

This is the peer reviewed version of the following article:

Convolutional Neural Network for Background Removal in Close Range Photogrammetry: Application on Cultural Heritage Artefacts / Bici, M.; Gherardini, F.; de Los Angeles Guachi-Guachi, L.; Guachi, R.; Campana, F.. - (2023), pp. 780-792. ( International Joint Conference on Mechanics, Design Engineering and Advanced Manufacturing, JCM 2022 ita 2022) [10.1007/978-3-031-15928-2\_68].

Springer Science and Business Media Deutschland GmbH

*Terms of use:*

The terms and conditions for the reuse of this version of the manuscript are specified in the publishing policy. For all terms of use and more information see the publisher's website.

19/12/2025 04:22

# Convolutional Neural Network for Background Removal in Close Range Photogrammetry: Application on Cultural Heritage Artefacts

Michele Bici<sup>1</sup>[0000-0002-7744-2152], Francesco Gherardini<sup>2</sup>[0000-0002-9275-4314], Lorena de Los Angeles Guachi-Guachi<sup>3,4</sup>[0000-0002-8951-8150], Robinson Guachi<sup>4</sup>[0000-0002-0476-6973], Francesca Campana<sup>1</sup>[0000-0002-6833-8505]

<sup>1</sup> Department of Mechanical and Aerospace Engineering, DIMA – Sapienza University of Rome, 00184, Rome, Italy

<sup>2</sup> Department of Engineering “Enzo Ferrari”, University of Modena and Reggio E., Modena, Italy

<sup>3</sup> Department of Mechatronics, Universidad Internacional del Ecuador - UIDE, 170411, Av. Simón Bolívar, Quito, Ecuador

<sup>4</sup> SDAS Research Group, Ibarra, Ecuador

michele.bici@uniroma1.it

**Abstract.** Post-processing pipeline for image analysis in reverse engineering modelling, such as photogrammetry applications, still asks for manual interventions mainly for shadows and reflections corrections and, often, for background removal. The usage of Convolutional Neural Network (CNN) may conveniently help in recognition and background removal. This paper presents an approach based on CNN for background removal, assessing its efficiency. Its relevance pertains to a comparison of CNN approaches versus manual assessment, in terms of accuracy versus automation with reference to cultural heritage targets. Through a bronze statue test case, pros and cons are discussed with respect to the final model accuracy. The adopted CNN is based on the U-NetMobilenetV2 architecture, a combination of two deep networks, to converge faster and achieve higher efficiency with small datasets. The used dataset consists of over 700 RGB images used to provide knowledge from which CNNs can extract features and distinguish the pixels of the statue from background ones. To extend CNN capabilities, training sets with and without dataset integration are investigated. Dice coefficient is applied to evaluate the CNN efficiency. Results obtained are used for the photogrammetric reconstruction of the Principe Ellenistico model. This 3D model is compared with a model obtained through a 3D scanner. Moreover, through a comparison with a photogrammetric 3D model obtained without the CNN background removal, performances are evaluated. Although few errors due to bad light conditions, the advantages in terms of process automation are consistent (over 50% in time reduction).

**Keywords:** Reverse Engineering, Close Range Photogrammetry, CNN, U-Net, MobilenetV2, Cultural Heritage Preservation.

## 1 Introduction

Image analysis and motion tracking are two fields of research where background and foreground distinction are necessary to assess the scene and to distinguish the object to be analysed [1,2]. Fields of application are video surveillance, image processing for medical purposes and quality control, image vision and reconstruction such as photogrammetry applications. Major problems to be faced are change of points of view, shadows and lighting, foreground movements, and calibrations. As stated in [3], from the computer vision field a large set of approaches are present in literature and many of them may be classified as mature approaches [4]. In aerial and close-range photogrammetry, automation of the process is extremely important due to the large amount of data to be processed. Machine learning and deep learning are two approaches with many investigations devoted to pipeline automation [5,6]. Concerning deep learning, the relevance of the training dataset on the final robustness of the net represents the major drawback. Although shared databases and addition of synthetic scenes for the training sets may help to enlarge the generality of these approaches, as highlighted in [7], the computation efforts may rapidly increase, shifting the bottleneck from the lack of automation to the lack of computational resources [8]. In recent years, Convolutional Neural Networks (CNN) have been increasingly used in image analysis and image segmentation, but only nowadays, they started to be used for the improvement of semantic photogrammetry [9], for increasing the level of automation in close-range applications [10], or for background removal with the aim of detecting moving objects [11].

