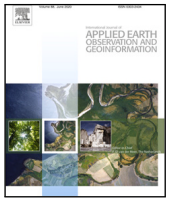
International Journal of Applied Earth Observation and Geoinformation 120 (2023) 103311

Contents lists available at ScienceDirect

International Journal of Applied Earth Observation and Geoinformation

journal homepage: www.elsevier.com/locate/jag



3DCentripetalNet: Building height retrieval from monocular remote sensing imagery

Qingyu Li [a], Lichao Mou [a], Yuansheng Hua [a], Yilei Shi [b], Sining Chen [a], Yao Sun [a], Xiao Xiang Zhu [a,*]

[a] Data Science in Earth Observation, Technical University of Munich, Munich, 80333, Germany
[b] Remote Sensing Technology, Technical University of Munich, Munich, 80333, Germany

ARTICLE INFO

Keywords:
Building height retrieval
Monocular imagery
Building footprint generation
Height estimation

ABSTRACT

Three-dimensional (3D) building structures are vital to understanding urban dynamics. Monocular remote sensing imagery is a cost-effective data source for large-scale building height retrieval when compared to LiDAR data and multi-view imagery. Existing methods learn building footprints and height maps per pixel via either a multi-task network or two separate networks, however, failing to consider the information of neighboring pixels that belong to the identical building. Therefore, we propose learning a novel representation for 3D buildings, namely 3D centripetal shifts, a unified representation of individual building instances. Our method is termed as 3DCentripetalNet and learns the 3D centripetal shift representation that incorporates planar and vertical structures of buildings. Afterward, a decoupling module is devised to learn building corner points. Finally, a 3D modeling module is designed to retrieve building height from the learned 3D centripetal shift map and corner points. We investigate the proposed 3DCentripetalNet on two datasets with different spatial resolutions, i.e., the ISPRS Vaihingen dataset (9 cm/pixel) and the Urban 3D dataset (50 cm/pixel). Experimental results suggest that 3DCentripetalNet is able to preserve sharp building boundaries, largely alleviate false detections, and significantly outperform other competitors. Thus, we believe that 3DCentripetalNet is a robust solution for the task of building height retrieval from monocular imagery.

1. Introduction

As one of the most important terrestrial objects, the building enables a comprehensive understanding of urban development (Li et al., 2022a,b). Due to the lack of available three-dimensional (3D) data, most of the existing studies investigate urban morphology by concentrating on analyzing two-dimensional (2D) building footprints, while neglecting building heights that characterize the vertical structure of urban form and are essential for understanding the urban process (Cao and Huang, 2021). Therefore, we aim to retrieve building heights that facilitate a wide range of applications.

Remote sensing technologies that hold huge potential for building height retrieval, become a fundamental way for 3D building modeling. There are two commonly used data types: light detection and ranging (LiDAR) (Cao et al., 2020) data and remote sensing imagery (Chen et al., 2021). LiDAR data allows highly accurate scene geometry measurement. However, applying LiDAR data in a wide range of areas, e.g., the whole city, is highly expensive. In contrast, remote sensing imagery is more cost-effective to provide building information with large spatial coverage. Although some studies (Frantz et al., 2021; Li et al., 2020a) estimate building heights at continental or global scales, the retrieved building heights are acquired at aggregated spatial scales because of the coarse spatial resolution of imagery. Note in our study, the building heights at the individual buildings are of interest. This is because the heterogeneity of individual buildings can offer a more detailed analysis of urban structure (Li et al., 2022c). Some existing methods estimate building heights from multiple-view remote sensing imagery with very high resolution, as they can impose geometric constraints (Hepp et al., 2018; Liu and Ji, 2020). However, the deployment of these methods is restricted because multi-view imagery needs to be captured within a short period. In contrast, estimating building heights from monocular imagery can break through this limitation and is able to allow building height retrieval in sparsely imaged regions (Mahmud et al., 2020).

Building height retrieval from monocular imagery can be decomposed into two sub-tasks: (1) building footprint generation (Lin et al., 2019) and (2) height estimation (Mou and Zhu, 2018). The goal of

* Corresponding author.
E-mail addresses: qingyu.li@tum.de (Q. Li), lichao.mou@tum.de (L. Mou), yuansheng.hua@szu.edu.cn (Y. Hua), yilei.shi@tum.de (Y. Shi), sining.chen@dlr.de (S. Chen), yao.sun@dlr.de (Y. Sun), xiaoxiang.zhu@tum.de (X.X. Zhu).

https://doi.org/10.1016/j.jag.2023.103311
Received 27 October 2022; Received in revised form 21 February 2023; Accepted 15 April 2023
Available online 15 May 2023
1569-8432/© 2023 The Author(s). Published by Elsevier B.V. This is an open access article under the CC BY license (http://creativecommons.org/licenses/by/4.0/).

building footprint generation is to assign each pixel a semantic label "building" or "non-building", while height estimation focuses on measuring each pixel's height. Both two sub-tasks learn pixel-wise output representations, i.e., building footprints and height maps. However, these representations deliver only the characteristics of individual pixels, while neglecting the information about neighboring pixels that belong to the same building. We have observed that pixels within the same building show more similar features (e.g., color and height) than those from another building. In this work, we want to learn a unified representation that encodes geometrical structures for individual building instances. Therefore, we propose a novel representation of 3D buildings, which is termed 3D centripetal shifts. The 3D centripetal shift representation is composed of spatial offsets from building roof points to the visual center of a building in 3D space (x, y, and z directions). By doing so, the planar and vertical properties of individual building instances can be captured.

Based on 3D centripetal shift representation, we propose a network, namely 3DCentripetalNet, to learn this representation for each building instance. Moreover, we note that the detected buildings show distorted and blurred boundaries. Therefore, we devise a decoupling module that learns to explicitly decouple the centripetal shift in z-direction into low frequency and high frequency components. As high frequency information indicates building boundaries where corner points are located, this module enables the network to better learn building corners. By connecting the detected corner points, sharper building boundaries could be preserved. Three contributions of this research are:

(1) We propose a novel type of representation for building instances in 3D space: *3D centripetal shift representation*. Both planar and vertical information is conveyed in 3D centripetal shift representation, characterizing well the structures of building instances in 3D space.

(2) In our research, a novel network, termed as 3DCentripetalNet, is proposed to learn 3D centripetal shift representation and building corners, which are further utilized to retrieve building height. Results show that 3DCentripetalNet outperforms other competitors.

(3) We propose to utilize a decoupling module for deriving a high spatial frequency map that contributes to detecting corner points of buildings. Afterward, building boundaries can be obtained by connecting corner points, and are endowed with good geometry property.

2. Related work

2.1. Building footprint generation

Building footprint generation refers to solving pixel-level labeling problems; the class labels are "building" or "non-building".

Early efforts rely heavily on complex handcrafted feature engineering and high human intervention. They are generally classified into four types: classifier-, segmentation-, index-, and geometrical primitive-based approaches. In classifier-based methods (Senaras et al., 2013), classifiers use the features of every pixel to determine its label. Segmentation-based methods (Karantzalos and Argialas, 2009) utilize over-segmentation algorithms to partition an image into various segments, in order to extract those corresponding to buildings. Index-based methods (Huang and Zhang, 2011) aim at designing an index that is capable of discriminating buildings from the background. In this way, buildings can be extracted by selecting an empirical threshold. In geometrical primitive-based methods (Cote and Saeedi, 2012), building corners or edges are extracted and grouped to form closed polygons for individual buildings. However, these methods suffer from a common limitation that the use of manually designed rules and handcrafted features often leads to poor results.

In the past decades, deep learning techniques have achieved remarkable results, as they are able to automatically and adaptively learn discriminative features from raw input. Based on deep learning architectures, recent studies are capable of providing impressive building extraction results by utilizing semantic segmentation networks (Lin et al., 2019; Wei et al., 2019; Xu et al., 2021). Fully convolutional networks (FCNs) (Long et al., 2015) and encoder–decoder (e.g., U-Net Ronneberger et al., 2015) are commonly used network architectures.

