An image-based Deep Learning workflow for 3D Heritage point cloud semantic segmentation

Eugénio Pellis, Arnadi Murtiyoso, Andrea Masiero, Grazia Tucci, Michele Betti, Pierre Grussenmeyer

AN IMAGE-BASED DEEP LEARNING WORKFLOW FOR 3D HERITAGE POINT CLOUD SEMANTIC SEGMENTATION

E. Pellis [1,3], A. Murtiyoso [2], A. Masiero [1], G. Tucci [1], M. Betti[1], P. Grussenmeyer [3]

[1] Department of Civil and Environmental Engineering (DICEA), University of Florence, 50139 Florence, Italy - (eugenio.pellis, andrea.masiero, grazia.tucci, michele.betti)@unifi.it
[2] Forest Resources Management Group, Institute of Terrestrial Ecosystems, Department of Environmental Systems Science, ETH Zürich, Switzerland - arnadidhestaratri.murtiyoso@usys.eth.ch
[3] Université de Strasbourg, INSA Strasbourg, CNRS, ICube Laboratory UMR 7357, Photogrammetry and Geomatics Group, 67000 Strasbourg, France – pierre.grussenmeyer@insa-strasbourg.fr

Commission II

KEY WORDS: 3D Point Cloud, Deep Learning, Semantic Segmentation, Cultural Heritage

ABSTRACT:

The interest in high-resolution semantic 3D models of historical buildings continuously increased during the last decade, thanks to their utility in protection, conservation and restoration of cultural heritage sites. The current generation of surveying tools allows the quick collection of large and detailed amount of data: such data ensure accurate spatial representations of the buildings, but their employment in the creation of informative semantic 3D models is still a challenging task, and it currently still requires manual time-consuming intervention by expert operators. Hence, increasing the level of automation, for instance developing an automatic semantic segmentation procedure enabling machine scene understanding and comprehension, can represent a dramatic improvement in the overall processing procedure. In accordance with this observation, this paper aims at presenting a new workflow for the automatic semantic segmentation of 3D point clouds based on a multi-view approach. Two steps compose this workflow: first, neural network-based semantic segmentation is performed on building images. Then, image labelling is back-projected, through the use of masked images, on the 3D space by exploiting photogrammetry and dense image matching principles. The obtained results are quite promising, with a good performance in the image segmentation, and a remarkable potential in the 3D reconstruction procedure.

1. INTRODUCTION

In recent years, the high level of automation achieved by the latest 3D acquisition technologies, like laser scanner or photogrammetry, allows to collect a large amount of data in a short time, and to obtain a high level of details in shape, geometry and semantic information (Shan and Toth, 2018). Thanks to these continuous technological developments, the use of such surveying methods had a significant increase, in particular for the study of heritage buildings, with the main aim of documentation, interpretation, protection, conservation and restoration of cultural heritage (Yastikli, N., 2007)). Certain of the most advanced applications also deals with virtual reality and Building Information Modelling (H-BIM) (López et al., 2018).

Given the huge amount of 3D data that are usually collected nowadays to properly describe the sites of interests, automatic processing procedures shall be preferable in order to reduce the processing times. Nevertheless, automatic raw data processing for the creation of as-built 3D models still faces several significant challenges, often causing the need of time-consuming manual intervention, in particular to deal with the complexity of heritage buildings.

One of the key points to enable automation in 3D data processing is the development of a semantic segmentation procedure. This procedure involves the task of classifying each point of a 3D point cloud into classes or categories according to its semantic meaning, for example according to the constructive element typology (e.g., wall, columns, vaults, etc.) (Matrone et al., 2020). Certain of the main issues related to working with 3D point clouds are: the large data size, which implies long computing time, the unordered structure of point clouds, and the low availability of shared datasets and tools, which may be the basis of a common processing strategy, e.g. based on the use of properly developed artificial intelligence tools (Malinverni et al., 2019).

As a basic step for scene understanding and comprehension, several semantic segmentation techniques have already been tested in a wide range of different applications. Nowadays machine learning (ML) and deep learning (DL) techniques are the most extensively investigated and the most promising (Zhang et al., 2019). The main approaches to face semantic segmentation of 3D point clouds can be divided in two main groups: (i) projection-based methods and (ii) point-based methods.

