Y-SHAPED CONVOLUTIONAL NEURAL NETWORK FOR 3D ROOF ELEMENTS EXTRACTION TO RECONSTRUCT BUILDING MODELS FROM A SINGLE AERIAL IMAGE

F. Alidoost [1, *], H. Arefi [2], M. Hahn[1]

[1] Photogrammetry and Geoinformatics, Faculty of Geomatics, Computer Science and Mathematics, Hochschule für Technik Stuttgart, Germany – (fatemeh.alidoost, michael.hahn)@hft-stuttgart.de
[2] School of Surveying and Geospatial Engineering, College of Engineering, University of Tehran, Iran – hossein.arefi@ut.ac.ir

KEY WORDS: Building Reconstruction, Deep Learning, Single Image, Convolutional Neural Networks (CNNs), Rooflines

ABSTRACT:

Fast and efficient detection and reconstruction of buildings have become essential in real-time applications such as navigation, 3D rendering, augmented reality, and 3D smart cities. In this study, a modern Deep Learning (DL)-based framework is proposed for automatic detection, localization, and height estimation of buildings, simultaneously, from a single aerial image. The proposed framework is based on a Y-shaped Convolutional Neural Network (Y-Net) which includes one encoder and two decoders. The input of the network is a single RGB image, while the outputs are predicted height information of buildings as well as the rooflines in three classes of eave, ridge, and hip lines. The extracted knowledge by the Y-Net (i.e. buildings' heights and rooflines) is utilized for 3D reconstruction of buildings based on the third Level of Detail (LoD2). The main steps of the proposed approach are data preparation, CNNs training, and 3D reconstruction. For the experimental investigations airborne data from Potsdam are used, which were provided by ISPRS. For the predicted heights, the results show an average Root Mean Square Error (RMSE) and a Normalized Median Absolute Deviation (NMAD) of about 3.8 m and 1.3 m, respectively. Moreover, the overall accuracy of the extracted rooflines is about 86%.

1. INTRODUCTION

Buildings are the most prominent objects in urban scenes, thus measuring and analyzing 3D shapes and positions of buildings are essential for many applications such as 3D map updating, urban management, smart cities, monitoring, navigation and mapping, civil infrastructure inspection, and scene understanding. Hence, a considerable number of researches is dedicated to automatic building detection, localization, and reconstruction in photogrammetry and remote sensing.

As a general categorization, current algorithms for 3D Building Reconstruction (3DBR) can be divided into three basic methods: data-driven (Awrangjeb et al., 2018; Cheng et al., 2011; Kim and Shan, 2011; Sampath and Shan, 2010; Yan et al., 2017), model-driven (Huang et al., 2011; Partovi et al., 2015; Zhang et al., 2014; Zheng et al., 2017), and hybrid methods (Wang et al., 2016; Xiong et al., 2015). The differences between the data-driven and model-driven methods have been discussed in previous studies (Tarsha-Kurdi et al., 2006; R. Wang et al., 2018).

The remotely sensed data such as stereo aerial and satellite images or LiDAR data are the main sources to extract 3D information of urban objects using photogrammetry techniques. However, these data sources are not available everywhere and generation of updated Digital Surface Models (DSMs) needs a considerable amount of effort, time, and cost, especially for large areas. On the other hand, sometimes, it is not possible to capture images from different views to reconstruct 3D models because of obstacles and occluded areas or the limited acquisition time.

To address this issue, many investigations are attempting to reconstruct 3D scenes from monocular images such as single satellite and aerial images as a low-cost solution for rapid 3D mapping and fast 3D visualization and rendering of urban scenes. As widely known 3D reconstruction from a single satellite or aerial image is a difficult ill-posed problem because of inherent ambiguities related to the scale and shape of the object. However, it is possible to extract structural information of the objects or measure the topology and geometry constraints from a single image in order to generate 3D models.

One of the state-of-the-art techniques to extract high-level information from a single remotely sensed image is based on deep learning-based algorithms and Convolutional Neural Networks (CNNs). The high-level information is semantic or geometric features such as depth or height of objects, land cover labels, textures, or camera exterior parameters that can be integrated to reconstruct the 3D shape of an object. However, buildings' heights or footprints are not sufficient for 3D reconstruction of buildings and geometric structures of building roofs such as planes and linear elements of roofs are required for 3DBR.

In this paper, the proposed approach for 3DBR is based on extracting the high-level knowledge from an RGB image and forming them to generate parametric models. The required knowledge for 3DBR includes the location of buildings, the linear elements of building roofs (i.e. rooflines) such as eave, ridge, and hip lines as well as the heights of buildings (e.g. normalized DSMs), which are effective to reduce the complexity of reconstruction. However, extracting 3D information from a single 2D image is impossible and under constraint theoretically. Therefore, a novel method including a Y-shaped Convolutional Neural Network (Y-Net) is employed to extract nDSM as well as segmented linear elements of building roofs, simultaneously, from single RGB images. This work's contributions are as follows.

- 3D parametric models of buildings (LoD2) can be constructed from a single RGB image contributing to a better understanding and interpreting the 3D scenes in real-time applications;

* Corresponding author

- Unlike the traditional photogrammetric techniques for 3DBR from a single image, the proposed method can extract the height information from non-oblique and nearly vertical single images using a CNN;
- Since nDSMs and rooflines share high-level features and representations of a building, the geometric structures of buildings can be learned efficiently during a two-stream network training.

2. RELATED WORK

There are several studies for 3D reconstruction of objects from a single image using photogrammetry techniques. These studies are mostly relying on detecting vanishing features (e.g. points and lines) as well as estimating the camera calibration parameters from oblique images. One of the earliest studies to restore 3D information from a single image is based on deriving geometric constraints such as image lines and object topologies during image interpretations (Van Den Heuvel, 1998). (Jizhou et al., 2004) proposed a framework to extract the height of buildings from an oblique UAV-based image. Their framework is based on the extraction of parallel lines and view angles of buildings. However, they also employed digital maps to calculate the scale of 3D models. (González-Aguilera et al., 2005) developed a software to extract 3D models based on vanishing points geometry of an oblique image. Later, they improved the accuracy of extracting vanishing points and lines using the RANSAC algorithm (Gonzalez-Aguilera and Gomez-Lahoz, 2008).

