Assembling an image and point cloud dataset for heritage building semantic segmentation

E. Pellis, A. Masiero, G. Tucci, M. Betti, P Grussenmeyer

ASSEMBLING AN IMAGE AND POINT CLOUD DATASET FOR HERITAGE BUILDING SEMANTIC SEGMENTATION

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

[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] Université de Strasbourg, INSA Strasbourg, CNRS, ICube Laboratory UMR 7357, Photogrammetry and Geomatics Group, 67000 Strasbourg, France - pierre.grussenmeyer@insa-strasbourg.fr

KEY WORDS: H-BIM, dataset, image segmentation, 3D point clouds, semantic segmentation, deep learning

ABSTRACT:

Creating three-dimensional as-built models from point clouds is still a challenging task in the Cultural Heritage environment. Nowadays, performing such task typically requires the quite time-consuming manual intervention of an expert operator, in particular to deal with the complexities and peculiarities of heritage buildings. Motivated by these considerations, the development of automatic or semi-automatic tools to ease the completion of such task has recently became a very hot topic in the research community. Among the tools that can be considered to such aim, the use of deep learning methods for the semantic segmentation and classification of 2D and 3D data seems to be one of the most promising approaches. Indeed, these kinds of methods have already been successfully applied in several applications enabling scene understanding and comprehension, and, in particular, to ease the process of geometrical and informative model creation. Nevertheless, their use in the specific case of heritage buildings is still quite limited, and the already published results not completely satisfactory. The quite limited availability of dedicated benchmarks for the considered task in the heritage context can also be one of the factors for the not so satisfying results in the literature.
Hence, this paper aims at partially reducing the issues related to the limited availability of benchmarks in the heritage context by presenting a new dataset for semantic segmentation of heritage buildings. The dataset is composed by both images and point clouds of the considered buildings, in order to enable the implementation, validation and comparison of both point-based and multiview-based semantic segmentation approaches. Ground truth segmentation is provided, for both the images and point clouds related to each building, according to the class definition used in the ARCHdataset, hence potentially enabling also the integration and comparison of the results obtained on such dataset.

1. INTRODUCTION

In the Cultural Heritage (CH) environment, Heritage Building Information Modelling (H-BIM) has gained particular attention in recent years, due to the growing interest in protection, conservation, and restoration of historical buildings (López et al., 2018; Volk et al., 2014). The process of 3D point cloud for the creation of three-dimensional and informative models is still a challenging task, it requires time consuming and manual intervention by specialized operators, and there is a lack of standard procedures and methodologies to manage and speed up this process (Rodriguez et al., 2016; Tang et al., 2010). A key factor in the automation of this workflow is the development of a reliable semantic segmentation technique, i.e. the ability of properly partitioning a 3D point cloud into classes based on a semantic interpretation of the scene (Macher et al, 2015). Indeed, semantic segmentation is a fundamental step in scene understanding and comprehension, and, thanks to the continuously increasing number of applications that can take advantage from machine understanding of a scene, it is a very active research field (Xie et al., 2020). In recent years, several techniques have been successfully used in a wide of vision application: algorithmic approach (Murtiyoso and Grussenmeyer, 2020), machine learning (ML) (Croce et al., 2021), Neural Networks (NN), and especially Deep Learning (DL) with the introduction of Convolutional Neural Networks (CNN) for 2D image processing (Zhang et al., 2019; Long et al., 2015). The remarkable results obtained by NN and DL in 2D image understating led to their application also in the heritage 3D point cloud semantic segmentation problem (Masiero et al., 2019; Grilli et al., 2017). The main approaches to face semantic segmentation of 3D point clouds are: (i) multiview-based, which leverages on an intermediate image representation, and (ii) point-based, which deals directly with point clouds. The use of point-based methods for the semantic segmentation of heritage buildings has been recently investigated (Matrone et al., 2020a; Grilli and Remondino, 2020). Instead, to the best of our knowledge, the use of multiview-based methods in this framework has not been experimented yet, probably due to the lack of a dedicated benchmark for 2D image semantic segmentation of cultural heritage buildings. Indeed, a large dataset is fundamental for training, validating, and tuning a NN. Furthermore, its availability can also enable the comparison between different machine learning algorithms. Despite multiview-based approaches may introduce some information loss, due to the use of an intermediate data representation, dealing with 2D images could still be an effective strategy. On one hand, this allows to exploit the well-established NN-based 2D image semantic segmentation techniques, and, on the other hand, nowadays LiDAR and photogrammetric surveys are often integrated, hence their combination can represent a viable way to transfer the semantic knowledge extracted from images to the corresponding point cloud. In addition, the integration with point-based methods may lead to a hybrid method that may improve the performance of both such approaches.