Projection based methods leverage on an intermediate representation of the cloud in order to face with the unordered structure of 3D points, then they apply standard 2D approaches to perform the segmentation (Badrinarayanan et al., 2015), (Chen et al., 2017), (Cordts et al., 2016), and finally they re-project the extracted features on the starting shape or point cloud (Su et al., 2015), (Boulch et al., 2018).

Differently, point based methods work directly with the 3D points and they leverage of the full use of the characteristic of the raw point cloud data, considering all the spatial and geometrical information (Qi et al., 2017a).

This paper focuses on the development of a workflow for the classification of 3D point clouds based on a multi-view approach, exploring and testing the use of state-of-the-art neural networks on heritage scenarios, with the main aim of improving automation in 3D heritage model generation.

2. RELATED WORK

Semantic analysis has become a central topic in several applications, such as computer vision, robotics, or remote sensing. Numerous approaches have been developed to automate this task, including algorithmic, machine learning and deep learning approaches. Despite the increasing demand in heritage survey data processing, few works focus on dealing with the classification of 3D heritage point clouds.

Malinverni et al. (2019) described a method to label and cluster automatically a point cloud based on a supervised Deep Learning approach, using a state-of-the-art neural network called PointNet++ (Qi et al., 2017b). The results are promising but they still face several challenges due to the complexity of the training scene. Further developments of this research should include the association to the Neural Netowork of a structured ontology for the semantic parsing, the future specification of more detailed classes and the creation of a synthetic dataset to overcome he problem of very poor-annotated training data.
Matrone et. al (2020) make a comparison between machine learning and deep learning methods working directly with point cloud, and then, they proposed a new architecture named DGCNN-Mod+3DFeat that combined the positive aspects and advantages of Machine Learning and Deep Learning techniques, and they test it on the ARCH dataset benchmark with promising results. Both ML and DL proved to be approximately equally valuable: each of the two techniques alternatively outperformed the other one.
In their work, Grilli et al. (2019) presented a research for the classification of 3D heritage data (point clouds or polygonal mesh models) using different supervised learning approaches, including Machine Learning and Deep Learning. Their method works on 2D data ("texture-based approach) or directly the 3D data ("geometric-based approach) depending on the needs and scope of classification. The method was applied and validated on four different archaeological scenarios, proving its reliability and replicability, being effective in providing metric information as well.
Murtiyoso et al. (2021) developed an approach to semantically segment building façades in 6 categories using Deep Learning on rectified images by deploying pre-existing and pre-trained networks. The network outcome is back-projected into the 3D space by exploiting a depth map to generate a semantically segmented point cloud. The obtained overall accuracy is quite promising, yielding to a value of 79,8%. A mask-based approach has also been presented in Murtiyoso et al. (2022), investigating the use of back-projection into 3D space to more complex scenarios.
This work proposes the combined use of deep learning based semantic image segmentation with photogrammetric 3D reconstruction. The main aim is to introduce the semantic classification at the beginning of the classical photogrammetric workflow. The initial results obtained in our tests are quite promising, experimentally proving the potential effectiveness of using projection-based techniques for the generation of semantically segmented building point clouds.

3. DEVELOPED METHODOLOGY

In this paper we propose a multi-view based methodology to perform semantic segmentation of the heritage 3D point clouds. This approach is based on the segmentation of a 2D intermediate representation of the cloud, and then on the re-projection of the extracted features on the initial cloud. Despite working directly with 3D data provides an opportunity for a better understanding of the spatial and geometrical information, dealing with 2D images is usually a quite effective strategy. On one hand, it allows to exploit the tried-and-tested results in 2D image processing, in particular with Convolutional Neural Networks (CNNs), and, on the other hand, it allows to develop an automatic procedure for the creation of a directly-segmented cloud starting from photogrammetric images. In addition, at this time, a multi-view approach for heritage buildings semantic segmentation has never been tested.
Our methodology is composed of two main steps: (i) at first, the images are processed by a neural network to perform the semantic segmentation directly on the images, and (ii) secondly, a back-projection procedure allows to transfer the image features on the 3D point cloud.