2.2. Height estimation

Approximating a height value for each pixel is the goal of height estimation from monocular remote sensing imagery.

Early studies exploit shape from shading (SFS) (Pentland, 1988) to estimate height values. SFS refers to the theory that the 3D surface shape can be recovered from the gradual variations in the shading (Horn, 1990). Moreover, a sparse digital terrain model (DTM) or a few control points are required as auxiliary information in these methods (Chen et al., 2012; Rajabi and Blais, 2004).

Recently, several approaches propose to use deep learning techniques for this challenge, which were motivated by the developments in monocular depth estimation from the computer vision community. These methods have two types: encoder–decoder architectures and generative adversarial networks (GANs) (Goodfellow et al., 2014). Encoder–decoder networks regress height values for individual pixels (Amirkolaee and Arefi, 2019; Mou and Zhu, 2018), while GANs simulate a height map on target scenes (Ghamisi and Yokoya, 2018; Paoletti et al., 2020).

2.3. Building height retrieval

The task of building height retrieval aims to solve two problems: (1) extracting building footprints and (2) modeling the height of each building.

Traditional methods first generate building footprint maps and then estimate building heights. These methods employ the shadow information and geometrical primitives as primary clues (Ok et al., 2012; Izadi and Saeedi, 2011). Meta information on remote cameras is also needed for height estimation. Besides, their complex procedures are performed on specific data, leading to poor generalization.

Recent developments in deep learning-based methods directly learn semantic masks and height maps via a multi-task network (Mahmud et al., 2020; Li et al., 2021a; Chen et al., 2021; Srivastava et al., 2017; Zheng et al., 2019; Elhousni et al., 2021). By doing so, the performance of two sub-tasks can be boosted during the simultaneous optimization process. However, most of these above methods ignore the integrity of building instances, which may degrade network capabilities of height estimation.

3. Methodology

The pipeline of 3DCentripetalNet is first described. Then, our proposed 3D building representation and decoupling module are introduced in detail. Afterward, the details of network learning are explained, and the 3D modeling module is introduced in the last subsection.

3.1. Pipeline

The pipeline of 3DCentripetalNet is illustrated in Fig. 1, which consists of two main components: (1) a multi-task deep neural network that learns 3D centripetal shift map and corner map, and (2) a 3D modeling module that integrates the network outputs to retrieve building heights. A remote sensing image $\boldsymbol{I} \in \mathbb{R}^{C\times W\times H}$ is taken as input, where H and W are the height and width, and C is the number of spectral bands. An encoder–decoder architecture is first adopted to extract features that are denoted as $\boldsymbol{F} \in \mathbb{R}^{K\times W\times H}$ with K being the number of channels. A convolutional layer is then attached to learn 3D centripetal shift map $\boldsymbol{D} \in \mathbb{R}^{3\times W\times H}$ where each channel encodes semantic and geometric

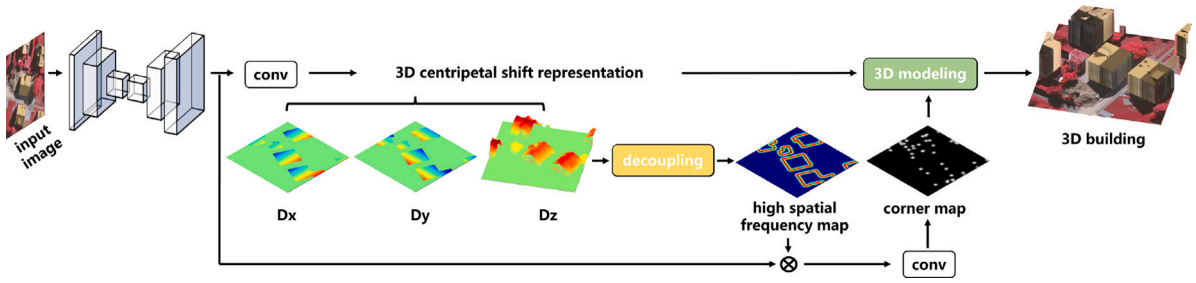


Fig. 1. Overview of the proposed framework.

information in one direction. For the sake of clarity, we term the three channels of $\boldsymbol{D}$ as $\boldsymbol{D}_x \in \mathbb{R}^{1\times W\times H}$, $\boldsymbol{D}_y \in \mathbb{R}^{1\times W\times H}$, and $\boldsymbol{D}_z \in \mathbb{R}^{1\times W\times H}$, respectively. Afterward, a decoupling module is appended on top of $\boldsymbol{D}_z$, aiming to learn a high spatial frequency map, which highlights pixels alongside building boundaries. By allocating the learned high frequency map to $\boldsymbol{F}$, we can obtain enhanced feature maps. Considering that building corners lie on building boundaries, these enhanced features can also help to detect building corners. In this regard, a corner map indicating the candidates for building corner points can be learned from them. Finally, a 3D modeling module is exploited to jointly leverage the learned 3D centripetal shift map and the corner map for retrieving high-quality building height information.

3.2. 3D building representation - 3D centripetal shift representation

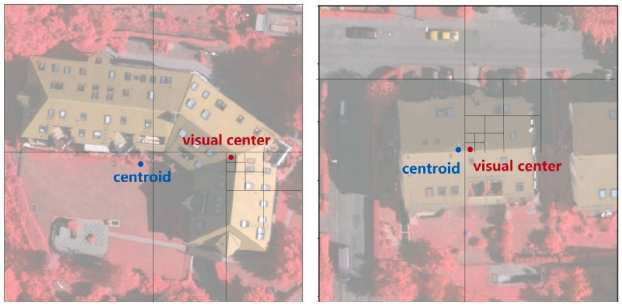
For building height retrieval, one commonly used method is to design a multi-task network that learns semantic masks and height maps, separately, which indicate the semantic category and height value of each pixel. However, the performance of this type of methods is restricted due to its incapability to relate to the structural integrity of individual building instances. Therefore, we propose a novel representation, namely 3D centripetal shift representation, where each building instance is taken as a unity. For each pixel belonging to a building, 3D centripetal shift measures the distance from it to the visual center of its corresponding building along x-, y-, and z- axes.

3.2.1. Definition of visual center

The visual center of every building is defined as the most distant internal point from the building boundary. Note that the visual center is different from the centroid, which might be outside of a building (see Fig. 2(a)). Specifically, the visual center of a building is extracted with an iterative grid-based algorithm. Initial square grid cells are first generated to fully cover the building instance and then split into 4 children cells that are put into a priority queue. This cell queue is sorted by the potential distance between the cell center and building boundaries. For each iteration, the grid cell with the maximum potential distance is selected as the best grid cell and will be split into 4 children cells that are added to the queue. The iteration stops when the size of the best grid cell is reduced to 1×1 pixel. Finally, this pixel is selected as the visual center of this building.

3.2.2. Learning 3D centripetal shift representation

As shown in Fig. 3, 3D centripetal shift consists of three components, which are the spatial offsets from the roof point to the visual center in x-, y-, and z-directions. Formally, a building instance B can be represented by its roof points and denoted as $B = \{\mathbf{p}^1, \mathbf{p}^2, \ldots, \mathbf{p}^m\}$ with m being the number of roof points. $\mathbf{p}^j = (p_x^j, p_y^j, p_z^j)$, where p_x^j, p_y^j, p_z^j are x-, y-, and z- coordinates, respectively. The building visual center is denoted as $\mathbf{v} = (v_x, v_y, v_z)$, where v_x, v_y, v_z are its x-, y-, and z-coordinates, respectively. Note that we set v_z to 0 in our research, as we measure the visual center in 2D aerial imagery. Afterward, 3D



(a) (b)

Fig. 2. (a) and (b) show visual centers of buildings obtained by the grid-based algorithm.

centripetal shift of $\mathbf{p}^j$, denoted as $\mathbf{d}^j$, is defined as the offsets between $\mathbf{p}^j$ and $\mathbf{v}$, and computed as

$$\mathbf{d}^j = \mathbf{v} - \mathbf{p}^j . \tag{1}$$

Concerning that inputs of layers in a deep network often have fixed sizes, we model 3D centripetal shifts of all pixels and form a 3D centripetal shift map $\boldsymbol{D} \in \mathbb{R}^{3\times W\times H}$ with respect to $\boldsymbol{I}$. To be more specific, we enumerate Eq. (1) over all building instances, while assigning non-building pixels $[0, 0, 0]^T$. With this design, 3D centripetal shift representations are supposed to deliver both planar and vertical information of individual building instances in a single representation, drawing a more comprehensive picture of building structures in 3D space.

