AUTOMATIC POINT CLOUD NOISE MASKING IN CLOSE RANGE PHOTOGRAMMETRY FOR BUILDINGS USING AI-BASED SEMANTIC LABELLING

A. Murtiyoso[1]* and P. Grussenmeyer[2]

[1] Forest Resources Management Group, Institute of Terrestrial Ecosystems, Department of Environmental Systems Science, ETH Zürich, Switzerland – arnadidhestaratri.murtiyoso@usys.ethz.ch
[2] Université de Strasbourg, CNRS, INSA Strasbourg, ICube Laboratory UMR 7357, Photogrammetry and Geomatics Group, Strasbourg, France - pierre.grussenmeyer@insa-strasbourg.fr

Commission II

KEY WORDS: Photogrammetry, AI, Semantic segmentation, Image masks, Automation, Point cloud cleaning, 3D reconstruction.

ABSTRACT:

The use of AI in semantic segmentation has grown significantly in recent years, aided by developments in computing power and the availability of annotated images for training data. However, in the context of close-range photogrammetry, although working with 2D images, AI is still used mostly for 3D point cloud segmentation purposes. In this paper, we propose a simple method to apply such methods in close range photogrammetry by benefitting from deep learning-based semantic segmentation. Specifically, AI was used to detect unwanted objects in a scene involving the 3D reconstruction of a historical building façade. For these purposes, classes e.g., sky, trees, and electricity poles were considered as noise. Masks were then created from the results which would then constraint the dense image matching process to only the wanted classes. In this regard, the resulting dense point cloud essentially projected the 2D semantic labels into the 3D space, thus excluding noise and unwanted object classes from the 3D scene. Our results were compared to manual image masking and managed to achieve comparable results while requiring only a fraction of the processing time when using a pre-trained DL network to do the task.

1. INTRODUCTION

In close range photogrammetry, the use of image masks is an optional step which could nevertheless help the user in generating cleaner end results (e.g., point clouds). The image masks serve the purpose of excluding objects deemed unnecessary on the scene, such as sky, person, etc. which would otherwise create noise on the resulting point clouds, 3D model, as well as textures. While less common in aerial and surveying-oriented photogrammetry, the addition of this masking step can be very useful in close range photogrammetry as it reduces greatly the time required to clean the final 3D product from noisy elements (Verhoeven, 2011).

The common way to create masks is via image processing software, which may use algorithms to easily detect the background of the object. In large-scale close-range photogrammetry applied to small objects this is one of the reasons why the use of single colour backgrounds (e.g., green screen) and rotating platforms is popular as it greatly facilitates the creation of masks. However, this technique is neither directly applicable nor very practical when dealing with outdoor scenes such is the case in building façade 3D reconstruction.

The application of AI-based semantic segmentation to help 3D reconstruction has recently seen many discussions, both in the form of indirect segmentation (Murtiyoso et al., 2021) and direct segmentation in the 3D space (Boonpook et al., 2021). It has also been applied for remote sensing purposes (Song, 2020). In the field of close range terrestrial photogrammetry, the proposed automatic creation of image can also be found in similar previous studies, namely Stathopoulou and Remondino (2019a) with further development in Stathopoulou and Remondino (2019b). The problem of semantic segmentation of images involves the automatic labelling of image pixels (Kirillov et al., 2019). From such segmented images, a binary mask can be quite easily created by simply considering required classes and excluding the others. An example of this method was presented by Grilli et al. (2021) in the process of detecting fruits in an agricultural setting. Kernel (2018) proposed a similar method in the context of aerial photogrammetry to exclude both the sky and water classes in an MVS-based 3D mesh model.

In this paper, we propose a method involving deep learning (DL)-based semantic segmentation to mask object noise in the context of close-range photogrammetry. The method was particularly tested on the case of the 3D reconstruction of historical buildings which is a common task encountered in our research. The following Section 2 will describe the proposed method in more detail, while Section 3 shall showcase and discuss the results obtained in one of our case studies. Finally, Section 4 will summarise the findings and determine possible improvements to the technique.

