POLITECNICO DI TORINO
Repository ISTITUZIONALE

Multiclass semantic segmentation for digitisation of movable heritage using deep learning techniques

Original
Multiclass semantic segmentation for digitisation of movable heritage using deep learning techniques / Patrucco, Giacomo; Setragno, Francesco. - In: VIRTUAL ARCHAEOLOGY REVIEW. - ISSN 1989-9947. - ELETTRONICO. - 12:25(2021), pp. 85-98. [10.4995/var.2021.15329]

Availability:
This version is available at: 11583/2912832 since: 2021-07-14T14:51:55Z

Publisher:
Editorial Universitat Politècnica de València

Published
DOI:10.4995/var.2021.15329

Terms of use:

This article is made available under terms and conditions as specified in the corresponding bibliographic description in the repository

Publisher copyright

(Article begins on next page)

12 October 2024

Virtual Archaeology Review, **12**(25): 85-98, 2021
https://doi.org/10.4995/var.2021.15329
© UPV, SEAV, 2015

Received: March 26, 2021
Accepted: May 4, 2021

MULTICLASS SEMANTIC SEGMENTATION FOR DIGITISATION OF MOVABLE HERITAGE USING DEEP LEARNING TECHNIQUES

SEGMENTACIÓN SEMÁNTICA MULTICLASE EN LA DIGITALIZACIÓN DEL PATRIMONIO MUEBLE UTILIZANDO TÉCNICAS DE APRENDIZAJE PROFUNDO

Giacomo Patrucco[a,*], Francesco Setragno[b]

[a] 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

[b] Volta® A.I. – Via Roberto Lepetit 34, 21040 Gerenzano, Italy. francesco@volta.ai

Highlights:

- In the framework of movable heritage digitisation processes, many procedures are very time-consuming, and they still require the operator's substantial manual involvement.
- This research proposes using deep learning techniques to enhance the automatism level in the generation of exclusion masks, improving the optimisation of the photogrammetric procedures.
- Following this strategy, the possibility of performing a multiclass semantic segmentation (on the 2D images and, consequently, on the 3D point cloud) is also discussed, considering the accuracy of the obtainable results.

Abstract:

Digitisation processes of movable heritage are becoming increasingly popular to document the artworks stored in our museums. A growing number of strategies for the three-dimensional (3D) acquisition and modelling of these invaluable assets have been developed in the last few years. Their objective is to efficiently respond to this documentation need and contribute to deepening the knowledge of the masterpieces investigated constantly by researchers operating in many fieldworks. Nowadays, one of the most effective solutions is represented by the development of image-based techniques, usually connected to a Structure-from-Motion (SfM) photogrammetric approach. However, while images acquisition is relatively rapid, the processes connected to data processing are very time-consuming and require the operator's substantial manual involvement. Developing deep learning-based strategies can be an effective solution to enhance the automatism level. In this research, which has been carried out in the framework of the digitisation of a wooden maquettes collection stored in the 'Museo Egizio di Torino', using a photogrammetric approach, an automatic masking strategy using deep learning techniques is proposed, to increase the level of automatism and therefore, optimise the photogrammetric pipeline. Starting from a manually annotated dataset, a neural network was trained to automatically perform a semantic classification to isolate the maquettes from the background. The proposed methodology allowed the researchers to obtain automatically segmented masks with a high degree of accuracy. The workflow is described (as regards acquisition strategies, dataset processing, and neural network training). In addition, the accuracy of the results is evaluated and discussed. Finally, the researchers proposed the possibility of performing a multiclass segmentation on the digital images to recognise different object categories in the images, as well as to define a semantic hierarchy to perform automatic classification of different elements in the acquired images.

Keywords: close-range photogrammetry; deep learning; semantic segmentation; automatic masking; movable heritage; cultural heritage documentation

Resumen:

Los procesos de digitalización del patrimonio mueble son cada vez más populares a la hora de documentar las obras de arte almacenadas en nuestros museos. En los últimos años se han desarrollado un número creciente de estrategias de adquisición y modelado tridimensional (3D) de estos activos de valor incalculable, que responden de manera eficiente a esta necesidad de documentación, además de contribuir a profundizar en el conocimiento de las obras maestras estudiadas constantemente por investigadores que operan en muchos trabajos de campo. Hoy en día, una de las soluciones más efectivas está relacionada con el desarrollo de técnicas basadas en imágenes, generalmente conectadas a un enfoque fotogramétrico de estructura-y-movimiento (SfM). Sin embargo, si bien la adquisición de las imágenes es relativamente rápida, son los procesos relacionados con el procesamiento de los datos los que consumen mucho tiempo y requieren una participación manual sustancial del operador. El desarrollo de estrategias basadas en el aprendizaje profundo puede ser una solución eficaz para mejorar el nivel de automatismo. En la presente investigación, que se ha llevado a cabo en el marco de la digitalización de una colección de maquetas de madera almacenadas en el 'Museo Egizio di Torino' mediante un enfoque fotogramétrico, se propone una estrategia de enmascaramiento

*Corresponding author: Giacomo Patrucco, giacomo.patrucco@polito.it

automático mediante técnicas de aprendizaje profundo; el objetivo de ello es incrementar el nivel de automatismo y, por tanto, optimizar el flujo fotogramétrico. A partir de un conjunto de datos anotados manualmente, se ha entrenado una red neuronal que realiza automáticamente una clasificación semántica, aislando así las maquetas del fondo. La metodología propuesta ha permitido obtener más caras segmentadas automáticamente con alto grado de precisión. Se describe el flujo de trabajo desarrolado (en cuanto a estrategias de toma, procesamiento del conjuntos de datos y entrenamiento de las redes neuronales). Además, se evalúa y discute la exactitud de los resultados. Finalmente, se propone la posibilidad de realizar una segmentación multiclase sobre las imágenes digitales que permitan reconocer diferentes categorías de objetos en las imágenes y definir una jerarquía semántica que clasifique automáticamente diferentes elementos en la toma de las imágenes.

Palabras clave: fotogrametría de objeto cercano; aprendizaje profundo; segmentación semántica; enmascaramiento automático; patrimonio mueble; documentación del patrimonio cultural

1. Introduction

Movable heritage assets (or movable cultural property, as they were called in the proceedings of the UNESCO General Conference of 1978) (UNESCO, 1979) is rightly considered invaluable evidence and legacy of our past. In order to properly document these artworks, statues, or museum collections, it is extremely important for the researchers operating in the field of geomatics and, specifically, of cultural heritage documentation, to develop new strategies and tools able to effectively acquire and model the digital replicas of these cultural assets (Malik & Guidi, 2018). For this reason, the digitisation processes of these museum collections represent a very interesting challenge in terms of acquisition strategies, 3D modelling, achievable accuracy, and the usability of the obtained models. Nowadays, this last aspect is highly relevant, due to the versatility of the digital products that are possible to obtain thanks to the technological revolution we witness in recent years. Actually, 3D models of heritage assets are used for a large number of applications, from documentation (Chiabrando, Sammartano, Spanò, & Spreafico, 2019) to the management (for example, thanks to HBIM platforms, which in recent years have become very useful tools for architects, restorers, archaeologists, and several other experts) (Salvador-García, Viñals, & García-Valldecabres, 2020), dissemination and sharing through online 3D viewers (Minto & Remondino, 2014), and 3D print (Balletti & Ballarin, 2019; Balletti, Ballarin, & Guerra, 2017) aimed at improving the accessibility to the heritage to visually impaired people and for many other purposes.

Thus, the digitisation of movable heritage is a very actual and relevant topic that in recent years has been attentively investigated by the researchers involved in the fieldwork of cultural heritage 3D metric documentation, as evidenced by several experiences aimed at acquiring and model movable heritage assets (Giuffrida et al., 2019; Patrucco, Chiabrando, Dondi, & Malagodi, 2018) using strategies and techniques offered by the recent technological developments in the field of geomatics.

The need to properly document these assets is due to their intrinsic fragility (a common characteristic of all the heritage constantly exposed to many hazards that could cause the deterioration or destruction of these precious objects that preserve our history). On the other hand, it is necessary to provide the opportunity to contribute to the dissemination of these objects that are often stored in warehouses and, therefore, not freely accessible to the public. Nowadays, the opportunities provided by the improvements of the modern technologies allow new experiences mainly connected to 3D visualisation and virtual or augmented reality (Barbieri, Bruno, & Muzzupappa, 2018; Kersten, Tschirschwitz, & Deggim, 2017), which can contribute to enriching the traditional museum experience and the knowledge of the exhibits. Besides, as stated above, these 3D databases could represent a valuable tool for management and catalogue in the framework of museum administration.

