Reconstruction of Building LoD2 Wireframe Models Using Semantic Segmentation

Faezeh Soleimani Vostikolaei [1], Shabnam Jabari [1]*

[1] Department of Geodesy & Geomatics Engineering, University of New Brunswick, Fredericton, NB, Canada- (fsoleima, sh.jabari)@unb.ca

Keywords: 3D Building Modeling, Sematic Segmentation, Multi-object Deep Learning, 3D Building Wireframe, Convolutional Neural Network.

Abstract

LoD2 building models can be used in different digital twin-related applications such as urban planning, disaster management, optimizing green energy efficiency, and solar panel recommendation. Existing technology for 3D modelling of buildings still relies on a large amount of manual work due to the irregular geometries of different roof types. Wireframes have shown to be an effective representation for 3D building especially in LoD2 format. Due to the complexity and diversity of roof types in urban areas, 3D building modeling remains a challenging task. In this paper, we propose a new framework for generating 3D wireframes to model different roof types. While high-resolution airborne images can be utilized to exploit the fine details of the roofs, they have difficulties in areas with poor contrast or shadows. The proposed framework incorporates the Digital Surface Model (DSM) as an auxiliary data source to address this limitation. In this work, we focus on the extraction of roof geometrical components including lines and planes of individual buildings to achieve a consistent LoD-2 building reconstruction. The proposed methodology is divided into two phases: (1) jointly predicting building lines and roof planes from the RGB imagery and DSM and (2) generating 3D wireframes of buildings using the extracted roof planes and lines. Subsequently, height values from the point clouds are used to derive 3D wireframes. Experiments with 1,620 buildings from Fredericton, the capital of New Brunswick in eastern Canada, demonstrate an IoU of 0.9337, an F1-score of 0.939, and an F2-score of 0.9378 for the roof geometrical components detection phase, as well as an RMSE of around 0.2-0.8 meter for the final 3D building model compared to the original LiDAR data was achieved.

1. Introduction

In today's world, there is an increasing demand for accurate and precise 3D city models to support the sustainable management of the growing urban population. These models provide a comprehensive understanding of a city's layout and infrastructure, empowering urban planners and decision-makers with a detailed view of the city's current state. This holistic perspective supports decisions that align with long-term goals and objectives. Beyond urban planning, 3D building modeling enables informed decisions in sustainable development, land-use planning, transportation analysis, flood risk assessment, green energy optimization, climate change mitigation, disaster preparedness, and real estate development (Biljecki et al., 2015; Doulamis & Preka, 2016; Peters et al., 2021).

The existence of a large amount of information on urban areas and the diversity of building roof types makes 3D modelling of roofs and further LoD2 3D modelling of buildings an active research area in photogrammetry, remote sensing, and computer vision.

To accurately reconstruct 3D models of building wireframes, it is necessary to identify building geometrical components, which can be achieved using deep learning, particularly through segmentation networks. The goal of semantic segmentation using deep learning in our work is to classify each pixel in images into specific categories that represent different roof geometric components, such as roof lines and planes. With the advent of multi-object segmentation, these models are gaining more attention due to their shared backbone that extracts key features, which are then used in various objects within the model. Multi-object segmentation not only boosts training speed by consuming less memory and computational power compared to using separate architectures for each object, but also can learn a generalized representation of data features that benefit multiple objects, potentially leading to better performance on each (Benjamin Bischke et al., 2019; Crawshaw, 2020).

Semantic segmentation often struggles to distinguish objects in areas with poor contrast and shadows in RGB images.

Using digital elevation models is an effective way to overcome this limitation since building planes, out-lines, and in-lines differ in height from the pixels of the background. Therefore, point clouds can be used as an independent source of information for building geometrical component extractions (Soleimani Vostikolaei & Jabari, 2023).

Although several studies have performed LoD2 3D modeling of building from aerial sensor data, only a few of them use deep learning segmentation and to the best of our knowledge there are no studies in the literature which can detect the roof geometrical components including footprints, lines, and planes of each roof by fusing the RGB and DSM features. Hence, in this work, we propose a bimodal multi-object network that combines deep RGB and height features of each building. Our proposed method aims to increase the accuracy of LoD2 3D modeling of building wireframes by improving the overall accuracy of roof geometrical components detection.

The presented methodology consists of two phases. The first phase focuses on segmenting roof geometrical components.

