ADVANCED APPROACH FOR AUTOMATIC RECONSTRUCTION OF 3D BUILDINGS FROM AERIAL IMAGES

D. Yu [1], S. Wei [1], J. Liu[1], S. Ji [1,*]

[1] School of Remote Sensing Information and Engineering, Wuhan University, Wuhan 430079, China - (yudawen, wei_sq, liujinwhu, jishunping)@whu.edu.cn

Commission II, WG II/4

KEY WORDS: 3D Building Reconstruction, DSM, 2D Building Boundary Delineation, Multi-View Aerial Images.

ABSTRACT:

In this work, a novel automatic 3D building reconstruction approach is proposed to extract accurate LoD1 building models from multi-view aerial images. The proposed approach consists of three main parts. The first step is to generate digital surface models (DSMs) from aerial images, which is implemented with the Smart3D software and can be replaced by other open-source multi-view stereo (MVS) algorithms as well. The second step is to produce structured 2D building footprints using combined deep learning and regularization. The initial building segmentation maps are obtained by the multi-scale aggregation fully convolutional network (MA-FCN), which takes both the images and DSM as input, through supervised learning. The initial segmentation maps are further refined with another segmentation maps that are derived from the DSM. After that, the contour extraction and regularization technology are applied to produce structured building footprints. In the last step, the elevations of the top and base of each individual building are reliably estimated by adopting an adaptive terrain generation approach and the neighbourhood buffer analysis. The georeferenced building footprint polygons and the elevations of building top and base form the watertight LoD1 building models. The qualitative and quantitative evaluations in Jinghai District, Tianjin, China demonstrate the robustness and effectiveness of the proposed approach.

1. INTRODUCTION

Building is the most representative entity of a city on the Earth. Three-dimensional (3D) building reconstruction from overhead images or lasers is one of the most key tasks nowadays for smart city construction, urban planning, population density analysis, mobile telecommunication, and disaster management (Bulatov et al., 2014). Although fully automatic building reconstruction systems had been envisioned from decades past, there are no mature algorithms or commercial software yet up-to-now (Xiong et al., 2015). 3D building reconstruction is still a challenging and unsolved task considering the complexity and diversity of building styles and entangled complex backgrounds in urban scenes.

The required levels of details of the reconstructed 3D building models vary for different applications. Levels of Details (LoDs), varying from zero to four, have been a widely accepted standard to represent the accuracy and completeness of the reconstructed 3D buildings (Kolbe et al., 2005). The coarsest LoD0 model is a 2.5D digital terrain model (DTM) overlapped with an orthophoto image or map, while a LoD4 model has detailed interior structures like rooms, doors, stairs, etc.

Extensive methods have been presented in previous literatures to generate 3D building models of different LoDs semi-automatically or automatically, from various data sources (Kada and McKinley, 2009; Akmalia et al., 2014; Kedzierski and Fryskowska, 2014; McClune et al., 2016; Rubinowicz, 2017; Alidoost et al., 2019). However, fully automatic and reliable 3D building model reconstruction is not possible beyond LoD2 (Tack et al., 2012; Moreira et al., 2013). Many studies require data that are difficult to access or go out-of-date quickly, e.g., 2D building ground plans from cadastral datasets (He et al., 2012; Tack et al., 2012; Sugihara et al., 2015), or data that are expensive, e.g. very dense LiDAR cloud with high-accurate georeferencing (Akmalia et al., 2014; Kedzierski and Fryskowska, 2014; Rubinowicz, 2017). Compared to the LiDAR-based or 2D building vector map assisted methods, there are fewer studies that started from easily accessed multi-view aerial images. The latter is more challenging that elevation information and building footprints are both unavailable.