In this paper, the usage of a CNN is investigated to evaluate benefits in close-range photogrammetry by automatic background extraction. Its relevance pertains to a better understanding of CNN's accuracy and processing time (including also training) with reference to cultural heritage targets, where free form geometry and morphological complexity may be relevant as the environment conditions and the manoeuvrability of the subjects to be acquired. A relevant example of this is provided by some cultural heritage preservation problems. Ancient bronze statues represent a niche sector that periodically asks for maintenance mainly due to the problem of material corrosion and exhibition changes. Commonly, during ordinary and extraordinary checks, photographic archives represent one of the major sources of information, helping in controls and maintenance. Furthermore, currently, the usage of close-range photogrammetry for the obtainment of a reverse engineered model of the retrieval represents an outstanding procedure in terms of speed, accuracy, and costs. 3D models by reverse engineering support investigations and exhibition set-up as demonstrated by the works recently made on the Vittoria Alata of Brescia [12] and the periodic maintenance of ancient statues of the Palazzo Massimo site in Rome, part of the Museo Nazionale Romano [13,14]. Among them, the Principe Ellenistico was recently acquired through close-range photogrammetry and the results were compared with laser scanner techniques, finding that the adoption of commercial cameras and software may provide good results also with reflective surfaces if a virtual cage for camera orientation is set [15]. Moreover, authors highlighted the necessity of an automatic background removal to speed up the reverse engineering pipeline, thus in this paper we are going to apply CNN to this case-study, so that a comparison with standard elaboration may be done. The paper

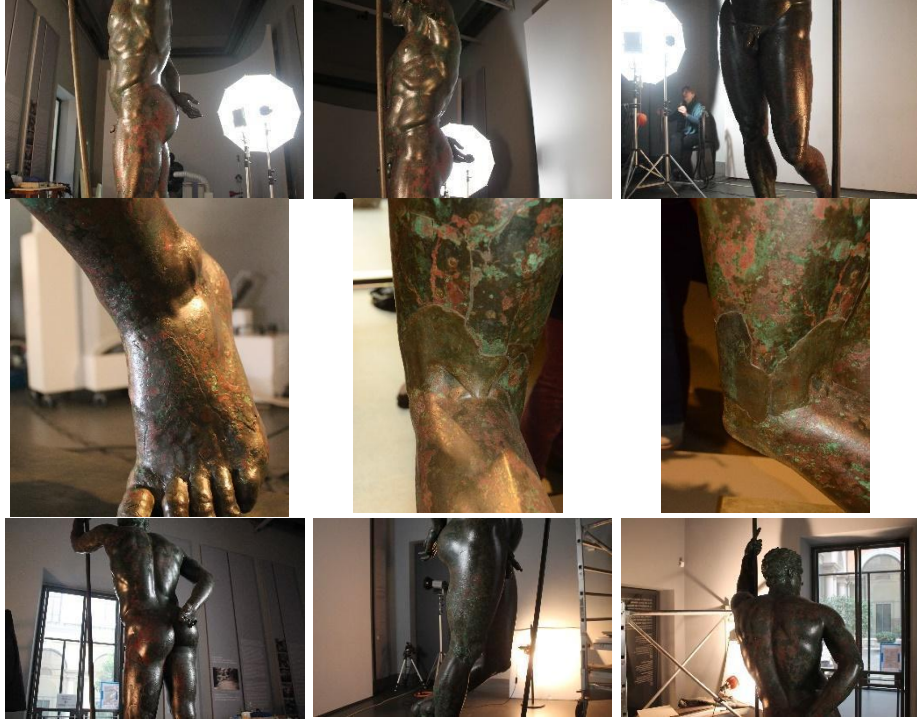
describes the methodology (Section 2), through 3 sub-sections: dataset, CNNs and the evaluation metric to assess the reliability of the results. Then, the experimental results are presented in Section 3 subdivided in CNN results and photogrammetry reconstruction through automatic background removal. Finally, Discussion (Section 4) about pros and cons and Conclusions (Section 5) are outlined.