* Corresponding author

This paper aims at presenting a new image-point cloud dataset for semantic segmentation of heritage buildings. Such dataset is composed by thousands of images and tens point clouds of buildings. Ground truth segmentation on both point clouds and images is available for each case study: such dataset shall be considered as a new *benchmark* for the development of machine and deep learning methods in the heritage sector, i.e., it can be used for the learning and validation phases, and for the comparison of new and already existing approaches. To be more specific, this work focuses on the description of the characteristics of such dataset and of the semi-automatic procedure used for producing the ground-truth data.

2. RELATED WORKS

Several datasets are available to assess the performance of different algorithms and networks architectures in different applications (Yu et al., 2018). In this section a summary of some of the most used datasets for segmentation purposes is provided.

PASCAL VOC – Visual Object Classes (Vicente et al., 2014) is one of the most popular datasets and the images are annotated for 5 different tasks, classification, segmentation, detection, action recognition and person layout. For the segmentation task there are 21 classes of object labels. This dataset is divided in two sets, training and validation, with 1,464 and 1,449 images, respectively.

MS COCO – Microsoft Common Object in Context (Lin et al., 2014) is a large-scale collection of images for object detection and segmentation. It is composed by 328k images with a total of 91 object types and 2.5 million labelled instances, mainly representing everyday scenes and common object in their natural contexts.

ADE20K (Zhou et al., 2017) is a scene parsing benchmark with 150 objects and stuff classes. There are 20,210 images in the training set, 2000 images in the validation set, and 3000 images in the test set.

The Cityscapes Dataset (Cordts et al., 2016) is a collection of diverse set of stereo video sequences recorded in street scene from 50 cities, mainly focus on semantic understanding of urban scenarios. It consists of 30 classes grouped in 8 categories in 5k fine annotated images.

CMP Façade Database is a dataset of façade images assembled at the Centre for Machine Perception, which includes 606 rectified images of façade from various sources, which have been manually annotated. The façades are from different cities around the world and diverse architectural styles, labelled in 12 classes.

In the context of heritage environment, few datasets are available, and a specific dataset for semantic segmentation of image of historical buildings (Fiorucci et al., 2020) is still missing. The most remarkable heritage datasets are reported below.

ArCH – Architectural Cultural Heritage dataset (Matrone et al., 2020b) is a benchmark for large scale heritage point cloud semantic segmentation. It is composed of 17 annotated scenes, derived from the union of several scans and their integration with photogrammetric surveys. The point clouds are labelled in 10 classes, including the main BIM standard elements.

CHAS – Cultural Heritage Architectural Segmentation (Pavia et al., 2019) is a point cloud dataset from cultural heritage aimed to provide data for semantic segmentation techniques. The data were generated by terrestrial laser scanning and UAV photogrammetric surveys. The dataset comprises relevant buildings representing religious and colonial Brazilian architecture.

AHE – Architectural Heritage Elements (Llamas et al., 2017) is an image dataset developed for the task of classification of architectural heritage images. The dataset consists of 10235 RGB images classified in 10 categories, including some construction elements like Domes, Altars or Bell towers. Most of the images have been obtained from Flickr and Wikimedia Commons, all of them under creative common license.

MonuMAI – Monument with Mathematics and Artificial Intelligence (Lamas et al., 2021) is a public image dataset labelled using two annotation types, which make it useful for several tasks, such as monument style classification, for the detection of key elements, and other potential applications. It contains 1514 RGB images grouped in four architectural styles. Some key elements are also identified using bounding boxes, which report element names and locations.

3. DATASET

The dataset presented in this paper is currently under construction: at the current stage, the dataset is composed by partial scenes of four buildings. Nevertheless, extending the dataset, reaching nine heritage building scenes, is a short-term goal of our project. In addition to such extension, the integration of other buildings and scenes will also be considered in our future work, in order to enable more in depth investigations and more reliable results of deep learning-based methods.
The nine buildings are located in Tuscany (Italy), they are built in different historical periods, and they are characterized by different architectural styles. Nevertheless, they share some common features, such as the presence of loggia, the presence of classic order, and the proportion between the various elements. These characteristics are typical of the renaissance style. Other buildings will be integrated in the dataset in the future in order to properly introduce also some other architectural elements and styles in the dataset.
Figure 1 shows the four buildings already present in the dataset.

- *(1_SC)* Spedale del Ceppo, Pistoia
- *(2_OSA)* Ospedale di Sant'Antonio, Lastra a Signa
- *(3_SSA)* Basilica Santissima Annunziata, Firenze
- *(4_CG)* Certosa del Galluzzo, Firenze

Figure 1. The first four building of the dataset.