Due to the significant advances in deep learning, the convolutional neural network (CNN) based semantic segmentation methods have increasingly been used for building footprint extraction from remote sensing images in recent years (Huang et al., 2016; Alshehhi et al., 2017; Yuan, 2017, Wei et al., 2020), and shown significant advantages against traditional methods. Image dense matching technologies, including conventional and machine learning based methods (Haala et al., 2015; Kendall et al., 2017), have also gotten development and can generate high-accurate digital surface models (DSMs). The two advanced methods provide a new chance for automatic 3D building reconstruction of urban scenes from multi-view remote sensing images directly.

This work presents an automatic and robust 3D building modeling approach which takes the multi-view aerial images as input and outputs the LoD1 building models with full automation and high accuracy, no 2D building vector maps or operator intervention is required. Generally, the basic idea is to combine the image dense matching derived DSM and the CNN-based building segmentation algorithm to extract the 2D building polygons first, and then the elevations of the base and

* Corresponding author

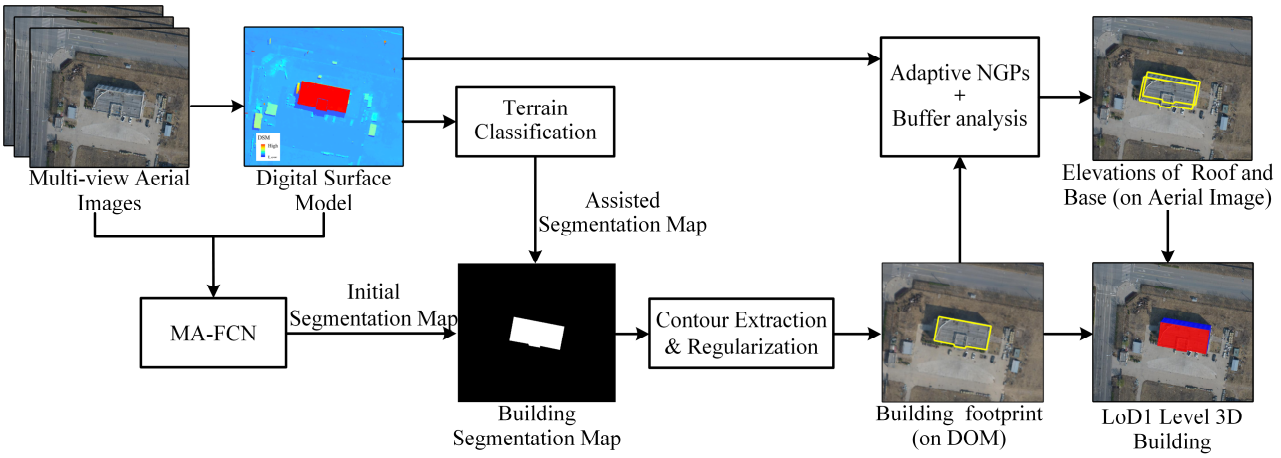


Figure 1. The workflow of our proposed 3D building reconstruction approach.

the top for each individual building are estimated by adopting the adaptive terrain generation approach and the neighbourhood buffer analysis. The located building footprints, the elevations of top and base form a LoD1 level building model in 3D space. The reconstructed 3D entities were qualitatively and quantitatively evaluated according to completeness and robustness, the effectiveness of the proposed approach was demonstrated.

2. METHODOLOGY

Our proposed approach can be divided into three main stages and the corresponding workflow is shown in Figure 1. The first step is generating DSM from multi-view aerial images, which were captured from a 5-view oblique camera, which is implemented with the Smart3D software in our study. It can also be replaced by other conventional or recent deep learning based algorithms. The corresponding depth maps of the down-looking aerial images are generated as well, as the elevation information of each pixel on images would be used to help the CNN-based method in 2D building segmentation in the second step.

The second step is to produce structured 2D building footprints. The multi-scale aggregation fully convolutional network (MA-FCN) (Wei et al., 2020) is adopted to extract the initial building segmentation maps from the down-looking aerial images and the corresponding depth maps. The original MA-FCN predicts building segmentation maps from images by concatenating the feature maps from four convolution layers at different scales, which has been proven effective for remote sensing building extraction. Considering the depth maps can provide supplementary topographic information for further improving the performance of image-based building segmentation, we take the depth map as an extra input channel (i.e., concatenated to the red, green, and blue channels).