## 2 Methodology

### 2.1 Dataset

CNNs support the automation of feature recognition and segmentation with improved results thanks to the opportunity of analysing large dataset and adopting training. In this application, the segmentation process consists in isolating the Principe Ellenistico details from the background. The Principe Ellenistico is an ancient bronze probably of the II century A.C., found in Rome nearby the ancient Constantin's thermae in 1885. It is 2.04 m tall, thus its acquisition requires many photos, with different angles of view, as reported in [15], lighting conditions (thus possible reflection) and of course background details, considering that it is not allowed to move or isolate the statue from the exhibition.

The Principe Ellenistico dataset for training and testing the proposed statue segmentation approach, was created by capturing the statue's appearance from various points of view, which were subsequently prepared and processed to generate the pixel labels (masks also named ground truth) in a fully supervised fashion. The dataset consists of over 700 RGB images ranging in size and resolution from 2499x2779 to 4272x2848 pixels, and they are used to provide knowledge from which CNNs can extract features and learn to distinguish the pixels of the statue from those of the background. Fig. 1 depicts some examples of statue images, showing the wide range of light and type of views that may occur. For training purposes, the dataset was divided into two sets with a percentage ratio of 85% and 15 % for training and validation, respectively. In order to match the available computing power, images needed to be resized. Due to this, the CNN's input shape was set to 256x256x3. The training was performed in 500 iterations using a batch size of 32. Experiments are carried out running the software on a 2.3GHz Intel Core i7 processor with 20 GB of RAM memory.

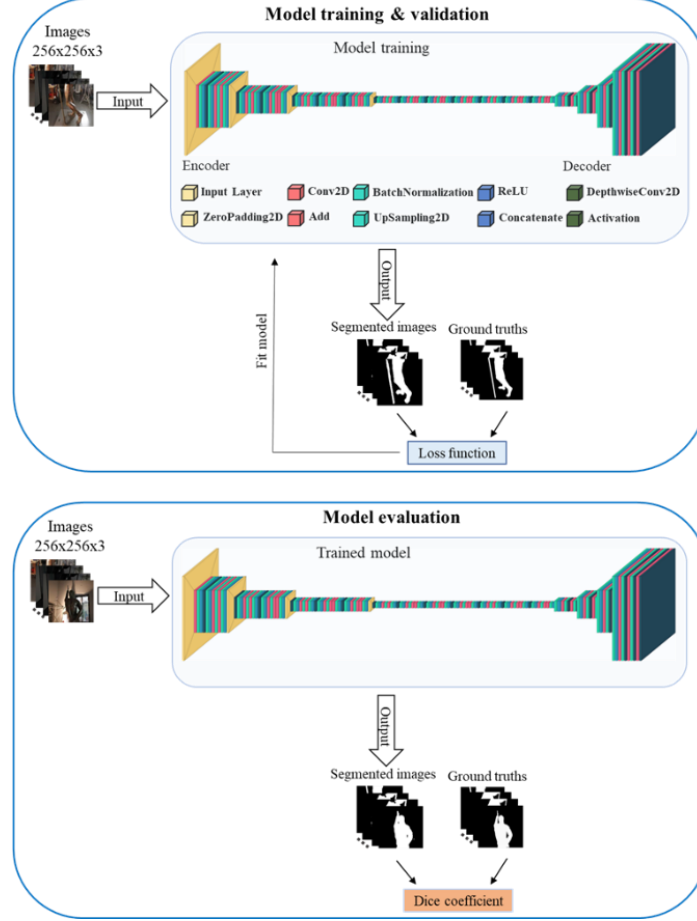


**Fig. 1.** Some statue images from the created dataset to feed the CNN. The dataset contains images from different angles, as well as a variety of illumination and background conditions.

## 2.2 CNNs set-up

The segmentation is performed by combining architectures capable of working with fewer training images while still producing accurate results, such as U-Net [16] and MobilenetV2 [17] convolutional neural networks.

On one hand, MobilenetV2 serves as a pre-trained encoder for the feature extraction task. It was pre-trained on the ImageNet dataset [18] to achieve higher performance and faster convergence than non-pre-trained models. On the other hand, U-Net is composed of a contraction and expansion path for semantic segmentation. The resulting combined architecture is illustrated in Fig. 2 and Table 1 describes its layers.



