FROM MULTI-VIEW TO POINT CLOUD SEGMENTATION: THE CASE STUDY OF VILLA ROBERTI BRUGINE

A. Masiero[1,*], A. Guarnieri[2], U. Coppa[3], G. Tucci[1]

[1] University of Florence, Italy - (andrea.masiero, grazia.tucci)@unifi.it
[2] University of Padova, Italy - alberto.guarnieri@unipd.it
[3] Vesuvius Observatory - National Institute of Geophysics and Volcanology, Naples, Italy - coppa@ov.ingv.it

Commission II

KEY WORDS: Point cloud segmentation, Back-projection, Semantic Segmentation, UAV

ABSTRACT:

Point cloud semantic segmentation is a key step for automatically deriving an informative building model from the 3D data reconstruction obtained by 3D surveying tools, such as laser scanners and photogrammetry. Such representation increases the richness of the information available of the represented building, leading to an at least rough interpretation of the scene, and, in particular, a discrimination between the different constitutive elements of the building. The growing interest in semantic building models recently motivated the development of several approaches aiming at obtaining an automatic semantic segmentation of a building point cloud. Such methods are usually either based on the direct segmentation of the point cloud, or on the segmentation of images of the building, then back-projecting the obtained segmentation on the point cloud. Similarly to the latter approach, this work assumes that a proper neural network is available in order to compute the semantic segmentation of building images, and it compares two different strategies for transferring such semantic information from the 2D images to the 3D point cloud. The results obtained in the case study of villa Roberti Brugine (Brugine, Padua, Italy) show that transferring the semantic information can be done quite effectively with the proposed, even when dealing with a certain amount of misclassified points. In particular, best results are obtained in our tests when determining a point class as the most popular classification of such point once projected on all the images where it is visible.

1. INTRODUCTION

Automatic point cloud semantic segmentation is a quite challenging problem that has been recently investigated in a number of research works. The semantic segmentation approaches proposed in the literature can roughly be divided in two categories, depending on if they are based on either 2D or 3D features (Grilli and Remondino, 2019). The approaches based on image analysis typically fall in the first category, whereas those directly working on 3D point clouds are within the second one.

Methods based on (usually multi-view) image analysis can exploit the consolidated results in image segmentation obtain during the last decade thanks to the development of convolutional neural networks and, more in general, deep learning techniques (Iandola et al., 2016). Such kind of techniques require to properly segment and classify the available images and to transfer such information to the 3D point cloud space (Murtiyoso et al., 2021).

Differently, the second category relies on the direct application of machine or deep learning methods to point clouds. Among such methods, it is possible to distinguish a first subcategory explicitly exploiting geometric features, such as eigenvalues of the 3D covariance matrix (Weinmann et al., 2015), and a second one using deep learning classifiers on the point cloud (Matrone et al., 2020, Maalek et al., 2018), often derived from PointNet and its new variants (Garcia-Garcia et al., 2016, Qi et al., 2017).

A full overview of the literature is out of the scope of this paper, hence the reader is referred to (Grilli et al., 2017, Grilli and Remondino, 2019) for quite recent and comprehensive reviews of the works on this topic.

Despite several works already considered this kind of problem, it still can be considered open, as the obtained results, in particular in the case of heritage buildings, are expected to be improvable: high rate recognition accuracy has been obtained in certain quite specific cases, however the obtained results in general cases are still not completely satisfactory.

Among the two categories previously described, this work falls in the first category, that of methods relying on multi-view image segmentation. In particular, it is thought for those cases where a 3D point cloud of a building can be generated by means of an Unmanned Aerial Vehicle (UAV) photogrammetric survey (Nex and Remondino, 2014), which nowadays is a quite common case (Masiero et al., 2019).

To be more specific, a UAV photogrammetric survey of the building of interest is assumed to be already available, and an image semantic segmentation tool is assumed to be available as well (the readers are referred for instance to (Pellis et al., 2021) for the description of a possible choice for such tool).