Each element in the 3D centripetal shift representation has three components, i.e., offsets from the visual center to it along x-, y-, and z- axes. To this end, 3D centripetal shift representation might be viewed as three learnable 2D feature maps. Therefore, the learning of the 3D centripetal shift representation is considered a dense prediction problem and solved by training a semantic segmentation network. Specifically, U-Net (Ronneberger et al., 2015) is exploited, as multi-scale skip connections of U-Net (Ronneberger et al., 2015) can effectively exploit low-level visual cues (e.g., building boundaries), which helps to improve the learning of 3D centripetal shifts.

3.3. Learning refined building boundary-decoupling module

We have observed that buildings directly predicted from CNNs tend to have blob-like shapes and blurred boundaries. Nevertheless, buildings are man-made objects, and often have straight edges and sharp corners. To preserve such properties, we characterize a building instance by a polygon comprising a sequence of building corner points and propose to predict well-shaped building boundaries by detecting and connecting their corners.

To increase the accuracy of corner detection, we devise a decoupling module to discard low spatial frequency component while keeping

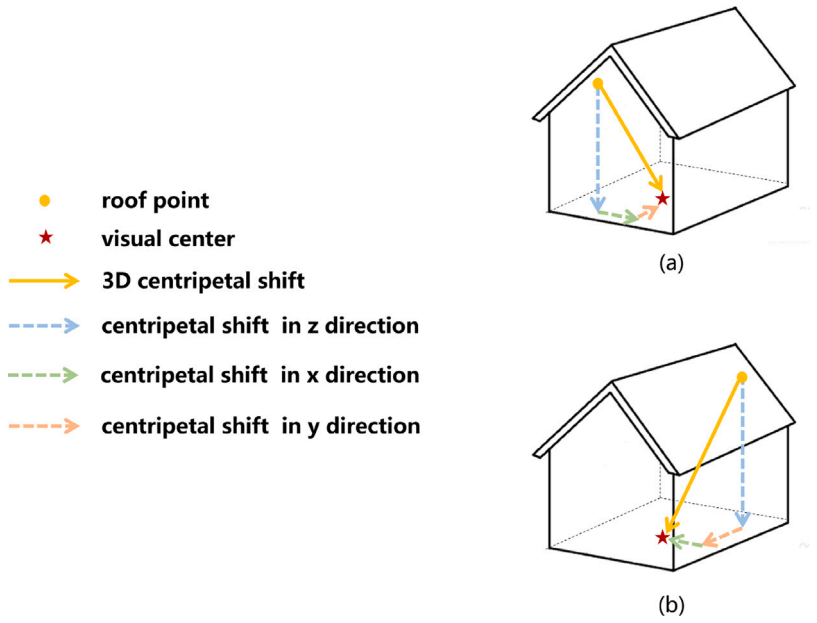


Fig. 3. Two examples of 3D centripetal shifts.

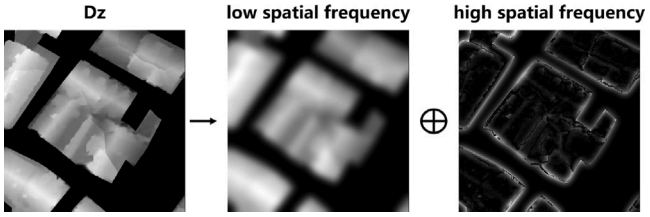


Fig. 4. The motivation of the decoupling module that decouples D_z into low frequency and high frequency. The decomposition is done by first implementing a Gaussian filter for a low frequency part, and then subtracting from the original input for a high frequency part. ⊕ means the operation of addition.

only high spatial frequency part. This is because the low frequency information represents structures with smooth change (Li et al., 2020b), while a high spatial frequency component indicates abrupt changes in the image, and usually refers to object corners and edges (Li and Yang, 2008). We note that in $\boldsymbol{D}_z$, background clutters (e.g., tree and road) are suppressed while building regions are more discernible (c.f. Fig. 4). Hence, we decompose $\boldsymbol{D}_z$ into two components, the high frequency component of which helps to detect building corners. Since the high frequency part is hard to directly learn, the proposed module first encodes low frequency representation $\boldsymbol{D}_z^{low}$, and then explicitly subtracts $\boldsymbol{D}_z^{low}$ from $\boldsymbol{D}_z$ to obtain the high spatial frequency component $\boldsymbol{D}_z^{high}$, which is termed as high spatial frequency map.

$$\boldsymbol{D}_z^{high} = \boldsymbol{D}_z - \boldsymbol{D}_z^{low}. \tag{2}$$

Fig. 5 illustrates the flowchart of the decoupling module, which consists of two steps. First, a flow-based method is utilized to generate low frequency representation $\boldsymbol{D}_z^{low}$. Afterward, the boundary attention map is generated by explicitly subtracting $\boldsymbol{D}_z^{low}$ from $\boldsymbol{D}_z$.

In the following, how the first step work is discussed in detail. A commonly used technique to acquire the low frequency term is degrading resolution. Specifically, strided convolutions are first applied to downsample $\boldsymbol{D}_z$ into the low resolution map which is then upsampled to the same size of $\boldsymbol{D}_z$ via bilinear upsampling. Afterward, this low resolution map is concatenated with $\boldsymbol{D}_z$, and fed into a convolutional layer for predicting a flow map $\boldsymbol{M} \in \mathbb{R}^{2\times W\times H}$. The flow map $\boldsymbol{M}$ represents the offsets of the same point in two images, i.e., $\boldsymbol{D}_z$ and its lower resolution variant. Finally, we can generate low frequency representation $\boldsymbol{D}_z^{low}$ by warping $\boldsymbol{D}_z$ with $\boldsymbol{M}$. The warping procedure is implemented as:

$$\boldsymbol{D}_z^{low}(\mathbf{s}^u) = \sum_{q\in\boldsymbol{D}_z} G(\mathbf{s}^q, \mathbf{s}^u + \boldsymbol{M}(\mathbf{s}^u))\boldsymbol{D}_z(\mathbf{s}^q), \tag{3}$$

where $\mathbf{s}^u$ denotes the spatial location (x- and y- coordinates) of the u_{th} pixel in $\boldsymbol{D}_z$, $\mathbf{s}^q$ enumerates all spatial locations in $\boldsymbol{D}_z$, and $G(.,.)$ is a bilinear interpolation kernel.