The following buildings are shortly going to complete the dataset: *(5_CB)* Cappella Buontalenti, Firenze, *(6_PV)* Palazzo Vecchio, Firenze, *(7_CGR)* Ca' Granda, Milano, *(8_PP)*, Palazzo Pitti, Firenze, *(9_GA)* Galleria dell'Accademia, Firenze. A portion of the dataset was collected in an educational context by the GECO research group (Geomatics and Conservation group of the Department of Civil and Environmental Engineering, University of Florence). Terrestrial Laser Scanner (TLS) data are already available for all buildings. The TLS datasets are being integrated with close-range photogrammetric surveys, four of them being already completed. This paper anticipates on the already completed acquisitions the data structure and characteristics of the overall datasets.

3.1 Class Definition

The main aim of this dataset is to support the development tools for the automatic determination of heritage architectural elements in the context of Building Information Modelling (BIM). Therefore, segmentation classes are chosen according to the standards of the main BIM-based design software.
However, the differences between BIM and H-BIM are remarkable: they mainly arise from the uniqueness and peculiarity of the historical constructive elements, causing data enrichment, information storage and 3D interoperability to be still open challenges in this context (Quattrini et al., 2017). During the last years, several works tried to tackle these issues, proposing some workflows to exploit BIM in the CH domain.
The segmentation categories considered in this dataset are structured following the conventions defined in ARCHdataset, which refers in particular to the Industry Foundation Class (IFC) data model, being the latter the main open file format for BIM interoperability. Since IFC is a standard for new buildings, and it is not properly suited to represent complex heritage elements, other two standards have also been taken into account in the ARCHdataset: CityGML (LOD3/4), and AAT (Art and Architecture Thesaurus) of Getty Institute (see also Figure 2).

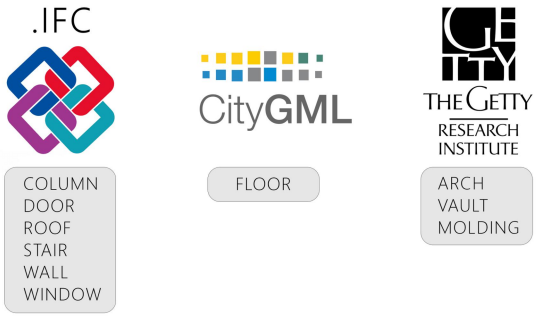


Figure 2. The standards used to determine the classes of architectural elements to be considered.

Hence, the dataset has been segmented in 10 classes, which correspond to the BIM constructive categories, including wall, floor, roof, column, moulding, vault, arch, stair, window/door and other. Due to the heterogeneity of the heritage architectural elements, ARCHdataset also provides the detailed guidelines to correctly annotate the scene.
Despite ARCHdataset is composed only by point clouds, its classes can be easily identified also in images, hence their use can be naturally extended to images, e.g. Figure 3 shows an example of image segmented according to the ARCHdataset classes.

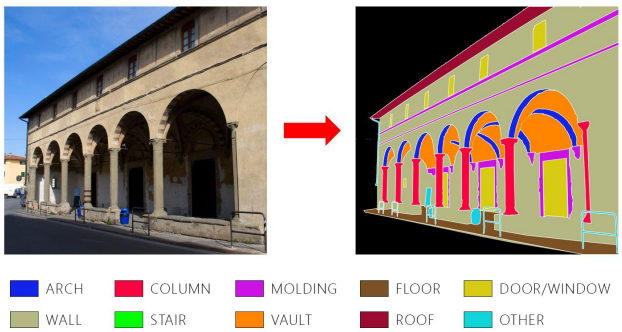


Figure 3. Segmented image according to the ARCHdataset classes.

ARCHdataset is currently the only benchmark realized to deal with point cloud-based machine and deep learning tools in the heritage field, and it is promoting crowdsourcing to enrich the already annotated scene. Consequently, the choice of maintaining the same class definition of ARCHdataset can be convenient for: (i) comparing the performance of an approach in both the datasets, (ii) enabling a potential integration of the two datasets, hence increasing the number and typologies of labelled buildings. Nevertheless, this dataset may be distributed in the future also with annotations made according to different class definitions.

3.2 Labelling Procedure

A semantic segmentation benchmark, usable to train and test segmentation with neural networks, requires assembling a large dataset, e.g. thousands of different labelled images and several point clouds.
In particular, the availability of an automatic image labelling procedure is fundamental to reduce the time to produce the ground truth labelling, and hence minimize the manual time-consuming operations.
To such aim, a semi-automatic procedure, based on the availability of an already labelled point cloud, has been

developed, aiming at properly segmenting the images acquired during photogrammetric surveys (Figure 4).
First, both a TLS and a photogrammetric georeferenced 3D reconstruction of the same building are assumed to be available.
After some pre-processing steps, such as cleaning, denoising and down sampling, the TLS cloud, which usually is more accurate, is manually segmented according to the ARCHdataset guidelines.
Despite the TLS and photogrammetric point clouds can already be georeferenced, the alignment between them can usually be still improved by means of the Iterative Closest Point algorithm. Thanks to the refined alignment between the two clouds, the labels previously set on the TLS cloud can be easily transferred to the photogrammetric reconstruction, based on a closest-point criteria (step 3 in Figure 4). Furthermore, the photogrammetric point cloud can be automatically cleaned and denoised as well by setting up a maximum admissible distance threshold between them. Finally, step 4 in Figure 4, the labelled points are reprojected on the initial images: the outcome of this operation, once properly regularized, corresponds to the automatic segmentation of the images involved in the photogrammetric reconstruction.