The implementation process of the building segmentation is summarized as follows. The parameters of the MA-FCN are pre-trained on a large open remote sensing building dataset (Ji et al., 2018) firstly. Then, the parameters are adjusted by using a certain amount of aerial images and manually labelled building footprints at the study area. Note that the corresponding depth maps are fed into the network only at the fine adjustment and testing stages.

After the binary building segmentation maps are produced by the MA-FCN, initial building contours are extracted from these segmentation maps and projected to the ground (DSM) with interior and exterior parameters for further adjustment. However, the elevation information provided by the image dense matching unavoidably contains some errors on building edges, the elevation of each corner point of the contours needs to be judged before projecting. In our work, a thresholding mechanism is adopted, if the elevation of the roof corner point is close to the ground (the difference is small than the predefined threshold), it would be replaced with an average elevation of its inward neighbourhood window.

After that, the initial building segmentation result is further modified with another assistant building segmentation map that is derived by classifying the points of DSM into buildings and ground through the scanline method. The scanline window of each building is first located, which is obtained by expanding the building bounding box which was generated by applying the pre-trained Mask R-CNN (He et al., 2017). Though NVDI is unavailable in our study due to lack of the infrared band, the trained CNN from optical images can distinguish buildings from vegetation well and effectively eliminate the impact of tall vegetation. Then each pixel in the scanline window is classified into ground and non-ground (building) by the directional scanline-based terrain filtering approach (Mousa et al., 2017). The point with the second minimum elevation on each scanline is selected as the initial ground point. The lowest elevation value may be caused by exceptional cases, such as mismatches or shade, therefore the second one is selected to avoid such cases. After that, the remaining points are determined as ground points if their elevation differences with the initial ground points are smaller than a predefined empirical threshold.

In general, the building segmentation maps from the MA-FCN provide more accurate building boundaries, while the assistant segmentation maps derived from the DSM can distinguish the ground regions from non-ground regions accurately but are often blurred at the building boundaries (due to the smoothing constraints during image matching). The morphological filter open operation is performed to remove the false building boundaries on the assisted segmentation maps. Then, depending on the assistant segmentation maps, the missing building regions and the misclassified ground areas in the initial segmentation maps are complemented and removed, respectively. After that, the contour extraction operation, and polygon regularization (Wei et al., 2020), are performed to produce the structured 2D building footprints. The Douglas-Peucker (Douglas and Peucker, 1973) polygonal approximation

method is utilized to simplify the building contours and the distance threshold ε is set to 0.2m in our case study.

In the last step, the elevations of individual building's top and base are estimated by combining the DSM and the structured 2D building footprints. The non-ground points of each building's neighbourhood on the DSM, which would impact the correct estimation of the height of the building base, are filtered using a modified DTM extraction approach, we named it as adaptive NGPs that developed from (Mousa et al., 2017). The filtering window is determined adaptively by expanding the bounding rectangle of each individual building footprint.

By applying the adaptive NGPs, a series of buffers with different distances from the building boundary are set up. The buffers are used to count the stable ground elevation surrounded the buildings considering possible errors in the DSM especially at those areas with dense buildings, shadows, or tall vegetation. The minimum elevation of each buffer (i.e. the elevation of the bin where the elevation histogram has been accumulated to the frequency of predefined threshold (5% in our case study) from low to high) is recorded and used for fitting the local terrain. Considering that most buildings are located on a plane or slope, a simple linear function is adopted to fit the local terrain. The base elevation of each individual building is obtained according to the fitted local topographic surface and the building footprint position. The highest elevation inside the building boundary is recorded as the top elevation after removing some obvious outliers. Combining the building footprints and the elevations of building top and base, the watertight LoD1 building models are generated.