The second phase of this work focuses on 3D modeling of building wireframes using a data-driven approach. This part can be divided into three steps: (1) vectorizing the roof planes and lines, (2) simplifying them with the Douglas Peucker algorithm, and (3) deriving 3D wireframes by using height values from LiDAR point clouds.

* Corresponding (presenting) author

In summary, this paper contributes to the literature by introducing a novel framework for 3D city modeling and focusing on creating 3D wireframe models of buildings from bimodal data sources.

2. Related Works

2.1 Semantic Segmentation

Building footprint detection using satellite or aerial images with Deep Convolutional Neural Networks (DCNN), is mostly assumed as image semantic segmentation. Deep convolutional neural networks have attracted great attention in recent research due to their exceptional performance in image segmentation. DCCNs can detect, segment, or classify objects, including buildings, roads, trees, or building roof types (Buyukdemircioglu et al., 2021). Here, we summarize the latest studies on DCNN, focussing on fully convolutional networks (FCN), a branch of CNN.

Fully convolutional networks are recognized as a seminal approach to semantic segmentation. Liu et al. (2019), developed a new FCN with a spatial residual inception (SRI) module. The SRI model is proposed to capture and aggregate multi-scale contexts for semantic understanding by successively combining multi-level features. Sang & Minh (2018) also, utilized FCN neural networks based on the backbone ResNet101 with additional up-sampling skip connections to preserve spatial resolution, improve gradient flow. They achieved an overall accuracy of 91% for building detection. The common loss functions for image segmentation purposes are cross-entropy loss, dice loss, and focal loss Jadon (2020). To address multitask learning, class imbalance, or specific task requirements, custom loss functions need to be defined. Benjamin Bischke et al. (2019) developed a novel multi-task FCN loss to preserve semantic segmentation boundaries in high-resolution satellite imagery. In the new loss, the network is biased to focus more on pixels near boundaries using multiple output representations of the segmentation mask.

Similarly, U+-Net-based semantic segmentation models have mostly been used in recent research due to their exceptional performance. Guo et al. (2020) and Robinson et al. (2022), developed U-Net architecture to extract building footprints from multispectral imagery. Similarly, in the (Robinson et al., 2022) study, an open-source web-based tool is developed for collecting polygon labels over a given imagery scene, and a framework is designed by integrating mobile-U-Net, with a generative adversarial network (GAN). In a CNN-based comprehensive study, Jewell et al. (2019) compared U-Net, U-Net++, FCN, and DeepLabv3 performance in extracting building footprint extraction in satellite imagery. The results show that DeepLabv3 with Resnet-101 backbone has a better accuracy compared to the other state-of-art networks. Besides, using a pre-trained model can improve the accuracy of building footprint detection. Li et al. (2019) and Schuegraf et al. (2024) used U-Net network to fuse satellite images with other datasources such as OpenStreetMap, Google Maps, and MapWorld to detect building footprints using the U-Net network. Their network achieved a total F1-score of 0.704.

2.2 LoD2 Reconstruction

Conventional methods of 3D city model generation can be divided into three categories: model-driven, data-driven, and hybrid techniques.

(1) Model-driven approaches, which are also known as top-down approaches, excel in accurately modeling buildings that exist in their predefined library. These approaches select the model that best fits the building data from a library of models (Krafczek & Jabari, 2022). The studies of (Lafarge et al., 2010) and (H. Huang et al., 2013) proposed a method to reconstruct buildings from a digital surface model (DSM). This process involved breaking the building footprints down into components either manually or automatically and then utilizing a Gibbs model to fit the 3D block models onto the building footprints. A Bayesian decision was taken to find the most appropriate roof primitives from the pre-defined library that would represent the point clouds by utilizing a Markov Chain Monte Carlo sampler and original proposition kernels.

To model the buildings outside the predefined library, data-driven approaches come into play.