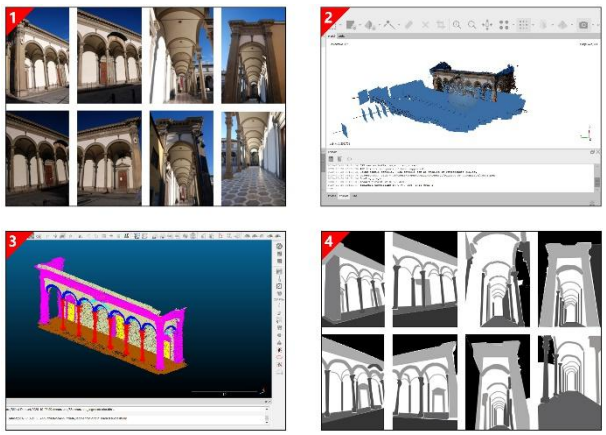
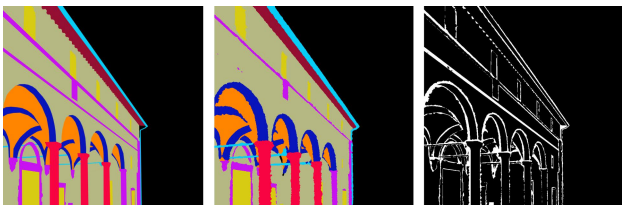


Figure 4. The semi-automatic labelling procedure.

Several tests were performed on the transfer labelling procedure from the cloud to the images, in order to guarantee the highest quality level and as many pixels as possible correctly annotated. The obtained results have shown that some issues can be caused by: (i) low density of the point cloud, (ii) local noise of the cloud, (iii) missing pixels in some parts of the building, (iv) presence of objects and obstacles in front of the buildings, (v) local difference between the LiDAR and the photogrammetric cloud. These issues can be at least partially tackled by optimizing the automatic labelling procedure settings, e.g. the point distance threshold between the two clouds, the point obstruction neighbourhood area in images. Removing cloud geometric noise and reducing gaps, if any, in the 3D clouds can also be useful to improve the automatic labelling procedure results.
Figure 5 shows an example of manually segmented image (Figure 5a), the corresponding automatically segmented image (Figure 5b) and the overlay-difference between the two: in black the pixels classified in the same way, whereas pixels wrongly classified by the automatic labelling procedure are shown in white (Figure 5c).



a) b) c)

Figure 5. a) Fine-manually segmented image, b) automatic segmented image, c) overlay between the two images.

The percentage of correctly annotated pixels by the automatic label transferring procedure is currently more than 97% on the considered testing images. Since different labelling results can be obtained by changing the settings and the strategy of the automatic procedure, the influence of such factors on the training of deep learning classifiers will be assessed as well.
The most tedious and time-consuming step in the entire procedure is the manual segmentation of the initial cloud. However, the automatic label transferring procedure provides a viable way to easily expanding the labelled image dataset, even exploiting the existence of certain already segmented clouds (e.g. ARCHdataset).

3.3 Dataset Structure

Designing a large-scale dataset requires taking multiple decisions, e.g on the data preparation, annotation protocol and data splitting. Our dataset has been created following the structure of the main image semantic segmentation datasets. We provide a set of RGB images, and the corresponding set of labelled images with the same size, in the *.png* file format. The images were acquired with different cameras, and the starting size was set to 2592x3872 pixel with a resolution of 300 dpi. Some pre-processing operations were initially performed to make the images more homogeneous. First, some images were rotated by 90 degrees, to maintain the right verticality of the building. Then, they were cropped in a square format, in order to obtain, for all the images, the same pixel height and width. Finally, the images were downscaled to an appropriate size for deep learning purposes. Currently, the downscaled image size is 720x720 pixel, but other options may also be implemented in the future.
Hence, the final structure of each image in the benchmark, usable for deep learning training or testing, is 720×720×3, whereas the size its corresponding ground truth file is 720×720×1.
Labels in the ground-truth file are compatible with those of ARCHdataset: 0 arch, 1 column, 2 moulding, 3 floor, 4 door/window, 5 wall, 6 stair, 7 vault, 8 roof, 9 other.
Differently from point clouds, background is always present in the images, hence a new class was introduced: it includes all the pixels that cannot be classified as part of the previously defined classes. Such class is conventionally named "background", and it is labelled with the index 10.