Given the working conditions mentioned above, this work focuses on transferring the semantic information from the 2D image space to the 3D point cloud. More specifically, it aims at comparing two semantic information transferring methods, and in particular it aims at investigating the effects of the image semantic segmentation errors on the point cloud segmentation accuracy.

[*] Corresponding author

This contribution has been peer-reviewed.
https://doi.org/10.5194/isprs-archives-XLVI-2-W1-2022-359-2022

2. CASE STUDY AND METHODS

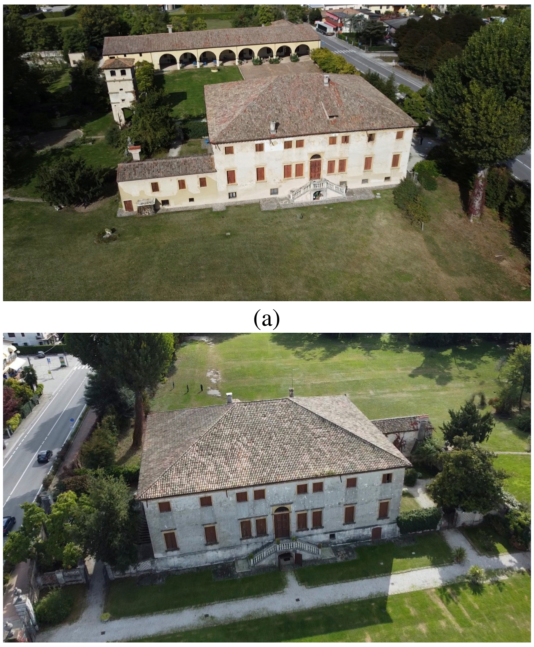
This work considers the case study of Villa Roberti Brugine, a historical building located in the small town Brugine (close to Padua, Italy), in order to assess the point cloud segmentation performance reachable by means of multi-view-based methods, and, more specifically, to make an investigation on the propagation of the image segmentation errors on the point cloud segmentation.

A 3D point cloud of Villa Roberti Brugine has been obtained by means of a UAV photogrammetric survey, by using a DJI Mavic Mini 2. DJI Mavic Mini 2 is an affordable ($\leq$ \$ 500), lightweight drone ($\leq$ 250 g), hence authorized to fly over urban areas in Europe, with provided with a 12 Mpixel camera. Table 1 summarizes the main camera characteristics.

Sensor	1/2.3" CMOS, 12 Mpixel
Focal length (equiv., 35 mm format)	24 mm
Field of View	83°
Max. image resolution	4000 pix × 3000 pix
Max. video resolution	3840 pix × 2160 pix

Table 1. DJI Mini 2 camera characteristics.

Two UAV views of Villa Roberti Brugine are shown in Figure 1.



(a)

(b)

Figure 1. UAV views of Villa Roberti Brugine.

The point cloud semantic segmentation procedure considered here assumes the use of an image semantic segmentation deep learning network (see for instance (Pellis et al., 2021)). Such network typically provides low resolution segmented images of the considered object.

Then, the segmentation information shall be transferred to the point cloud, with a back-projection-like operation (Murtiyoso et al., 2021, Murtiyoso et al., 2022).

Since segmentation of the same 3D area are typically available on several images, determining a proper procedure to effectively deal with such information is a key step for obtaining a reliable back-projection of the image segmentation results.

In this paper two different methods are compared to make such back-projection:

A) back-projecting just the information of the "optimal view" (closest and close-to-normal direction) of each area,

B) back-projecting the obtained classification from all the images from which an area is visible. Then, each 3D point of the area is classified according to the class with the highest number of occurrences (among the considered images) for such point.

In the first strategy, the choice of the "optimal view" for each area is motivated by the reasonable assumption that such image shall be the ideal one in order to obtain the most reliable one-image-based segmentation of such area.

Instead, the second strategy is based on the rationale that on each image view certain pixels are wrongly classified, but the majority of the pixels shall be correctly classified. Consequently, such approach is expected to be more robust with respect to mistakes in the segmented images, at least up to a certain extent.