Nowadays, deep learning algorithms have shown remarkable performances in the automatic 3D reconstruction of objects from single RGB images in computer vision applications (Fan et al., 2016; Henderson and Ferrari, 2019; J. Wang et al., 2018; Wu et al., 2017). In photogrammetry and remote sensing, CNNs can be employed to extract height information such as DSMs from single aerial or satellite-based images (Amini Amirkolaee and Arefi, 2019; Ghamisi and Yokoya, 2018), as well as building detection and footprints extraction (Aamir et al., 2019; Wu et al., 2018; Xu et al., 2018; Yang et al., 2018). (Li et al., 2019) used two independent CNNs for land cover classification and building height estimation from single satellite images. The CNN for height estimation task is a fully connected network and estimates a fixed height value for each building block for 3D reconstruction in LoD1. (Tripodi et al., 2019) employed the U-Net to extract the building footprints from single satellite images. Since the footprints have no extra information about the shapes of the building roofs, the final 3D models are in LoD1 only.

3. PROPOSED METHOD

As shown in Figure 1, the proposed framework for 3DBR based on the Y-Net includes three main steps as data preparation, CNN training, and 3D reconstruction. First, a training dataset is generated for height prediction and roofline extraction. Next, a Y-shaped CNN is designed which includes one encoder block to extract features from input images and two decoder blocks to convert extracted features to nDSMs as well as rooflines. After training the Y-Net using the generated training dataset, it is applied to a test image to extract the essential knowledge of 3DBR. In the third step of the proposed approach, predicted rooflines and nDSMs are combined together in order to generate parametric models of buildings in LoD2, according to the CityGML Standard. The summary of each step and their main components are given in the following sub-sections.

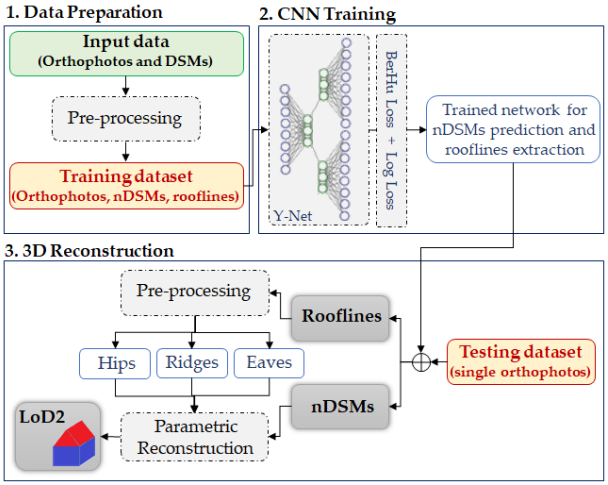


Figure 1. The flowchart of the proposed method

3.1 Data Preparation

The main data used in this study include aerial orthophotos and the corresponding DSMs. On the other hand, the required training dataset for the proposed framework should be composed of RGB images (Figure 2, a) and corresponding nDSMs (Figure 2, b) as well as rooflines (Figure 2, c). Therefore, the Digital Terrain Models (DTMs) are first generated from DSMs by employing the progressive TIN densification algorithm (Axelsson, 2000), and then nDSMs are calculated by subtracting DTMs from DSMs. The nDSMs include the absolute height values of urban objects from the bare Earth. To generate corresponding rooflines, the aerial orthophotos are manually digitized for linear elements of individual roofs into three classes of eave, ridge, and hip lines. Next, the vector-based data are converted to raster images including three RGB channels for three classes of rooflines (i.e. R for eave lines, G for ridge lines, and B for hip lines), as shown in Figure 2, c.

In the pre-processing step, several image tiles are cropped from the generated training dataset and resized to the size of 224×224×n, so that n is equal to 3 for orthophotos and rooflines, and 1 for nDSM tiles. Moreover, the number of training samples increases using different data augmentation techniques such as scaling, rotating, and flipping operations.

a b c

Figure 2. A sample of generated training data including a: the RGB image; b: the nDSM; c: rooflines

3.2 CNN Training

In this paper, a novel convolutional-deconvolutional network (Y-Net) is proposed to extract the height data and rooflines, simultaneously from a single image. The network includes one encoder and two decoders. The structure of the Y-Net is shown in Figure 3. The encoder extracts the high levels of features from RGB images.