(2) Data-driven techniques are used to detect the roof planes and extrude roof shapes based on geometrical components such as lines, edges, and points (Park & Guldmann, 2019; Schuegraf et al., 2024). There are various methods for segmenting the LiDAR point clouds and determining roof planes, such as edge-based methods (Jiang & Bunke, 1994), region-growing methods (Alharthy & Bethel, 2004), random sample consensus (RANSAC) methods (Hartley & Zisserman, 2002), and clustering methods (Shan & Toth, 2009), or the combination of two or more algorithms (Dorninger & Pfeifer, 2008). (H. Huang et al., 2011) proposed generative modeling of building roofs with an assembly of primitives allowing overlapping using the Reversible Jump Markov Chain Monte Carlo algorithm. (J. Huang et al., 2022) presented a methodology for reconstructing 3D models of buildings from airborne LiDAR point clouds using a data-driven approach. In both works, they segmented point clouds into planar patches. Then, a 3D optimization is used to create a topologically consistent 3D building model from its compositional primitives.

(3) Due to the limitations of model-driven approaches and the complexity of data-driven approaches, hybrid methods have been developed. (Pepe et al., 2021) and (Tripodi et al., 2020) used stereo satellite imagery to build a digital surface model and extract the height of each object using the DSM. The latter used Deep Learning to extract the contour polygons of the buildings and the digital train model. (Zhao et al., 2021) proposed the reconstruction framework to reconstruct a 3D model containing a complete shape and accurate scale from a single image. The proposed method involves using two convolutional neural networks to create watertight mesh models and optimizing them using another CNN network. (Krafczek & Jabari, 2022) proposed a decision-tree-based methodology for generating LoD2 3D city models. They decomposed the building footprints into building primitives to have a better estimation of height for each building's parts. Some works create 3D city models directly from 3D point clouds. These methods are based on using Terrestrial Laser Scanners (Akmalia et al., 2014) to generate dense point clouds from it and then perform segmentation to detect building façade and features. (Kada, 2022) extracted the geometrical features from buildings utilizing a deep learning network.

To model all types of building roofs, even complex types, we investigate data-driven approaches in this research.

3. Methodology

As shown in

Figure 1, the proposed method is divided into two phases. Phase 1 focuses on the segmenting of roof geometric components, i.e. roof planes and lines. Phase 2 of this work involves the reconstruction of a wireframe for the LoD2 model of buildings. The following subsections provide details about each phase.

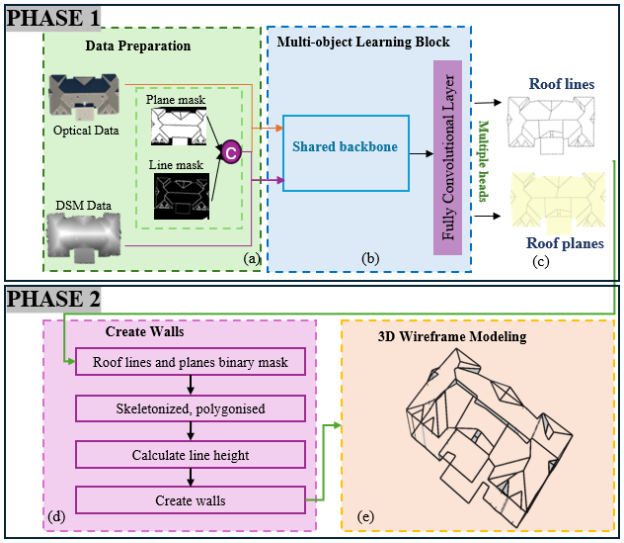


Figure 1. The overall workflow of the proposed network includes: (a) data preparation, (b) building geometrical components segmentation by fusing optical and height information of each building (c) roof lines and planes detection (d) 3D walls modelling, and (e) 3D wireframe reconstruction.

3.1 Phase 1: Roof Geometrical Components Segmentation Using a Semantic U-Net Based Network

For the task of building line and roof plane segmentation, U-Net architecture with ResNet-50 blocks has been utilised. U-Net has an encoder-decoder structure, in which the contracting path represents the encoder, and the expanding path represents the decoder. As a result of this design, the information is encoded in a compressed form and then decoded to reconstruct the original format which is crucial for the accuracy of image segmentation (H. Huang et al., 2020).
Our approach involves the following detailed steps:

3.1.1. Data Preparation:

The roofline and roof plane masks were concatenated to create a single combined mask for each image. This can be represented as:
$Mask_{combined} = \text{concat}\ (Mask_{roof\_lines}, Mask_{roof\_planes})$
where concat (X, Y) denotes the concatenation of tensors X and Y along the specified axis.

3.1.2. Network Architecture:

RGB Network: The RGB images and concatenated masks were input into a U-Net network with ResNet-50 serving as the encoder backbone.
DSM Network: Similarly, the DSM images and concatenated masks were input into another U-Net network with the same ResNet-50 backbone.