2. METHODOLOGY

In this paper, we propose a method for automatically extracting image masks to exclude unwanted information in the context of close-range photogrammetry applied to building facades. The developed method is based on the use of artificial neural networks trained on a database of mobile mapping images using the deep learning approach.

* Corresponding author

This contribution has been peer-reviewed.
https://doi.org/10.5194/isprs-archives-XLVI-2-W1-2022-389-2022

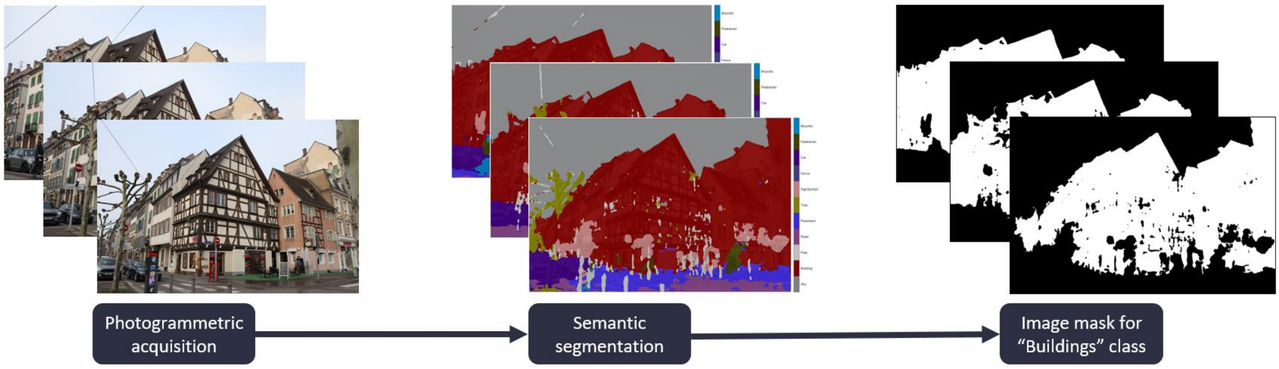


Figure 1. General workflow of the proposed method by creating 2D masks excluding unwanted classes to be used during dense image matching.

Specifically in this paper, an implementation of the DeepLab V3+ architecture (Chen et al., 2018) pre-trained from a ResNet-18 network was used. The neural network was further enhanced via transfer learning with the addition of the CamVid dataset (Brostow et al., 2009). In the original implementation of the neural network, eleven classes were identified, namely "Sky", "Building", "Pole", "Road", "Pavement", "Tree", "SignSymbol", "Fence", "Car", "Pedestrian" and "Bicyclist". However, for the purposes of this study these classes were further binarised into “Building” and “Noise”. The “Noise” class comprises therefore of pixels belonging to the other ten classes.

The AI-based semantic labelling process automatically and rapidly classifies the images to buildings and noises. Subsequently an algorithm generates image masks from the result (Figure 1). The masks specifically exclude everything which is not considered as building façade on the scene, thus producing a cleaner dense point cloud in the hope of reducing greatly traditional requirements of manual point cloud cleaning. These masks were then used during the 3D reconstruction process, especially during the dense image matching step in order to generate dense point clouds. For the purposes of this paper, the photogrammetric reconstruction and image masks were performed using the software Agisoft Metashape. Also, within the context of this study, the images were previously oriented, scaled and refined using bundle adjustment. Although this process will not be detailed in this paper, the assumption taken for all practical purposes is that the image orientation and scaling was performed with a satisfactory level of precision.

