokkinos, I., Murphy, K., Yuille, A. L., 2017. Deeplab: Semantic Image Segmentation with Deep Convolutional Nets, Atrous Convolution, and Fully Connected CRFs. *IEEE transactions on pattern analysis and machine intelligence*, 40(4), 834–848.

de Melo, C. M., Torralba, A., Guibas, L., DiCarlo, J., Chellappa, R., Hodgins, J., 2021. Next-generation deep

learning based on simulators and synthetic data. *Trends in Cognitive Sciences*, 26(2), 174-187.

Fan, Q., Yang, J., Hua, G., Chen, B., Wipf, D., 2018. Revisiting deep intrinsic image decompositions. *Proceedings of the IEEE Conference of Computer Vision and Pattern Recognition*, 8944-8952.

Garcia-Garcia, A., Orts-Escolano, S., Oprea, S., Villena-Martinez, V., Garcia-Rodriguez, J., 2017. A survey on deep learning techniques for image and video semantic segmentation. *Applied Soft Computing*, 70, 41-65.

Granshaw, S. I., 2020. Photogrammetric terminology: fourth edition. *The Photogrammetric Record*, 35(170), 143-288.

Grilli, E., Remondino, F., 2019. Classification of 3D Digital Heritage. *Remote Sensing*, 11(7), 847.

Gu, J., Wang, Z., Kuen, J., Ma., L., Shahroudy, A., Shuai, B., Chen, T., 2018. Recent advances in convolutional neural networks. *Pattern Recognition*, 77, 354-377.

Hatır, E., Korkanç, M., Schachner, A., İnce, I., 2021. The deep learning method applied to the detection and mapping of stone deterioration in open-air sanctuaries of the Hittite period in Anatolia. Journal of Cultural Heritage, 51, 37-49.

Kirillov, A, Mintun, E., Ravi, N., Mao, H., Rolland, C., Gustafson, L., Xiao, T., Whitehead, S., Berg, A. C., Lo, W.-Y., Dollár, P. & Girshick, R. (2023). Segment Anything. *arXiv preprint*, arXiv:2304.02643v1.

Lo Turco, M., Piumatti, P., Rinaudo, F., Tamborrino, R., González-Aguilera, D., 2018a. B.A.C.K. TO T.H.E. F.U.T.U.RE. – BIM acquisition as cultural key to transfer heritage of ancient Egypt for many uses to many users replayed. In Bertocci, S. (Ed.), *Programmi Multidisciplinari Per L'internazionalizzazione Della Ricerca. Patrimonio Culturale, Architettura e Paesaggio*, 107–109, DIDA Press.

Lo Turco, M., Piumatti, P., Rinaudo, F., Calvano, M., Spreafico, A., Patrucco, G., 2018b. The digitisation of museum collections for research, management and enhancement of tangible and intangible heritage. *3rd Digital Heritage International Congress (DigitalHERITAGE) held jointly with 24th International Conference on Virtual Systems & Multimedia (VSMM 2018)*, San Francisco, CA, USA.

Matrone, F., Lingua, A., Pierdicca, R., Malinverni, E. S., Paolanti, M., Grilli, E., Remondino, F., Murtiyoso, A., Landes, T., 2020. A benchmark for large-scale heritage point cloud semantic segmentation. *Int. Arch. Photogramm. Remote Sens. Spatial Inf. Sci.*, XLIII-B2-2020, 1419–1426.

Mikołajczyk, A., Grochowski, M., 2018. Data augmentation for improving deep learning in image classification problem. *IEEE 2018 international interdisciplinary PhD workshop*, 117-122.

Mishra, M., Barman, T., Ramana, G. V., 2022. Artificial intelligence-based visual inspection system for structural health monitoring of cultural heritage. *Journal of Civil Structural Health Monitoring*.

Patrucco, G., Setragno, F., 2021. Multiclass semantic segmentation for digitisation of movable heritage using deep learning techniques. *Virtual Archaeology Review*, 12(25), 85-98.

Ridnik, T., Ben-Baruch, E., Noy, A. & Zelnik-Manor, L. (2021). Imagenet-21k pretraining for the masses. *arXiv preprint*, arXiv:2104.10972.

Tremblay, J., Prakash, A., Acuna, D., Brophy, M., Jampani, V., Anil, C., To, T., Cameracci, E., Boochoon, S., Birchfield, S., 2018. Training Deep Learning Networks with Synthetic Data: Bridging the Reality Gap by Domain Randomization. *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*, 1082-1090.

Wang, J., Perez, L., 2017. The effectiveness of data augmentation in image classification using deep learning. *Convolutional Neural Networks Vis. Recognit*, 11, 1-8.

Zia, T., Bashir, N., Ullah, M., A., Murtaza, S., 2022. SoFTNet: A concept-controlled deep learning architecture for interpretable image classification. *Knowledge-Based Systems*, 468, 317-223.