To be more precise, a reference segmentation of the villa 3D reconstruction is assumed to be already available. A view of such segmentation of the 69 Mpoints of the cloud describing Villa Roberti Brugine is shown in Figure 2.

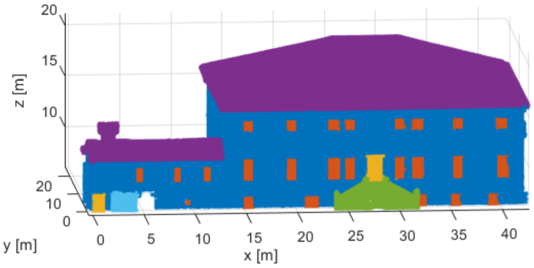


Figure 2. Reference segmentation of Villa Roberti Brugine.

Then, the semantic information of the point cloud segmentation is projected on a set of 101 images, distributed approximately along a circle centered in the villa. The projection allows to obtain the semantic segmentation of such 101 images (see for instance Figure 3).

Then, the obtained semantically segmented images are used to (re)compute and assess a semantic segmentation of the 3D point cloud, comparing the previously described approaches A) and B).

It is worth to notice that since the segmented images mentioned above have been obtained by projecting the reference point cloud segmentation onto the images, working with them can be considered as a quite ideal working condition.

In fact, any 2D semantic segmentation in a real application is affected by some classification/segmentation errors, which hence shall affect in some way the consequent 3D point cloud segmentation.

Hence, in order to make an analysis of the effect of image classification errors on the obtained segmented point cloud, the set

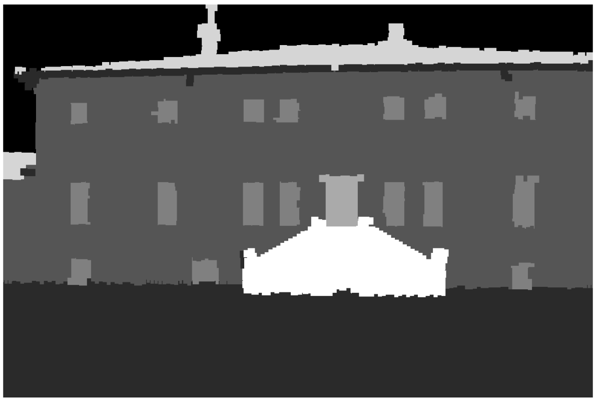


Figure 3. Example of labeled image of building in Figure 1.

of originally segmented images is progressively recomputed increasing the level of classification errors, introducing classification errors related to each other, e.g. spatially correlated. Figure 4 shows some examples of synthetic images obtained from Figure 3 by progressively increasing the classification error (approximately at the following levels: 2%, 5% and 7%). These figures have been obtained by introducing randomness in the parameters of a multi-resolution representation of such images.

The obtained results are shown in the following section.

3. RESULTS AND DISCUSSION

The proposed method has been tested, as previously described, on a subsampled version of the Villa Roberti Brugine point cloud, in particular, randomly selecting 3.6 Million points. These points have been used to generate the segmented images (see for instance Figure 3 and 4), and, finally, the segmented point clouds have been obtained either using the method A) or B).

First, Table 2 compares the 3D point classification performance with the approaches A) and B). The results obtained with the two method are also graphically represented in Figure 5 and 6, where the two segmented point clouds are shown.

	Case A)	**Case B)**
Accuracy [%]	86.6	92.3

Table 2. Classification results using the original synthetic labeled images.

It is worth to notice that the results shown in Table 2 consider only the classified points. Indeed, a certain amount of the 3D points have not been classified, because not visible in any of the selected 101 images.

Clearly, the number of views where a point is visible can be quite different from point to point. Figure 7 and 8 aim at investigating the role that the number of views from which a point is visible have on the effectiveness of its classification. To be more specific, Figure 7 shows the number of 3D points correctly (and incorrectly) classified in the approach A) as a function of the number of views from which the points are visible. Similarly, Figure 8 shows the same kind of results but for approach B).