An experimental test was performed on a medieval timber frame building located within the city centre of Strasbourg (France), itself part of a UNESCO World Heritage site. Twelve terrestrial images were taken using a Canon R5 mirrorless camera with a 24 mm lens. The building was chosen as a first case study due to several reasons, notably the fact that it presents an object typical to be found in the case of heritage documentation while at the same time presenting an additional challenge to the neural network due to the half-timbered style of the facades. The existence of wooden elements on the half-timber house presents an interesting contrast to the usually smooth and unicoloured modern buildings used in many training data. Furthermore, this particular house is located on a road with occasional traffic of cars, pedestrians and bikes. Some vehicles were also parked near the structure. The presence of trees and light posts complete the scene with potential “noise” to be detected and excluded by our proposed method.

In this case study, the proposed workflow was deployed, and its results were assessed. A twofold analysis was thereafter performed: firstly, an assessment of the neural network’s performance on the 2D space i.e., semantic image segmentation was performed. Secondly, the masks created from the semantically labelled images were used in creating a dense point cloud via photogrammetry. A metric assessment was then performed on the result of this 3D projection by comparing the proposed method to manual image masking. In this regard, the basic assumption is that the images were well oriented. The parameters commonly used to assess segmentation results were used in this study, namely the precision, recall, F1 score, and IoU (Intersection over Union).

3. RESULTS AND DISCUSSIONS

A visual illustration of the result generated by the proposed method may be consulted in Figure 2. In this figure, a visual comparison was made between three scenarios: one without any noise masking, one with manual masking and another with the proposed AI-based noise masking. Note that the figure only shows one image as an illustration, but the process was repeated for all input images (in this case twelve close-range images). As can be observed in Figure 2, a direct dense matching without image masking generated a substantial amount of noise in the scene in the form of vehicles, trees, roads and people. Using manual masking, not only was the processing time faster, but most of these unwanted objects were also automatically excluded from the final point cloud. The proposed automatic masking method also generated good results when considered visually, as only the building was mostly reconstructed. In order to have a better assessment on the results obtained, a quantitative analysis was conducted.

As has been previously mentioned, for quantitative analysis the precision, recall, F1 and IoU scores were computed. First, analysis was conducted to determine the quality of the 2D semantic segmentation. In this regard the quality parameters were computed for all twelve images separately using a manually labelled ground truth dataset. Results for this assessment show that the AI-based method worked well in performing the 2D semantic segmentation, with an average F1 score of 86.38% and IoU of 80.75% for the twelve input images (Figure 3). The result seems to worsen slightly in very complex scenes such as image number 7, although the decrease in quality is small and with only twelve samples no systematic error can be detected yet. A larger sample of image datasets will be considered in the future to investigate this point further.