3. EXPERIMENT AND ANALYSIS

An experimental area located in the Jinghai District, Tianjin, China, was chosen for evaluating our proposed approach, which covers an area of around 3.96 km^2 and various residential and factory buildings. As shown in Figure 2 (left), sub-area 1 and sub-area 3 are used to train the MA-FCN, and the sub-area 2 was used for accuracy verification of our proposed approach. The aerial images were captured from a five-view oblique camera-rig in April 2019 with 0.04 m GSD and a size of 7952×5304 pixels. The DSM with a resolution of 0.2 meters was generated from the Smart3D software with give ground control points. The 3D building ground truth was manually edited from the OSGB surface model generated by Smart3D and carefully checked, as shown in Figure 2 (right).

To evaluate the completeness and reliability of the reconstructed individual 3D building models, the evaluation criteria of the object-level rather than pixel-level is adopted. We took the precision and recall at the different 3D intersection over union (IOU) threshold as the final criteria. For example, the threshold of 3D IOU = 0.5 means when the 3D IOU between an individual 3D building model produced by our proposed approach and its corresponding ground truth model reaches or exceeds 50%, it was counted as a valid instance. The precision and recall were calculated using all reconstructed individual instances at a certain 3D IOU threshold, which varied from 0.5 to 0.9. Note that 0.9 is a very rigorous threshold, for example, for a 10×10×10 m^3 building model a maximum 0.18 m shift is allowed at each side to reach 0.9 IoU. Besides, we also calculated some intermediate criteria for comparing with other recent methods, including the building footprint extraction methods and height estimation strategies, as shown in table 1 and table 2.

Table 1 shows the performance of different 2D building footprint extraction methods. We compared the original MA-FCN (Wei et al., 2020) which represents the current state-of-the-art CNN-based building segmentation methods, the MA-FCN+ which takes the DSM as an extra input channel, and the MA-FCN++, in which the DSM is used for training first and then for modification by producing assist building segmentation maps. The images and corresponding depth maps were cropped into 512×512 tiles before they were fed into these networks. All three networks shared the same training configurations, the Adam algorithm was used for gradient optimization, a mini-batch contained 5 images/depth maps, and the weights were updated using the learning rate of 10^{-4}. All the experiments were conducted in a Windows PC equipped with an Nvidia GTX 1080 Ti 11G GPU. The performance of the DSM assisted MA-FCN (MA-FCN+) is improved greatly compared with the original MA-FCN, and the modification based on the terrain classification approach further improves the performance of MA-FCN+.

Table 2 shows four different building height estimation methods using the DSM and the ground-truth building footprints as the input. The multi-directional and slope dependent method (MSD) (Perko et al., 2015) and the original network of ground points (NGPs) method (Mousa et al., 2017) first generate the digital terrain model (DTM) and then subtract the DTM from the DSM to generate the normalized DSM (nDSM). Combining the building footprints and the nDSM, the building heights are obtained directly. The elevation difference model (EDM) (Zeng et al., 2014) estimated building heights based on neighbourhood analysis. Our method combines an adaptive NGPs and buffer analysis strategy to fit local terrain and then calculate the building heights. From Table 2, our method performs better than the other three methods on estimating the building heights.

Table 3 shows the precision and recall of our proposed 3D building reconstruction workflow from the input aerial images. Our proposed approach can reach an average precision of 0.6672 and an average recall of 0.6226 when the 3D IOU threshold varies from 0.5 to 0.9, which is very inspiring without any requirement of extra 2D map data and operator intervention.

The qualitative results of the reconstructed 3D buildings are shown in Figure 3. We projected the facades and roofs of the reconstructed 3D buildings to the original aerial images with light yellow masks and light cyan masks respectively, to visualize the degree of agreement between the reconstructed models and the real buildings. It is observed that the masks of the reconstructed 3D buildings can cover the whole buildings accurately in most cases, indicating our method is robust and reliable. The tall vegetation and building shadows caused some inaccurate edges of buildings and several failed reconstructions (e.g. the one in the red dotted box of the first row, third column in figure 3), which need to be further improved in future work.