3.1 The dataset

To train the various networks we used an ongoing image-based dataset, proposed by Pellis et al. (2021), specifically designed for heritage semantic segmentation. Datasets play an indispensable role in the Neural Network training phase and few datasets are freely available in heritage scenarios. Two of the most important datasets in this context are Architectural Elements Dataset (AHE dataset), a collection of 10,000 images classified in 10 types of architectural elements for the task of classification, and ARCH dataset (Matrone et al., 2020b), a benchmark for large scale heritage point cloud semantic segmentation, composed of 17 annotated scenes in 10 categories.
The proposed dataset (Pellis et al., 2021) will be composed by 9 heritage buildings, built in different period and characterized by different architectural style. It is an ongoing project, and, at the current stage, the dataset is composed by 5 buildings: *(1_SC)* Spedale del Ceppo, Pistoia, *(2_OSA)* Ospedale Sant'Antonio, Lastra a Signa, *(3_SSA)* Basilica della Santissima Annunziata, Firenze e *(4_CG)* Certosa del Galluzzo, Firenze, *(5_CB)* Cappella Buontalenti, Firenze. The segmentation categories considered in this dataset are structured following the guidelines defined in the ARCH dataset, which refers to the Industry Foundation Class (IFC) file format, to the CityGML (LOD3/4) and to the Art and Architecture Thesaurus (AAT). Hence, the dataset has been segmented in 10 classes, which correspond to certain BIM constructive elements, including arch, column, moulding, floor, door/window, wall, stair, vault, roof, other and background. Table 1 shows the percentage of pixels in each of the ten classes. For each building, a set of representative images from various perspective and angles, and their corresponding pixel-level ground truth are available. The total number of available images is currently about 4,700. All such images were collected during photogrammetric surveys, hence their alignment information is available as well, and it is possible to calculate the 3D point corresponding to each pixel.

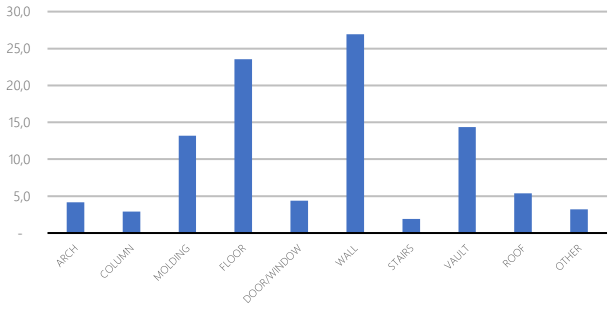

Table 1: General workflow of the developed methodology.

3.2 Image Segmentation

The first step of the workflow consists in the segmentation of the intermediate images according with the conventions of the ARCHdataset. Image segmentation is a key topic in a lot of computer vision application and various algorithms have been developed in literature including thresholding methods, k-means clustering, region growing and other. However, over the past few years, deep learning networks have yielded a new generation of models, achieving the highest accuracy on the most popular benchmarks. The most performing architecture are Fully

Convolutional Networks (FCN), U-Net, SegNet and the DeepLab family.
In this paper for the task of image segmentation we used one of the most popular network architectures, DeepLabv3+. This network is an improvement of the previous DeepLab and it employs atrous convolution with upsampled filters to extract dense feature maps and to capture long range context. The network has been implemented with MATLAB using ResNet18 as a base classification architecture. Training a deep learning model from scratch is often unfeasible due to the dataset size required and due to long time for the initial tuning of the network. For these reasons we used transfer learning, a technique that allows to grab a model trained on a certain task and dataset and employ it on a different task. We used a pretrained version of ResNet18 trained on ImageNet dataset, one of the most popular and largest benchmarks for image semantic segmentation.

3.3 Back-projection procedure

The second step involves the back-projection of the classified pixels into the 3D point cloud. This is the most challenging step of the procedure because feature propagation on 3D space involves inevitably a loss of geometrical and spatial information, a loss of detail of shapes and the propagation of errors. In our proposed methodology we leverage on the photogrammetry principles using a masking method for the creation of a directly labelled point cloud starting from the images. The back-projection workflow starts from the initial segmentation of the survey images performed by a trained neural network. Secondly, according with the segmentation results, for each image and for each segmentation class a binary mask is created. The masked images were then employed during dense image matching in order to constrain the point cloud creation process. The results of this process are a separate dense 3D point clouds for each class.
In the next section we are going to show some results of the developed procedure on our ongoing dataset.

4. RESULTS

4.1 Image Semantic Segmentation Results

This section reports some results on the semantic segmentation of the 2D images of historical buildings.
It is worth to notice that deep neural network training require thousands of images and case studies in order to ensure a proper and reliable performance. Given the current size of the ongoing dataset and the mentioned requirements, several quite simple tests with different data combinations are considered.