https://doi.org/10.5194/isprs-archives-XLVI-2-W1-2022-389-2022

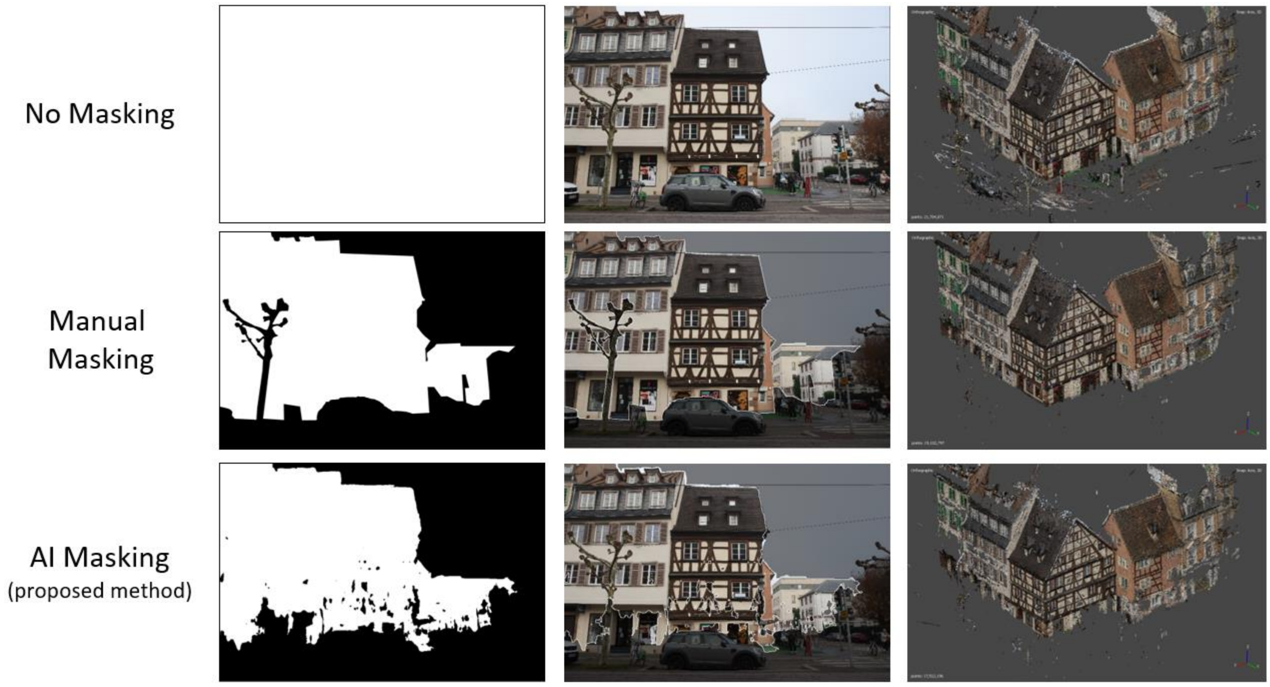


Figure 2. Illustration of the proposed workflow and comparison with traditional manual masking of unwanted objects. The third column shows the resulting dense point clouds.

Image No.	Precision (%)	Recall (%)	F1 Score (%)	IoU (%)
1	94.57	84.75	89.39	84.75
2	94.77	86.68	90.55	86.68
3	95.84	82.65	88.76	82.65
4	91.64	82.40	86.77	82.40
5	92.99	82.49	87.43	82.49
6	93.43	81.06	86.81	81.06
7	88.39	75.43	81.40	75.43
8	96.62	85.85	90.92	85.85
9	96.30	80.06	87.43	80.06
10	89.02	76.02	82.01	76.02
11	89.63	74.36	81.28	74.36
12	91.56	77.27	83.81	77.27
Mean	**92.90**	**80.75**	**86.38**	**80.75**