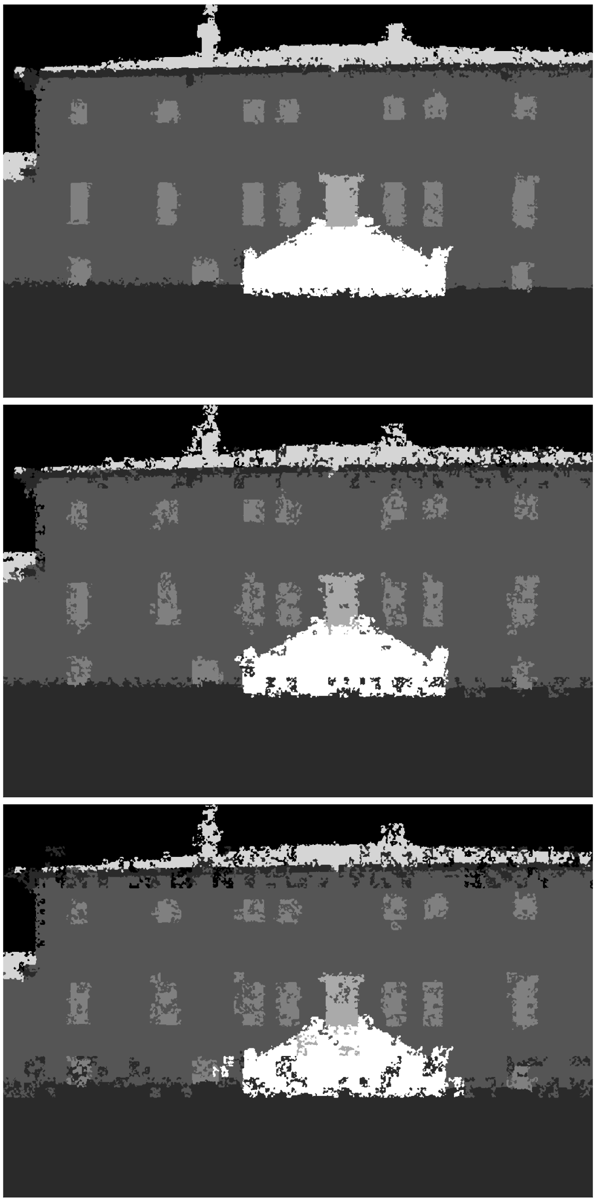


Figure 4. Example of synthetic labeled images derived from Figure 3 increasing the level of random noise.

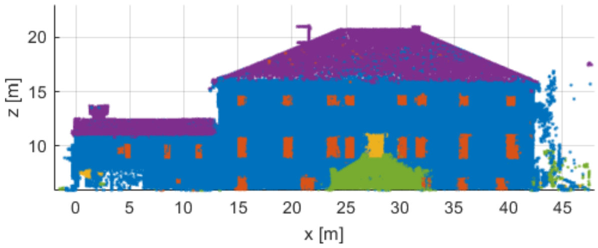


Figure 5. Predicted semantic segmentation of Villa Roberti Brugine: best view approach.

Intuitively, the higher number of views for a point, the higher the chance of having certain good views, which should correspond to good image segmentations. Despite such intuition is confirmed by Figure 7 and 8, a comparison between the two figures also shows that the number of incorrect classifications decreases faster for case B), which hence should better exploit the availability of more views of a point.

Then, the effect of incorrect classifications in the 2D images

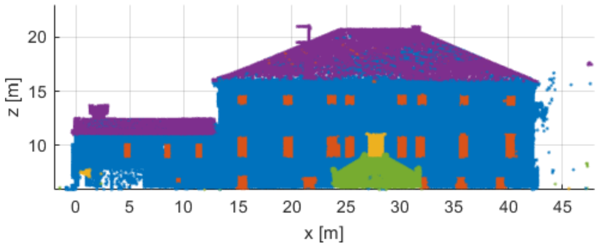


Figure 6. Predicted semantic segmentation of Villa Roberti Brugine: most popular vote approach.