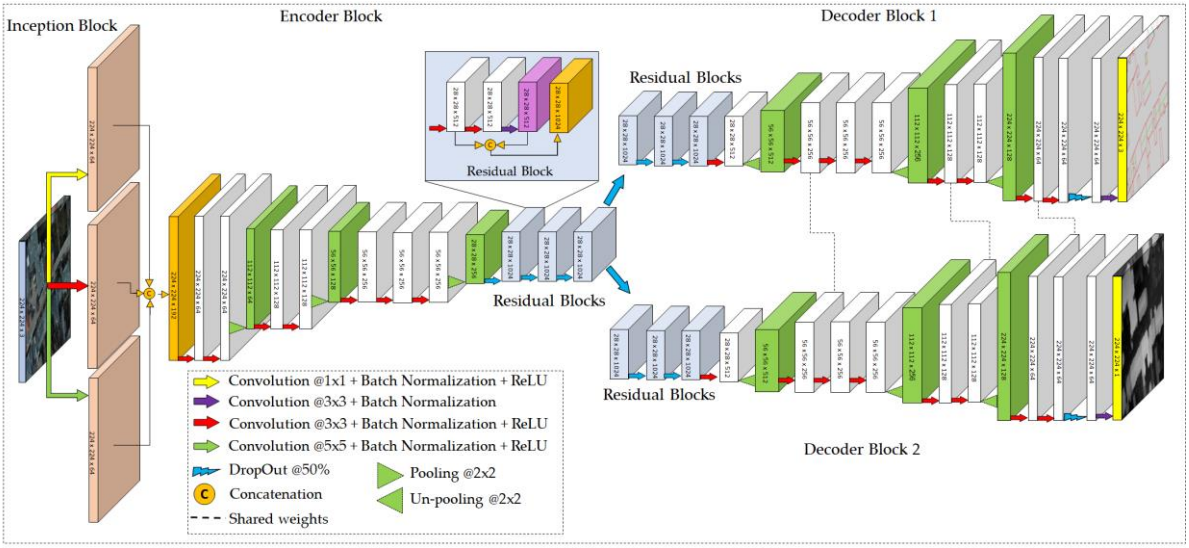


Figure 3. The proposed Y-Net

The first part of the encoder is an inception-based module with three different sizes of filters (i.e. 1×1, 3×3, and 5×5) which allows the network to take advantage of multi-level features extraction and improves the generalization capability of the network. For instance, it extracts general (5×5) and local (1×1) features at the same time. Next, there are 7 convolutional layers followed by Batch Normalization (BN) and Rectified Linear Unit (ReLU) layers to generate feature maps, as well as three max-pooling layers to reduce the size of feature maps by a factor of 2. The convolutional layers include 3 × 3 kernels with a stride of 1 and the max-pooling layers include 2 × 2 kernels with a stride of 2. In the last part of the encoder, there are three modified residual blocks. The ideas of skip connections and residual blocks are first introduced in the study by (He et al., 2016). However, the ReLU layers perturb the data flowing through identity connections. Therefore, compared to the original residual block, in the proposed architecture, the ReLU layers are removed after addition in order to boost the performance of the network.

Y-Net includes two decoders which are exactly the opposite of the encoder. One of the decoders contains the parameters of a regression-based model to convert high-level features into height values of objects (nDSMs), while the other one is for a segmentation-based problem and converts features into rooflines. Since nDSMs and rooflines share the same high-level features and representations of buildings, we applied a weight-sharing constraint between three convolutional layers of two decoders. By sharing features between two decoders, the network is able to estimate more accurate nDSMs for rooflines which are important for 3DBR. The size of the input is 224×224×3, while the output sizes are 224×224×1 and 224×224×3 for predicted nDSMs and rooflines, respectively.

To train Y-Net, random initial values are considered for training parameters. Moreover, the berHu loss function (Laina et al., 2016) is applied for nDSM prediction, given by Equation 1. While the logistic log loss is used for roofline segmentation, given by Equation 2. The combination of the loss functions is utilized as Equation 3 and the network is trained using the ADAM optimizer (Kingma and Ba, 2015).

$$L_1(x) = \begin{cases} |x| & |x| \le c \\ \dfrac{x^2 + c^2}{2c} & |x| > c \end{cases} \quad (1)$$

where, $\boldsymbol{x}$ is the difference between the predicted and ground truth values, and $\boldsymbol{c}$ is 20% of the maximal per-batch error.

$$L_2(x, c) = \log(1 + \exp(-c.x)) \quad (2)$$

where, $\boldsymbol{c}$ is a binary attribute of ground truth values in (+1, -1). Here, +1 denotes the presence of an attribute, and -1 denotes its absence.

$$L = L_1(x) + \alpha\, L_2(x, c) \quad (3)$$

where, $\boldsymbol{\alpha}$ is a scale factor for combining two loss functions and equals to 0.001, in this study.

3.3 3D Reconstruction

The proposed approach for 3DBR from a single image relies on extracting the essential geometrical knowledge of buildings such as nDSMs (Figure 6, b) and rooflines (Figure 6, c) by applying the trained Y-Net to a test image (Figure 6, a), as shown in Figure 6. The predicted rooflines in three classes of eave, ridge, and hip lines are used to define the locations and orientations of individual building parts. In this approach, most of the building blocks are decomposed into the individual building parts including flat, gable or hip buildings by analysing of the predicted rooflines. In the first step of proposed approach for 3DBR, the predicted rooflines are pre-processed to remove all small and noisy segments. Next, binary polygons of building blocks (Figure 6, e) are generated using the first channel of rooflines which is mostly composed of eave lines (Figure 6, d). The Minimum Bounding Rectangle (MBR)-based technique (Arefi and Reinartz, 2013) is then employed to enhance the binary polygons and convert them to the regularized and approximated polygons (Figure 6, f). The approximated binary polygons are initial primitives for the prismatic models of building blocks (i.e. LoD1). Next, a rule-based search technique (Alidoost et al., 2019) is utilized to decompose the building blocks into individual buildings (Figure 6, g). To this end, the approximated binary polygons and eave lines are rotated based on the main orientation of the building block. Next, for each binary polygon, all vertical or horizontal eave lines inside the polygon and with the endpoints on the boundary of the polygon

are searched. These lines are separator lines which divide building blocks into the individual roofs (Figure 4).

The second channel of the predicted rooflines contains the ridge lines (Figure 6, h), which is utilized to generate parametric models of buildings (i.e. LoD2). The individual ridge line for each individual building is extracted by analyzing the predicted ridge lines inside each binary polygons (Alidoost et al., 2019). A polyline that is parallel to the main orientation of the individual roof and crossing the center of the polygon is the main ridge line, as shown in Figure 5, b. Then, an optimized line is fitted to the candidate polyline to generate the regularized ridge line for the roof (Figure 5, c). The ridge line is extended if the distances between the endpoints and the eave lines are less than 3 m.

Finally, the hip lines can be reconstructed by connecting the endpoints of ridge lines to the vertexes of approximated polygons and the median height values of the eave, ridge and hip lines are then extracted from the predicted nDSMs to generate the final 3D models (Figure 6, k).