3.4. Network learning

3DCentripetalNet learns 3D centripetal shift map and corner map from remote sensing imagery in a supervised manner with an objective function L being devised:

$$L = \lambda \cdot L_D + L_E, \tag{4}$$

where L_D and L_E are two loss functions for optimizing the 3D centripetal shift map and corner map, respectively. λ is a hyperparameter controlling the relative weights of two terms. For L_D, a Smooth L1 loss function is used:

$$L_D = \sum_{q\in\boldsymbol{I}} \begin{cases} 0.5(\hat{\boldsymbol{D}}(\mathbf{s}^q) - \boldsymbol{D}(\mathbf{s}^q))^2 & \text{if } |\hat{\boldsymbol{D}}(\mathbf{s}^q)) - \boldsymbol{D}(\mathbf{s}^q)| < 1 \\ |\hat{\boldsymbol{D}}(\mathbf{s}^q) - \boldsymbol{D}(\mathbf{s}^q)| - 0.5 & \text{otherwise} \end{cases}, \tag{5}$$

where $\hat{\boldsymbol{D}}(\mathbf{s}^q)$ and $\boldsymbol{D}(\mathbf{s}^q)$ are the predicted and ground reference 3D centripetal shift map of buildings at spatial position $\mathbf{s}^q$, respectively. For L_E, cross-entropy loss function is used:

$$L_E = \sum_{q\in\boldsymbol{I}} \begin{cases} -log(\hat{\boldsymbol{E}}(\mathbf{s}^q)) & \text{if } \boldsymbol{E}(\mathbf{s}^q) = 1 \\ -1og(1 - \hat{\boldsymbol{E}}(\mathbf{s}^q)) & \text{if } \boldsymbol{E}(\mathbf{s}^q) = 0 \end{cases}, \tag{6}$$

where $\hat{\boldsymbol{E}}(\mathbf{s}^q) \in [0, 1]$ represents the value of output probability at $\mathbf{s}^q$, and $\boldsymbol{E}(\mathbf{s}^q)$ is its corresponding ground reference. For values in $\boldsymbol{E}$, 1 denotes building corner, and 0 represents background.

3.5. 3D modeling module

The goal of the 3D modeling module is to retrieve building height using the learned 3D centripetal shift map and corner map. Fig. 6 shows the workflow of our 3D modeling module. To solve duplicate predictions of building corners, an algorithm called ExtractPeak (Zhou

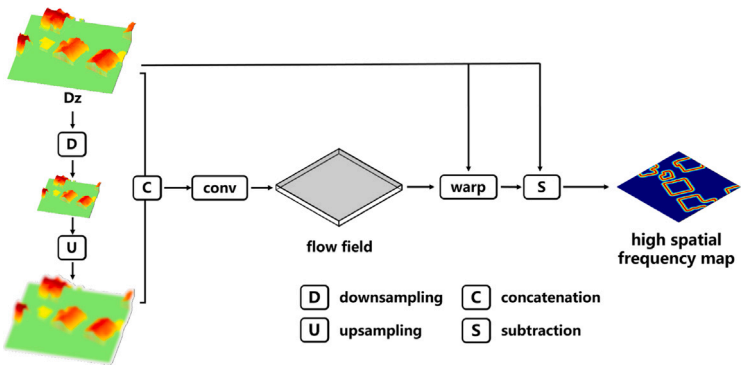


Fig. 5. The flowchart of the decoupling module.

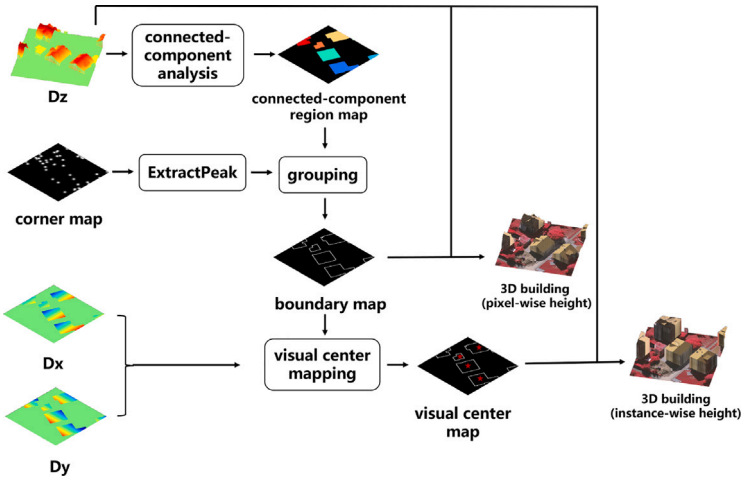


Fig. 6. The flowchart of the 3D modeling module.

et al., 2019) is utilized to select best-fitting corner points by detecting all peaks. More specifically, we first discard pixels with confidence scores smaller than a threshold τ. Here, we set $\tau = 0.06$. Afterward, we select pixels being local maximums in respective 3×3 windows as final corner points.

To group corners belonging to the same building, we apply connected component labeling (CCL) on $\boldsymbol{D}_z$ to obtain a connected-component map. Specifically, each pixel is connected to its positive neighbors within a 3×3 window, forming multiple connected components in an image. Each set of connected pixels represents a building instance. By sequentially connecting corner points in each connected component, a boundary map can be generated. By doing so, building boundaries are sharper as corner points can well characterize the geometry of buildings.

Our method is capable of creating two different types of building height information, which is instance-wise height and pixel-wise height. The former model describes the height of each building by only one value, while the latter model has the height value of every roof point. Pixel-wise height can be achieved by overlaying the boundary map and $\boldsymbol{D}_z$. As to instance-wise height, we need to find the visual centers of different building instances. Given that roof points within the same building should share one visual center, an intuitive idea is to utilize $\boldsymbol{D}_x$ and $\boldsymbol{D}_y$, which depict the planar offsets between the visual center and roof points. For each building, every roof point provides an estimated position of the visual center, and the average of all estimated values is taken as the visual center. Afterward, we assign the height value of the roof point located at the visual center to the corresponding building.

4. Experiment

4.1. Dataset

The performance of 3DCentripetalNet is explored on ISPRS Vaihingen dataset (ISPRS, 0000) and Urban 3D dataset (Goldberg et al., 2017). Note that remote sensing images in both datasets are orthophotos.

4.1.1. ISPRS Vaihingen dataset

The ISPRS Vaihingen dataset involves 33 aerial imagery that is collected in the city of Vaihingen. Every aerial image has near-infrared, red, and green bands at 9 cm/pixel. Their corresponding ground truth land cover data includes six class types. nDSM that indicates the height of all elevated objects is also provided for each aerial imagery. In this work, we only use *building* as the positive class, while other land cover types are merged into *non-building*. The data split follows (Mahmud et al., 2020; Mou et al., 2019), 11 tiles are for training, and 5 tiles (tile id: 11, 15, 28, 30, and 34) are used to test models.

4.1.2. Urban 3D dataset

The Urban 3D dataset is composed of satellite images and nDSM profiles that are collected over two cities, Jacksonville and Tampa. Each image has red, green, and blue bands at 50 cm/pixel. The corresponding ground reference data provides 74000 building masks. In this paper, the dataset is separated following the set up in Mahmud et al. (2020), 130 imagery for training and 44 imagery for testing.

Table 1
Numerical results of various methods on the ISPRS Vaihingen dataset (spatial resolution: 9 cm/pixel).

Method	Pixel-wise height (m)		Instance-wise height (m)		Footprint (%)	
	RMSE	MAE	RMSE	MAE	IoU	F1 score
Mou and Zhu	2.30	1.74	–	–	–	–
Mahmud et al.	1.93	1.43	–	–	–	–
Lin et al.	–	–	–	–	88.12	93.69
Wei et al.	–	–	–	–	88.01	93.62
Xu et al.	–	–	–	–	87.28	92.85
Srivastava et al.	3.39	2.65	–	–	88.49	93.89
Elhousni et al.	3.66	2.80	–	–	88.68	93.98
Chen et al.	–	–	1.44	1.10	88.36	93.82
3DCentripetalNet	**1.87**	**1.40**	**1.33**	**1.07**	**89.71**	**94.57**

4.2. Implementation details

Our experiments are conducted within a Pytorch framework on an NVIDIA Tesla P100 GPU with 16 GB of memory. Details on implementing 3DCentripetalNet are as follows.

4.2.1. Data preprocessing

For both ISPRS Vaihingen and Urban 3D datasets, the original remote sensing imagery, as well as the corresponding ground reference building footprints and height maps, are cropped into the size of 512×512 pixels.