Figures 5, 6, 7, 8 show the photogrammetric point clouds of the four initial buildings, and the corresponding ground truth segmentation.

1_SC - Spedale del Ceppo

a)

b)

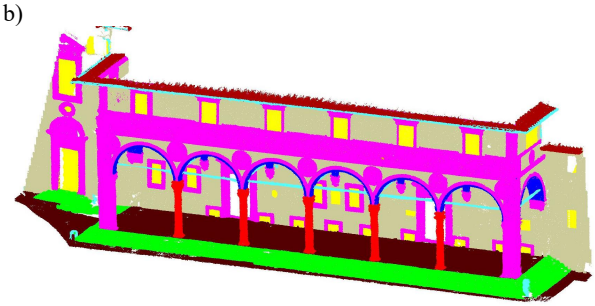


Figure 5. a) RGB photogrammetry point cloud and b) corresponding labelled cloud of Spedale del Ceppo.

2_OSA – Ospedale Sant’Antonio

a)

b)

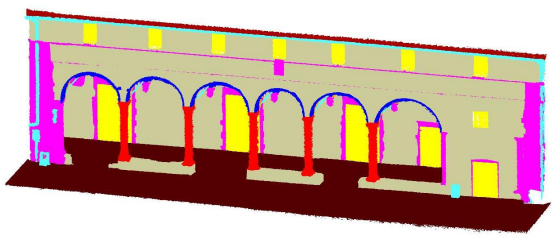


Figure 6. a) RGB photogrammetry point cloud and b) corresponding labelled cloud of Ospedale Sant’Antonio.

3_SSA – Basilica Santissima Annunziata

a)

b)

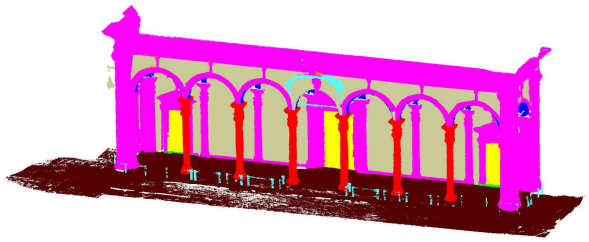


Figure 7. a) RGB photogrammetry point cloud and b) corresponding labelled cloud of SS Annunziata

4_CG – Certosa del Galluzzo

a)

b)

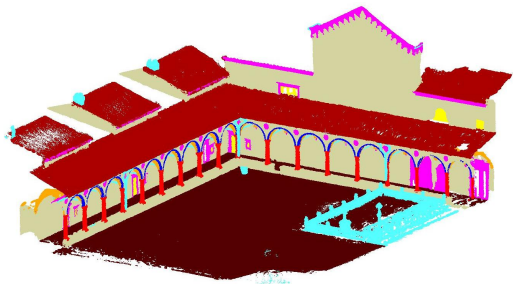


Figure 8. a) RGB photogrammetry point cloud and b) corresponding labelled cloud of Certosa del Galluzzo.

Histogram in Figure 9 shows the percentage of points in the considered classes, for both the TLS and photogrammetric clouds, considering all the four buildings. Figure 9 shows also an overall significant imbalance of the classes.

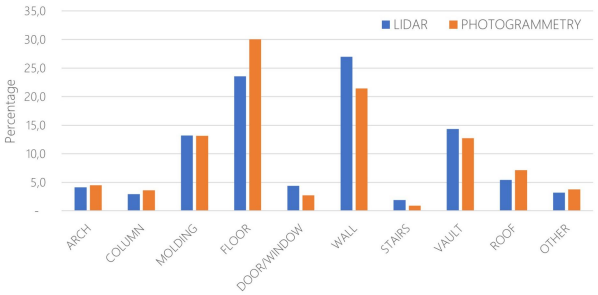


Figure 9. Distribution of the classes in the two clouds.

Imbalance of the classes is a common issue in several semantic segmentation datasets (Ren et al., 2020), and, if not properly handled, it can be detrimental to the learning process: indeed, in

this case learning results will be biased in favour of dominant classes. Class weighting and the use of the correct evaluation metric are mainly used to properly face this problem. In addition, future acquisitions could privilege buildings with the prevalence of low percentage classes.
The total number of the images used for the construction of the initial clouds are 3,078 and, after the pre-processing operation, the current number of generated images composing the dataset is 6,156. Despite the large amount of data, which is quite enough to train a deep network, future integration will be fundamental to increase the capabilities of the dataset with new architectural styles, constructive elements, and object typologies, enabling networks to learn and generalize new scenes.
Figure 10 shows some examples of labelled images in the benchmark dataset. All the photogrammetric survey images are included in the portion of the image dataset corresponding to such building survey. Typically they are thousands of different views of the building, at different distances, angles, and perspectives.