Traditionally, digital photogrammetry has been highly effective for the generation of these kinds of 3D models (usually dense point clouds or complex 3D mesh). This approach is connected to image-based acquisition techniques and the processing of digital images using photogrammetric SfM-based (Structure-from-Motion) algorithms.

One of the advantages offered by this digitisation strategy is the possibility to achieve, besides an accurate and high-detailed 3D model, a high-resolution photographic texture that provides important information about the consistency of the object.

In the last few years, the use of an image-based methodology has allowed reaching remarkable results during different research experiences (Adami et al., 2015; Giuffrida et al., 2019; Guidi, Malik, Frischer, Barandoni, & Paolucci, 2018). However, many issues connected to a massive digitisation process should be considered, in particular regarding time consumption implied in the operations connected to the processing of the images. Despite the rapidity of acquisition reached by new image-based sensors, the time needed from the processing phases of the collected data is often remarkably high, and many procedures require a strong and sometimes challenging manual involvement by the operator.

Machine learning can represent a promising solution to solve this problem. In particular, deep learning appeared to be effective in different signal processing tasks, such as image recognition, semantic segmentation, audio processing, text processing, etc. (Goodfellow, Bengio, & Courville, 2016). The availability of a large amount of data and the increasing computing power of new machines allow the training of neural network models that can automatically perform many tasks with a high degree of accuracy.

Semantic segmentation is one of the most known cases of study in the deep learning field, as neural networks have been proved to outperform traditional methods (Garcia-Garcia et al., 2018). However, training a neural network for semantic segmentation requires an initial effort related to data collection and labelling. Like other supervised learning algorithms, it is necessary to acquire a large amount of data and perform manual annotation. This process represents a relevant issue for many

researchers who do not have the resources to build an adequate dataset. For this reason, different techniques for data augmentation are often studied and applied (Shorten & Khoshgoftaar, 2019).

1.1. Deep learning and cultural heritage

Convolutional neural networks (CNN) represented state of the art in the context of image processing. They proved to outperform traditional methods in different contexts (image classification, object detection, semantic segmentation, etc.) (Gu et al., 2018).

Traditional machine learning methods rely on the manual definition of visual features that are extracted from images. These features are particularly meaningful for humans, but they are not guaranteed to be the optimal representation for a particular task. Instead, neural networks learn to extract representative features in an iterative process and use them to solve the task they are designed for.

In the field of image processing, the most critical operation that neural networks perform on images is convolution. With the convolution, the image is filtered using a kernel (or filter) of a specific size and shape, and a new image is produced. In a particular layer of a CNN, several kernels are used to process the image in different ways. The outputs of the kernels are stacked and fed to the next layer, where other convolutions are applied. Multiple layers allow modelling features with a high level of abstraction.

Depending on how the kernels are designed, the output of each convolution is different. During the training phase, the kernels are continuously modified to minimise a cost function that has to be defined appropriately. In other words, kernels are trainable parameters of a neural network (Goodfellow et al., 2016).

Training neural networks require the availability of a labelled dataset. In the case of semantic segmentation, every image of the dataset must be provided together with the segmentation mask that assigns each pixel to a specific class. A well-known problem is the size of the dataset. Many deep learning models require a large number of images to be properly trained. This is particularly true with large models where the number of parameters (kernels) is very high.

As underlined in the following sections, this problem can be overcome with an accurate design of the neural network architecture and an augmentation procedure.

The use of deep learning-based techniques in different frameworks is significantly promising. It has been explored in a large number of disciplines and working scenarios (Vargas, Mosavi, & Ruiz, 2018). The versatility and potentialities of a fully automated approach are undeniably intriguing. It should be underlined that the possibility of automating a series of processes, which are often highly onerous and time-consuming, represents a very interesting opportunity that requires to be explored.

Of course, in the field of heritage documentation, conservation, and valorisation, there is a particular focus on developing these increasingly efficient solutions, as evidenced by numerous studies and research carried out in the last few years (Fiorucci et al., 2020; Felicetti, Paolanti, Zingaretti, Pierdicca & Malinverni, 2020). In fact, recently, deep-learning strategies have been often applied by researchers operating in the fieldwork of geomatics and heritage documentation, for example, to automatically recognise pre-established elements in digital images (Condorelli, Rinaudo, Salvadore, & Tagliaventi, 2020) or to perform classification of architectural heritage (Llamas, Lerones, Medina, Zalama, & Gómez-García-Bermejo, 2017).

Another most investigated topic regards the possibility of performing semantic segmentation on different types of data or metric products. In most cases, the main aim is to carry out an automatic clustering or regrouping of the considered elements in classes characterised by certain common characteristics. More and more frequently, the researchers are exploring these strategies and establishing new methodologies to perform a semantic classification both on images (Knyaz, Kniaz, Remondino, Zheltov, & Gruen, 2020; Stathopoulou & Remondino, 2019) and 3D data, such as 3D models (George, Xie, & Tam, 2018) or point clouds (Grilli, Farella, Torresani, & Remondino, 2019; Grilli, Özdemir, & Remondino, 2019; Pierdicca et al., 2020).

In the research presented in the current paper, a deep learning-based approach for the automatic semantic segmentation of movable heritage images is proposed (Fig. 1).

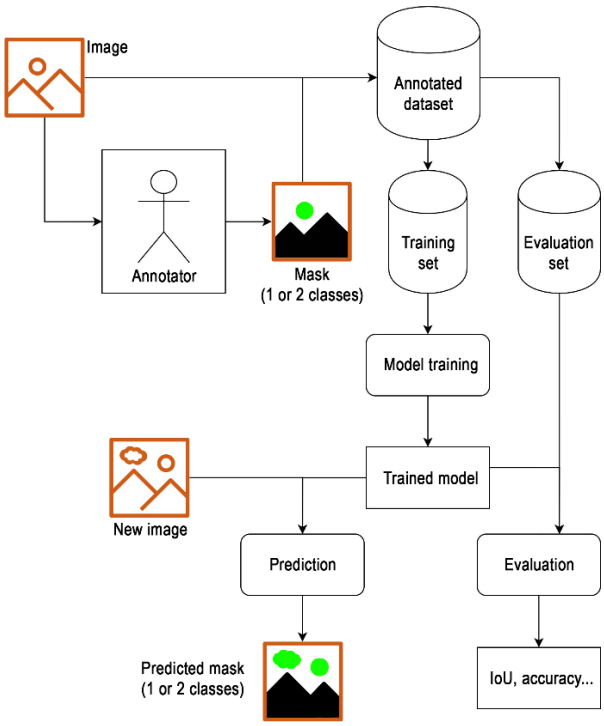


Figure 1: Flowchart of the deep learning approach followed during the current research.

The study started from a manually annotated dataset (composed of almost 3000 high-resolution images) and applied augmentation techniques to increase the diversity and, therefore, the efficiency of the subsequent training. It trained a neural network starting from the labelled images to perform the masking task considering two different cases: when just the main object of interest is segmented and when other objects in the frames are

included and treated as a different class. In this way, it is possible to observe that with proper training, using a deep learning approach, it is possible to recognise objects belonging to different categories and perform automatic semantic segmentation of the considered images, possibly also for 3D reconstruction and classification purposes, by defining a semantic hierarchy between them classes.

2. Digitisation of a collection of wooden maquettes: An image-based approach

The acquisition of the primary data used during this experience has been performed in the framework of 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) research project of Polytechnic University of Turin (Lo Turco, Piumatti, Rinaudo, Tamborrino & González-Aguilera, 2018).

One of the aims of the research (started in 2017) was the digital reconstruction of the collection of 'Expedition models of Egyptian architectures' (Mafrici & Giovannini, 2020) (a collection of wooden maquettes of the beginning of the 19th century, composed of a total of 26 pieces representing ancient Nubian temples) for a subsequent enrichment process of information content and online sharing for dissemination purposes.

The different issues regarding the 3D modelling of the maquettes have been considered, and the wooden models have been acquired using several approaches and different sensors (Patrucco, Rinaudo, & Spreafico, 2019; Turco et al., 2018).