4.2.2. Training

For the training set, 3D centripetal shift maps and corner maps are first produced using the ground reference building footprints and height maps for network learning. 3DCentripetalNet is then trained on the ground reference 3D centripetal shift maps and corner maps from the training set. In our method, the extracted feature map F has the size of $64 \times 512 \times 512$ pixels. λ in Eq. (4) is set as 0.1 empirically. Models are trained 200 epochs by the optimizer of stochastic gradient descent (SGD) with a learning rate of 0.0001. The training batch size is 4.

4.2.3. Inference

For each remote sensing imagery with the size of 512×512 pixels in the test set, we first input it into the trained model to get the predicted 3D centripetal shift maps and corner maps. Afterward, the predictions are merged into large-scale outputs of the original size. Finally, we retrieve building heights from the learned 3D centripetal shift maps and corner maps by using the 3D modeling module.

4.3. Accuracy assessment

The capability of models is assessed from two aspects: footprint prediction accuracy and height estimation precision.

4.3.1. Footprint prediction accuracy

Intersection over union (IoU) and F1 score are computed with the following equations:

$$\text{F1 score} = \frac{2 \times \text{precision} \times \text{recall}}{\text{precision} + \text{recall}}, \tag{7}$$

$$\text{IoU} = \frac{TP}{TP + FP + FN}, \tag{8}$$

$$\text{precision} = \frac{TP}{TP + FP}, \tag{9}$$

$$\text{recall} = \frac{TP}{TP + FN}, \tag{10}$$

where FN, FP, and TP indicate the numbers of false negatives, false positives, and true positives, respectively. F1 score and IoU are derived based on pixels rather than instances.

4.3.2. Height estimation precision

Root mean squared error (RMSE) and mean absolute error (MAE) are exploited as two evaluation criteria to evaluate the network performance in building height estimation. They are defined as:

$$\text{RMSE} = \sqrt{\frac{\sum_{i=1}^{Q} (h_i - \hat{h}_i)^2}{Q}}, \tag{11}$$

$$\text{MAE} = \frac{\sum_{i=1}^{Q} |h_i - \hat{h}_i|}{Q}, \tag{12}$$

where $\hat{h}_i$ and h_i are height values of the ith building instance/pixel in the predicted results and ground reference, respectively. The evaluation target is individual building pixels for pixel-wise building height, while refers to individual building instances for instance-wise building height. Q denotes the number of building instances or pixels. Note that non-building pixels or instances in the ground reference are not taken into account for these two metrics.

5. Results

For a comprehensive evaluation, we compare 3DCentripetalNet with other competitors from two perspectives, height estimation precision as well as footprint prediction accuracy. Specifically, as to the former, we compare with height estimation models, Mou and Zhu, and Mahmud et al.. Regarding the latter, we perform comparisons with building footprint generation networks, Lin et al., Xu et al., and Wei et al.. Moreover, we take the approaches proposed by Srivastava et al., Chen et al., and Elhousni et al. as competitors for evaluating the comprehensive performance in both height estimation and building footprint generation.

5.1. Performance on ISPRS Vaihingen dataset

Table 1 reported accuracy metrics on the ISPRS Vaihingen dataset. 3DCentripetalNet surpasses all competitors with regard to both footprint prediction accuracy and height estimation precision. Specifically, in comparison with height estimation methods (Mahmud et al., 2020; Mou et al., 2019), our method achieves lower RMSE and MAE, which validates the satisfactory performance of 3DCentripetalNet in height estimation. Besides, when compared to building footprint generation methods (Lin et al., 2019; Wei et al., 2019; Xu et al., 2021), 3DCentripetalNet obtains increments of above 1.5% in IoU. This indicates that our method is effective in extracting building footprints. Furthermore, we make comparisons with building height estimation methods. Compared to Srivastava et al. (2017) and Elhousni et al. (2021), our approach not only reaches improvements of above 1% in IoU but also reduces the pixel-wise height RMSE error by about 50%. 3DCentripetalNet shows higher building footprint prediction and instance-wise height estimation accuracies when compared to Chen's method (Chen et al., 2021). The results find clear support for the benefits of 3DCentripetalNet in estimating building height information. This is due to the fact that the proposed 3D centripetal shift representation can encode planar and vertical structures of individual building instances, while other

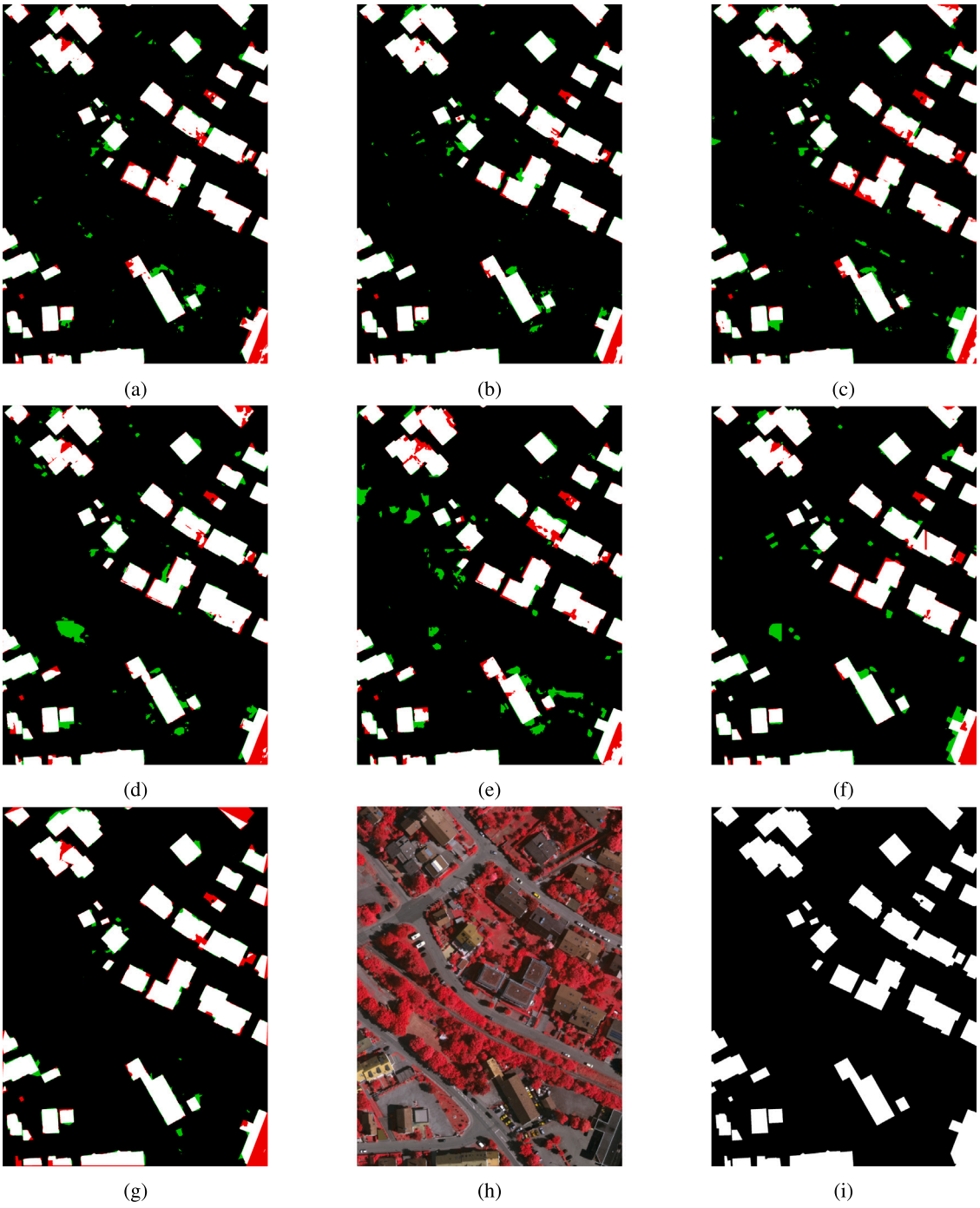


Fig. 7. Building footprints obtained by (a) Srivastava et al., (b) Elhousni et al., (c) Lin et al., (d) Wei et al., (e) Xu et al. (f) Chen et al., and (g) 3DCentripetalNet. (h) and (i) are the corresponding aerial imagery and ground truth from the ISPRS Vaihingen dataset (spatial resolution: 9 cm/pixel). Pixel-based false negatives, false positives, and true positives are illustrated in red, green, and white, respectively. (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

approaches learn the category and height with respect to each pixel separately. This indicates that 3DCentripetalNet is able to better model 3D building structures by learning 3D centripetal shift representation.