4.1.1 Test 1

In the first test, a small portion of the dataset is used, including just one building, in this case Spedale del Ceppo. For the purpose of training, all the labelled images of the building were randomly divided in 60% for the training (896 images), 20% as validation test (300 images), and 20% as test set (300 images). The training was performed for 30 epochs, yielding a validation accuracy of 94.5%. Once trained, the network was deployed to obtain the prediction on the entire test set. Table 2 reports some of the most significant results. Since just one building was used, the images of training and test set are quite similar, hence this task is very simple and not really significant for the network generalization.

Global Accuracy	Mean Accuracy	Mean IoU	Weighted IoU	Mean BFScore
93,4%	94,1%	82,8%	87,1%	84,2%

Table 2: DeepLabv3+ results on the test set (Test 1).

The performance of the network on the test set is remarkable, yielding to an overall accuracy of 93.4% and an Intersection Over Union (IoU) of 82.8%.
Figure 1 shows a comparison between the input images (A), the ground-truth (B) and the prediction (C) obtained on three of the test set images. Figure 2 shows the confusion matrix.

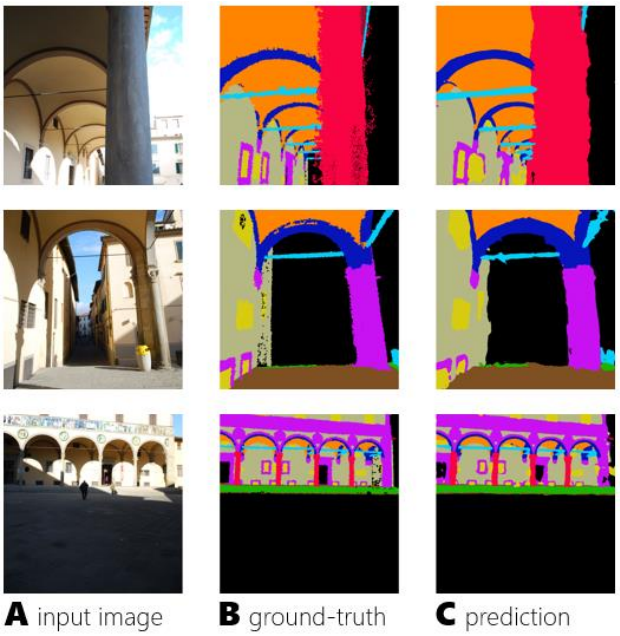

Figure 1: DeepLabv3+ results on the test set (Test 1): A) input image, B) ground-truth, C) prediction.

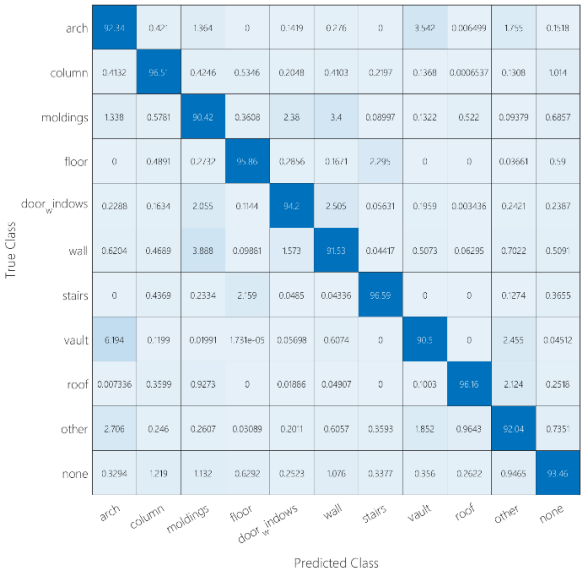

Figure 2: Confusion Matrix for Test 1 - Spedale del Ceppo

This test was repeated for all the other buildings of the dataset, obtaining the results reported in Table 3.

	Global Accuracy	Mean Accuracy
2_OSA	75%	66%
3_SS	91%	76%
4_CG	80%	55%
5_CB	74%	49%

Table 3: DeepLabv3+ results on the test set for all the other buildings (Test 1).