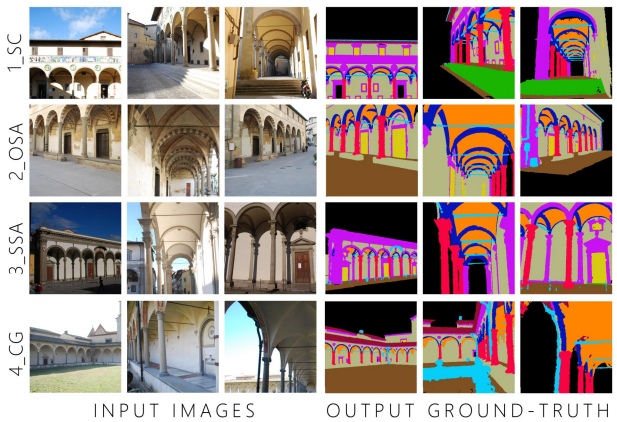


Figure 10. Examples of labelled images in the dataset.

The histogram in Figure 11 shows the distribution of the classes on the total images after the labelling projection, and with the introduction of the class “background”. The classes are still imbalanced, with a predominance of the pixels labelled as “background”. Differently from the point cloud case, a specific image selection could be performed to harmonize (e.g. balancing) the classes, and to reduce the amount of background pixel that may negatively influence the network learning procedure.

Figure 11. Distribution of the classes in the images.

Finally, Table 1 reports the current total number of points for each class, both for LiDAR and photogrammetric clouds, and their point percentage distribution. In addition, also the total number of pixels for each class and the percentage on the total number of pixels is reported.

DATA TYPE							
INDEX	CLASS	N° POINTS	% TOTAL	N° POINTS	% TOTAL	N° PIXELS	% TOTAL
0	ARCH	1,276,521	4,1	1,954,038	4,5	175,694,000	7,5
1	COLUMN	901,765	2,9	1,567,339	3,6	152,427,000	6,5
2	MOLDING	4,064,406	13,2	5,706,566	13,1	31,3395,000	13,3
3	FLOOR	7,250,888	23,5	13,039,116	30,0	308,524,000	13,1
4	DOOR/WINDOW	1,346,016	4,4	1,187,204	2,7	68,303,900	2,9
5	WALL	8,302,865	27,0	9,296,059	21,4	689,156,000	29,3
6	STAIRS	592,046	1,9	401,379	0,9	16,167,880	0,7
7	VAULT	4,423,300	14,4	5,524,446	12,7	476,757,000	20,2
8	ROOF	1,663,267	5,4	3,099,932	7,1	57,528,700	2,4
9	OTHER	983,015	3,2	1,644,076	3,8	98,035,800	4,2
10	BACKGROUND	-	-	-	-	784,020,000	33,3
	TOTAL	30,804,089		43,420,155		2,355,989,280	

Table 1. Summary of the number of points and pixels of the current dataset.

Another important aspect in a benchmark design is data splitting. The optimal proportion depends on the full size of the dataset. Nevertheless, the commonly used strategy allocates 60% of cases on training set, 20% on validation set, and 20% on the test set (Dobbin and Simon, 2011).
Since the dataset is still incomplete, the final splitting will be performed in the future. The splitting strategy will be designed in such a way to test in particular the network ability in properly deal with unseen scenes.

4. FUTURE DEVELOPMENT

The dataset presented in this paper is still under construction: some actions are still needed before making it freely available to the research community. First, some buildings shall be added to the dataset, being their TLS surveys already available, whereas their photogrammetric surveys and the image analysis/labelling is still to be completed. Once completed, the dataset will be split in a training set, composed of 8 buildings, and a test set composed of one building. According with the generalization and the capabilities obtained by the first training test, other new study cases could be added in the future.
Although the label transferring procedure has achieved a high accuracy in our tests, it may still be improved and tested on new cases.
Once the dataset will be completed, the following two goals will be pursed: (i) testing of existing neural networks on the new benchmark to evaluate its solidity and robustness (ii) developing a reliable procedure to project the image labels on the point cloud.
For the first point, the main state-of-the-art deep networks will be tested on the new dataset: DeepLabv3+ (Chen et al., 2017), U-Net (Ronnenberg et al., 2015), SegNet (Badrinarayanan et al., 2015) and FCN (Long et al., 2015).
For the second point, since the final aim of our research is to obtain the segmentation of 3D point cloud, and to speed-up the construction of BIM-based model, a reliable procedure to link 2D labelled pixels with 3D points is crucial. Several works are facing this problem (Wang et al., 2019; Lertniphonohan et al., 2018; Murtiyoso et al., 2021) and in our future research we are going to test existing approaches and to develop new methods that could be suitable on our heritage scenarios.