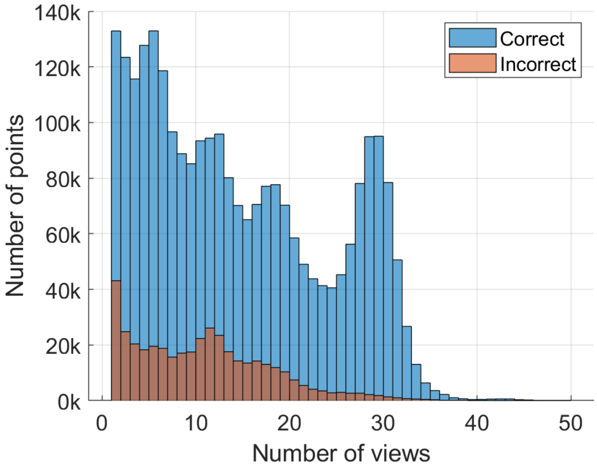


Figure 7. Counts of correctly and incorrectly classified points as functions of the number of views from which they are visible.

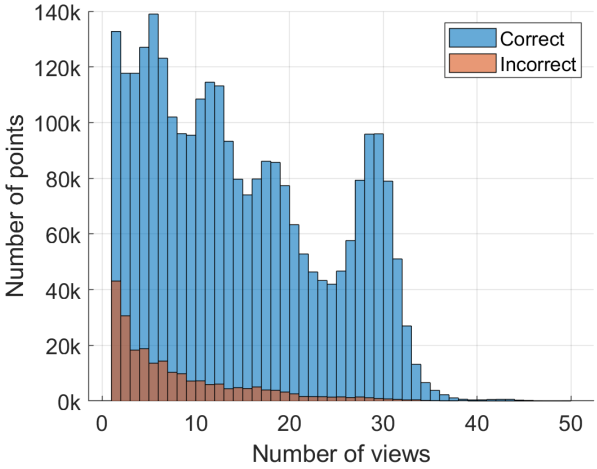


Figure 8. Counts of correctly and incorrectly classified points as functions of the number of views from which they are visible.

on the 3D segmentation is evaluated varying the average percentage of points incorrectly classified from 2% to 7%. The obtained 3D point classification results are shown in Table 3.

	Case A)	Case B)
Accuracy [%] with 2% noise	86.0	91.9
Accuracy [%] with 5% noise	84.0	91.0
Accuracy [%] with 7% noise	82.8	90.7

Table 3. Classification results using the labeled noisy images, varying the level of noise.

Furthermore, the front views of the obtained segmented villa for all the six combinations of segmentation transferring methods (A) and B)) and of incorrect 2D segmentation percentage (2%, 5%, 7%) are shown in Figure 9 and 10. Figure 9 shows the results for case A), increasing the average level of incorrectly classified 2D points from the top to the bottom of the figure. Similarly, Figure 10 shows the corresponding results for case B).

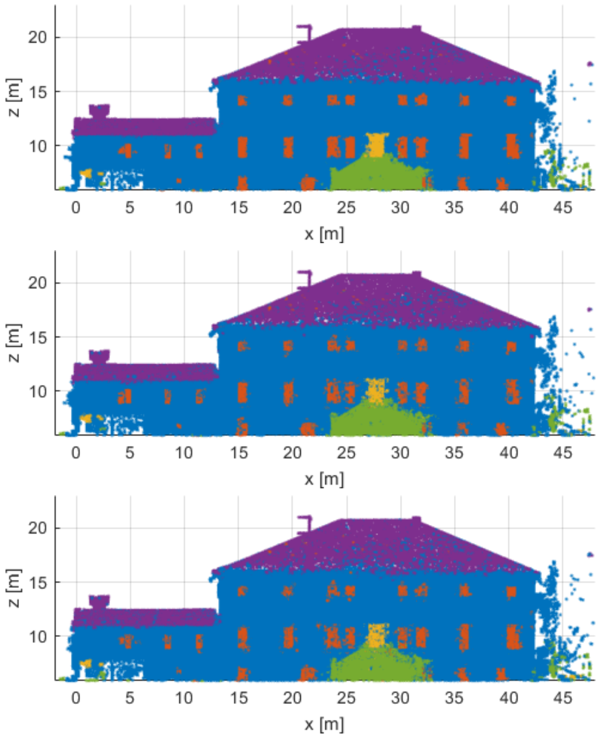


Figure 9. Predicted semantic segmentation of Villa Roberti Brugine: best view approach (case A)) increasing (from the top to the bottom row) the noise level in the synthetic labeled images.