**Fig. 2.** General workflow for training, validation and testing of the U-netMobilenetV2 architecture for statue segmentation from digital RGB images.

All parameters have been optimised to achieve the highest level of segmentation accuracy. We use the Nesterov-accelerated Adaptive Moment Estimation (Nadam) optimization algorithm, which is an extension to the Gradient Descent Optimization algorithm and mainly used for noisy gradients. To fit the U-NetMobileNetV2 model, we use the Dice coefficient to calculate the similarity between spatially matched pixels as shown in Equation (1).

$$Dice\ loss\ (I_{gt}, I_s) = 1 - (2 I_{gt} I_s + 1) / (I_{gt} + I_{s+1}) \quad (1)$$

where  $I_{gt}$  is the ground truth (image segmented correctly) and  $I_s$  is the segmented image (output of the CNN model).

The proposed approach is developed through python software routines using TensorFlow, Keras and scikit-image.

**Table 1.** Details of CNN U-NetMobilenetV2 for statue segmentation.

Layer name	Description	# of layers	Kernel size (encoder)	Kernel size (decoder)
Input	Input for the model	1		
Conv2D	Computes a 2-D spatial convolution	36	1x1, 3x3	3x3
BatchNormali- zation	Normalises the activations of the previous layer for each batch	48	-	-
ReLU	It is a piecewise linear function that directly generates the input if it is positive; otherwise, it generates zero.	27	-	-
Depthwise- Conv2D	Applies a single convolutional filter to each input channel	13	3x3	3x3-
ZeroPad- ding2D	Adds rows and columns of zeros at the top, bottom, left and right side of each input.	3	-	-
Upsam- pling2D	Repeats the rows and columns of the input	4	-	-
Concatenate	Concatenates a list of inputs	4	-	-

### 2.3 Evaluation metric for segmentation

From the CNN point of view, the ability of the proposed approach to segment the statue of the Hellenistic Prince is measured in terms of the Dice coefficient given by Equation 2.

$$Dice\ coefficient = [(2 \times TP) / ((2 \times TP) + FP + FN)] \times 100 \quad (2)$$

The Dice coefficient is a spatial overlap index and a reproducibility validation metric calculated between the binary segmented images from U-NetMobilenetV2 output and the ground truth. The Dice coefficient value ranges from 0 to 1, with 0 indicating no spatial overlap and 1 indicating complete overlap [19], where TP is the number of statue pixels correctly identified as statue pixels, TN is the number of background pixels correctly identified as background pixels, FP is the number of statue pixels erroneously identified as background pixels, while FN is the number of background pixels erroneously identified as statue pixels.

### 3 Results

#### 3.1 CNN results

Table 2 shows the Dice coefficient computed through Equation (2) and, in the case of the original dataset, it is of about 94%. To investigate the effect of dataset expansion via different images, training with different dataset was also made and reported in Table 2. It can be seen that one of the major advantages of this work is the high dice coefficient achieved by the proposed approach (higher than 93%) for all evaluated datasets. Thus, the addition of existing datasets related to the problem domain can improve the overall performance of the model. However, a wrong addition could significantly reduce the model's performance.

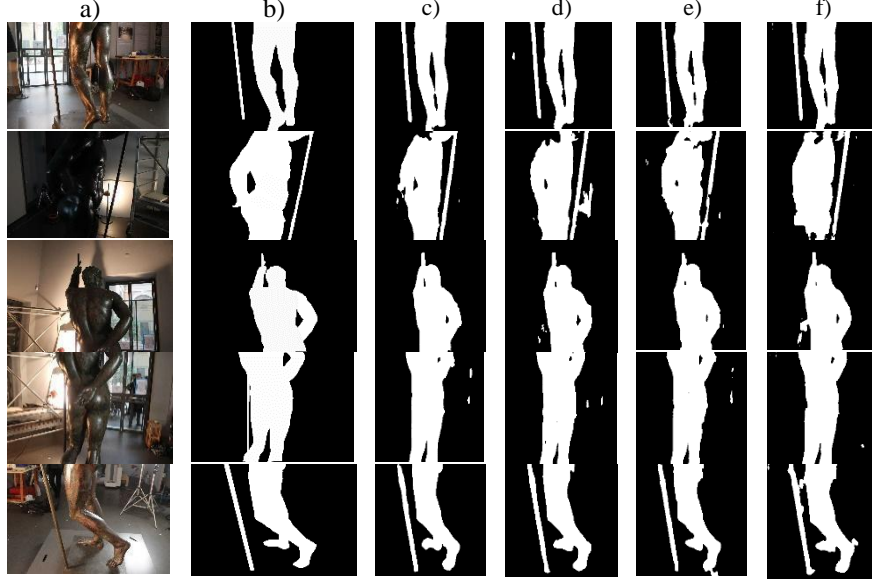
**Table 2.** Dice coefficients.