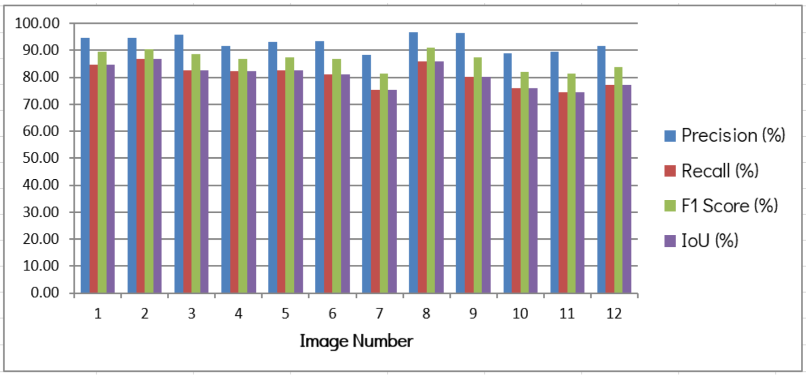
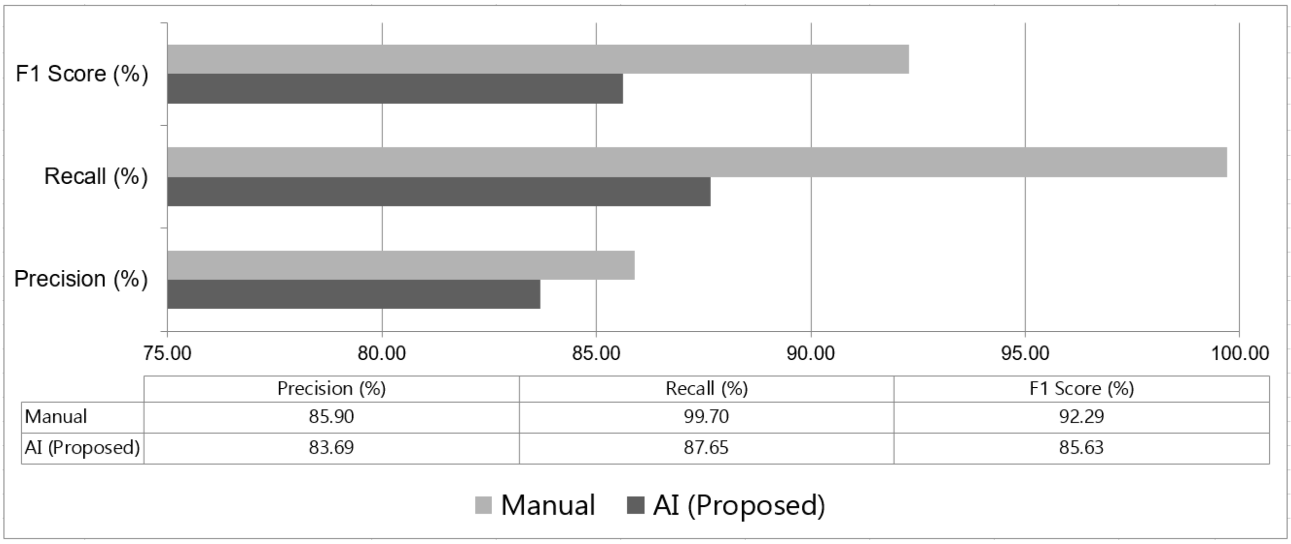
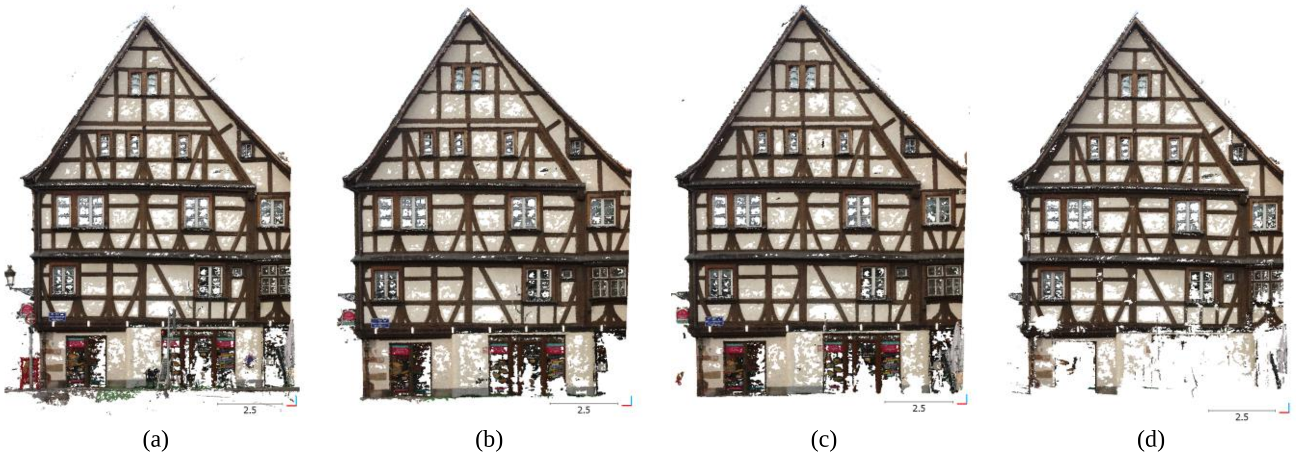


Figure 3. Results of the AI-aided 2D semantic segmentation for the “Buildings” class on the input close range images showing four metrics (precision, recall, F1 score, and IoU) commonly employed in pattern recognition.