5. CONCLUSION

In this work, we presented the assembling of a new dataset for benchmarking semantic segmentation in the scenario of heritage buildings. We described the dataset structure and characteristics, focusing in particular on the four buildings already inserted in the dataset. Currently the corresponding image part of the

dataset is composed by more than 6,000 labelled images, and, in the upcoming months other five buildings are going to increase the dataset size. Once a segmentation of the building point cloud is available, the implemented automatic label transferring procedure can be used to quickly label images as well.
The main aim of the benchmark is to offer the possibility to implement and compare multi-view approaches on heritage building scenarios and leverage on the existing 2D segmentation architecture to ease the development of new classification machine learning and deep learning techniques.
TLS cloud and the photogrammetric clouds, both segmented following the same class definition used for the images, are available in addition to the images for each building. These multiple-source data can useful to perform comparisons and assessments such as: (i) compare the accuracy of point-based and multi-view based methods on the same dataset, (ii) compare the accuracy of multi-view based approach on heritage benchmark with that obtained on standard buildings, (ii) assess the accuracy of point-based networks on two types (TLS and photogrammetric) of point cloud data. Hence, the presented dataset can be (i) integrated with ARCHdataset, (ii) used to tailor existing network architectures on the CH building case, (iii) exploited to develop new hybrid networks that can leverage on both images and point clouds.

REFERENCES

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

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. arXiv:1606.00915.

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.

Croce, V., Caroti, G., De Luca, L., Jacquot, K., Piemonte, A., et al., 2021. From the Semantic Point Cloud to Heritage-Building Information Modeling: A Semiautomatic Approach Exploiting Machine Learning. *Remote Sensing, MDPI,* 13 (3), pp. 461. https://doi.org/10.3390/rs13030461.

Dobbin, K. K., Simon, R.M., 2011. Optimally splitting cases for training and testing high dimensional classifiers. *BMC Med Genomics* 4, 31. https://doi.org/10.1186/1755-8794-4-31.

Fiorucci, M., Khoroshiltseva M., Pontil, M., Traviglia A., Del Bue A., James, S., 2020. *Pattern Recognition Letters*, 133, pp. 102-108. https://doi.org/10.1016/j.patrec.2020.02.017.

Grilli, E., Menna, F., Remondino, F., 2017. A review of point cloud segmentation and classification algorithms. Int. Arch. Ph. Remote Sens. Spatial Inf. Sci., XLII-2/W3, 339–344. https://doi.org/10.5194/isprs-archives-XLII-2-W3-339-2017.

Grilli, E., Remondino, F., 2020. Machine Learning Generalisation across Different 3D Architectural Heritage. *International Journal of Geo-Information*, 9(6), 379. https://doi.org/10.3390/ijgi9060379.

Lamas, A., Tabik, S., Cruz, P., Mintes, R., Martinez-Sevilla A., Cruz, T., Herrera, F., 2021. MonuMAI: Dataset, deep learning pipeline and citizen science based app for monumental heritage taxonomy and classification. *Neurocomputing,* volume 420, pp. 266-280. https://doi.org/10.1016/j.neucom.2020.09.041.

Lertniphonphan, K., Komorita, S., Tasaka K., Yanagihara, K., 2018. 2D to 3D Label Propagation For Object Detection In Point Cloud. *IEEE International Conference on Multimedia & Expo Workshops (ICMEW)*, pp. 1-6, doi: 10.1109/ICMEW.2018.8551515.

Lin, T.Y., Maire, M., Belongie, S., Bourdev, L., Girshick, R., Hays, J., Perona, P., Ramanan, D., Zitnick, C.L., Dollar, P., 2014. Microsoft coco: common objects in context. arXiv preprint arXiv:1405.0312.

Llamas, J.; M. Lerones, P.; Medina, R.; Zalama, E.; Gómez-García-Bermejo, J., 2017. Classification of Architectural Heritage Images Using Deep Learning Techniques. *Appl. Sci.*, *7*, 992. https://doi.org/10.3390/APP7100992.

Long, J., Shelhamer, E., Darrell, T., 2015. Fully convolutional networks for semantic segmentation. *Proceedings of the IEEE conference on computer vision and pattern recognition*, 3431–3440. arXiv:1411.4038

Long, J., Shelhamer, E., Darrell, T., 2015. Fully convolutional networks for semantic segmentation. *Proceedings of the IEEE conference on computer vision and pattern recognition*, 3431–3440. arXiv:1411.4038.

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.

Macher, H., Landes, T., Grussenmeyer, P., 2015. Point clouds segmentation as a base for as-built BIM creation. *ISPRS Annals of the Photogrammetry, Remote Sensing and Spatial Information Sciences, Volume II-5/W3*.

Masiero, A., Chiabrando, F., Lingua, A. M., Marino, B. G., Fissore, F., Guarnieri, A., Vettore, A., 2019. 3D Modeling of Girifalco Fortress. *International Archives of the Photogrammetry, Remote Sensing & Spatial Information Sciences*, vol. 42W9, pp. 473-478, doi:10.5194/isprs-archives-XLII-2-W9-473-2019.