3.1.3. Feature Fusion:

The encoder outputs from the RGB and DSM networks were concatenated to combine optical and height features. This operation is expressed as:

$F_{fused} = \text{concat}\ (F_{RGB}, F_{DSM})$

where F_{RGBF} and F_{DSM} represent the feature maps from the RGB and DSM networks, respectively. The fused feature map F_{fused} was then processed through a Fully Convolutional Layer (FCL) and trained.

3.1.4. Segmentation Output:

The FCL output was passed to multiple segmentation heads, each responsible for generating masks for roof lines and roof planes which can be used for 3D building modelings.
The architecture of a shared backbone is shown in Figure 2.

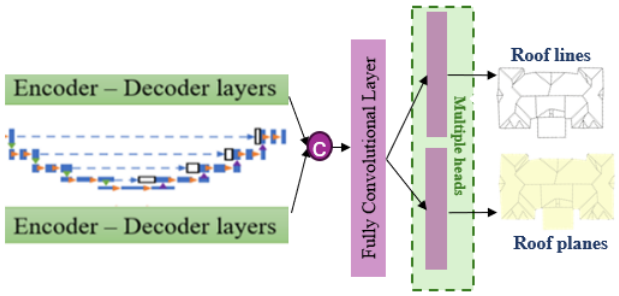


Figure 2. Shared backbone architecture

3.2 Phase 2: LoD2 3D Wireframe Reconstruction

The methodology for 3D wireframe reconstruction of buildings involves several key steps, which are outlined below:

3.2.1. Conversion to Vector Format:

The raster planes and lines are first converted to polygons required for LoD2 reconstruction. This involves identifying pixels with a value of 1 and converting them into the LineString format.

3.2.2. Simplification Using Douglas-Peucker Algorithm:

The converted vector files are simplified using the Douglas-Peucker algorithm. This algorithm works by recursively dividing a polyline into segments and retaining points that significantly deviate from a straight line connecting the endpoints, thereby reducing complexity while preserving essential geometrical features (Douglas and Peucker, 1973).

3.2.3. Transformation to Polygonal Structure:

The modified roof lines are transformed into a polygonal structure, where the boundaries of the polygons represent the roof eaves. The height of the roof eave line segments is determined using LiDAR point cloud data. By assigning zero elevation to the roof eave, the building footprints are established.

3.2.4. 3D Wireframe Modeling:

The edges of the eave and footprint are identified and corresponding edges sharing the same x and y coordinates are connected. This process forms the structural basis for 3D wall modeling (Figure 1, section e). Using accurate height measurements of each line segment derived from the LiDAR data, the 3D frames of the buildings are reconstructed. This ensures that each segment accurately represents the building's height, contributing to the overall precision of the 3D model.

4. Experiments

In this study, we used the first returns of LiDAR point clouds with a point density of six to create a digital surface model and also high-resolution orthoimages with spatial resolution of 72 mm. The orthoimages were manually digitized, modified to extract optical datasets from each building. Then, roof geometrical component masks including line and plane masks were created using optical and DSM datasets of 1624 buildings. 70% of the data was used for training and validation purposes while 30% of data were used for testing the network. The details of the input data are provided in Table 1.

First, we designed a unimodal line segmentation network to extract line masks from orthoimages as a baseline network. Then, in order to improve the line segmentation network, we used DSM as a secondary data source and converted the unified network to a multi-object segmentation network with a shared backbone between objects. To extract meaningful optical and height features of each building in our proposed network, we separately fed the optical images and DSM data in addition to the concatenated lines and plane masks to the two ResNet-50 encoder and decoder layers in our backbone. Next, we extracted two sets of descriptors from these two encoder-decoder layers. These descriptors are concatenated together using the proposed bimodal feature fusion network.

City	Data	Resolution	Date	Source	Number of buildings
Fredericton, Canada	Orthophoto	72 mm	10/ 2015	GeoNB http://www.snb.ca/geonb	1624
	LiDAR	6 p/m^2	09/ 2019		

Table 1. Specification of the data used in this study

The descriptor or boosted feature map is followed by two segmentation heads which are related to each object resulting in roof line and plane masks.