Among the various investigated strategies, the approach that has turned out to be adequate for the final goals of the research project was the one connected to the use of digital photogrammetry. In fact, thanks to the modern computer vision and image matching techniques and algorithms, nowadays it is possible to achieve not only a highly detailed point cloud (from which the user can triangulate a complex 3D mesh characterised by a very high spatial resolution, level of detail and accuracy) but also a high-resolution photographic texture, which provides high radiometric contents (obviously if the acquisition has been performed following the well-known photogrammetric criteria, and the collected digital images are characterised by the correct level of illumination, sharpness, and depth of field, without blur or shadows) (Dall'Asta, Bruno, Bigliardi, Zerbi, & Roncella, 2016). As it is well known, the texture that is possible to generate using photogrammetry can provide valuable information about the material, consistency, and conservation status of a digitally reconstructed object. For this reason, an image-based approach has been considered more appropriate for the aims of the project.

2.1. Acquisition of the digital images

The digital images of the maquettes have been collected during several acquisition campaigns carried out in 2018 and 2019. The DLSR camera used during the survey is a Canon EOS 5DSR equipped with a Zeiss 50 mm macro lens. The main specifications of this sensor are reported in Table 1.

Table 1: Main specifications of the used camera.

Model	Canon EOS 5DSR
Sensor	CMOS 50.3 [Mpx]
Sensor size	36 x 24 [mm]
Image size	8688 x 5792 [px]
Lens	Zeiss Milvus 50mm f/2M
Focal lenght	50 [mm]

During the acquisition, a tripod and a remote release device have been employed (to improve the camera's stability while shooting), and the maquettes have been artificially enlightened using two led panels equipped with diffusers (to provide homogeneous illumination and avoid shadows). In addition, three or four scale bars have been positioned alongside the maquettes to provide a metric reference and, therefore, to properly scale the 3D model during the data processing (Fig. 2).

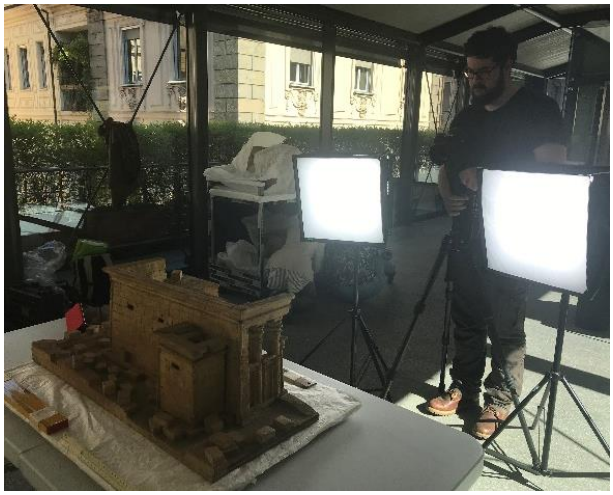
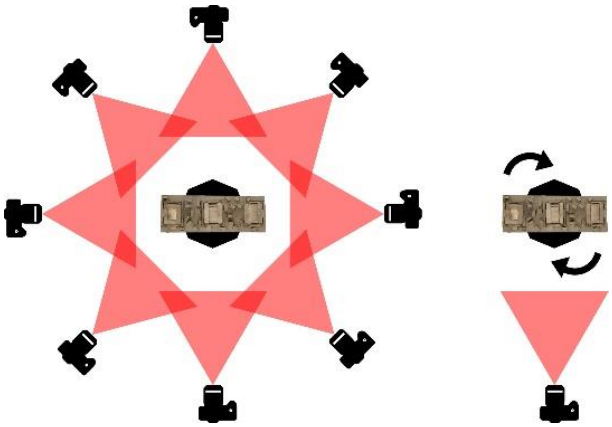


Figure 2: Acquisition stage.

One of the issues that have been considered in this phase regards the optimisation of the acquisition procedures both in terms of rapidity and efficiency. Quite frequently, it is not possible to bring the artworks and the museums' assets in dedicated spaces for acquisition placed outside the museum buildings, both for logistic reasons (for example, the impossibility to move the artefacts) (Dall'Asta et al., 2016) and security reasons. Therefore, often the survey is carried out in spaces where the difficulties in organising a proper acquisition set are high (because of the presence of visitors and museum operators, the suboptimal illumination –which almost always requires the use of additional artificial lighting systems to avoid opening the aperture of the camera and increasing the camera's ISO– or still the limited size of the spaces).

Usually, as regards these kinds of surveys, the operator moves the camera around the object, acquiring convergent and overlapping images. However, as described in Patrucco et al., 2019, in this case, each model (and, of course, also the scale bars) have been positioned on a rotating platform to optimise the time required from the repositioning of the tripod and the focus procedures (Fig. 3). In this way, it was possible to

significantly reduce the time required to acquire the images and, therefore, increase the survey's rapidity. Following this strategy, the acquisition of the images required approximately 30 minutes for each maquette.



(a) (b)

Figure 3: a) Traditional multi-view photogrammetric acquisition scheme; b) Acquisition carried out with a fixed camera and a rotating platform (solution adopted in the current research).

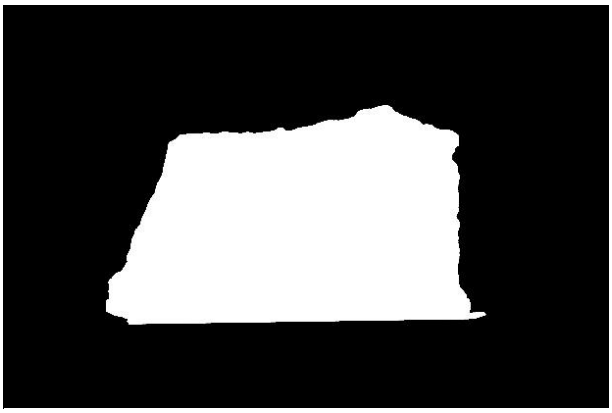
For each maquette, several circular acquisitions of images with different camera assets and orientation have been collected to cover the higher number of surfaces (avoiding as many occlusions caused by recesses as possible) to ensure a proper digital reconstruction of the geometries of the wooden models.

The images have been processed using the well-known SfM-based software Agisoft Metashape v. 1.5. A standard pipeline has been applied (orientation of the images; tie-points extraction; dense cloud generation; 3D mesh generation) but, in this case, it was preliminarily necessary to apply exclusion masks (generated by the operator with manual or semi-automatic procedures) to all the processed images (Fig. 4).

In fact, following the aforementioned strategy, the acquisition phase finishes fast, but, as a consequence, the data processing procedures require more time since it is necessary to apply the exclusion masks to remove the background (otherwise, it would not be possible to perform a proper tie-points extraction and the relative orientation of the frames, because of the movement of the rotating platform from the background). In this way, at the end of this procedure, the area of the digital images related to the maquettes (and to the metric bars) have been manually isolated from the background.

Collectively, to acquire all the 26 pieces of the Nubian Temples maquettes collection, a total of 2908 images have been collected. Considering that the time required to generate the exclusion mask is approximately 1-2 minutes, the considerable investment of time required from the operator is underlined. In the case of the current research, as described in Section 3.3, 90% of the 2908 masks have been used as primary data for the neural network training while the remaining 10% has been used for quality control. As a result of this operation, the frames have been

(a)

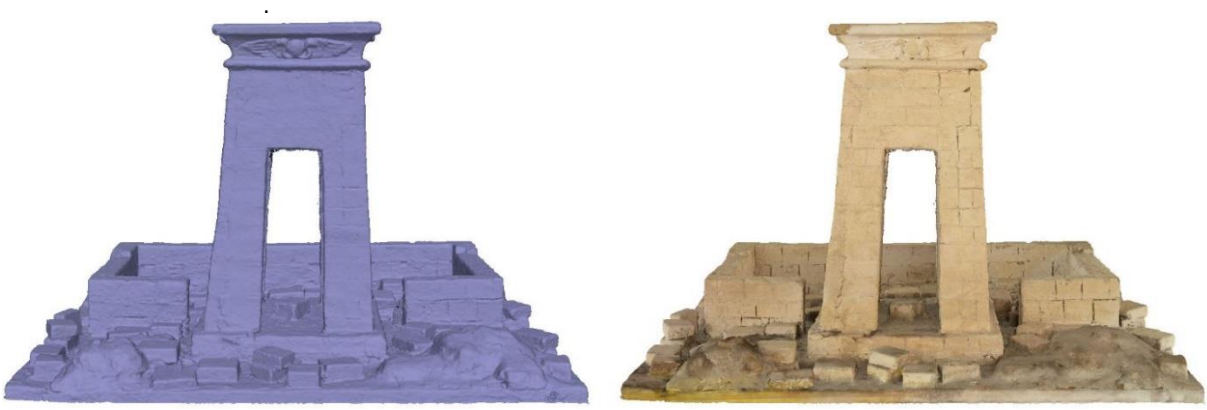


(b)

Figure 4: Data acquisition and preprocessing: a) Example of one of the wooden maquettes acquired; b) Corresponding exclusion mask (generated with manual and semi-manual procedures).

successfully oriented, and a sparse cloud of tie points has been extracted (Fig. 5). A dense point cloud has been generated for each dataset, and then a texturised 3D mesh has been triangulated (Fig. 6).