Both the numerical (Table 3) and graphical results (Figure 9 and 10) confirms that case B) is more robust to the presence of incorrect segmentations of the 2D images.

In particular, Table 3 shows that an increase of 5% of the percentage of incorrectly classified pixels leads in this case study to an increase of the incorrectly classified 3D points of less than 2%. Hence, this confirms the quite good robustness of approach B) with respect to incorrectly classified 2D pixels. Nevertheless, despite the numerical decrease of performance is not so large, its impact on the graphical results is quite apparent (for instance, compare Figure 6, with top and bottom rows in 10).

Overall, the approach B) performed better than A) in all the considered tests, showing a higher effectiveness both in numerical and graphical results.

Despite the obtained results on the comparison between approach A) and B) seem to be quite conclusive, a more extensive comparison should be done in order to obtain more reliable results. Indeed, the generalization of the results obtained in just one case study and one camera network configuration (converging views, with cameras distributed along a circle centered in the building) may lead to not so reliable conclusions.

Furthermore, the implemented multi-resolution method for generating synthetic labeled images affected by incorrect pixel

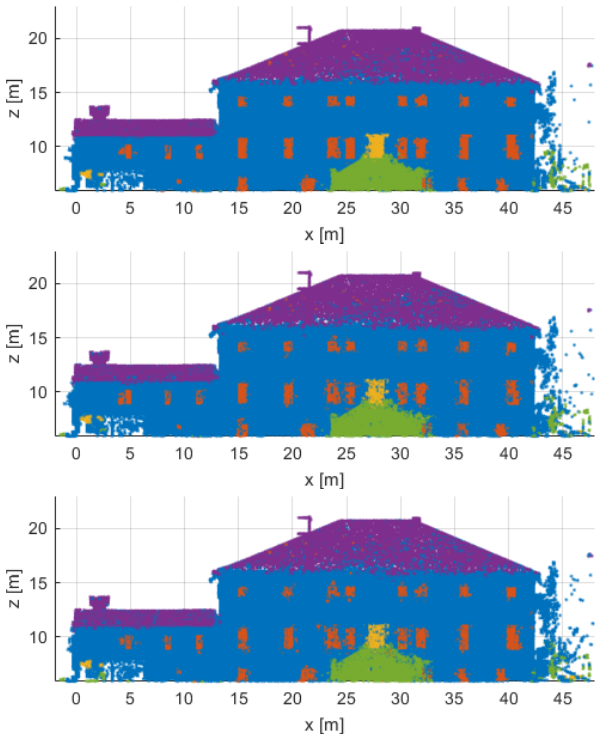


Figure 10. Predicted semantic segmentation of Villa Roberti Brugine: most popular vote approach (case B)) increasing (from the top to the bottom row) the noise level in the synthetic labeled images.

classifications should be improved in order to more naturally modify the original images, reproducing incorrect classifications more similar to the typical neural network errors on this task (e.g. mostly errors on the object edges).

4. CONCLUSIONS

This paper considered the problem of semantic segmentation of 3D point clouds. In particular, a multi-view approach has been adopted, assuming that a tool for properly identifying the objects of interest is already available, this paper focused in particular on assessing the performance of the considered strategies in properly transferring the 2D segmentation information to the 3D point cloud.

Furthermore, the classification performance has also been investigated considering the results of the strategies as functions of the image segmentation error level. The segmented images involved in this test have been obtained by modifying the original ones in a multi-resolution representation. Such representation should also be considered and improved, for instance exploiting a wavelet representation of each image (Mallat, 1999), in order to make the results more consistent with a natural view of the considered objects.

According to the obtained result, incorrectly classified 2D points impacted much more on the overall performance than the semantic information transferring approach A).