Figs. 7 and 9 present two examples of building footprints obtained by different methods. Other competitors easily misclassify impervious surfaces or cars as buildings, whereas our approach suffers less from false alarms. This is mainly because the proposed 3D centripetal shift representation can effectively convey useful geometrical properties for individual building instances in the image. As illustrated in Fig. 9, some methods fail to achieve sharp building boundaries, but building boundaries produced by our algorithm are more rectilinear and precise. These observations suggest that our model really benefits from learning building corners, enabling the gain of more geometric details of buildings. Figs. 8 and 10 are two examples to illustrate the pixel-wise and instance-wise building height maps estimated by 3DCentripetalNet.

A thorough view of the performance of 3DCentripetalNet for building height retrieval is provided in Fig. 11 where 3D building models are generated on the ISPRS Vaihingen dataset. Fig. 11(a) and (c) refer to the level of detail (LOD)-1 model, as they represent instance-wise building heights. Fig. 11(b) and (d) illustrate the building models with

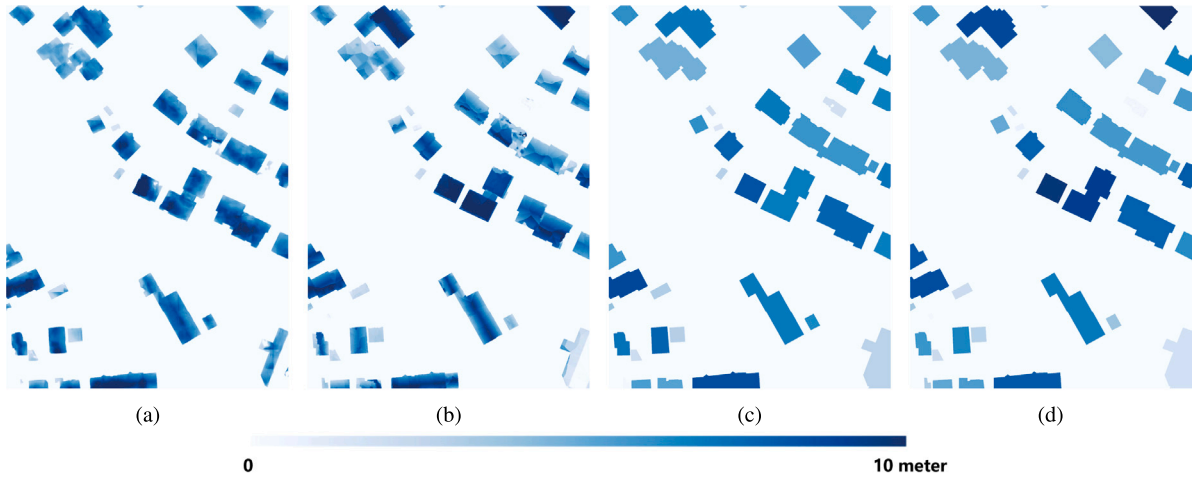


Fig. 8. (a) And (c) are the pixel-wise and instance-wise building height maps estimated by 3DCentripetalNet. (b) and (d) are the corresponding ground reference height maps. The example is from the ISPRS Vaihingen dataset (spatial resolution: 9 cm/pixel).

Table 2
Numerical results of various methods on the Urban 3D dataset (spatial resolution: 50 cm/pixel).

Method	Pixel-wise height (m)		Instance-wise height (m)		Footprint (%)	
	RMSE	MAE	RMSE	MAE	IoU	F1 score
Mou and Zhu	6.62	2.35	–	–	–	–
Mahmud et al.	6.15	**2.34**	–	–	–	–
Lin et al.	–	–	–	–	66.38	79.79
Wei et al.	–	–	–	–	65.83	79.40
Xu et al.	–	–	–	–	63.30	77.53
Srivastava et al.	10.40	5.00	–	–	65.79	79.27
Elhousni et al.	10.59	5.56	–	–	63.76	77.86
Chen et al.	–	–	2.63	1.60	66.55	79.99
3DCentripetalNet	**5.80**	**2.34**	**2.46**	**1.43**	**66.86**	**80.14**

Table 3
Ablative results on the ISPRS Vaihingen dataset (spatial resolution: 9 cm/pixel).

Method	Pixel-wise height (m)		Instance-wise height (m)		Footprint (%)	
	RMSE	MAE	RMSE	MAE	IoU	F1 score
Without 3D centripetal shift representation	1.91	1.46	–	–	88.88	94.11
Without decoupling module	1.90	1.47	1.45	1.11	86.99	93.04
3DCentripetalNet	**1.87**	**1.40**	**1.33**	**1.07**	**89.71**	**94.57**

Table 4
Ablative results on the Urban 3D dataset (spatial resolution: 50 cm/pixel).

Method	Pixel-wise height (m)		Instance-wise height (m)		Footprint (%)	
	RMSE	MAE	RMSE	MAE	IoU	F1 score
Without 3D centripetal shift representation	6.75	2.78	–	–	60.94	75.73
Without decoupling module	7.12	3.22	3.38	2.41	63.51	77.68
3DCentripetalNet	**5.80**	**2.34**	**2.46**	**1.43**	**66.86**	**80.14**

pixel-wise height, offering more details than the LOD-1 model (Biljecki et al., 2016).

5.2. Performance on urban 3D dataset

To further validate the proposed method, we report experimental results on the Urban 3D Dataset. Numerical results are shown in Table 2. 3DCentripetalNet contributes to the reduction of 0.82 m and 0.35 m in RMSE with respect to height estimation methods (Mou and Zhu, 2018; Mahmud et al., 2020). Our method also brings increments in the footprint prediction accuracy compared to building footprint generation approaches (Lin et al., 2019; Wei et al., 2019; Xu et al., 2021). For the performance of joint footprint and height prediction, 3DCentripetalNet is also superior to other building height estimation methods (Srivastava et al., 2017; Elhousni et al., 2021; Chen et al., 2021). Specifically, 3DCentripetalNet shows the lowest RMSE errors, 5.80 m and 2.46 m for pixel-wise and instance-wise height error, respectively. Furthermore, our approach obtains the best IoU and F1 score, 66.86% and 80.14%. This again demonstrates that our approach is effective and robust in retrieving building heights.