Figure 5: Image orientation and extraction of a sparse cloud of tie-points.



(a) (b)

Figure 6: Photogrammetric outputs: a) 3D model of one of the modelled wooden maquettes (triangular mesh generated from the photogrammetric dense point cloud); and b) Photorealistic 3D model.

3. Deep learning-based approach for semantic segmentation

In this section, a deep learning-based method is illustrated to address the aforementioned problem of generating exclusion masks for the acquired images. A neural network has been trained to perform the segmentation of the images and obtain an exclusion mask (Fig. 7). This method automatically performs a time-consuming task in a short time with a high degree of accuracy.

As mentioned in the previous sections, both a 1-class scenario and a 2-classes scenario have been considered. In the first case, the network is trained to segment only the object of interest (the maquette placed in the centre of the frame), referred to as Class 1. In the second case, the network is trained to segment both the object of interest and the metric bars (used as reference objects during the photogrammetric process) included in the images and referred to as Class 2. An evaluation of the network performance is proposed for both the observed cases.

3.1. Dataset annotation and processing

Ground truth is needed to train the segmentation models. For each image, ground truth is given by an image of the same size where each pixel has a value corresponding to the class it belongs to (e.g., in the current paper, 0 is used for background, 1 for Class 1, and 2 for Class 2).

The ground truth used during this research experience has been derived by the exclusion masks that have been mentioned in Section 2. Agisoft Metashape allows exporting the manually segmented masks in PNG format. In this case, a script has been written to convert the PNG masks into a properly formatted ground truth.

The object of interest (Class 1) and the metric bars (Class 2) was segmented into different classes. Class 2 was then ignored in the one-class segmentation training.

The dataset included a total number of 2908 images.

In the context of semantic segmentation, it is crucial to ensure that the classes in the dataset are balanced (Vargas et al., 2018). In this case, the objects belonging to Class 1 were contained in every image, while objects belonging to Class 2 were present in approximately 90% of the images. As far as the total number of pixels is concerned, the average percentage of pixels in the images is 65% for background, 30% for Class 1, and 5% for Class 2. However, as the results presented in Section 3.6 indicate, this imbalance between Class 1 and Class 2 is not high enough to prevent the neural network from being trained properly.

Typically, neural networks require a large number of images to be trained. This is particularly true with semantic segmentation, where many of the network architectures are very large. In this case, a network with a small number of parameters is employed, and the segmentation task includes only 1/2 classes. With a proper augmentation design for the dataset, these conditions can lead to an accurate model, even if the number of images is small (Shorten & Khoshgoftaar, 2019) .

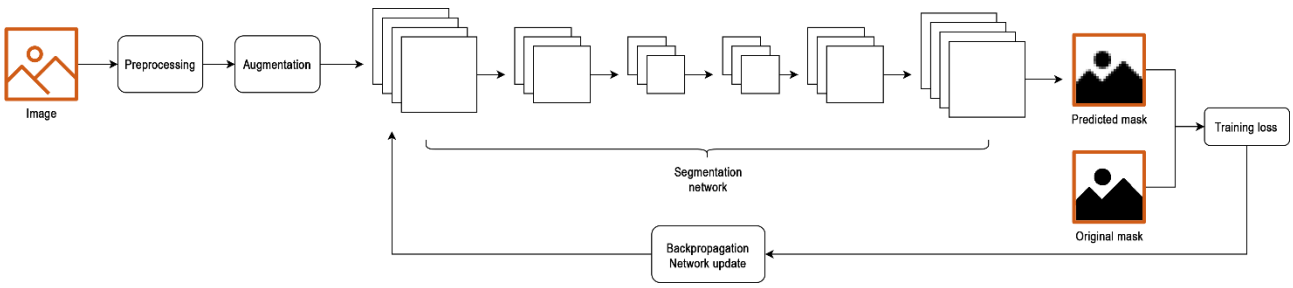


Figure 7: Flowchart representing the workflow applied during the network training phase.

Augmentation refers to dynamically modifying the input images during the training in such a way that the diversity of the training data is wider. Traditional methods for image augmentation are as follows: random rotation, random cropping, horizontal and vertical flipping, random brightness change, and random occlusions. In this particular case, the main aim was to isolate the maquette from the background. Since the background content is not relevant, and the network is not supposed to learn features in the background, another augmentation step was added. For each image fed to the network during the training phase, the background is randomly modified or substituted with another one. This process highly enhances the diversity of the training data. An example of the augmentation techniques that have been used is depicted in Figure 8.

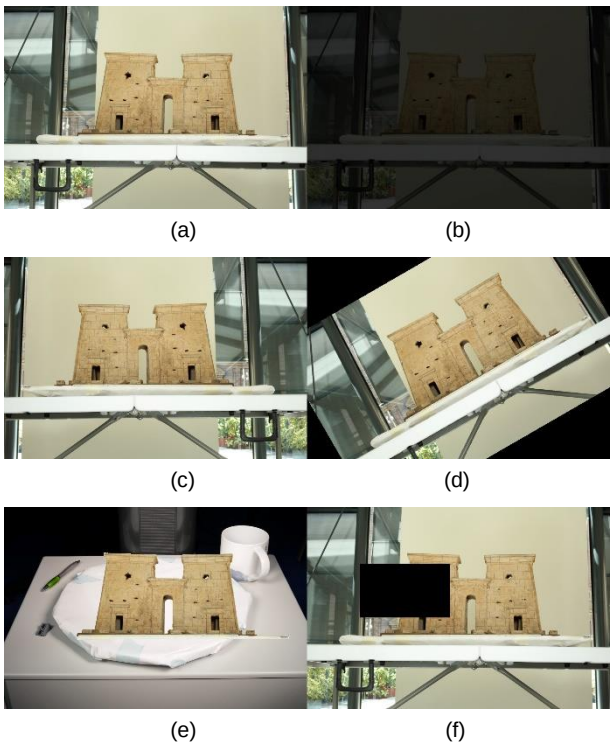


Figure 8: Examples of augmentation strategies followed during the presented research: a) original image, b) random brightness change, c) horizontal flip, d) random rotation, e) random background change, f) random occlusion. When the coordinates of the object of interest are modified, the occlusion mask is modified accordingly.

The original images are characterised by a very high resolution (8688x5792 pixels). Although the use of high-resolution digital images has become a standard in the framework of digital photogrammetric reconstruction (to achieve very detailed 3D results), this represents a potential issue when training a deep learning model. The number of parameters that have to be processed depends on the network architecture and the input size. The amount of data produced by each layer of the network represents another determinant of the number of parameters. If the size is too large, out-of-memory errors may occur during the training phase. For this reason, the images were resized to a resolution of 512x512 pixels to be successfully processed. After the segmentation process, the output masks were subjected to a post-processing step to go back to the original size (this procedure is further explained in Section 3.5).

3.2. Neural network architecture

The deep learning research community is particularly active in semantic segmentation, which represents one of the most important topics of research.

Many approaches have been undertaken, but in general, a semantic segmentation network is composed of two parts: an encoder, which takes an input image and extracts relevant features from it, and a decoder, which takes the extracted features and builds an output mask. The encoder and the decoder can be designed using different architectures.

One of the most known models is SegNet, which was presented in 2015 and proved to outperform previous methods for semantic segmentation (Badrinarayanan, Kendall, & Cipolla, 2017).