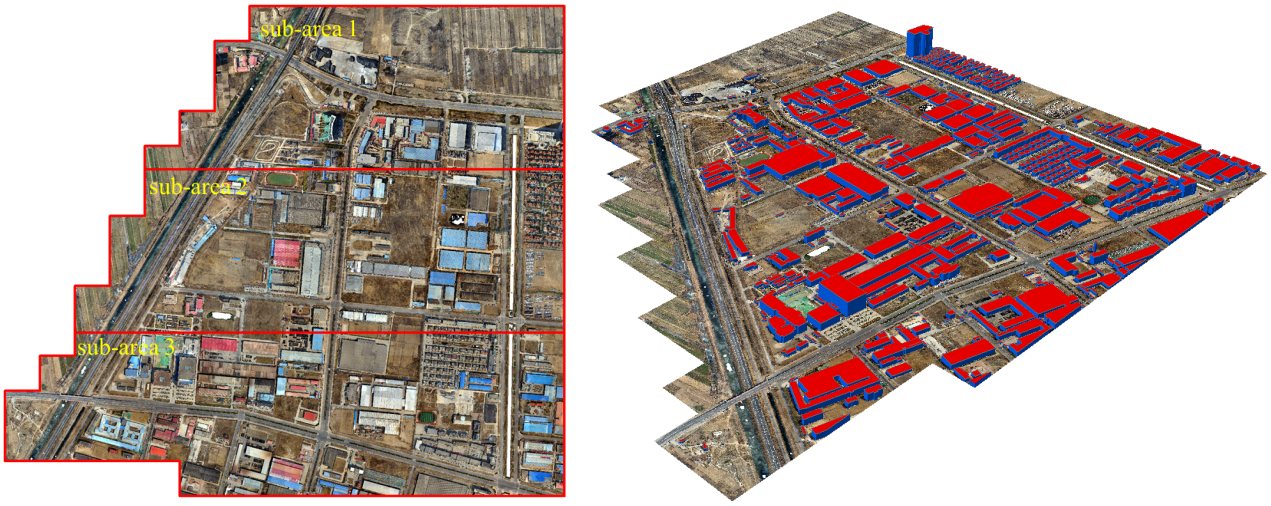


Figure 2. The digital orthophoto image of the study area (left) and the 3D building ground truth (right) which has been carefully edited and reviewed by the operators.

Method	Input	Criterion	2D IOU=0.5	2D IOU=0.6	2D IOU=0.7	2D IOU=0.8	2D IOU=0.9	Average
MA-FCN	Aerial Images	Precision	0.7134	0.6743	0.6219	0.5078	0.2614	0.5558
		Recall	0.7063	0.6676	0.6156	0.5027	0.2587	0.5502
MA-FCN+	Aerial Images + DSM	Precision	0.8450	0.8207	0.7792	0.6779	0.4200	0.7086
		Recall	0.8097	0.7863	0.7466	0.6495	0.4024	0.6789
MA-FCN++	Aerial Images + DSM	Precision	0.8761	0.8511	0.8102	0.7284	0.4528	0.7438
		Recall	0.8176	0.7943	0.7561	0.6797	0.4226	0.6941