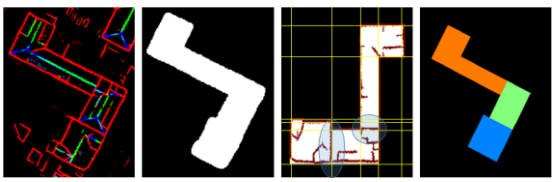


Figure 4. The rule to detect individual building roofs: a: the predicted rooflines; b: the binary polygon of eave lines; c: The approximated binary polygon, corresponding eave lines, and horizontal and vertical lines; d: individual building parts

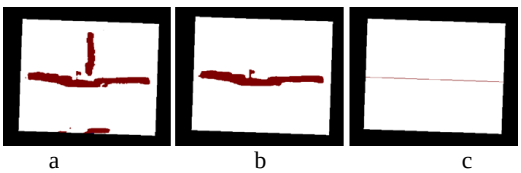


Figure 5. The ridge line detection strategy: a: original ridge lines; b: the ridge line parallel to the main orientation of the polygon; c: the best fitted ridge line

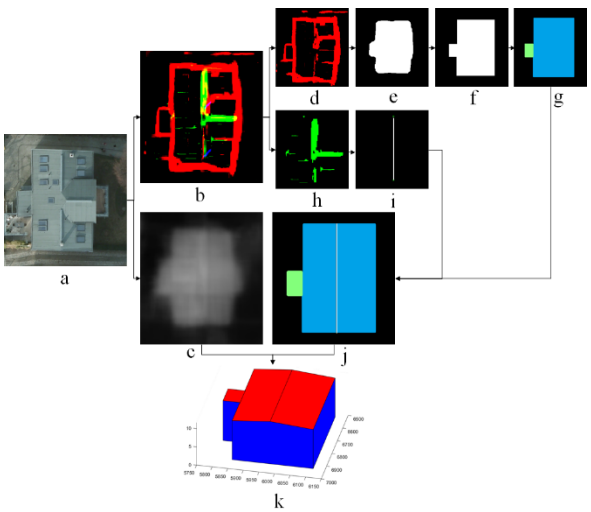


Figure 6. 3D reconstruction of individual building parts: a: the RGB image; b: the predicted rooflines; c: the predicted nDSM; d: the predicted eave lines, e: the binary polygon; f: the approximated polygon; g: the decomposed polygons; h: the predicted ridge lines; i: the best fitted ridge line; j: The final roof elements; k: 3D models of building parts

4. EXPERIMENTS AND RESULTS

To evaluate the performance of the proposed approach, an airborne dataset from Potsdam, Germany, provided by ISPRS (ISPRS, 2018), is used which consists of very high-resolution true orthophoto tiles with a ground sampling distance (GSD) of 5 cm and corresponding DSMs derived from dense image matching techniques. Two non-overlapping areas of this dataset are selected for training the Y-Net and 3D reconstruction, as shown in Figure 7. The training dataset includes 4,800 tiles of RGB images, nDSMs, and rooflines which are increased to 24,000 tiles with a size of 224×224 after data augmentation. The Y-Net was trained using the training dataset on a single NVIDIA GTX 1080 Ti with a batch size of 10 for 100 epochs. The learning rate, beta 1, beta 2, and epsilon parameters are selected as 0.01, 0.9, 0.999, and 1 × 10−8 for the Adam optimizer.

Figure 7. Overview of training and testing datasets

In addition to test data from Potsdam (e.g. Areas 1-4 in Table 1), the second dataset from Zeebrugge, Belgium (IEEE, 2015) consisting of a true ortho-photo with a GSD of 5 cm and LiDAR data with a 10 cm point spacing is also employed to assess the transferability of the trained network (e.g. Area 5 in Table 1).

The trained Y-Net is applied to the testing RGB images and the predicted nDSMs and rooflines are shown in Figure 8, compared to the ground truth data. The accuracy of the estimated nDSMs is evaluated based on standard metrics such as Mean Error (ME), Standard Deviation (SD), Root Mean Square Error (RMSE), Relative Error (REL), and Root Mean Squared Logarithmic Error (RMSLE), as well as robust statistical metrics such as Median Error (MeE), Normalized Median Absolute Deviation (NMAD), Quantile 68.3% (Q68.3), and Quantile 95% (Q95), as reported in Table 1. Also, the results are compared to other studies for the nDSM prediction task for testing areas in Table 2. Since there are random noises, outliers, and systematic errors in the predicted nDSMs, robust metrics are useful to have accurate and reliable assessments.

Metric [m]	Testing Areas					Ave.
	Area1	Area2	Area3	Area4	Area5	All
ME	1.62	1.59	1.49	2.27	1.95	1.78
SD	1.48	1.15	1.39	1.46	1.09	1.31
RMSE	3.84	3.46	3.75	4.24	3.51	3.76
RMSLE	0.03	0.03	0.03	0.04	0.34	0.09
REL [%]	0.07	0.07	0.07	0.08	0.72	0.20
MeE	1.15	1.38	1.02	2.02	1.92	1.50
NMAD	1.25	1.14	1.00	1.66	1.32	1.27
Q68.3	2.00	1.99	1.66	2.99	2.58	2.24
Q95	4.86	3.95	4.67	4.95	3.74	4.43

Table 1. The accuracy of the predicted nDSMs

According to the Table 1, the average RMSE of the predicted nDSMs for all areas is about 3.76 m, while the SD is about 1.31 m. As a result, the distribution of errors is not normal and there are outliers or systematic errors in the predicted nDSMs. Therefore, the NMAD, which is about 1.27 m, is more reliable metric to report the accuracy of the results.

Methods	Metrics		
	RMSE [m]	RMSLE [m]	REL [%]
GAN (Ghamisi and Yokoya, 2018)	3.89	-	-
FCRN (Amini Amirkolaee and Arefi, 2019)	3.47	0.26	0.57
Proposed Y-Net	**3.76**	**0.09**	**0.20**

Table 2. Comparison between the proposed Y-Net and the state-of-the-art methods for nDSM prediction over Potsdam dataset

As shown in Figure 8, not only the linear elements of roofs are extracted appropriately, but the buildings are also classified and distinguished from non-building objects such as trees and roads. The accuracy and quality of the predicted rooflines are calculated using the standard quality measures of completeness (or recall), correctness (or precision), quality (McGlone and Shufelt, 1994; McKeown et al., 2000), the F1 score, and Overall Accuracy (OA), given by Equation (4).

$$Comp. = \frac{TP}{TP + FN};\ Corr. = \frac{TP}{TP + FP};$$
$$Qual. = \frac{TP}{TP + FN + FP};\ F1 = 2.\frac{Corr. \times Comp.}{Corr. + Comp.} \quad (4)$$

where, ***TP*** is the true positive, ***FP*** is the false positive, and ***FN*** is the false negative. The quality measures of testing areas for each class of rooflines (e.g. eave, ridge, and hip lines) are presented in Table 3.