In 2017, the Fully-Convolutional DenseNet (FC-DenseNet) was presented as a novel network for semantic segmentation (Jegou, Drozdzal, Vazquez, Romero, & Bengio, 2017). FC-DenseNet is based on the DenseNet architecture, originally designed for image classification. DenseNet is innovative because each layer of the network receives information from all the previous layers. This network allows using fewer channels for each layer, thus having fewer training parameters and a smaller network. FC-DenseNet leverages this idea and applies it to a convolutional segmentation network. The encoder part is based on the DenseNet architecture, while the decoder reconstructs the output image with a series of convolutions.

The high accuracy, the small number of parameters, and the possibility of further reducing the network size lead to the choice of FC-DenseNet as the reference model for this task. Indeed, a small training dataset having too many parameters would result in overfitting. This issue has to be avoided.

3.3. Neural network training and validation

The training workflow was the same for both cases (one class segmentation and two classes segmentation).

The dataset was divided into a training set (90%) and a validation set (10%). The images belonging to the training set are the only ones used during the training, while the validation set is used to measure the model's performance during the training.

The network was trained using the RMSProp (Root Mean Square Propagation) optimisation algorithm (Yazan & Talu, 2017), with a learning rate of 0.0001 and a decay of 0.995. These parameters regulate the variation speed of the network parameters during the training. The loss function to be minimised by the optimiser was the mean cross-entropy. For a single pixel, the cross-entropy is defined as

$$L_{CE} = -\sum_{i=1}^{n} t_i \, log(p_i)$$

where n is the number of classes, t is the true labels, and p is the class probabilities predicted by the network. The cross-entropy is computed for each pixel and then averaged over the whole image, obtaining the mean

cross-entropy. A perfect model has a mean cross-entropy loss of 0.

Before each training iteration, the training set was shuffled. The accuracy and the Intersection over Union were monitored during the training, both for the training set and the validation set. The details of these measures are explained in Section 3.5.

The training was stopped when these measures stopped growing and reached a plateau. This criterion is known as Early Stopping, and it is useful to prevent overfitting and maintain a good generalisation capability (Caruana, Lawrence, & Giles, 2000). In this case, the training stopped after 8 epochs (i.e., 8 iterations of the whole training set).

So far as the time required (both for the training and for the generation of the segmented images) is concerned, the training took approximately 10 h, while the inference time (for each image) was estimated as less than 1 s. Of course, the time required for both the procedures varies depending on the performance of the employed workstation.

Two commonly used measures were employed to evaluate the performance of the network. The first one was the global Accuracy (A), which measured the percentage of pixels in the image that has been correctly classified by the model. The second one was the Intersection over Union (IoU), which was computed by taking the intersection between the predicted mask and the original one and dividing it by the union between them. In the ideal case, the IoU should be 1. The IoU is also known as the Jaccard Index.

The accuracy and the IoU provide a quantitative measure of the model's performance. For qualitative analysis, a 3D reconstruction of a model is presented in the next sections, using the automatically segmented images as exclusion masks.

3.4. Post-processing

A post-processing stage has been designed to scale the output masks back to the original image size. This stage has the purpose of smoothing the generated mask and removing the noise produced by the segmentation process, merging small, isolated clusters of pixels, and obtaining a uniform masking.

The post-processing includes the following steps:

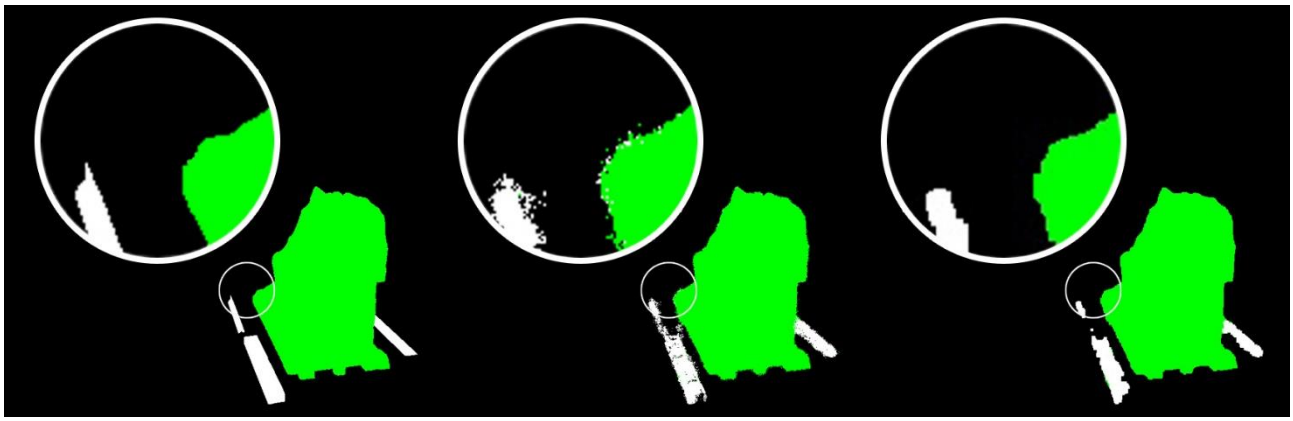
- Morphological opening, an operation that opens up a gap between objects connected by a thin bridge of pixels and eliminates small noisy or misclassified clusters.
- Morphological closing fills small holes in the regions while keeping the initial region size.
- Resizing the mask to the original size.
- Median filtering with a kernel of 9 pixels to depixelate the resized image and obtain a smoothed version.

It is possible to observe an example of the designed post-processing procedure –which has been applied to the entire processed dataset after the automatic generation (Figure 9).

3.5. Results

The performance of the trained models is evaluated for the validation set, which includes images that have never been used during the training phase. In addition to the A and the mean IoU used to monitor the performance during the training, the F1-score is evaluated, as suggested in Badrinarayanan et al. (2017), since it provides a better quality measure than IoU for this kind of task. In the context of semantic segmentation, the F1-score is defined as two times the overlap between the predicted mask and the original one, divided by the total number of pixels.

The results can be observed in Table 2, both for the case with one class and two classes. Regarding the classification capability, the model with two classes (the maquette and the metric bars) performs better than the one with one class. This may seem surprising as the segmentation task is more complex. However, the strong contrast between the metric bars and the background and the regular shape of the metric bars makes it easy for the model to identify the objects that belong to that category.



(a) (b) (c)

Figure 9: Post-processing procedures: a) Original mask (Class 1 in green, Class 2 in white); b) Automatically generated mask; c) Generated mask after morphological operations. These operations fill the small holes produced by the segmentation, smoothen the shapes' edges and remove misclassified pixels.

Table 2: Comparison of Accuracy, Mean IoU, and F1-score between 1 Class training and 2 Classes training.

	1 Class	*2 Classes*
Accuracy	92%	97%
Mean IoU	85%	80%
F1 score	92%	97%

The IoU scores show a different trend, being lower in the case of the two classes. This could be explained by the fact that in this case, the generated masks exhibit more jagged contours, resulting in a lower IoU score. However, this problem can be efficiently addressed with the post-processing step that we introduced in Section 3.4.

Finally, to perform a qualitative analysis on the usability of these results in the framework of a 3D reconstruction and to verify the suitability of the segmented images as exclusion masks during the photogrammetric process, the 148 images of the Model of the Portals of the Temple of Debod (not belonging to the training set) have been reprocessed (Fig. 10) using the automatically generated masks.

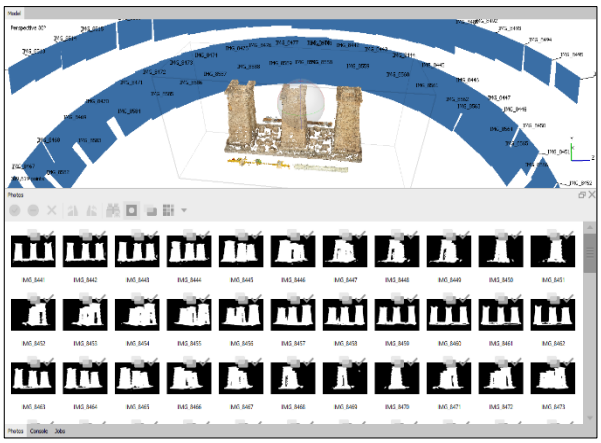
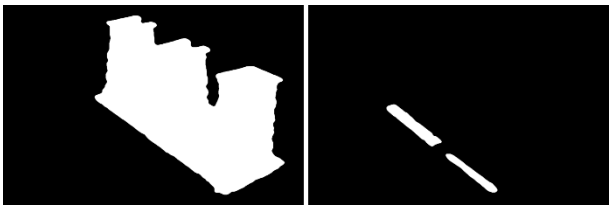


Figure 10: 3D photogrammetric reconstruction of the geometry of the maquette. In this case, the automatic exclusion masks have been imported in Metashape and used for the 3D reconstruction.