4.1.2 Test 2

Test 2 considers all the five buildings currently available in the dataset, picking 500 images for each building. The images were randomly shuffled and divided in three sets: 60% as training test (1400), 20% as validation test (466), and 20% as test set (466). As expected, in this case the performance of the network is lower than in Test 1, yielding to an overall accuracy of 88.1% and a IoU of 71.8% (Table 4). Test and training set are still quite similar, but it is remarkable that the model could well generalize the solution even when dealing with several buildings.

Global Accuracy	Mean Accuracy	Mean IoU	Weighted IoU	Mean BFScore
88,1%	88,6%	71,8%	80,9%	72,1%

Table 4: DeepLabv3+ results on the test set (Test 2).

Figure 3 shows a comparison between the input images, the ground-truth and the prediction obtained on three test set images. Figure 4 shows the confusion matrix obtained in Test 2.

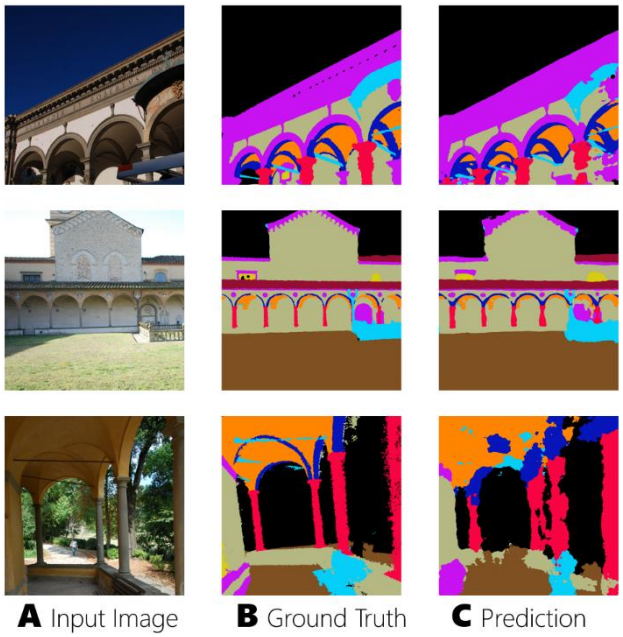

Figure 3: DeepLabv3+ results on the test set (Test 2): A) input image, B) ground-truth, C) prediction.

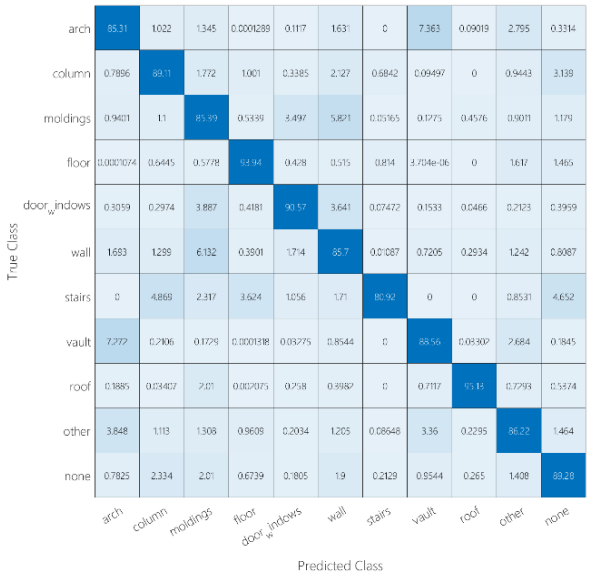

Figure 4: Confusion Matrix for Test 2.

4.1.3 Test 3

Finally, in Test 3 the network is tested on the prediction of an unseen scene. Four buildings are used to train the network, picking 500 images for each building, whereas the remaining building is used as test set (1_SC). Being more challenging, the performance in this test is clearly worse than in Test 1 and 2: the model reached a global accuracy of 54.2% and a IoU of 27.7%.

Global Accuracy	Mean Accuracy	Mean IoU	Weighted IoU	Mean BFScore
54,2%	41,1%	27,7%	37,3%	34,0%

Table 5: DeepLabv3+ results on the test set (Test 3).