	Precision (%)	Recall (%)	F1 Score (%)
Manual	85.90	99.70	92.29
AI (Proposed)	83.69	87.65	85.63

Figure 4. A histogram representation of the classification statistics (precision, recall, and F1 score) for the “Broglie” dataset’s “buildings” class in the 3D space, comparing the proposed method with manual mask creation.



(a) (b) (c) (d)

Figure 5. Dense point cloud results from (a) raw dense matching without noise cleanup, (b) manual cleaning on CloudCompare, (c) automatic noise exclusion via manual image masks and (d) automatic noise exclusion via AI-created image masks.

Nevertheless, the results as presented in Figure 3 generally matches the quality parameters of the neural network training. The DeepLab v3+ network trained using CamVid images attained an overall training accuracy of 83.53% and validation accuracy of 86.23%.

These segmented images were then used to generate masks which were then used as constraints during the dense matching phase in the close-range photogrammetric workflow. Two sets of masks were created, one created automatically by the AI, and another created manually as a control. As a ground truth for this second quantitative assessment, the resulting dense point cloud created without masks were cleaned manually in the 3D space. This involves the segmentation and suppression of noisy elements in the dense point cloud, for which the software CloudCompare was utilised. The results of the dense matching with both sets of image masks were then compared to the ground truth reference, yielding promising results for our proposed method with 85.63% F1 score compared to 92.29% obtained using manually generated masks (Figure 4).

As may be inferred from Figure 2 and Figure 4, image masking was shown to be very useful in reducing dense point cloud noise. At the same time, the constraint on the input images during dense image matching also provided the added benefit of reducing processing time since the reconstructed region is limited by them. Using manual masking, most noisy points were automatically excluded from the dense matching (with precision of 85.90% and recall up to 99.70% for the "Buildings" class). The proposed AI-based method also showed promising results with a respectable precision of 83.69% and a recall value of 87.65%. There are however some shortcomings on the results in the 3D space.

In Figure 5, four images of the northern façade of the building are shown. Figure 5 (a) depicts the raw point clouds created by dense image matching without any noise masking or cleaning; notice the presence of noisy elements on the ground floor. In Figure 5 (b), the point cloud was generated without any image masking. Noise cleaning was then performed directly in the 3D space. This data was used as reference in the computation of values found in Figure 4. Figures 5 (c) and (d) show the automatic noise exclusion using manual and AI-based image masking respectively. First, the result of the dense image matching in general encounters the problem of lack of texture on some parts of the wall. This is a common problem in photogrammetric dense image matching and is evidenced by the holes which can be found throughout Figure 5. Taking this fact out of the equation still shows that manual image masks were very useful in noise exclusion, although some minor noises were still present on the scene. These minor artefacts are however very small.

The AI-based method on the other hand, while succeeding in reducing noise in the final point cloud, also excluded some other objects. Notably, some parts of the building openings on the ground level e.g., doors were considered as noise and therefore excluded. This may be partially explained by the fact that many complex elements were present at the ground level of the scene. The semantic segmentation process was therefore not precise enough to distinguish between the building and the other classes considered as noise, such as pedestrians, vehicles, and roads.

It should also be noted that the neural network used in the semantic segmentation process was trained on a dataset comprising of lower-resolution mobile mapping images. Indeed, the CamVid dataset consists of frames extracted from videos. A possible improvement in this regard could therefore be the use of other training datasets which may be more appropriate for close-range photogrammetry cases. Furthermore, presently no other DL-based architecture has been tested. This is however a feasible idea for future study in order to find the best architecture which suits this particular task of noise exclusion.