Dataset	U-NetMobilenetV2
Hellenistic Prince	<b>94.01%</b>
Hellenistic Prince + Sculpture 6K [20]	94.07%
Hellenistic Prince + PedCut2013 [21]	93.87%
Hellenistic Prince + FMD [22]	<b>94.21%</b>
All datasets	93.88%

Based on the assumption that increasing the number of training samples related to the study problem improves model accuracy, the training set for the new Hellenistic Prince dataset was expanded with related images such as Sculptures (Sculpture 6k [20]), Humans (PedCut2013 [21]), and materials (FMD [22]). Some of the results obtained for the referred datasets are depicted in Fig. 3.

It can be observed that the U-NetMobilenetV2 trained with Hellenistic Prince + FMD dataset leads to less noisy results. It is attributed to information provided to the model during the training stage that are mainly related to the problem domain. Hellenistic Prince dataset includes bronze segmentation while the FMD dataset includes data for fabric, foliage, glass, metal, paper, plastic, stone, water and wood. Nonetheless, this combination of both datasets is sensible in dark environments. On the other hand, the U-NetMobilenetV2 trained with only the Hellenistic Prince dataset produces less misclassified results in most of the cases.

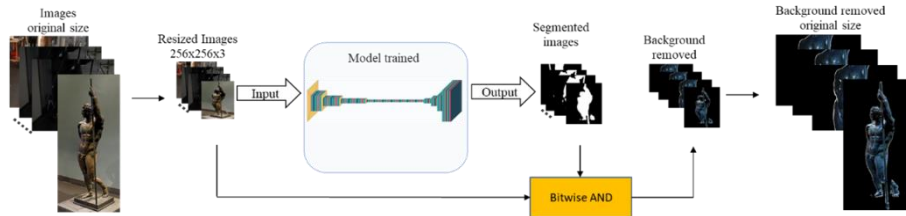




**Fig. 3.** Results related to: a) Input frames; b) ground truths; c) results obtained by model trained with dataset of the Hellenistic Prince Statue; d) results obtained by model trained with Hellenistic Prince + Sculpture 6K [20] datasets; e) results obtained by model trained with Hellenistic Prince + PedCut2013 [21] datasets; f) results obtained by model trained with Hellenistic Prince + FMD [22] datasets.

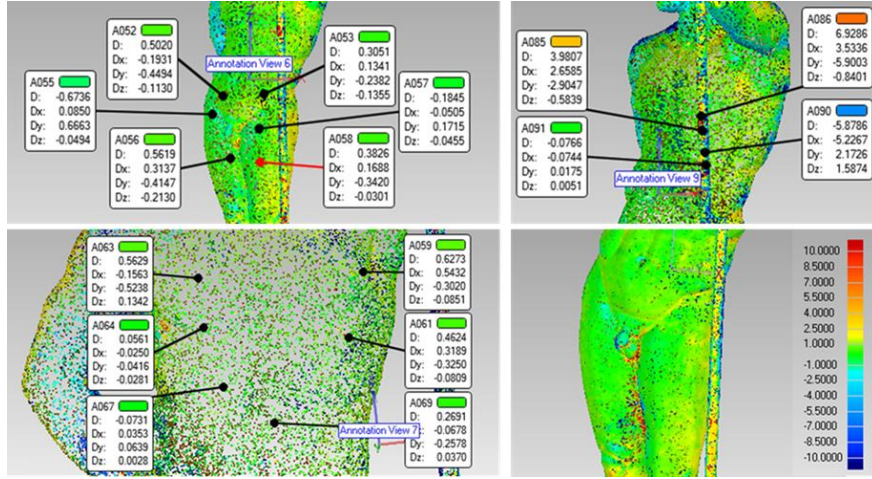
### 3.2 Reconstruction based on automatic background removal