Metric	Roof -line	Testing Areas					Ave
		Area1	Area2	Area3	Area4	Area5	All
TP	eave	395882	116240	398803	251229	88667	-
	ridge	67239	32620	58738	38707	21643	-
	hip	29282	19068	19919	32301	7191	-
FN	eave	28715	5486	36367	20055	5057	-
	ridge	49265	4966	66074	21705	6400	-
	hip	10311	2857	14446	11276	5640	-
FP	eave	24945	6690	43957	21705	6357	-
	ridge	15497	4946	20052	11335	4072	-
	hip	47849	1673	52878	23625	6668	-
Comp. [%]	eave	93.2	95.5	91.6	92.6	94.6	93.5
	ridge	57.7	86.8	47.1	64.1	77.2	66.6
	hip	73.9	86.9	57.9	74.1	56.0	69.8
Corr. [%]	eave	94.1	94.5	90.0	92.0	93.3	92.8
	ridge	81.3	86.8	74.5	77.3	84.2	80.8
	hip	37.9	91.9	27.4	57.7	51.9	53.4
Qual. [%]	eave	88.1	90.5	83.2	85.7	88.6	87.2
	ridge	50.9	76.7	40.5	53.9	67.4	57.9
	hip	33.5	80.8	22.8	48.1	36.9	44.4
F1 [%]	eave	93.6	95.0	90.8	92.3	93.9	93.2
	ridge	67.5	86.8	57.7	70.1	80.5	72.5
	hip	50.2	89.4	37.2	64.9	53.9	59.1
OA		84.8	92.6	85.0	80.3	87.3	86.0

Table 3. The accuracy of the predicted rooflines

The results in Table 3 show that the eave lines are estimated with higher precision (about 92.8%) than ridge and hip lines (about 80.8% and 53.4%, respectively). Accordingly, the trained Y-Net is able to distinguish between building and non-building objects better and there are some misclassification errors in ridge and hip lines.

Finally, the extracted nDSMs and rooflines are employed for 3D reconstruction of buildings. Since the predicted nDSM includes some outliers as well as systematic errors, the median of height values is considered for modeling of each individual roof. On the other hands, the rooflines, which are extracted completely and correctly, are only utilized to generate approximated binary polygons. Accordingly, the 3D parametric models of buildings can be reconstructed by assigning the height values to the binary polygons and ridge lines, as shown in Figure 9. The geometrical accuracy of generated 3D models is measured based on the 3D coordinates of roof planes' vertexes, compared to the ground truth, and reported as RMSxy and RMSz measures. Experimentally we found that the RMSxy value of 3D models is less than 0.5 m, while the RMSz value is about 3.8 m which is within the accuracy range of the predicted nDSMs. In addition, the quality measures of building footprints in Figure 9 are reported in Table 4, inspired by the ISPRS guideline for evaluation of building reconstruction (Rottensteiner, 2013). The average values for completeness, correctness, quality and F1 score are 97.4%, 91.8%, 89.3%, and 81.9%, respectively.

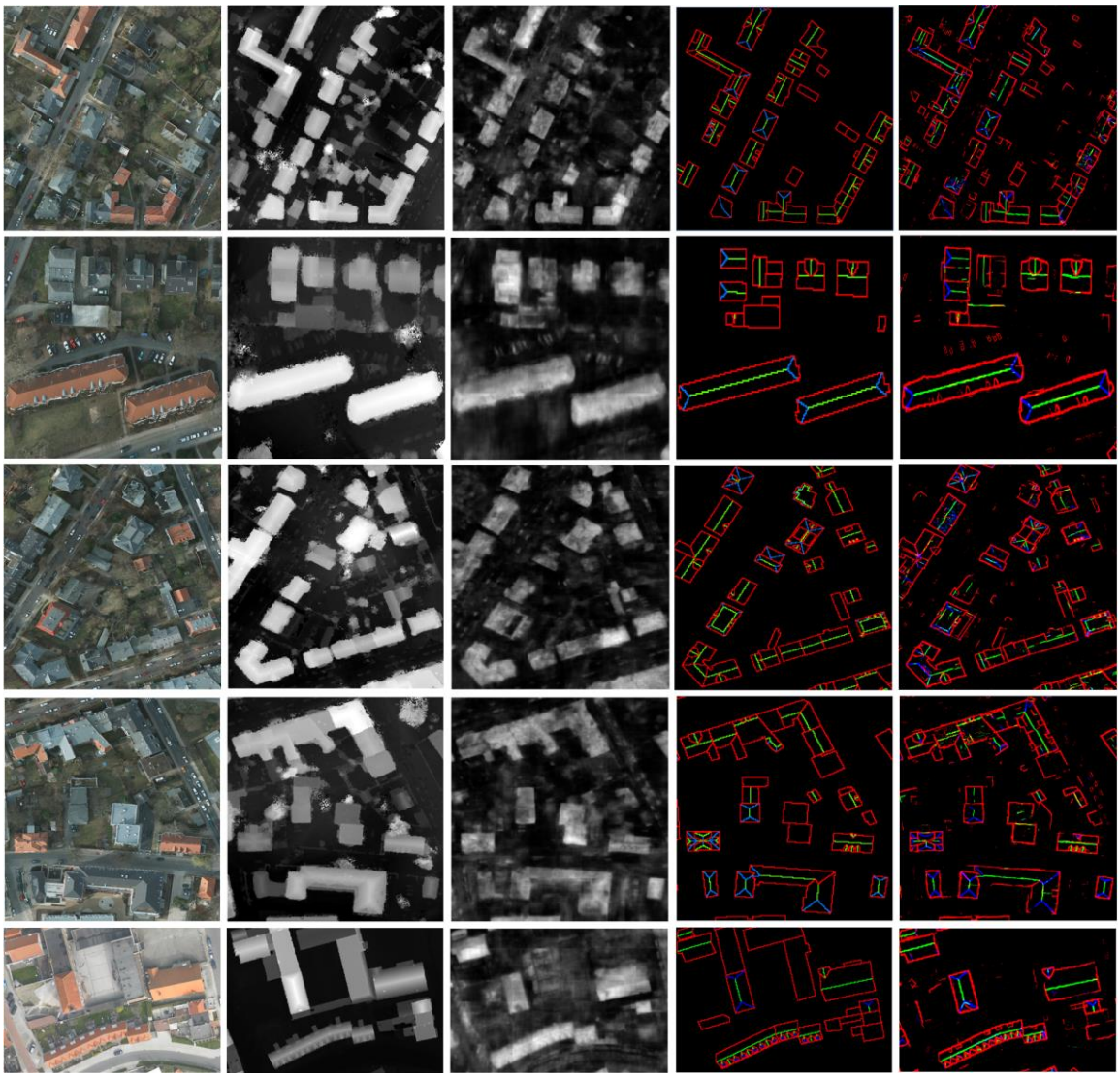
Metric	Testing Buildings						
	B1	B2	B3	B4	B5	B6	B7
TP	446263	239275	80112	241759	124842	44612	175248
FN	19445	6199	1364	0	13831	0	4
FP	11325	23422	12009	54927	631	2997	14209
Comp. [%]	95.8	97.5	98.3	100	90.0	100	100
Corr. [%]	97.5	91.1	86.9	81.5	99.5	93.7	92.5
Qual. [%]	93.5	88.9	85.7	81.5	89.6	93.7	92.5
F1 [%]	96.7	94.2	92.3	89.8	94.5	96.7	96.1