Despite these shortcomings, the proposed method still provides a crucial benefit when it comes to processing time. The creation of 2D masks on input images was the most time consuming, as it may take up to two or three minutes for each input image. This time duration will increase further the more input photos are concerned in the photogrammetric processing. Point cloud cleaning in the 3D space can be faster or slower depending on the object to be reconstructed. In the case of simpler objects and buildings, this can be done fairly fast. However, with complex structures as often seen in historical buildings and architecture, this can be much slower and tedious. As a point of reference, the cleaning of the point cloud of the Broglie building used as a reference in this study took more or less ten minutes to perform correctly. Using the AI-based methodology on the other hand, semantic segmentation and mask creation was done very rapidly in less than one minute. The training part was the longest, taking a few hours using a computer with an NVIDIA GeForce RTX 3060 GPU.

4. CONCLUSIONS AND FURTHER WORK

In this paper, we attempted to use DL-based semantic segmentation in order to exclude unwanted point clouds noises automatically. A 2D semantic segmentation was performed on the input images, dividing pixels into either buildings or noise classes. Based on this process, image masks were automatically created to be used during dense image matching. One major advantage to this approach, apart from the automatic noise exclusion, is the reduction in processing time due to constrained reconstruction zone.

Based on the findings described in this paper, we argue that the proposed method at its current state already showed promising results in automating noise exclusion from photogrammetric point clouds. Specifically, the paper presented a proof of concept of what may be achieved by such methods in the case of heritage buildings. Some shortcomings may still be observed, notably on the quality of the semantic segmentation itself. Further investigation is required to improve this aspect, mainly the choice and/or augmenting of training data quality and quantity as well as tests on other DL architectures. Nevertheless, the obtained F1 score of 85.63% for the final building point cloud already shows the potential of this method and that more research to improve it is warranted.

More experiments and assessments are also planned, including application on other types of building architecture. It is also interesting to note that while the proposed method was applied on European timber-frame houses in this method, it may be easily adapted into other scenarios depending on the availability of training data.

ACKNOWLEDGEMENTS

The authors wish to thank Bastien Wirtz of INSA Strasbourg for his help during data acquisition.

REFERENCES

Boonpook, W., Tan, Y., Xu, B., 2021. Deep learning-based multi-feature semantic segmentation in building extraction from images of UAV photogrammetry. International Journal of Remote Sensing 42, 1–19.

Brostow, G.J., Fauqueur, J., Cipolla, R., 2009. Semantic object classes in video: A high-definition ground truth database. Pattern Recognition Letters 30, 88–97.

Chen, L.C., Zhu, Y., Papandreou, G., Schroff, F., Adam, H., 2018. Encoder-decoder with atrous separable convolution for semantic image segmentation. Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics) 11211 LNCS, 833–851.

Grilli, E., Battisti, R., Remondino, F., 2021. An advanced photogrammetric solution to measure apples. Remote Sensing 13, 3960.

Kernell, B., 2018. Improving Photogrammetry using Semantic Segmentation. Master of Science Thesis in Electrical Engineering, Department of Electrical Engineering, Linköping University, Sweden.

Kirillov, A., He, K., Girshick, R., Rother, C., & Dollar, P., 2019. Panoptic segmentation. Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition, 9396–9405.

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, in: The International Archives of the Photogrammetry, Remote Sensing and Spatial Information Sciences. pp. 201–206.

Song, A., 2020. Semantic segmentation of remote-sensing imagery using heterogeneous big data: International Society for Photogrammetry and Remote Sensing Potsdam and cityscape datasets. ISPRS International Journal of Geo-Information 9, 601.

Stathopoulou, E.K., Remondino, F., 2019a. Semantic photogrammetry - boosting image-based 3D reconstruction with semantic labeling, in: ISPRS Annals of the Photogrammetry, Remote Sensing and Spatial Information Sciences. pp. 685–690.

Stathopoulou, E.K., Remondino, F., 2019b. Multi-view stereo with semantic priors, in: International Archives of the Photogrammetry, Remote Sensing and Spatial Information Sciences - ISPRS Archives. pp. 1135–1140.

Verhoeven, G.J., 2011. Taking Computer Vision Aloft - Archaeological Three-dimensional Reconstructions from Aerial Photographs with PhotoScan. Archaeological Prospection 18, 67–73.