Matrone, F., Lingua, A., Pierdicca, R., Malinverni, E., Paolanti, M., Grilli, E., Remondino, F., Murtiyoso, A., Landes, T., 2020b. A benchmark for large-scale heritage point cloud semantic segmentation. *ISPRS - International Archives of the Photogrammetry, Remote Sensing and Spatial Information Sciences. XLIII-B2*. 1419-1426. https://doi.org/10.5194/isprs-archives-XLIII-B2-2020-1419-2020.
Matrone, F.; Grilli, E.; Martini, M.; Paolanti, M.; Pierdicca, R.; Remondino, F., 2020a. Comparing Machine and Deep Learning Methods for Large 3D Heritage Semantic Segmentation. ISPRS Int. J. Geo-Inf., 9, 535. https://doi.org/10.3390/ijgi9090535.

Murtiyoso, A., Grussenmeyer, P., 2020. Virtual Disassembling of Historical Edifices: Experiments and Assessments of an Automatic Approach for Classifying Multi-Scalar Point Clouds into Architectural Elements. *Sensors,* 20, 2161.https://doi.org/10.3390/s20082161.

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 & Spatial Information Sciences*, vol. XLIII-B2-2021. https://doi.org/10.5194/isprs-archives-XLIII-B2-201-2021.

Paiva, P., Cogima, C. K., Dezen-Kempter, E., & Carvalho, M. A., 2019. CHAS - Cultural Heritage Architectural Segmentation dataset. *Zenodo*. https://doi.org/10.5281/zenodo.2609498.

Quattrini, R., Pierdicca, R., Morbidoni, C., 2017. Knowledge-based data enrichment for H-BIM: Exploring high-quality models using the semantic-web. *Journal of Cultural Heritage*. Vol. 28, pp. 129-139. https://doi.org/10.1016/j.culher.2017.05.004.

Ren, Y., Zhang, X., Ma, Y., Yang, Q., Wang, C., Liu, H., Qi, Q., 2020. Full Convolutional Neural Network Based on Multi-Scale Feature Fusion for the Class Imbalance Remote Sensing Image Classification. *Remote Sensing*, 12, 3547. https://doi.org/10.3390/rs12213547.

Rodriguez, C., Reinoso, J. F., Rivas-Lopez, E., Gomez-Blanco, A., Ariza-Lopez F., Ariza, I., 2016. From point cloud to BIM : an integrated workflow for documentation, research and modelling of architectural heritage. *Survey Review,* 50(360), 1-20. https://doi.org/10.1080/00396265.2016.1259719.

Ronneberger, O., Fischer, P., Brox, T., 2015. U-Net: convolutional networks forbiomedical image segmentation. *Medical Image Computing andComputer-Assisted Intervention (MICCAI),* vol. 9351 of LNCS, pp. 234–241.

Tang, P., Huber, D., Akinci, B., Lipman, R., Lytle, A., 2010. Automatic reconstruction of as-built building information models from laserscanned point clouds: A review of related techniques. *Automation in construction*, 19(7), 829–843.

Vicente, S., Carreira, J., Agapito L., Batista, J., 2014. Reconstructing PASCAL VOC. *2014 IEEE Conference on Computer Vision and Pattern Recognition*. pp. 41-48. 10.1109/CVPR.2014.13.

Volk, R., Stengel, J., Schultmann, F., 2014. Building Information Modelling (BIM) for existing buildings - Literature review and future needs. *Automation in Construction*, 38, 109-127. https://doi.org/10.1016/j.autcon.2013.10.023.

Wang, B.H., Chao, W., Wang, Y., Hariharan, B., Weinberger, K.Q., Campbell, M., 2019. LDLS: 3-D Object Segmentation Through Label Diffusion From 2-D Images. *IEEE Robotics and Automation Letters*, vol. 4, no. 3, pp. 2902-2909. doi: 10.1109/LRA.2019.2922582.

Xie, Y., Tian, J. & Zhu, X., 2020. Linking Points with Labels in 3D: A Review of Point Cloud Semantic Segmentation. *IEEE Geoscience and Remote Sensing Magazine*. https://doi.org/10.1109/MGRS.2019.2937630.

Yu, H., Tan, Y., Wang, L., Yaonan, S., Sun, W., Mi, Yandong, M. T., 2018. Methods and Datasets on Semantic Segmentation: A review. *Neurocomputing*, 304. https://doi.org/10.1016/j.neucom.2018.03.037.

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.

Zhou, B., Zhao, H., Puig, X., Fidler, S., Barriuso, A., Torralba, A., 2017. Scene parsing through ade20k dataset. *Proceedings of the IEEE conference on computer vision and pattern recognition.*