Table 4. The accuracy of the building footprints

Although the accuracy of the predicted nDSMs from single images using deep learning techniques such as the Y-Net is not comparable to the high- resolution DSMs extracted from LiDAR data or image matching techniques, they are valuable information for specific applications such as real-time navigation, rapid 3D rendering, land cover classification using RGB-depth fusion techniques, urban growing, change detection and so on.

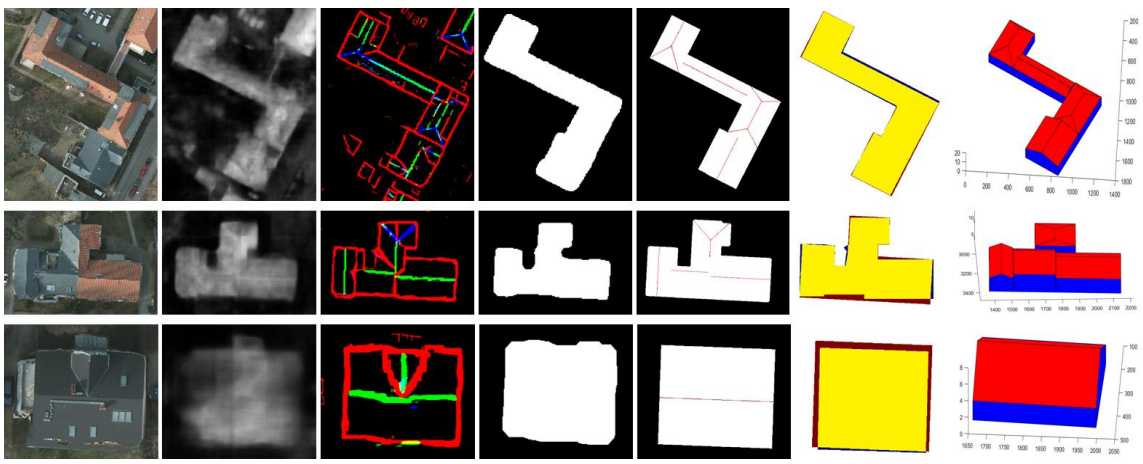
5. CONCLUSION

In this study, we presented a novel approach based on supervised deep learning techniques to extract nDSMs and rooflines of buildings from a single aerial image and generate the parametric models in LoD2.Unlike existing methods in photogrammetry and remote sensing that require both ortho images and high-resolution DSMs, the proposed method uses the single RGB images and the power of CNNs to extract the valuable information which is essential for 3D representations of buildings. Although we have some limitations to produce the proper training dataset for rooflines, the results show the reasonable performance of the proposed Y-Net to predict rooflines with the overall accuracy of 86%, and predict the nDSMs with the RMSE of 3.8 m for different test datasets.