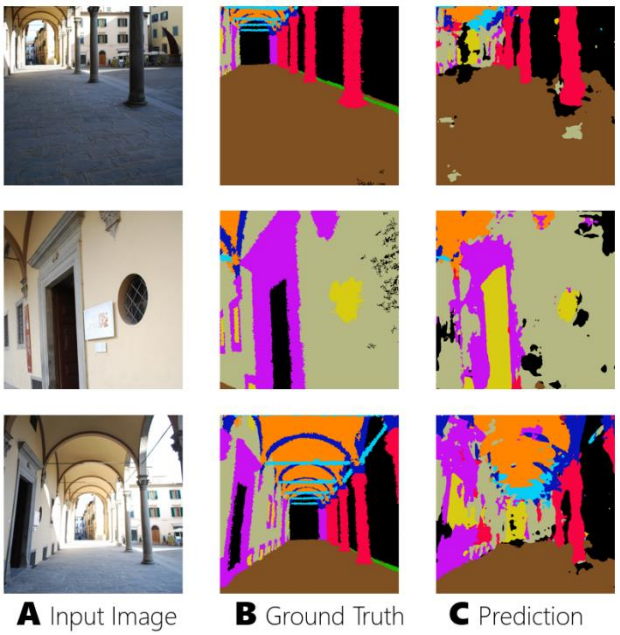

Figure 5: DeepLabv3+ results on the test set (Test 3): A) input image, B) ground-truth, C) prediction.

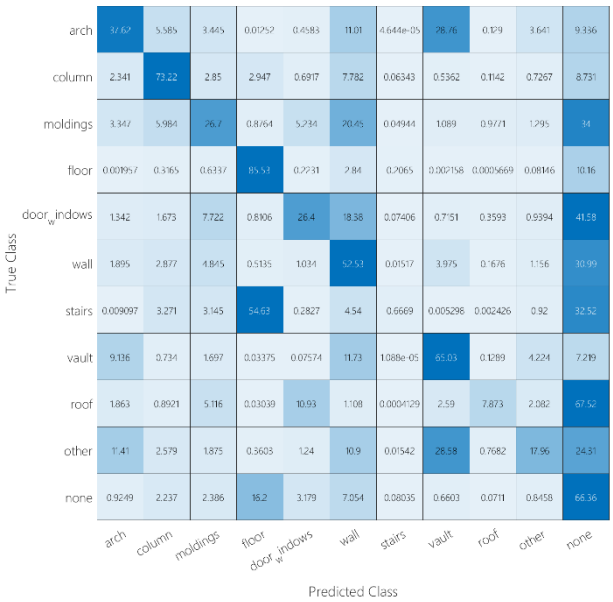

Figure 6: Confusion Matrix for Test 3.

The confusion matrix, in Figure 6, shows that the most common classes like “floor”, “vault”, “wall” and “column” are well segmented. This proves that the strong class imbalance, which characterize the ongoing dataset, is a critical factor for the network performance.

4.2 Back-projection procedure

Finally, this subsection shows the results of the back-projection procedure on two dataset buildings, *(1_SC)* Spedale del Ceppo and *(2_OSA)* Ospedale Sant'Antonio. Two back-projection cases are considered for comparison.
First, a point cloud is built using a set of masks created directly from the ground truth images used for the training phase of the network. Secondly, a point cloud is built using the set of masks created from the prediction output by the network. In both the cases the image prediction is obtained by means of the neural network trained in Test 1.
The comparison between such two reconstructions allows to evaluate the correct functioning of the procedure and the influence of the prediction accuracy on the final result.
Figure 8 shows A) the ground-truth point cloud, B) the point cloud reconstruction from the image ground-truth, and C) the point cloud reconstruction from the image prediction output by the neural network for the Spedale del Ceppo building.

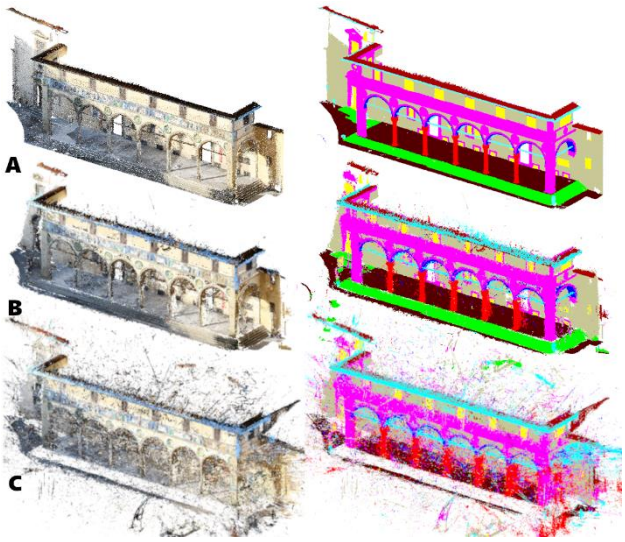

Figure 8: Back-Projection results for 1_SC: A) ground-truth cloud, B) image ground-truth reconstruction, C) prediction reconstruction.