To experimentally test the ability of the U-NetMobileNetV2 model trained with the Principe Ellenistico + FMD datasets to automatically segment images to remove the background for cultural heritage modelling, some additional experiments were performed with 716 images taken from different viewpoints. The methodology for automatic background removal first resizes all images to the expected size of the U-NetMobileNetV2 model (256x256x3). Secondly, the resized images are segmented using the trained U-NetMobileNetV2 model. Then, the bitwise and operator is applied between each resized image and its corresponding segmented image to remove all pixels when the segmented image has a zero-pixel value. Finally, the results are resized to the original, as shown in Fig. 4.



**Fig. 4.** General workflow of the proposed methodology for automatic background removal for modelling of cultural heritage.

After this removal operation, the set of worked images has been used for obtaining a new virtual model, through Agisoft Metashape Pro. This model, called CNN-based model, similarly to what was done in previous tests, reported in [15], has been compared to a model obtained from 3D scanning, assumed as reference. The 3D scanner used for the development of the reference point cloud was an Artec Eva portable scanner, with an accuracy of 0.1 mm and a resolution of about 0.2 mm. The CNN-based photogrammetric model showed, in general, good correspondence with the scanned one.



**Fig. 5.** Comparison between the scanned model and the CNN-based model (the colour scale shown in the bottom is the same in all the figures).

Fig. 5 highlights the distance analysis between the two models taken on the upper part of the body and the legs. In the best conditions, differences are averagely below 0.70 mm. The spear is more affected by errors as well as the legs intersection and the shoulders, as discussed in the next section, where a comparison of the CNN-based model and the previous structured reconstruction, reported in [15], is discussed.

## 4 Discussion

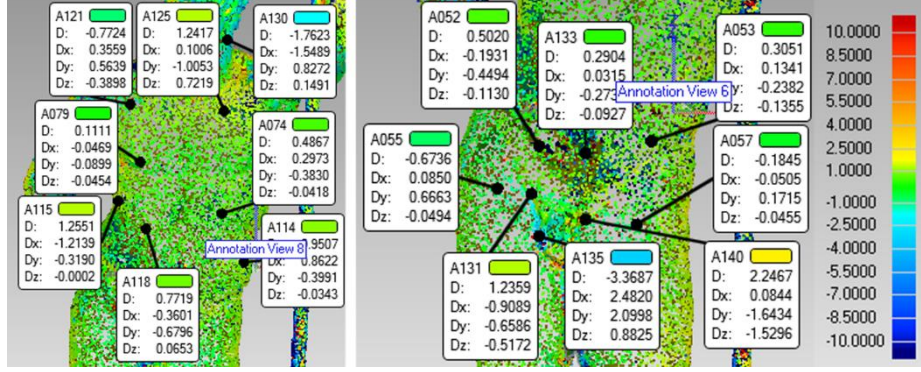
We were able to compare the results obtained in the photogrammetric reconstruction made without the CNN-based background removal, described in [15]. The previous reconstruction was based on the same original dataset of 716 photos. Inside that procedure, background removal operations were evaluated and done, when needed, case by case, through the Agisoft Metashape dedicated tool, depending on the user's choices.

The comparison shows an equivalence about the percentage of aligned images (84.64% in the previous reconstruction, 84.91% in the CNN-based), and substantial improvements in terms of time consumption and memory usage. The processing time is considerably reduced, especially in the generation phase of the depth maps (6h 42' vs 1h 47'), but also the other parameters of Table 3 are halved as well. The shorter time necessary for the generation of depth maps, as well as for the alignment, is linked to the fact that in the reconstruction without masking, the software reconstructs environment points (and sometimes noise), which require time and computational resources even if they are not part of the final reconstructed model.