The same standard workflow reported in Section 2.1 has been followed, and the exact parameters of the previous processing have been used. The obtained 3D products are coherent with the results achieved using the manually generated masks instead of the automatic ones (in terms of number of extracted tie points, number of points of the dense point cloud, level of detail, noise, and the resolution of the generated 3D mesh).

In addition, another interesting aspect that should be underlined regards the possibility to label the obtained 3D point cloud according to the class segmented by the neural network on the images to perform the same semantic segmentation observed on 2D data also on the 3D products such as point clouds.

In this case, after the relative orientation of the images and the tie-points extraction, the dense point cloud has been generated firstly using masks with only the elements labelled as Class 1. In contrast, the second time, the point cloud has been generated using the masks with only Class 2 elements (Fig. 11).



(a) (b)

Figure 11: Automatic multiclass masks: a) Class 1; b) Class 2.

Following this strategy, it is possible to generate distinct point clouds for the different elements of each class (wooden model and metric bars). In both cases, outliers not belonging to the predicted class have been generated. In the case of the Class 1 point cloud, the unrelated outliers correspond to 2% of the total points, and therefore, they can be considered insignificant. However, in the Class 2 point cloud, a higher number of points has been erroneously generated (both noise and unrelated outliers), corresponding to approximately 15% of the total points. Nevertheless, considering that it was the first attempt of 3D classification following this strategy, the results have been considered adequate to pursue the next part of the research (the generation of a merged point cloud, containing – embedded in each point – the information of the original class). Therefore, before merging the two different point clouds, an additional scalar field (CLASS) has been created, and the point clouds labelled with values 1 (for the points of the maquette) and 2 (for the points of the metric bars). The open-source software CloudCompare v. 2.11 alpha Anoia) has been used to complete this procedure. As it is possible to observe in Figure 12, the embedded information related to the semantic classification of the different elements is stored in the same point cloud –created by the merging of the previous two– in addition to the original RGB value.

4. Discussion

As evidenced in the previous sections, in this work, the problem of manually generating exclusion masks has been addressed with the implementation of a deep learning model –trained starting from the images of the maquettes and the corresponding labelled masks, following the strategies described in the previous sections– that performs semantic segmentation on movable heritage images, providing masks both for the object of interest and, optionally, elements from different categories (in this case, the metric bars).

Considering the low inference time (<1 s) for each mask's generation and the high accuracy in elements recognition achieved after the neural network training, the presented methodology represents a very effective solution to generate the exclusion masks automatically and classify the elements belonging to different classes. In the framework of the debate connected to the automatisation of the heritage documentation processes (specifically, focusing on time-consuming and repetitive procedures), the pipeline described in this paper can be exploited.

One of the aspects that still remains open for further research is the generalisation capability of the model, currently limited to recognising the wooden maquettes (or similar objects) and not extensible to other assets

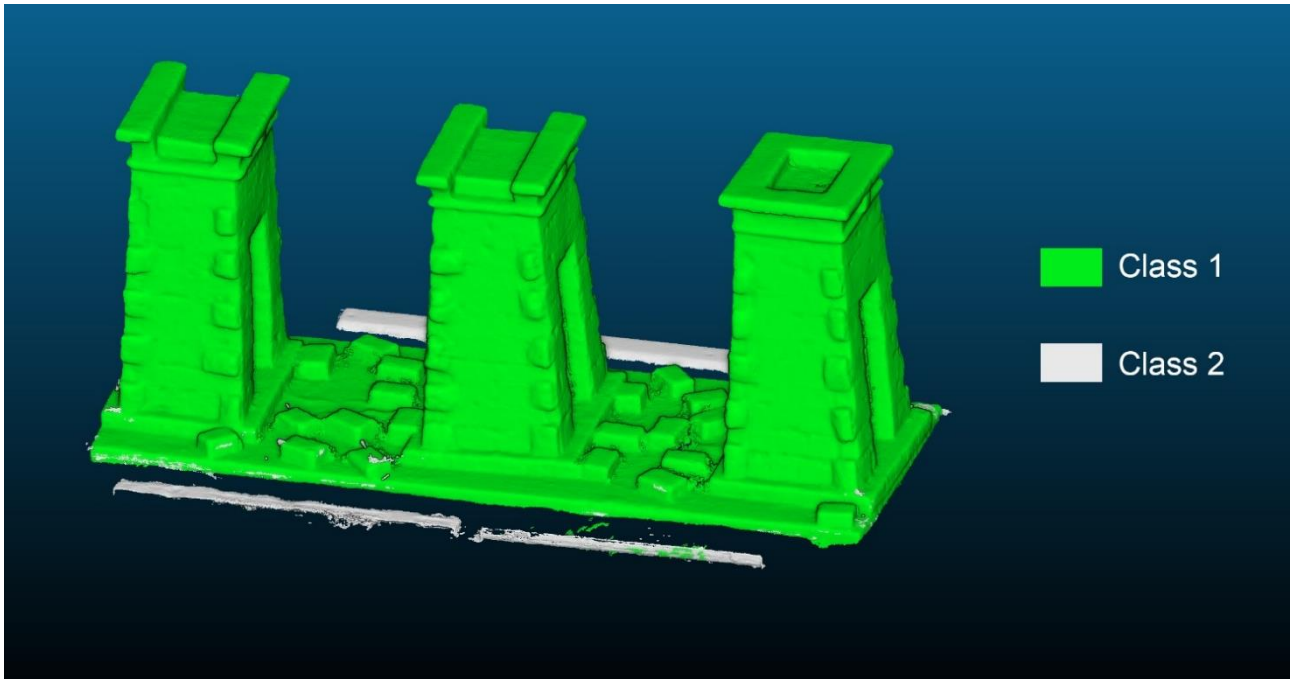


Figure 12: 3D segmented point cloud. A new scalar field (CLASS) has been created, and the corresponding value (1 or 2) has been labelled for the points of each class. In this case, the green points correspond to Class 1 (the maquette) and the white points to Class 2 (the metric bars).

characterised by a less homogeneous appearance. The acquisition procedure described in Section 2 can be exploited to increase the generalisation capability of the model. When multiple images are acquired at slightly different angles (as done during the survey of movable heritage, where convergent frames are collected following photogrammetric criteria to perform a 3D reconstruction), and when the rotating platform strategy is employed, the change in the foreground (i.e., the object to be segmented) is higher than the change in the background. If subsequent images are combined, this difference can be used to distinguish the object of interest from the background, regardless of the homogeneity between the acquired assets. For this reason, new techniques will be investigated by the authors to produce exclusion masks starting from a sequence of images (and therefore, not just considering each image one by one). In this way, the neural network will rely not only on the nature of the object to be segmented but also on the clues given by the background and the motion of the object and its position inside the image.

Regarding the further application of the described strategy, the proposed approach shows that the object of interest can be separated from the background with an adequate degree of accuracy (Table 2). It should be underlined that the strategy which has been described in the paper is connected to a specific case (the digitisation of movable heritage and, specifically, the automatic recognition and segmentation of the observed wooden models' collection – case study of this research). However, the described methodology is extremely flexible and characterised by many potentialities. It is not just connected to movable heritage. For example, this methodology can be successfully applied to built heritage documentation starting from an adequate dataset. In connection to this, it would be possible to automatically segment the different elements (in the case of a heritage building: walls, windows, roofs, columns, decorative elements, etc.) included in the object of interest (Stathopoulou & Remondino, 2019). This last application is potentially very interesting, in particular considering the recent developments carried out in the framework of parametric modelling and HBIM, where the need for new efficient and effective strategies for smart management of information (both in terms of geometric and semantic contents) represents a critical but still open issue.

As regards the achieved accuracies, it should be underlined that despite the good performance metrics that have been obtained in this work, there is a margin for improvement. Deep learning-based semantic segmentation is an active field of research and new techniques are constantly being studied in order to improve the performances.