Traditionally, ResNets are initialized with random weight parameters, requiring substantial computational resources and extensive training data. To reduce these demands and avoid overfitting, we used a pre-trained ResNet on the ImageNet dataset. This model is then fine-tuned using transfer learning principles to adapt to our specific task of roof geometrical components segmentation (Bengio, 2012; Donahue et al., 2013). To find an optimal learning rate, we trained the network with a range of learning rates, including 0.1, 0.01, 0.001, 0.0001, 0.00001, and 0.000001, each for just four epochs. The learning rate that results in the minimum loss function is selected as the starting point, and then we fine-tune the learning rate by exploring three decimal places below and above the chosen value. The one that has the minimum lost function is chosen as an optimal learning rate. The network was trained for 100 epochs. To quantify the performance of the proposed segmentation network for building geometrical components detection, we used three metrics, namely Intersection over Union (IoU), Dice Coefficient (F1-Score), and F2-Score.

$$IoU = \frac{\text{Area of Overlap}}{\text{Area of Union}}$$

Where:

Area of Overlap is the pixel area common to both the predicted segmentation and the ground truth (correctly predicted building line and plane areas).

The Area of Union includes all areas predicted as building line and plane plus all ground truth building geometrical components, accounting for both correctly and incorrectly predicted areas.

Dice Coefficient or F1-Score, is similar to IoU but gives twice as much weight to the intersection part. It is defined as:

$$F1_{Score} = \frac{2 * \text{Area of Overlap}}{\text{Total Pixels of Ground Truth} + \text{Total Pixels of Prediction}}$$

This metric can balance the precision (how many of the predicted pixels for buildings are correct) and recall (how many of the actual building pixels were correctly predicted). Finally, F2-score is a variation of the F1-Score that adjusts the beta parameter to weight recall higher than precision. It is particularly useful when the cost of a false negative (failing to detect a part of a line or plane) is higher than that of a false positive (incorrectly marking non-line or non-plane areas as line or plane). The formula for the F2 Score is:

$$F2\_score = \frac{(1 + 2^2) * \text{Precision} * \text{Recall}}{(2^2 * Precison) + Recall}$$

Where:

Precision is the proportion of positive identifications that were correct.

Recall is the proportion of actual positives that were correctly identified.

It is also necessary to assess whether the proposed LOD2 3D wireframe modelling performs properly over the semantic roof geometrical components segmentation task. Thus, we needed to assess the final 3D wireframe model of buildings. The accuracy of the final 3D model depends on the accuracy of the roof geometrical components segmentation and building decomposition steps. While CityGML-3 does not prescribe any fixed values, according to the CityGML-2 standard, the geometric error of 3D models should not exceed two meters.

To evaluate the accuracy of the final 3D wireframe model, we used the digital surface model as the ground truth and calculated the root mean square error (RMSE) (Chai & Draxler, 2014) between each 3D building model and DSM. The formula for the RMSE is presented in Equation (3).

$$\text{RMSE} = \sum_{i=1}^{N} \sqrt{(x_i - \hat{x}_i)^2}$$

Where i is variable, N is the number of buildings, x_i is DSM value of each building and $\hat{x}_i$ is the 3D model of the building.

5. Results and Discussion

The results of this research are divided into two phases. The first phase is dedicated to the detection of the roof geometrical components, and the second phase deals with the creation of 3D wireframe modeling. We report the results of each phase and discuss them in the following sections.

5.1 Roof Geometrical Components Segmentation

The raster line and plane masks resulting from the proposed network are shown in Figure 3. White pixels in masks represent objects.

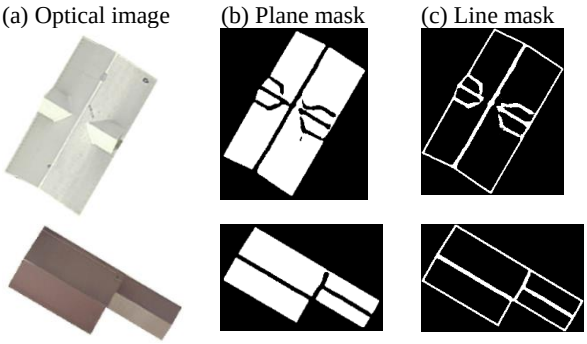


Figure 3. Plane and line masks resulting from the proposed network

The comprehensive performance analysis of optical-based line segmentation network compared to the multi-object bimodal semantic segmentation is presented in Table 2.
As we can see from Table 2, when the network is based solely on optical data, the IoU, F1, and F2 scores are 0.772, 0.7654, and 0.772, respectively. In contrast, using the proposed network