Table 6 reports the Global Accuracy, the Mean Intersection Over Union and the DICE coefficient for the two reconstructions.

TYPE OF RECONSRUCTION	Global Accuracy	Mean IoU	Mean DICE
Ground Truth Images	82,6%	56,7%	68,9%
Prediction Images	33,4%	25,6%	29,7%

Table 6: Evaluation metrics for the back-projection of 1_SC.

Figure 8 shows that the projection procedure works quite well with the ground-truth images, but its performance degrades when working with the prediction images.
Figure 9 shows the results for the Ospedale Sant'Antonio building.

Figure 9: Back-Projection results for 2_OSA: A) ground-truth cloud, B) image ground-truth reconstruction, C) prediction reconstruction.
Table 7 reports some of the most remarkable evaluation metrics.

TYPE OF RECONSRUCTION	Global Accuracy	Mean IoU	Mean DICE
Ground Truth Images	69,7%	49,7%	61,2%
Prediction Images	57,3%	37,7%	52,7%

Table 7: Evaluation metrics for the back-projection of 2_OSA.

In this second case, the projection procedure reached a lower overall accuracy, but the performance on the cloud built with the prediction images has less degradation compared with the previous case. As expected, the results strictly depend on the quality of image segmentation prediction and hence on the accuracy of the neural network.
Table 8 compares the image segmentation with the 3D point cloud segmentation results.

	Prediction	Global Acc	Mean IoU
1_SC	Images	93,4%	82,8%
	Point Cloud	33,4%	25,6%
2_OSA	Images	75,0%	66,0%
	Point Cloud	57,3%	37,7%

Table 8: Comparison between the neural network-based image segmentation results and the corresponding results on the 3D point cloud.

Despite the high value of the accuracy in the image prediction for the first case (1_SC) the Global Accuracy on the back-projection is quite low, with a percentage loss of 60%. On the second case the accuracy on the images is lower, but the percentage loss on the back-projection is less than 20%. Considering only these two case studies is not possible to correctly evaluate and generalize the performance of the procedure: future tests will be performed to this aim.
However, these issues are related at least to two critical aspects: (i) the lack of accuracy of the prediction, in particular for what concerns object edges, and (ii) the problem of label overlapping. Issue (i) is related also to the dataset characteristics: a larger number of buildings and images should be considered, and the class imbalance should be reduced. In addition to increasing the dataset size increasing the number of the buildings, using data augmentation and class weighting during training should be useful as well.

Instead, the label overlapping issue (ii) should be at least partially compensated by introducing a regularization step in the back-projection procedure.
Since the results obtained using the ground truth segmented images are promising (low percentage loss: 10% for Spedale del Ceppo and 6% for Spedale Sant'Antonio), an improvement in the automatic image segmentation shall lead to a significant increase of the overall point cloud classification performance.

5. CONCLUSIONS

This paper presented a procedure for the semantic segmentation of 3D point clouds of heritage buildings The procedure is composed by two main steps: (i) the labelling of the intermediate images, using a pre-trained (deep learning-based) neural network, and then (ii) the projection of the image features on the 3D point cloud, thanks to a masking method. The results are quite promising, but more in depth tests shall be performed to assess the general performance of the entire workflow. Furthermore, the foreseen generalization and size increase of the training dataset shall improve the performance of the automatic image segmentation step. Finally, an improvement of back-projection procedure, in particular to better deal with the label overlapping issue, will be investigated as well. Such changes are expected to improve the point cloud segmentation results.

REFERENCES

Badrinarayanan,V., Kendall, A., Cipolla, R., 2015. Segnet: a deep convolutional encoder-decoder architecture for scene segmentation. *IEEE Trans. PatternAnal. Mach. Intell., 39 (12)*, 2481–2495. 10.1109/TPAMI.2016.2644615.