Moreover, building footprints predicted by different approaches are presented in Figs. 12 and 14. As shown in Fig. 14, many buildings are

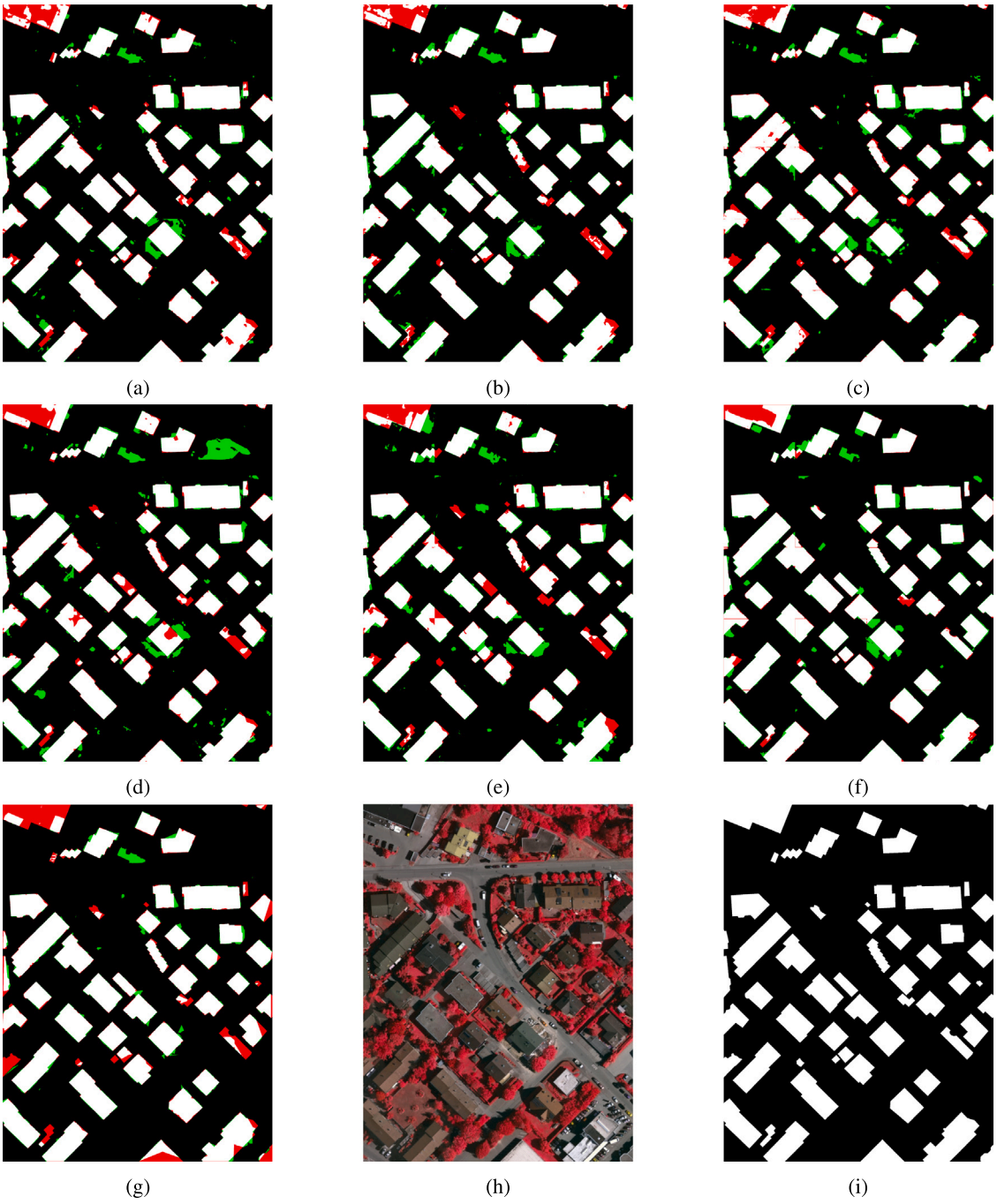


Fig. 9. Building footprints obtained by (a) Srivastava et al., (b) Elhousni et al., (c) Lin et al., (d) Wei et al., (e) Xu et al. (f) Chen et al., and (g) 3DCentripetalNet. (h) and (i) are the corresponding aerial imagery and ground truth from the ISPRS Vaihingen dataset (spatial resolution: 9 cm/pixel). Pixel-based false negatives, false positives, and true positives are illustrated in red, green, and white, respectively. (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

occluded by vegetation, semantic masks provided by other competitors show blob-like shapes. Note that 3DCentripetalNet is able to recover more precise building outlines. This is due to its decoupling module, which can preserve the high frequency component of given images for learning precise building corners. By connecting the detected corner points, building boundaries are naturally sharper compared to other competitors. Pixel-wise and instance-wise building height maps obtained by 3DCentripetalNet are shown in Figs. 13 and 15.

3D building models that are obtained by 3DCentripetalNet on the Urban 3D dataset are illustrated in Fig. 16, offering an overall view of the performance of our method in this task. Specifically, Fig. 16(a) and (c) illustrate the LOD-1 model, while Fig. 16(b) and (d) show building models with more details than the LOD-1 model (Biljecki et al., 2016).

5.3. Ablation studies

One contribution of this study is that we propose a novel type of representation for building instances in 3D space: 3D centripetal shift representation. We first carry out an ablation study to investigate the effectiveness of 3D centripetal shift representation. Specifically, we also derive the results of the proposed method where $\boldsymbol{D}_x$ and $\boldsymbol{D}_y$ are not learned. That is to say, only $\boldsymbol{D}_z$ and corner map are learned from networks. Note that in this case, we can only get pixel-wise building heights.

We carry out ablation studies on both ISPRS Vaihingen and Urban 3D Datasets. From numerical results in Tables 3 and 4, learning 3D

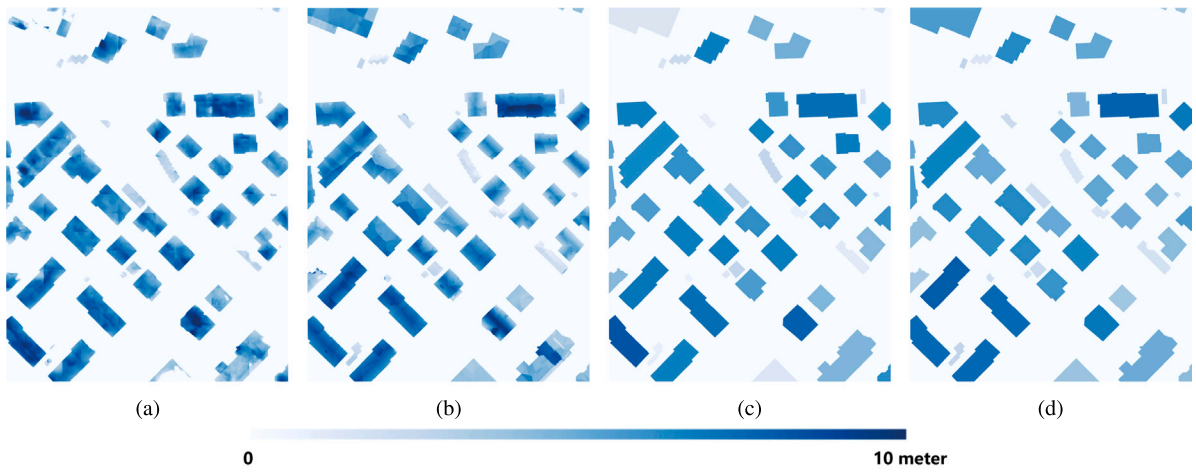


Fig. 10. (a) And (c) are the pixel-wise and instance-wise building height maps estimated by 3DCentripetalNet. (b) and (d) are the corresponding ground reference height maps. The example is from the ISPRS Vaihingen dataset (spatial resolution: 9 cm/pixel).