**Table 3.** Process parameters of the photogrammetric reconstruction phase.

Phase	Previous reconstruction [16]	CNN-based reconstruction	Comparison (%)
Images (automatically) aligned out of 716	606 (84.64%)	608 (84.92%)	+0.33
Point Cloud: Matching time	2 h 35 min	1 h 17 min	-50.32
Matching memory usage	830.35 MB	420.79 MB	-49.32
Point Cloud: Alignment time	10 min 39 s	21 min 36 s	+102.82
Alignment memory usage	816.55 MB	328.87 MB	-59.72
Depth maps generation parameters Processing time	6 h 42 min	1 h 47 min	-73.38
Dense cloud generation parameters Processing time	1 h 45 min	1 h 8 min	-35.24
Model processing time	42 min 44 s	13 min 18 s	-68.88
<i>Total time</i>	<i>~ 11 h 55min</i>	<i>~ 4 h 47 min</i>	<i>-59.86</i>

Table 3 shows that the point cloud alignment time is the only subphase that becomes more time consuming in the CNN-based reconstruction. It is principally due to the fact that, in the previous reconstruction, the presence of the image background may help the alignment, providing additional constraints. Nevertheless, this subphase represents just a little percentage of the entire time consumption: with CNN-based photogrammetry, about 22 minutes over 5 hours; without CNN 11 minutes over 12 hours. Through the alignment of these two models (the one obtained with the previous procedure and the CNN-based one) onto the scanned one assumed as reference in the previous Section, we were able to compare the results of the procedures.



**Fig. 6.** Comparison between the photogrammetric model and the CNN-based one (the colour scale shown on the right is the same in both the figures).

The comparisons of areas not affected by photographic issues show really limited differences, as shown in Fig.5. Areas that present photogrammetric problems like bad light conditions, reflection and accessibility issues, show higher distances (top of the head, spear, inner legs and feet). In the CNN-based model, the effect of these noise sources seems not reduced, probably due to the missing background that, in the previous case, mitigated the problems in some regions.

## 5 Conclusions

In this paper, a CNN approach for the automatic background removal in images to be used for close range photogrammetric reconstructions is presented. The U-NetMobilenetV2 architecture has been used on over 700 images related to a bronze statue. U-NetMobilenetV2 is the integration of two CNNs: U-Net and MobilenetV2. It is developed with the aim of reducing the training dataset without losing accuracy. Through Dice coefficient, the effect of dataset expansion, adding images from Sculpture 6k, Humans and materials original dataset, has been evaluated, finding comparable results with the original training dataset.

The images achieved after background removal have been used to build the 3D model of the statue via close-range photogrammetry post-processing through Agisoft Metashape Pro. This model has been compared to a model obtained from 3D scanning, used as reference. In the best conditions (areas from images with good light conditions and accessibility) we find a mean error of 0.7 mm. This value is consistent to those achieved by close-range photogrammetry without automatic background removal. Areas with higher errors are due to the quality limit of some of the images considered in the investigated dataset. From the elaboration point of view the adoption of the CNN highly reduced the time efforts (about -60%). Substantially, it can be said that the CNN-based procedure is able to speed up the process without a significant influence on the results in terms of reconstruction. So, it can be said that this represents the main pro of

the procedure, in addition to the easiness of usage of photogrammetry in reverse engineering.

Next improvements and developments will be connected to the usage of filters (integrated in optics and/or in post-processing of the images) for improving image quality and mitigating reflections also through dedicated light systems and measurements. In addition, to assess the robustness of the procedure, other applications onto bronze statues and other retrievals will be carried out, testing also different background conditions.