Networks	IoU	F1-Score	F2-Score
Optical-based line segmentation network	0.772	0.7654	0.7720
Multi-object bimodal segmentation network (proposed network)	0.937	0.9390	0.9378

Table 2. Performance analysis of the proposed network and the line segmentation network

results in improved IoU, F1, and F2 scores of 0.937, 0.939, and 0.9378, respectively. This indicates that using the proposed method leads to substantial increases of 16.5%, 17.4%, and 17% in the IoU, F1-score, and F2-score, compared to the baseline unified line segmentation network that relies solely on optical data. As a result, combining DSM data with orthoimages and using a shared backbone between two objects can increase the performance of the segmentation model by leveraging a shared line and plane feature map in addition to the optical and height features of each building.

5.2 LoD2 3D wireframe modeling

The final 3D city model is generated based on roof geometrical components extracted using the proposed multi-object segmentation method. Snapshots of the LoD2 3D wireframe model of buildings are presented in Figure 4.

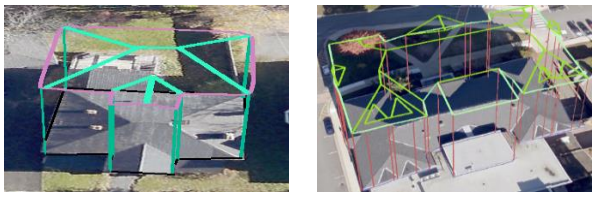


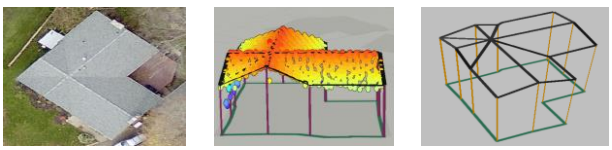
Figure 4. Snapshots of 3D wireframes of buildings

The RMSE of the 14 3D wireframe models of buildings with different roof structures are represented in Table 3. For the accuracy assessment of the final 3D wireframe, we used CityGML, an open standardized data model and exchange format for storing digital 3D models of cities and landscapes. Since version 3 of CityGML doesn't specify accuracy standards, we referred to version 2.

Roof type	RMSE (m)	Roof type	RMSE (m)
Complex	0.77	Flat	0.16
Cross-hip	0.36	Dutch	0.32
Cross-gable	0.32	Gable	0.31
Pyramid	0.29	Gambrel	0.19
Hip	0.23		

Table 3. RMSE result of the 3D wireframe models

According to the CityGML 2.0 standard, the accuracy of the 3D city model should be better than 2 meters. As shown in **Table 3**, the RMSE of the 3D wireframe models even with the complex roof structure is less than 1 meters. Therefore, the results demonstrate high accuracy for the presented model though its performance varies on different roof types (Table 3). Furthermore, for a better understanding of how the model correlates with the LiDAR point cloud, orthophoto, the overlay of the LiDAR point cloud on the 3D building model, and 3D wireframe are shown in Figure 5.



(a) Orthoimage (b) LiDAR Point clouds overlay on wireframe (c) Wireframe

Figure 5. Comparison of LiDAR point clouds to the 3D wireframe models

6. Conclusion

In this study, we proposed a multi-object semantic segmentation network to detect building roof lines and mask pixels using optical and DSM data and constructing a LoD2 3D wireframe of buildings. The methodology inputs high-resolution orthophotos, and LiDAR point cloud data and follows two different phases to create the 3D wireframes. In the first phase, a bimodal muti-object UNet-based network is used to segment roof geometrical components. In this way, the line and plane features are extracted in a shared backbone. Then, the geometrical components are recognized by fusing the optical and DSM features through a deep multi-object segmentation network. In the second phase, the 3D wireframe models of buildings are created using the building's geometry information, such as roof lines and planes. As shown in this paper, our roof geometrical components segmentation network confirmed that utilizing the optical and height features of each building besides using a shared backbone to extract roof geometrical components can improve the semantic segmentation performance, thereby enhancing the overall accuracy of 3D wireframe reconstruction.