For example, the authors in Cermelli, Mancini, Bulo, Ricci & Caputo (2020) introduce a technique to better model the background and separate it from the object of interest, even when a new training set is introduced with new classes. This could be beneficial from a generalisation point of view.

Relying on a well-defined and tested neural network architecture is a common practice in deep learning, as it allows to train of a model that has been extensively tested and benchmarked. However, new approaches are focused on automatically obtaining a network architecture that is optimized for the specific case of interest. In Lin, Sun, Cheng, Xie, Li, & Shi (2020), this approach is studied in the case of semantic segmentation, obtaining good results in terms of both accuracy and inference time.

In He, Shen, Tian, Gong, Sun, & Yan (2019), a knowledge distillation method is proposed in order to increase the computational efficiency of the

segmentation network without losing accuracy. This would allow to performing segmentation on larger images, obtaining a more accurate result in the original size mask.

In the light of these considerations, it should be underlined that the constant development of new strategies and techniques represents an evident opportunity to improve the performances of the trained models and, therefore, the overall quality of the obtainable results (especially in terms of achievable accuracies).

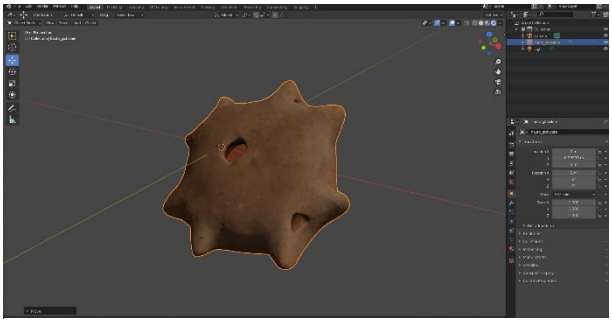
5. Conclusions and future perspectives

From all these considerations, it is possible to underline the extremely high potential of deep-learning-based techniques in the image-based heritage documentation field, where the development of new flexible strategies and solutions is always required. In this context, this work represents a basis for our future research, where segmentation will be more complex (e.g., very different backgrounds, heterogeneous objects, etc.), and deep learning is a very effective solution. The standard computer vision procedures are based on a human-designed processing workflow requiring some degree of a priori-knowledge of the image content, but when (as in this case) the main aim is to obtain a model capable of performing segmentation on images with different content in comparison to the images used during the training, the use of deep learning techniques represents an extremely powerful and effective solution. Moreover, deep-learning-based techniques have become a standard in several image processing fields. This motivates our choice to use this strategy.

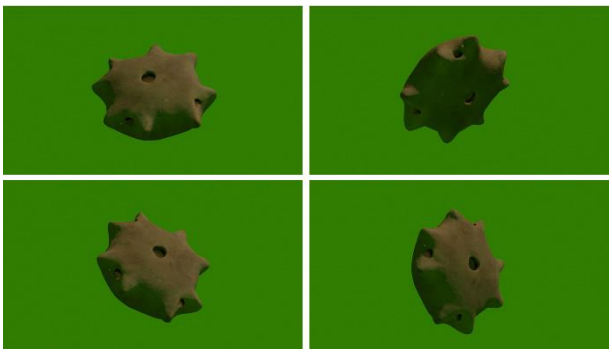
However, one of the main critical issues in this framework is strictly connected to the acquisition and processing of the primary data for neural network training. The labelling procedures are often manually performed. The sourcing of a dataset is composed of an adequate number of images. Since the availability of a suitable dataset of annotated images is a common issue in this field, it is necessary to improve and perform new augmentation. In addition, the authors are investigating the possibility of integrating the training dataset with ‘artificial’ images automatically generated in a virtual environment starting from photogrammetric 3D models. A first test has been carried out using the free and open modelling software Blender to automatically generate digital images starting from a reality-based model achieved with an image-based approach (Fig. 13).

In addition to the possibility to acquire frames of the model at different angles, the choice of software is due to the opportunity to modify the intensity, the colour, and the direction of the virtual light to increase the effectiveness of the adopted augmentation strategies. Furthermore, increasing the contrast between the 3D model and the virtual background, this approach automatically generates the masks for the neural network training without manual work.

Lastly, a final reflection focuses on the opportunity to increasingly automatise most of the processes (not only in the framework of heritage or geomatics but also in a very high number of fields and disciplines)



(a)



(b)

Figure 13: Automatic generation of ‘artificial’ images with an automatic approach starting from a photogrammetric 3D model: a) Graphical user interface of Blender platform; b) Examples of automatically generated digital images.

and the role of the operator in these kinds of procedures. Despite a generalised distrust in automatic procedures, mainly due to the low degree of control and the lack of tools for accuracy check observed in some software or instruments that make massive use of automatism –more or less connected to artificial intelligence– during the current research, it was possible to remark the utility of a deep-learning approach to developing a smart and time-saving workflow, and therefore, significantly decrease the manual workload required from the operator. However, it should be underlined that the role of the operator (from whom an increasingly high level of competence is required) remains central and irreplaceable in terms of data managing, accuracy checking (during these kinds of operations, close monitoring of the various phases of all procedures is almost mandatory), and, equally important, the interpretation of the obtained results.

Acknowledgements

The authors thank Volta® A.I. (and in particular Silvio Revelli) for the contribution to this work and for providing high-end hardware for neural network training.

In addition, they would like to thank Alessia Fassone of Museo Egizio di Torino and all the people involved in the B.A.C.K. TO T.H.E. F.U.T.U.RE. project (in particular, Fulvio Rinaudo, who coordinated the Geomatic team).

Finally, they wish to express their gratitude to Nannina Spanò and Filiberto Chiabrando for the helpful confrontation during the presented research.

References

Adami, A., Balletti, C., Fassi, F., Fregonese, L., Guerra, F., Taffurelli, L., Vernier, P. (2015). The bust of Francesco II Gonzaga: From digital documentation to 3D printing. *ISPRS Annals of the Photogrammetry, Remote Sensing and Spatial Information Sciences*, II-5/W3, 9–15. https://doi.org/10.5194/isprsannals-II-5-W3-9-2015

Badrinarayanan, V., Kendall, A., Cipolla, R. (2017). Segnet: A deep convolutional encoder-decoder architecture for image segmentation. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, *39*(12), 2481–2495. https://doi.org/10.1109/TPAMI.2016.2644615

Balletti, C., Ballarin, M., & Guerra, F. (2017). 3D printing: state of the art and future perspectives. *Journal of Cultural Heritage*, *26*,172–182. https://doi.org/10.1016/j.culher.2017.02.010

Balletti, C., & Ballarin, M. (2019). An application of integrated 3D technologies for replicas in Cultural Heritage. *International Journal of Geo-Information*, *8*(6), 285. https://doi.org/10.3390/ijgi8060285

Barbieri, L., Bruno, F., & Muzzupappa, M. (2018). User-centered design of a virtual reality exhibit for archaeological museums. *International Journal on Interactive Design and Manufacturing (IJIDeM)*, *12,* 561–571. https://doi.org/10.1007/s12008-017-0414-z

Caruana, R., Lawrence, S., & Giles, C. L. (2001). Overfitting in neural nets: Backpropagation, conjugate gradient, and early stopping. *Advances in Neural Information Processing Systems* (pp. 402–408).