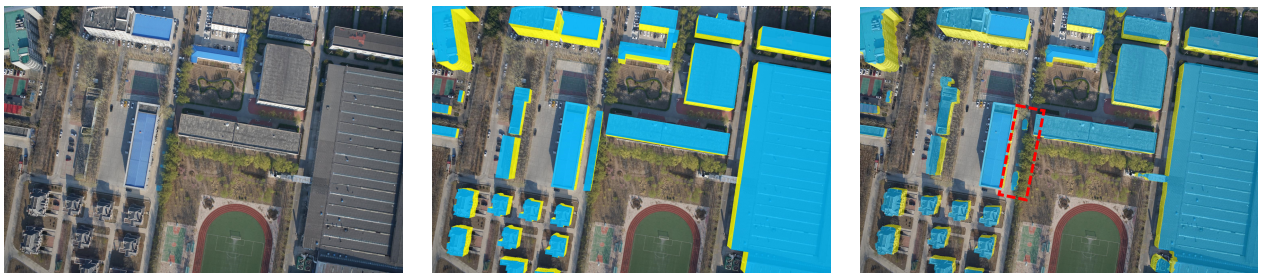
Table 1. The comparison of different building footprint extraction methods.

Method	Input	Mean Absolute Error	Standard Deviation	Root Mean Squared Error
MSD	DSM+Building Polygon	0.4047	0.7387	0.7805
NGPs	DSM+Building Polygon	0.3482	0.6963	0.7130
EDM	DSM+Building Polygon	0.3699	0.6552	0.6557
Ours	DSM+Building Polygon	0.3136	0.6382	0.6433

Table 2. The test results (in meter) of four different building height estimation methods.

Method	Input	Criterion	3D IOU=0.5	3D IOU=0.6	3D IOU=0.7	3D IOU=0.8	3D IOU=0.9	Average
Ours	Aerial images	Precision	0.8449	0.8108	0.7580	0.6233	0.2989	0.6672
		Recall	0.7884	0.7566	0.7073	0.5817	0.2789	0.6226

Table 3. The performance of our proposed 3D building reconstruction method.