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). https://doi.org/10.1088/1755-1315/18/1/012064

Alharthy, A., & Bethel, J. (2004). *DETAILED BUILDING RECONSTRUCTION FROM AIRBORNE LASER DATA USING A MOVING SURFACE METHOD*.

Bengio, Y. (2012). *Deep Learning of Representations for Unsupervised and Transfer Learning* (Vol. 27). http://www.causality.inf.ethz.ch/unsupervised-learning.php

Benjamin Bischke, Patrick Helber, Joachim Folz, Damian Borth, & Andreas Dengel. (2019). MULTI-TASK LEARNING FOR SEGMENTATION OF BUILDING FOOTPRINTS WITH DEEP NEURAL NETWORKS. *2019 IEEE International Conference on Image Processing (ICIP)*.

Biljecki, F., Stoter, J., Ledoux, H., Zlatanova, S., & Çöltekin, A. (2015). Applications of 3D city models: State of the art review. In *ISPRS International Journal of Geo-Information* (Vol. 4, Issue 4, pp. 2842–2889). MDPI AG. https://doi.org/10.3390/ijgi4042842

Buyukdemircioglu, M., Can, R., & Kocaman, S. (2021). Deep learning based roof type classification using very high resolution aerial imagery. *International Archives of the Photogrammetry, Remote Sensing and Spatial Information Sciences - ISPRS Archives*, *43*(B3-2021), 55–60. https://doi.org/10.5194/isprs-archives-XLIII-B3-2021-55-2021

Chai, T., & Draxler, R. R. (2014). Root mean square error (RMSE) or mean absolute error (MAE)? -Arguments against avoiding RMSE in the literature. *Geoscientific Model Development*, *7*(3), 1247–1250. https://doi.org/10.5194/gmd-7-1247-2014

Crawshaw, M. (2020). Multi-Task Learning with Deep Neural Networks: A Survey. *ArXiv Preprint ArXiv:2009.09796*. http://arxiv.org/abs/2009.09796

Donahue, J., Jia, Y., Vinyals, O., Hoffman, J., Zhang, N., Tzeng, E., & Darrell, T. (2013). *DeCAF: A Deep Convolutional Activation Feature for Generic Visual Recognition*. http://arxiv.org/abs/1310.1531

Dorninger, P., & Pfeifer, N. (2008). A comprehensive automated 3D approach for building extraction, reconstruction, and regularization from airborne laser scanning point clouds. *Sensors*, *8*(11), 7323–7343. https://doi.org/10.3390/s8117323

Doulamis, A., & Preka, D. (2016). *3D Building Modeling in LoD2 using the CityGML Standard*. https://www.researchgate.net/publication/309384841

Guo, M., Liu, H., Xu, Y., & Huang, Y. (2020). Building extraction based on U-net with an attention block and multiple losses. *Remote Sensing*, *12*(9). https://doi.org/10.3390/RS12091400

Hartley, R., & Zisserman, A. (2002). *Multiple view geometry in computer vision* (2003rd ed.).

Huang, H., Brenner, C., & Sester, M. (2011). *3D Building Roof Reconstruction from Point Clouds via Generative Models*.

Huang, H., Brenner, C., & Sester, M. (2013). A generative statistical approach to automatic 3D building roof reconstruction from laser scanning data. *ISPRS Journal of Photogrammetry and Remote Sensing*, *79*, 29–43. https://doi.org/10.1016/j.isprsjprs.2013.02.004

Huang, H., Lin, L., Tong, R., Hu, H., Zhang, Q., Iwamoto, Y., Han, X., Chen, Y.-W., & Wu, J. (2020). *UNET 3+: A FULL-SCALE CONNECTED UNET FOR MEDICAL IMAGE SEGMENTATION*.

Huang, J., Stoter, J., Peters, R., & Nan, L. (2022). City3D: Large-Scale Building Reconstruction from Airborne LiDAR Point Clouds. *Remote Sensing*, *14*(9). https://doi.org/10.3390/rs14092254

Jadon, S. (2020). *A survey of loss functions for semantic segmentation*. https://doi.org/10.1109/CIBCB48159.2020.9277638

Jewell, J., Ning, J., Ma, S., & Ahmadi, E. (2019). *A Comparative Study of Semantic Segmentation Models for Building Footprint Extraction Using Satellite Imagery*.

Jiang, X., & Bunke, H. (1994). *Fast segmentation of range images into planar regions by scan line grouping*.