The visibility of a point by a large number of views also showed to be a quite important factor to reach a high classification performance.

Even if approach B) showed to be quite robust, the availability of well-segmented 2D images clearly leads to better point cloud segmentation results. Consequently, improvement also on this direction should be taken into account in the authors' future work, for instance better exploiting the spatial characteristics of the images in order to improve the segmentation results (for instance exploiting also other segmentation and detection methods (Masiero et al., 2015)), or introducing in the 2D segmentation algorithm information also by other sensors mounted on the platform (Sankey et al., 2017, Zahran et al., 2018).

A more in-depth analysis will be performed by the authors in their future investigations in order to ensure a more reliable generalization of the obtained results.

REFERENCES

Garcia-Garcia, A., Gomez-Donoso, F., Garcia-Rodriguez, J., Orts-Escolano, S., Cazorla, M., Azorin-Lopez, J., 2016. Pointnet: A 3d convolutional neural network for real-time object class recognition. *2016 International joint conference on neural networks (IJCNN)*, IEEE, 1578–1584.

Grilli, E., Menna, F., Remondino, F., 2017. A review of point clouds segmentation and classification algorithms. *ISPRS - International Archives of the Photogrammetry, Remote Sensing and Spatial Information Sciences*, 42, 339.

Grilli, E., Remondino, F., 2019. Classification of 3D Digital Heritage. *Remote Sensing*, 11(7). https://www.mdpi.com/2072-4292/11/7/847.

Iandola, F. N., Han, S., Moskewicz, M. W., Ashraf, K., Dally, W. J., Keutzer, K., 2016. SqueezeNet: AlexNet-level accuracy with 50x fewer parameters and¡ 0.5 MB model size. *arXiv preprint arXiv:1602.07360*.

Maalek, R., Lichti, D. D., Ruwanpura, J. Y., 2018. Robust segmentation of planar and linear features of terrestrial laser scanner point clouds acquired from construction sites. *Sensors*, 18(3), 819.

Mallat, S., 1999. *A wavelet tour of signal processing*. Elsevier Academic Press, Burlington (U.S.A.).

Masiero, A., Chiabrando, F., Lingua, A., Marino, B., Fissore, F., Guarnieri, A., Vettore, A., 2019. 3D modeling of Girifalco fortress. *International Archives of the Photogrammetry, Remote Sensing & Spatial Information Sciences*.

Masiero, A., Guarnieri, A., Pirotti, F., Vettore, A., 2015. Semi-Automated Detection of Surface Degradation on Bridges Based on a Level Set Method. *ISPRS - International Archives of Photogrammetry, Remote Sensing and Spatial Information Sciences*, 40(3), 15–21.

Matrone, F., Lingua, A., Pierdicca, R., Malinverni, E., Paolanti, M., Grilli, E., Remondino, F., Murtiyoso, A., Landes, T., 2020. A benchmark for large-scale heritage point cloud semantic segmentation.

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. *The International Archives of Photogrammetry, Remote Sensing and Spatial Information Sciences*, 43, 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.

Nex, F., Remondino, F., 2014. UAV for 3D mapping applications: a review. *Applied Geomatics*, 6(1), 1–15.

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

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

Sankey, T., Donager, J., McVay, J., Sankey, J. B., 2017. UAV lidar and hyperspectral fusion for forest monitoring in the southwestern USA. *Remote Sensing of Environment*, 195, 30–43.

Weinmann, M., Jutzi, B., Hinz, S., Mallet, C., 2015. Semantic point cloud interpretation based on optimal neighborhoods, relevant features and efficient classifiers. *ISPRS Journal of Photogrammetry and Remote Sensing*, 105, 286–304.

Zahran, S., Mostafa, M., Masiero, A., Moussa, A., Vettore, A., El-Sheimy, N., 2018. Micro-RADAR and UWB aided UAV navigation in GNSS denied environment. *ISPRS - International Archives of the Photogrammetry, Remote Sensing and Spatial Information Sciences*, XLII-1, 469–476.