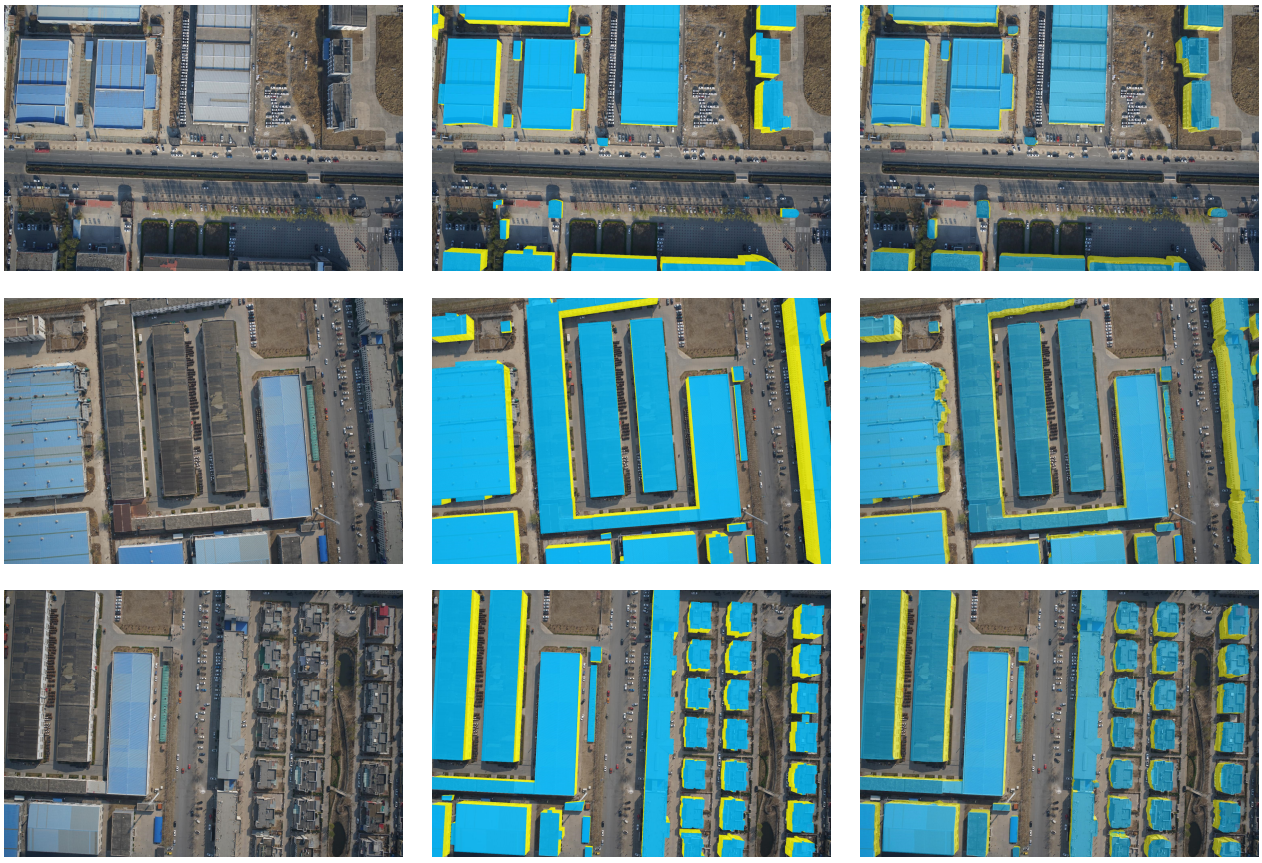


Figure 3. Samples of the reconstructed 3D buildings, which were projected on to the original aerial images. Left: the original aerial images, middle: the original aerial images covered with the projection masks of the building ground truth, right: the original aerial images covered with the reconstructed 3D building masks from our method. The light-cyan masks color the roofs and the light-yellow masks color the facades of the buildings.

4. CONCLUSION

This work presents a fully automatic and robust 3D building reconstruction workflow, which takes the multi-view aerial images as input and produces the accurate LoD1 3D building models. In summary, the proposed approach has shown a promising solution for 3D building reconstruction only starting from the aerial images, the accurate reconstruction results have also been demonstrated from the qualitative and quantitative evaluations. We also proved the intermediate results, including 2D building footprint extraction and building height estimation, exceeded the other recent relevant methods.

REFERENCES

Akmalia, R., Setan, H., Majid, Z., Suwardhi, D., Chong, A., 2014. TLS for generating multi-LOD of 3D building model. *IOP conference series: Earth and environmental science,*18(1): 012064.

Alidoost, F., Arefi, H., Tombari, F., 2019. 2D Image-To-3D Model: Knowledge-Based 3D Building Reconstruction (3DBR) Using Single Aerial Images and Convolutional Neural Networks (CNNs). *Remote Sensing,* 11, 2219.

Alshehhi, R., Marpu, P. R., Woon, W. L., Dalla Mura, M., 2017. Simultaneous extraction of roads and buildings in remote sensing imagery with convolutional neural networks. *ISPRS Journal of Photogrammetry and Remote Sensing*, 130, 139-149.

Bulatov, D., Häufel, G., Meidow, J., Pohl, M., Solbrig, P., Wernerus, P., 2014. Context-based automatic reconstruction and texturing of 3D urban terrain for quick-response tasks. *ISPRS Journal of Photogrammetry and Remote Sensing*, 93, 157-170.

Douglas, D. H., Peucker, T. K., 1973. Algorithms for the reduction of the number of points required to represent a digitized line or its caricature. *Cartographica: the international journal for geographic information and geovisualization*, 10(2), 112-122.

Haala, N., Rothermel, M., Cavegn, S., 2015. Extracting 3D urban models from oblique aerial images. *Joint Urban Remote Sensing Event*, pp.1-4.

He, K., Gkioxari, G., Dollár, P., Girshick, R., 2017. Mask R-CNN. *The IEEE international conference on computer vision,* pp. 2961-2969.