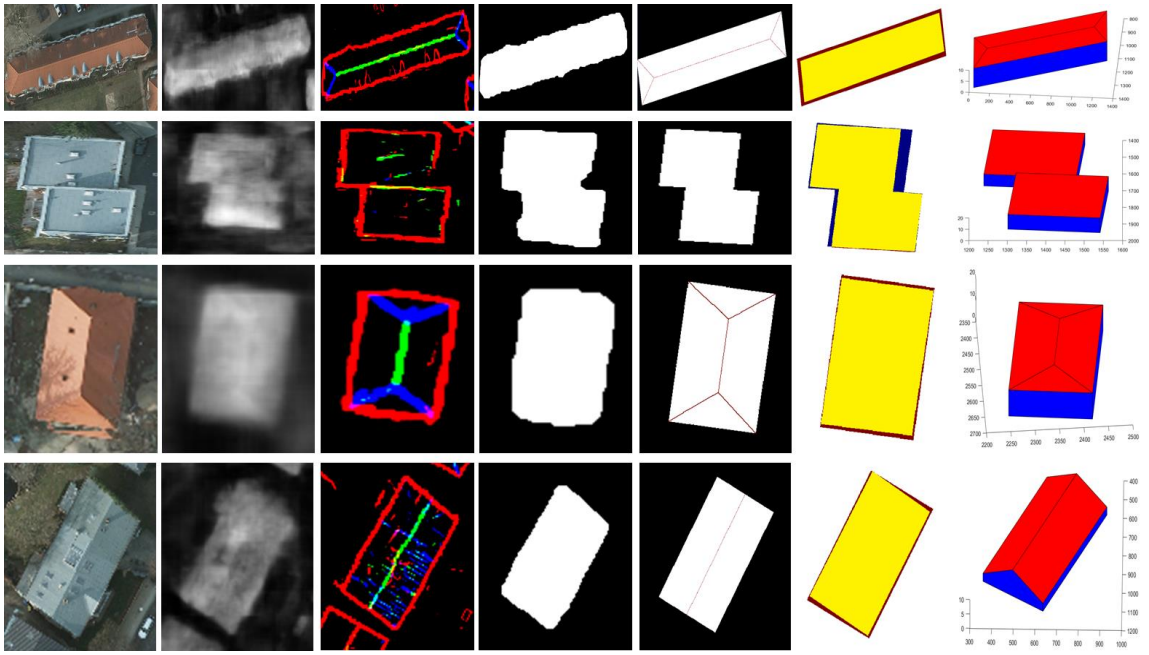


a b c d e

Figure 8. The results of the predicted nDSMs and rooflines: a: the input RGB images from two test datasets; b: the reference nDSMs; c: the predicted nDSMs; d: the reference rooflines; e: the predicted rooflines



This contribution has been peer-reviewed. The double-blind peer-review was conducted on the basis of the full paper.
https://doi.org/10.5194/isprs-annals-V-2-2020-321-2020 | © Authors 2020. CC BY 4.0 License.



a b c d e f g

Figure 9. The results of 3D building reconstruction: a: the input RGB image; b: predicted nDSMs; c: predicted rooflines; d: binary polygons, extracted from eave lines; e: approximated polygons and best fitted ridge lines; f: differences between the building footprints and the ground truth; g: 3D models of roofs in LoD2

REFERENCES

Aamir, M., Pu, Y.F., Rahman, Z., Tahir, M., Naeem, H., Dai, Q., 2019. A framework for automatic building detection from low-contrast satellite images. *Symmetry* (Basel). 11, 1–19. https://doi.org/10.3390/sym11010003

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 Sens.* 11(19), 2219. https://doi.org/10.3390/rs11192219.

Amini Amirkolaee, H., Arefi, H., 2019. Height estimation from single aerial images using a deep convolutional encoder-decoder network. ISPRS J. Photogramm. Remote Sens. 149, 50–66. https://doi.org/10.1016/j.isprsjprs.2019.01.013

Arefi, H., Reinartz, P., 2013. Building Reconstruction Using DSM and Orthorectified Images. *Remote Sens*. 5, 1681–1703. https://doi.org/10.3390/rs5041681

Awrangjeb, M., Ali, S., Gilani, N., 2018. An Effective Data-Driven Method for 3-D Building Roof Reconstruction and Robust Change Detection. *Remote Sens.* 10, 1–31. https://doi.org/10.3390/rs10101512

Axelsson, P.E., 2000. DEM generation from laser scanner data using adaptive TIN models, in: *International Archives of the Photogrammetry and Remote Sensing*. pp. 110–117.

Cheng, L., Gong, J., Li, M., Liu, Y., 2011. 3D Building Model Reconstruction from Multi-view Aerial Imagery and Lidar Data. Photogramm. Eng. *Remote Sens*. 77, 125–139.

Fan, H., Su, H., Guibas, L., 2016. A Point Set Generation Network for 3D Object Reconstruction from a Single Image, in: *CVPR2016*. pp. 1–4.

Ghamisi, P., Yokoya, N., 2018. IMG2DSM: Height Simulation From Single Imagery Using Conditional Generative Adversarial Net. IEEE Geosci. *Remote Sens. Lett.* 5, 794–798. https://doi.org/10.1109/LGRS.2018.2806945
Gonzalez-Aguilera, D., Gomez-Lahoz, J., 2008. From 2D TO 3D through modelling based on a single image. *Photogramm. Rec.* 23, 208–227. https://doi.org/10.1111/j.1477-9730.2008.00482.x

González-Aguilera, D., Gomez-Lahoz, J., Finat-Codes, J., 2005. SV3DVision: 3D Reconstruction and Visualization From a Single View, in: *International Archives of the Photogrammetry, Remote Sensing and Spatial Information Sciences*. p. 8.

He, K., Zhang, X., Ren, S., Sun, J., 2016. Deep Residual Learning for Image Recognition, in: *CVPR2016*. Las Vegas, NV, USA, pp. 1–9. https://doi.org/10.1109/CVPR.2016.90

Henderson, P., Ferrari, V., 2019. Learning Single-Image 3D Reconstruction by Generative Modelling of Shape, Pose and Shading. *Int. J. Comput. Vis.* https://doi.org/10.1007/s11263-019-01219-8

Huang, H., Brenner, C., Sester, M., Hannover, D.-, 2011. 3D Building Roof Reconstruction from Point Clouds via Generative Models Categories and Subject Descriptors, in: *19th ACM SIGSPATIAL International Conference on Advances in Geographic Information Systems (GIS)*. New York, NY, USA, pp. 16–24. https://doi.org/10.1145/2093973.2093977

IEEE, 2015. GRSS Data Fusion Contest. Available online: http://www.grss-ieee.org/community/technicalcommittees/data-

fusion. (accessed on 15 September 2017).

ISPRS, 2018. 2D Semantic Labeling Contest - Potsdam. ISPRS WG II/4. http://www2.isprs.org/commissions/comm3/wg4/2d-sem-label-potsdam.html

Jizhou, W., Zongjian, L., Chengming, L.I., 2004. Reconstruction of Buildings From a Single UAV Image, in: *International Society for Photogrammetry and Remote Sensing Congress*. pp. 100–103.

Kim, K., Shan, J., 2011. Building roof modeling from airborne laser scanning data based on level set approach. *ISPRS J. Photogramm. Remote Sens.* 66, 484–497. https://doi.org/10.1016/j.isprsjprs.2011.02.007

Kingma, D.P., Ba, J.L., 2015. Adam: a Method for Stochastic Optimization, in: *The 3rd International Conference on Learning Representations (ICLR)*. pp. 1–15.