Cermelli, F., Mancini, M., Bulo, S. R., Ricci, E., & Caputo, B. (2020). Modeling the background for incremental learning in semantic segmentation. *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 9233–9242. https://doi.org/10.1109/CVPR42600.2020.00925

Condorelli, F., Rinaudo, F., Salvadore, F., & Tagliaventi, S. (2020). A neural network approach to detecting lost heritage in historical video. *International Journal of Geo-Information, 9*(5), 297. https://doi.org/10.3390/ijgi9050297

Chiabrando, F., Sammartano, G., Spanò, A., & Spreafico, A. (2019). Hybrid 3D models: When Geomatics innovations meet extensive built heritage complexes. *International Journal of Geo-Information*, *8*(3), 124. https://doi.org/10.3390/ijgi8030124

Dall'Asta, E., Bruno, N., Bigliardi, G., Zerbi, A., & Roncella, R. (2016). Photogrammetric techniques for promotion of archaeological Heritage: the Archaeological Museum of Parma (Italy). *International Archives of the Photogrammetry, Remote Sensing and Spatial Information Sciences*, XLI-B5, 243–250. https://doi.org/10.5194/isprs-archives-XLI-B5-243-2016

Felicetti, A., Paolanti, M., Zingaretti, P., Pierdicca, R., & Malinverni, E. S. (2020). Mo.Se.: Mosaic image segmentation based on deep cascading learning. *Virtual Archaeology Review*, *12*(24), 25–38. https://doi.org/10.4995/var.2021.14179

Fiorucci, M., Khoroshiltseva, M., Pontil, M., Traviglia, A., Del Bue, A., & James, S. (2020). Machine Learning for Cultural Heritage: A Survey. *Pattern Recognition Letters*, *133*, 102–108. https://doi.org/10.1016/j.patrec.2020.02.017

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. https://doi.org/10.1016/j.asoc.2018.05.018

George, D., Xie, X., & Tam, G. K. (2018). 3D mesh segmentation via multi-branch 1D convolutional neural networks. *Graphical Models*, *96*, 1–10. https://doi.org/10.1016/j.gmod.2018.01.001

Giuffrida, D., Mollica Nardo, V., Giacobello, F., Adinolfi, O., Mastelloni, M. A., Toscano, G., & Ponterio, R. S. (2019). Combined 3D surveying and Raman Spectroscopy Techniques on artifacts preserved at Archaeological Musem of Lipari. *Heritage*, *2*(3), 2017–2027. https://doi.org/10.3390/heritage2030121

Grilli, E., Farella, E. M., Torresani, A., & Remondino, F. (2019). Geometric features analysis for the classification of Cultural Heritage point clouds. *International Archives of the Photogrammetry, Remote Sensing and Spatial Information Sciences*, XLII-2/W15, 541–548. https://doi.org/10.5194/isprs-archives-XLII-2-W15-541-2019

Grilli, E., Özdemir, E., & Remondino, F. (2019). Application of machine and deep learning strategies for the classification of Heritage point clouds. *International Archives of the Photogrammetry, Remote Sensing and Spatial Information Sciences*, XLII-4/W18, 447–454. https://doi.org/10.5194/isprs-archives-XLII-4-W18-447-2019

Goodfellow, I., Bengio, Y., & Courville, A. (2016). *Deep Learning*. MIT Press.

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. https://doi.org/10.1016/j.patcog.2017.10.013

Guidi, G., Malik, U. S., Frischer, B., Barandoni, C., & Paolucci, F. (2017). The Indiana University-Uffizi project: Metrological challenges and workflow for massive 3D digitization of sculptures. *23rd International Conference on Virtual System & Multimedia (VSMM)*, 1–8. https://doi.org/10.1109/VSMM.2017.8346268

He, T., Shen, C., Tian, Z., Gong, D., Sun, C., & Yan, Y. (2019). Knowledge adaptation for efficient semantic segmentation. *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 578–587. https://doi.org/10.1109/CVPR.2019.00067

Jégou, S., Drozdzal, M., Vazquez, D., Romero, A., & Bengio, Y. (2017). The one hundred layers tiramisu: Fully convolutional densenets for semantic segmentation. *Proceedings of the IEEE conference on computer vision and pattern recognition workshops* (pp. 11–19). https://doi.org/10.1109/CVPRW.2017.156

Kersten, T. P., Tschirschwitz, F., & Deggim, S. (2017). Development of a virtual museum including a 4D presentation of building history in Virtual Reality. *International Archives of the Photogrammetry, Remote Sensing and Spatial Information Sciences*, XLII-2/W3, 361–367. https://doi.org/10.5194/isprs-archives-XLII-2-W3-361-2017

Knyaz, A. V., Kniaz, V. V., Remondino, F., Zheltov, S. Y., & Gruen, A. (2020). 3D reconstruction of a complex grid structure combining UAS images and deep learning. *Remote Sensing*, 12(19), 3128. https://doi.org/10.3390/rs12193128

Lin, P., Sun, P., Cheng, G., Xie, S., Li, X., & Shi, J. (2020). Graph-guided architecture search for real-time semantic segmentation. *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 4203–4212. https://doi.org/10.1109/CVPR42600.2020.00426

Llamas, J., Lerones, P. M., Medina, R., Zalama, E., & Gómez-García-Bermejo, J. (2017). Classification of architectural heritage images using deep learning techniques. *Applied Science*, *7*(10), 992. https://doi.org/10.3390/app7100992

Lo Turco, M., Piumatti, P., Rinaudo, F., Tamborrino, R., & González-Aguilera, D., (2018). 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 S. Bertocci (Ed.), *Programmi Multidisciplinari Per L'internazionalizzazione Della Ricerca. Patrimonio Culturale, Architettura e Paesaggio* (pp. 107–109). DIDA Press.

Lo Turco, M., Piumatti, P., Rinaudo, F., Calvano, M., Spreafico, A., & Patrucco, G. (2018). 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. https://doi.org/10.1109/DigitalHeritage.2018.8810128

Mafrici, N., & Giovannini, E. C. (2020). Digitalizing data: From the historical research to data modelling for a (digital) collection documentation. In M. Lo Turco, E. C. Giovannini, , & N. Mafrici (Eds.), *Digital & Documentation. Digital Strategies for Cultural Heritage* (Vol. 2, pp. 38–51). Pavia University Press.

Malik, U. S., Guidi, G. (2018). Massive 3D digitization of sculptures: Methodological approaches for improving efficiency. *IOP Conference Series: Material Science and Engineering*, *364*. https://doi.org/10.1088/1757-899X/364/1/012015

Minto, S., & Remondino, F. (2014). Online access and sharing of reality-based 3D models. *SCIRES-IT-SCIentific RESearch and Information Technology*, *4*(2), 17–28. http://doi.org/10.2423/i22394303v4n2p17

Patrucco, G., Chiabrando, F., Dondi, P, & Malagodi, M. (2018). Image and range-based 3D acquisition and modeling of popular musical instruments. *Proceedings from the Document Academy*, *5*(2),9. https://doi.org/10.35492/docam/5/2/9

Patrucco, G., Rinaudo, F., & Spreafico, A. (2019). A new handheld scanner for 3D survey of small artifacts: The Stonex F6. *International Archives of the Photogrammetry, Remote Sensing and Spatial Information Sciences*, XLII-2/W15, 895–901. https://doi.org/10.5194/isprs-archives-XLII-2-W15-895-2019

Pierdicca, R., Paolanti, M., Matrone, F., Martini, M., Morbidoni, C., Malinverni, E. S., Frontoni, E., & Lingua, A. M. (2020). Point cloud semantic segmentation using a deep learning framework for Cultural Heritage. *Remote Sensing*, *12*(6), 1005. https://doi.org/10.3390/rs12061005

Salvador-García, E., Viñals, M. J., & García-Valldecabres, J. L. (2020). Potential of HBIM to improve the efficiency of visitor flow management in Heritage sites. Towards smart heritage management. *International Archives of the Photogrammetry, Remote Sensing and Spatial Information Sciences*, XLIV-M-1-2020, 451–456. https://doi.org/10.5194/isprs-archives-XLIV-M-1-2020-451-2020

Shorten, C., & Khoshgoftaar, T. M. (2019). A survey on image data augmentation for deep learning. *Journal of Big Data, 6(1)*, 1–48. https://doi.org/10.1186/s40537-019-0197-0

Stathopoulou, E. K., & Remondino, F. (2019). Semantic photogrammetry: Boosting image-based 3D reconstruction with semantic labeling. *International Archives of the Photogrammetry, Remote Sensing and Spatial Information Sciences*, XLII-2/W9, 685–690. https://doi.org/10.5194/isprs-archives-XLII-2-W9-685-2019

UNESCO. (1979). *Recommendation for the Protection of Movable Cultural Property, Records of the General Conference, 20th Session, I: Resolutions.* Paris: UNESCO.

Vargas, R., Mosavi, A., & Ruiz, R. (2018). Deep learning: A review. *Advances in Intelligent Systems and Computing*, *29*(8), 232–244. https://doi.org/10.20944/PREPRINTS201810.0218.V1

Yazan, E., & Talu, M. F. (2017). Comparison of the stochastic gradient descent based optimization techniques. *2017 International Artificial Intelligence and Data Processing Symposium (IDAP)*, 1–5. https://doi.org/10.1109/IDAP.2017.8090299