He, S., Moreau, G., Martin, J.-Y., 2012. Footprint-based generalization of 3D building groups at medium level of detail for multi-scale urban visualization. *International Journal on Advances in Software,* Volume 5, Number 3 & 4, 2012.

Huang, Z., Cheng, G., Wang, H., Li, H., Shi, L., Pan, C., 2016. Building extraction from multi-source remote sensing images via deep deconvolution neural networks. *IEEE International Geoscience and Remote Sensing Symposium*, pp.1835-1838.

Ji, S., Wei, S., Lu, M., 2018. Fully convolutional networks for multisource building extraction from an open aerial and satellite imagery data set. *IEEE Transactions on Geoscience and Remote Sensing*, 57(1), 574-586.

Kada, M., McKinley, L., 2009. 3D building reconstruction from LiDAR based on a cell decomposition approach. *International Archives of Photogrammetry, Remote Sensing and Spatial Information Sciences*, 38(Part 3), W4.

Kedzierski, M., Fryskowska, A., 2014. Terrestrial and aerial laser scanning data integration using wavelet analysis for the purpose of 3D building modeling. *Sensors*, 14(7), 12070-12092.

Kendall, A., Martirosyan, H., Dasgupta, S., Henry, P., Kennedy, R., Bachrach, A., Bry, A., 2017. End-to-end learning of geometry and context for deep stereo regression. *IEEE International Conference on Computer Vision*, pp. 66-75.

Kolbe, T. H., Gröger, G., Plümer, L., 2005. CityGML: Interoperable access to 3D city models. *Geo-information for disaster management*, 883-899.

McClune, A.P., Mills, J.P., Miller, P.E., Holland, D.A., 2016. Automatic 3d building reconstruction from a dense image matching dataset. *International Archives of the Photogrammetry, Remote Sensing & Spatial Information Sciences,* 41.

Moreira, J.M., Nex, F., Agugiaro, G., Remondino, F., Lim, N.J., 2013. From DSM to 3D building models: a quantitative evaluation. *International Archives of the Photogrammetry, Remote Sensing and Spatial Information Sciences,* 1, W1.

Mousa, A.-k., Helmholz, P., Belton, D., 2017. New DTM extraction approach from airborne images derived DSM. *International Archives of the Photogrammetry, Remote Sensing & Spatial Information Sciences*, 42.

Perko, R., Raggam, H., Gutjahr, K., Schardt, M., 2015. Advanced DTM generation from very high resolution satellite stereo images. *ISPRS Annals of Photogrammetry, Remote Sensing & Spatial Information Sciences,* 2.

Rubinowicz, P., 2017. Generation of CityGML LoD1 city models using BDOT10k and LiDAR data. *Przestrzeń i FORMA*.

Sugihara, K., Murase, T., Zhou, X., 2015. Automatic generation of 3D building models from building polygons on digital maps. *International Conference on 3D Imaging*, pp.1-9.

Tack, F., Buyuksalih, G., Goossens, R., 2012. 3D building reconstruction based on given ground plan information and surface models extracted from spaceborne imagery. *ISPRS Journal of Photogrammetry and Remote Sensing*, 67, 52-64.

Wei, S., Ji, S., Lu, M., 2020. Toward automatic building footprint delineation from aerial images using CNN and regularization. *IEEE Transactions on Geoscience and Remote Sensing,* 58(3), 2178-2189.

Xiong, B., Jancosek, M., Elberink, S. O., Vosselman, G., 2015. Flexible building primitives for 3D building modeling. *ISPRS Journal of Photogrammetry and Remote Sensing*, 101, 275-290.

Yuan, J., 2017. Learning building extraction in aerial scenes with convolutional networks. *IEEE transactions on pattern analysis and machine intelligence*, 40(11), 2793-2798.

Zeng, C., Wang, J., Zhan, W., Shi, P., Gambles, A., 2014. An elevation difference model for building height extraction from stereo-image-derived DSMs. *International journal of remote sensing*, 35(22), 7614-7630.