## References

1. Toyama, K., Krumm, J., Brumitt, B., Meyers, B.: Wallflower: Principles and practice of background maintenance. *Proceedings of the IEEE International Conference on Computer Vision*, 1, 255-261(1999).
2. Sajid, H., Cheung, S.-C.S.: Universal multimode background subtraction. *IEEE Transactions on Image Processing* 26 (7), art. no. 7904604, 3249-3260 (2017).
3. Garcia-Garcia, B., Bouwmans, T., Rosales-Silva, A.: Background Subtraction in Real Applications: Challenges, Current Models and Future Directions. *Computer Science Review*, Volume 35, (2020).
4. Guachi, L., Cocorullo, G., Corsonello, P., Frustaci, F., Perri, S.: A novel background subtraction method based on color invariants and grayscale levels *Proceedings of International Carnahan Conference on Security Technology*, (2014).
5. Qin, R., Gruen, A.: The role of machine intelligence in photogrammetric 3D modeling—an overview and perspectives. *International Journal of Digital Earth* 14 (1), 15-31 (2021).
6. Ghosh, M., Obaidullah, S.M., Gherardini, F., Zdimalova, M.: Classification of geometric forms in mosaics using deep neural network. *Journal of Imaging* 7(8), 149 (2021).
7. Lowe, D. G.: Distinctive Image Features from Scale-Invariant Keypoints. *International Journal of Computer Vision* 60 (2), 91–110 (2004).
8. Qin, R., Gruen, A.: The role of machine intelligence in photogrammetric 3D modeling—an overview and perspectives. *International Journal of Digital Earth* 14 (1), 15 – 31 (2021).
9. Stathopoulou, E.-K., Remondino, F.: SEMANTIC PHOTOGRAMMETRY - BOOSTING IMAGE-BASED 3D RECONSTRUCTION with SEMANTIC LABELING. *ISPRS Annals of the Photogrammetry, Remote Sensing and Spatial Information Sciences* 42 (2/W9), 685 – 690 (2019).
10. Eastwood, J., Sims-Waterhouse, D., Piano, S., Weir, R., Leach, R.K.: Autonomous close-range photogrammetry using machine learning. *Proceedings of 14th ISMTII*, (2019).
11. Elhabian, S., El-Sayed, K. M., Ahmed, S. H.: Moving Object Detection in Spatial Domain using Background Removal Techniques - State-of-Art. *Recent Patents on Computer Science* 2008 (1), 32-54 (2008).
12. Bici, M., Brini, A., Campana, F., Capoferri, S., Guarnieri, R., Morandini, F., Patera, A.: Design of the New Inner Frame for the Vittoria Alata di Brescia: How Engineering Design May Support Ancient Bronze Restoration. *Lecture Notes in Mechanical Engineering*, 951-962 (2022).
13. Bici, M., Campana, F., Colacicchi, O., D'Ercoli, G.: CAD-CAE methods to support restoration and museum exhibition of bronze statues: the "Principe Ellenistico". *IOP Conference Series: Materials Science and Engineering* 364 (1), (2018).
14. Cicconi, P., Bici, M., Colacicchi Alessandri, O., D'Ercoli, G., Campana, F.: A CAD-Based Framework for Interactive Analysis in the Restoration of Bronze Statues. *Lecture Notes in Mechanical Engineering*, 938-950 (2022).

15. Bici, M., Gherardini, F., Campana, F., Leali, F.: A preliminary approach on point cloud reconstruction of bronze statues through oriented photogrammetry: The “Principe Ellenistico” case. *IOP Conference Series: Materials Science and Engineering* 949 (1), (2020).
16. Weng, W., Zhu, X.: INet: Convolutional Networks for Biomedical Image Segmentation. *IEEE Access* 9, 16591–16603 (2021).
17. Sandler, M., Howard, A., Zhu, M., Zhmoginov, A., Chen, L.C.: MobileNetV2: Inverted Residuals and Linear Bottlenecks. *Proc. IEEE Comput. Soc. Conf. Comput. Vis. Pattern Recognit.*, 4510–4520 (2018).
18. IMAGENET, <https://image-net.org/update-mar-11-2021.php>, last accessed 2022/01/21.
19. Zijdenbos, A.P., Dawant, B.M., Margolin, R.A., Palmer, A.C.: Morphometric analysis of white matter lesions in MR images: method and validation. *IEEE Transactions on Medical Imaging* 13(4), 716–724 (1994).
20. Sculptures 6k datasets, <https://www.robots.ox.ac.uk/~vgg/data/sculptures6k/>, last accessed 2022/01/21.
21. PedCut: an iterative framework for pedestrian segmentation combining shape models and multiple data cues (2013), [http://www.gavrilanet.net/Datasets/Daimler\\_Pedestrian\\_Benchmark\\_D/Daimler\\_Pedestrian\\_Segmentation/daimler\\_pedestrian\\_segmentatio.html](http://www.gavrilanet.net/Datasets/Daimler_Pedestrian_Benchmark_D/Daimler_Pedestrian_Segmentation/daimler_pedestrian_segmentatio.html), last accessed 2022/01/21.
22. Flickr Material Database (FMD), <https://people.csail.mit.edu/ceiliu/CVPR2010/FMD/>, last accessed 2022/01/21.