Kada, M. (2022). 3D RECONSTRUCTION OF SIMPLE BUILDINGS FROM POINT CLOUDS USING NEURAL NETWORKS WITH CONTINUOUS CONVOLUTIONS (CONVPOINT). *International Archives of the Photogrammetry, Remote Sensing and Spatial Information Sciences - ISPRS Archives*, *48*(4/W4-2022), 61–66. https://doi.org/10.5194/isprs-archives-XLVIII-4-W4-2022-61-2022

Krafczek, M., & Jabari, S. (2022). Generating LOD2 city models using a hybrid-driven approach: A case study for New Brunswick urban environment. *Geomatica*, *75*(1), 130–147. https://doi.org/10.1139/geomat-2021-0016

Lafarge, F., Descombes, X., Zerubia, J., & Pierrot-Deseilligny, M. (2010). Structural approach for building reconstruction from a single DSM. *JOURNAL OF L A T E X CLASS FILES*, *32*(1), 135–147. https://doi.org/10.1109/TPAMI.2008.281ï

Li, W., He, C., Fang, J., Zheng, J., Fu, H., & Yu, L. (2019). Semantic segmentation-based building footprint extraction using very high-resolution satellite images and multi-source GIS data. *Remote Sensing*, *11*(4). https://doi.org/10.3390/rs11040403

Liu, P., Liu, X., Liu, M., Shi, Q., Yang, J., Xu, X., & Zhang, Y. (2019). Building footprint extraction from high-resolution images via spatial residual inception convolutional neural network. *Remote Sensing*, *11*(7). https://doi.org/10.3390/rs11070830

Park, Y., & Guldmann, J. M. (2019). Creating 3D city models with building footprints and LIDAR point cloud classification: A machine learning approach. *Computers, Environment and Urban Systems*, *75*, 76–89. https://doi.org/10.1016/j.compenvurbsys.2019.01.004

Pepe, M., Costantino, D., Alfio, V. S., Vozza, G., & Cartellino, E. (2021). A novel method based on deep learning, gis and geomatics software for building a 3d city model from vhr satellite stereo imagery. *ISPRS International Journal of Geo-Information*, *10*(10). https://doi.org/10.3390/ijgi10100697

Peters, R., Dukai, B., Vitalis, S., Van Liempt, J., & Stoter, J. (2021). *Automated 3D reconstruction of LoD2 and LoD1 models for all 10 million buildings of the Netherlands*.

Robinson, C., Ortiz, A., Park, H., Bank, W., Lozano, N., World Bank, G., Kaw, J. K., Sederholm, T., Dodhia, R., & Ferres, J. M. L. (2022). *Fast building segmentation from satellite imagery and few local labels*.

Sang, D. V., & Minh, N. D. (2018). Fully residual convolutional neural networks for aerial image segmentation. *ACM International Conference Proceeding Series*, 289–296. https://doi.org/10.1145/3287921.3287970

Schuegraf, P., Shan, J., & Bittner, K. (2024). PLANES4LOD2: Reconstruction of LoD-2 building models using a depth attention-based fully convolutional neural network. *ISPRS Journal of Photogrammetry and Remote Sensing*, *211*, 425–437. https://doi.org/10.1016/j.isprsjprs.2024.04.015

Shan, Jie., & Toth, C. K. (2009). *Topographic laser ranging and scanning : principles and processing*. CRC Press/Taylor & Francis Group.

Soleimani Vostikolaei, F., & Jabari, S. (2023). Large-Scale LoD2 Building Modeling using Deep Multimodal Feature Fusion. *Canadian Journal of Remote Sensing*, *49*(1). https://doi.org/10.1080/07038992.2023.2236243

Tripodi, S., Duan, L., Poujade, V., Trastour, F., Bauchet, J. P., Laurore, L., & Tarabalka, Y. (2020). Operational Pipeline for Large-Scale 3D Reconstruction of Buildings from Satellite Images. *International Geoscience and Remote Sensing Symposium (IGARSS)*, 445–448. https://doi.org/10.1109/IGARSS39084.2020.9324213

Zhao, C., Zhang, C., Yan, Y., & Su, N. (2021). A 3d reconstruction framework of buildings using single off-nadir satellite image. *Remote Sensing*, *13*(21). https://doi.org/10.3390/rs13214434