Boulch, A.; Guerry, J.; Le Saux, B.; Audebert, N., 2018. SnapNet: 3D point cloud semantic labeling with 2D deep segmentation networks. *Computers and Graphics (Pergamon)*, 71, 189–198.

Chen, L. C., Papandreou, G., Kokkinos, I., Murphy, K., Yuille, A. L., 2017. Deeplab: Semantic image segmentation with deep convolutional nets, atrous convolution, and fully connected CRFs. *IEEE transactions on pattern analysis and machine intelligence, 40(4)*, pp. 834–848.

Cordts, M., Omran, M., Ramos, S., Rehfeld, T., Enzweiler, M., Benenson, R., Franke, U., Roth, S., Schiele, B., 2016. The cityscapes dataset for semantic urban scene understanding. *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp 3213–3223.

Grilli, E., Remondino, F., 2019. Classification of 3D digital heritage. *Remote Sensing*, *11*(7).

Grilli, E., Dininno, D., Petrucci, G., Remondino, F., 2018. From 2D to 3D supervised segmentation and classification for cultural heritage applications. *International Archives of the Photogrammetry, Remote Sensing and Spatial Information Sciences - ISPRS Archives*, *42*(2), 399–406.

López, F.J., Lerones, P.M., Llamas, J., Gómez-García-Bermejo, J., Zalama, E., 2018. A Review of Heritage Building Information Modeling (H-BIM). *Multimodal Technologies and Interaction*, 2, 21. https://doi.org/ 10.3390/mti2020021.

Malinverni, E. S., Pierdicca, R., Paolanti, M., Martini, M., Morbidoni, C., Matrone, F., Lingua, A., 2019. Deep Learning for Semantic Segmentation of 3D Point Cluod. *International Archives of the Photogrammetry, Remote Sensing and Spatial Information Sciences - ISPRS Archives*, *42*(2/W15), 735–742.

Matrone, F., Grilli, E., Martini, M., Paolanti, M., Pierdicca, R., Remondino, F., 2020. Comparing machine and deep learning methods for large 3D heritage semantic segmentation. *ISPRS International Journal of Geo-Information*, *9*(9). https://doi.org/10.3390/ijgi9090535.

Murtiyoso, A., Lhenry, C., Landes, T., Grussenmeyer, P., Alby, E., 2021. Semantic segmentation for building façade 3D point cloud from 2D orthophoto images using transfer learning. *International Archives of the Photogrammetry, Remote Sensing and Spatial Information Sciences - ISPRS Archives*, *43*(B2-2021), 201–206.

Murtiyoso, A., Pellis, E., Grussenmeyer, P., Landes, T., Masiero, A., 2022. Towards semantic photogrammetry: generating semantically rich point clouds from close range photogrammetry. *Sensors*, 22.

Pellis, E., Masiero, A., Tucci, G., Betti, M., Grussenmeyer, P., 2021. Assembling an Image and Point Cloud Dataset for Heritage Buildings Semantic Segmentation. *The International Archives of the Photogrammetry, Remote Sensing and Spatial Information Sciences*, *XLVI-M-1–2021*, 539–546.

Qi, C. R., Su, H., Mo, K., Guibas, L. J., 2017a. Pointnet: Deep learning on point sets for 3D classification and segmentation. *In Proceedings of the IEEE conference on computer vision and pattern recognition* (pp. 652-660).

Qi, C. R., Yi, L., Su, H., Guibas, L. J., 2017b. Pointnet++: Deep hierarchical feature learning on point sets in a metric space. *arXiv preprint arXiv*:1706.02413.

Shan, J., Toth, C. K. (Eds.)., 2018. *Topographic laser ranging and scanning: principles and processing*. CRC press.

Su, H.; Maji, S.; Kalogerakis, E.; Learned-Miller, E., 2015. Multi-view convolutional neural networks for 3D shape recognition. *Proceedings of the IEEE International Conference on Computer Vision*, Vol. 2015 Inter, pp. 945–953.

Yastikli, N., 2007. Documentation of cultural heritage using digital photogrammetry and laser scanning. *Journal of Cultural heritage*, 8(4), 423-427.

Zhang, J., Zhao, X., Chen, Z., Lu, Z., 2019. A Review of Deep Learning-Based Semantic Segmentation for Point Cloud. *IEEE Access, vol. 7*, pp. 179118-179133. doi: 10.1109/ACCESS.2019.2958671.