Laina, I., Rupprecht, C.., Belagiannis, V., Tombari, F., Navab, N., 2016. Deeper Depth Prediction with Fully Convolutional Residual Networks, in: *IEEE International Conference on 3D Vision (3DV)*. https://doi.org/10.1109/3dv.2016.32

Li, S., Zhu, Z., Wang, H., Xu, F., 2019. 3D Virtual Urban Scene Reconstruction from a Single Optical Remote Sensing Image. *IEEE Access* 7, 68305–68315. https://doi.org/10.1109/ACCESS.2019.2915932

McGlone, J.C., Shufelt, J.A., 1994. Projective and object space geometry for monocular building extraction, in: *Proceedings of IEEE Conference on Computer Vision and Pattern Recognition CVPR-94*. Seattle, USA, pp. 54–61. https://doi.org/10.1109/CVPR.1994.323810

McKeown, D.M., Bulwinkle, T., Cochran, S., Harvey, W., McGlone, C., Shufelt, J.A., 2000. Performance evaluation for automatic feature extraction, in: *International Archives of the Photogrammetry , Remote Sensing and Spatial Information Sciences*. Amsterdam, The Netherlands, pp. 379–394.

Partovi, T., Huang, H., Krauß, T., Mayer, H., Reinartz, P., 2015. Statistical building roof reconstruction from worldview-2 stereo imagery, in: I*nternational Archives of the Photogrammetry, Remote Sensing and Spatial Information Sciences*. Munich, Germany, pp. 161–167. https://doi.org/10.5194/isprsarchives-XL-3-W2-161-2015

Rottensteiner, F., 2013. Evaluation of Building Reconstruction Results, ISPRS - Commission III - Photogrammetric Computer Vision and Image Analysis Working Group III / 4 - 3D Scene Analysis.

Sampath, A., Shan, J., 2010. Segmentation and reconstruction of polyhedral building roofs from aerial lidar point clouds. *IEEE Trans. Geosci. Remote Sens.* 48, 1554–1567. https://doi.org/10.1109/TGRS.2009.2030180

Tarsha-Kurdi, F., Landes, T., Grussenmeyer, P., Koehl, M., 2006. Model-Driven And Data-Driven Approaches Using Lidar Data : Analysis And Comparison, in: *International Archives of Photogrammetry, Remote Sensing and Spatial Information Sciences*. pp. 87–92.

Tripodi, S., Duan, L., Trastour, F., Poujad, V., Laurore, L., Tarabalka, Y., 2019. Automated chain for large-scale 3D reconstruction of urban scenes from satellite images, in: *International Archives of the Photogrammetry, Remote Sensing and Spatial Information Sciences*. pp. 243–250. https://doi.org/10.5194/isprs-archives-XLII-2-W16-243-2019

Van Den Heuvel, F.A., 1998. 3D reconstruction from a single image using geometric constraints. *ISPRS J. Photogramm. Remote Sens.* 53, 354–368. https://doi.org/10.1016/S0924-2716(98)00019-7

Wang, J., Sun, B., Lu, Y., 2018. MVPNet: Multi-View Point Regression Networks for 3D Object Reconstruction from A Single Image, in: *AAAI Conference on Artificial Intelligence*. pp. 8949–8956. https://doi.org/10.1609/aaai.v33i01.33018949

Wang, Q., Yan, L., Zhang, L., Ai, H., Lin, X., 2016. A semantic modelling framework-based method for building reconstruction from point clouds. *Remote Sens.* 8, 1–23. https://doi.org/10.3390/rs8090737

Wang, R., Peethambaran, J., Chen, D., 2018. LiDAR Point Clouds to 3D Urban Models : A Review. *IEEE J. Sel. Top. Appl. Earth Obs. Remote Sens.* 11, 606 – 627. https://doi.org/10.1109/JSTARS.2017.2781132

Wu, G., Shao, X., Guo, Z., Chen, Q., Yuan, W., Shi, X., Xu, Y., Shibasaki, R., 2018. Automatic building segmentation of aerial imagery usingmulti-constraint fully convolutional networks. *Remote Sens.* 10, 1–18. https://doi.org/10.3390/rs10030407

Wu, J., Wang, Y., Xue, T., Sun, X., 2017. MarrNet: 3D Shape Reconstruction via 2.5D Sketches, in: *NIPS 2017*. Long Beach, USA.

Xiong, B., Jancosek, M., Oude Elberink, S., Vosselman, G., 2015. Flexible building primitives for 3D building modeling. *ISPRS J. Photogramm. Remote Sens.* 101, 275–290. https://doi.org/10.1016/j.isprsjprs.2015.01.002

Xu, Y., Wu, L., Xie, Z., Chen, Z., 2018. Building extraction in very high resolution remote sensing imagery using deep learning and guided filters. *Remote Sens.* 10. https://doi.org/10.3390/rs10010144

Yan, Y., Gao, F., Deng, S., Su, N., 2017. A Hierarchical Building Segmentation in Digital Surface Models for 3D Reconstruction. *Sensors.* 17, 1–14. https://doi.org/10.3390/s17020222

Yang, H.L., Yuan, J., Lunga, D., Laverdiere, M., Rose, A., Bhaduri, B., 2018. Building Extraction at Scale Using Convolutional Neural Network: Mapping of the United States. *IEEE J. Sel. Top. Appl. Earth Obs. Remote Sens.* 11, 2600–2614. https://doi.org/10.1109/JSTARS.2018.2835377

Zhang, W., Wang, H., Chen, Y., Yan, K., Chen, M., 2014. 3D Building Roof Modeling by Optimizing Primitive's Parameters Using Constraints from LiDAR Data and Aerial Imagery. *Remote Sens.* 6, 8107–8133. https://doi.org/10.3390/rs6098107

Zheng, Yuanfan, Weng, Q., Zheng, Yaoxing, 2017. A Hybrid Approach for Three-Dimensional Building Reconstruction in Indianapolis from LiDAR Data. *Remote Sens.* 9, 1–24. https://doi.org/10.3390/rs9040310