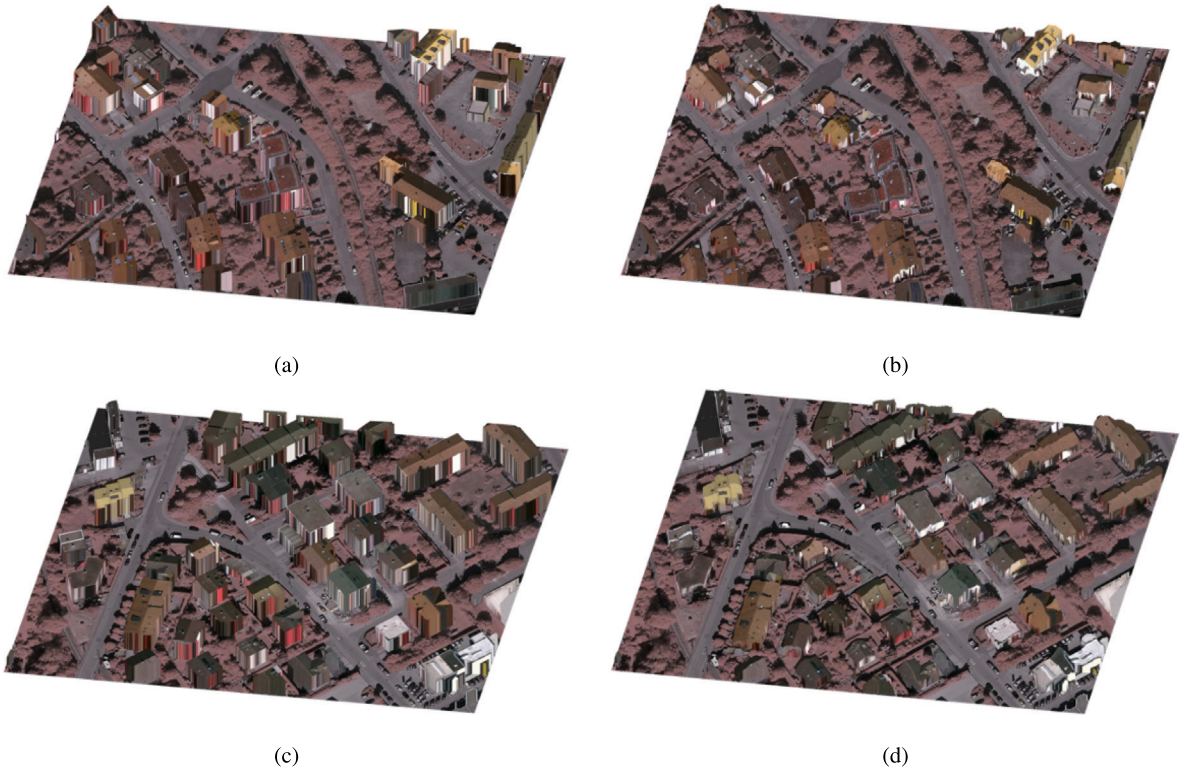


Fig. 11. (a) And (c) are the 3D building models with instance-wise heights estimated by 3DCentripetalNet. (b) and (d) are the 3D building models with pixel-wise heights estimated by 3DCentripetalNet. The example is from the ISPRS Vaihingen dataset (spatial resolution: 9 cm/pixel).

centripetal shift representation (i.e., $\boldsymbol{D}_x$, $\boldsymbol{D}_y$, and $\boldsymbol{D}_z$) can boost the performance on both datasets. Specifically, it can not only bring gains in IoU but also reduce RMSE errors. Fig. 17 presents the corresponding visual results on both datasets, respectively. The building masks obtained from the method with 3D centripetal shift representation are more adherent to the ground reference. We can thus conclude that learning 3D centripetal shift representation can contribute to building height retrieval.

The other contribution of the proposed method worthy of being highlighted is that we introduce a decoupling module that decouples the centripetal shift in z-direction into low frequency and high frequency information. The high frequency map is further taken as input for learning building corners. To investigate whether the decoupling module is effective, we also derive the results of the proposed method where the decoupling module is removed. That is to say, the centripetal shift in z-direction is taken as the input for learning building corners.

From numerical results on both datasets (see Tables 3 and 4), the decoupling module can bring an above 2% gain in IoU, which exerts a positive effect on building footprint generation. Fig. 17 presents the corresponding building segmentation masks. The results obtained

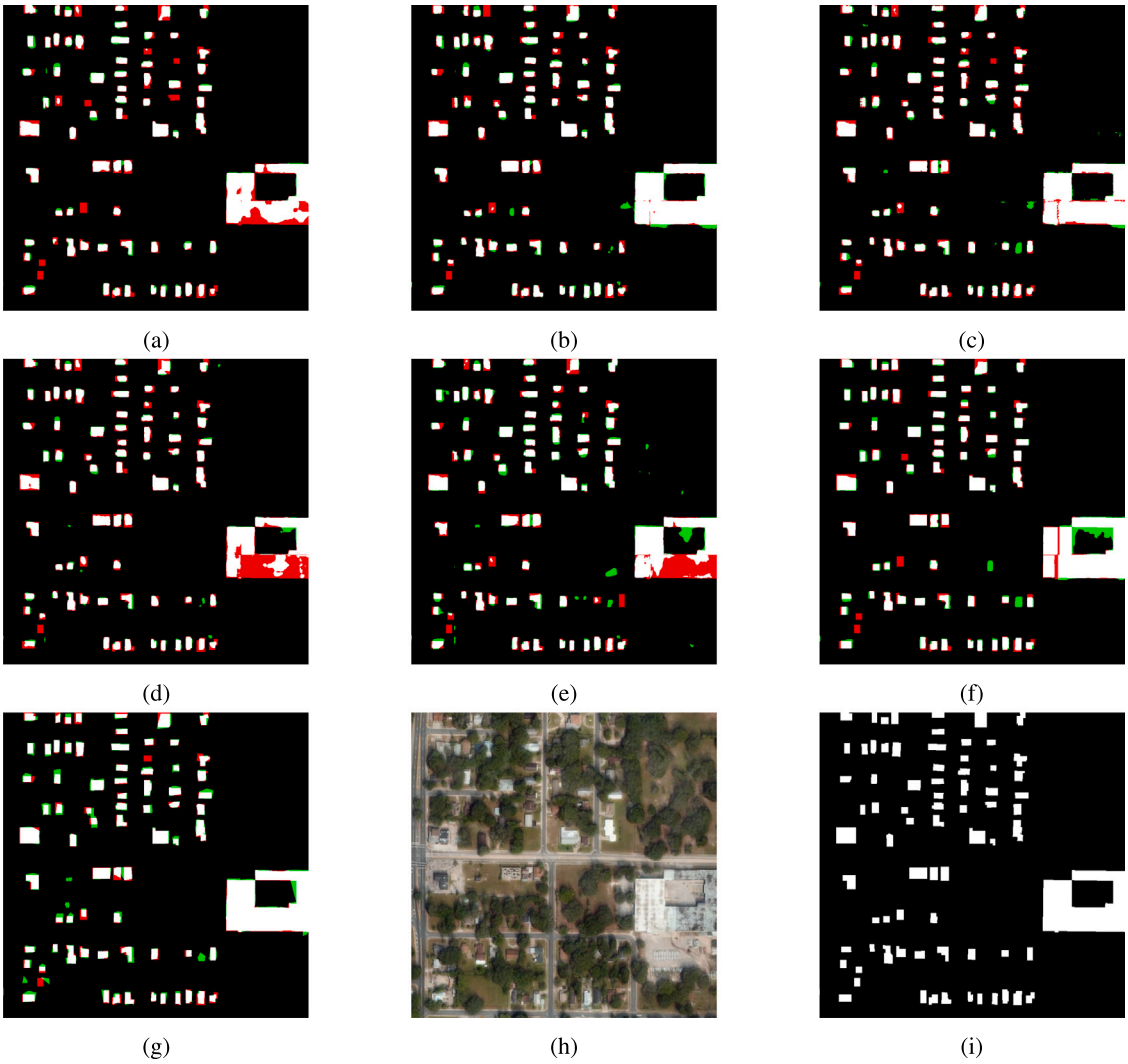


Fig. 12. Building footprints obtained by (a) Srivastava et al., (b) Elhousni et al., (c) Lin et al., (d) Wei et al., (e) Xu et al. (f) Chen et al., and (g) 3DCentripetalNet. (h) and (i) are the corresponding aerial imagery and ground truth from the Urban 3D dataset (spatial resolution: 50 cm/pixel). Pixel-based false negatives, false positives, and true positives are illustrated in red, green, and white, respectively. (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

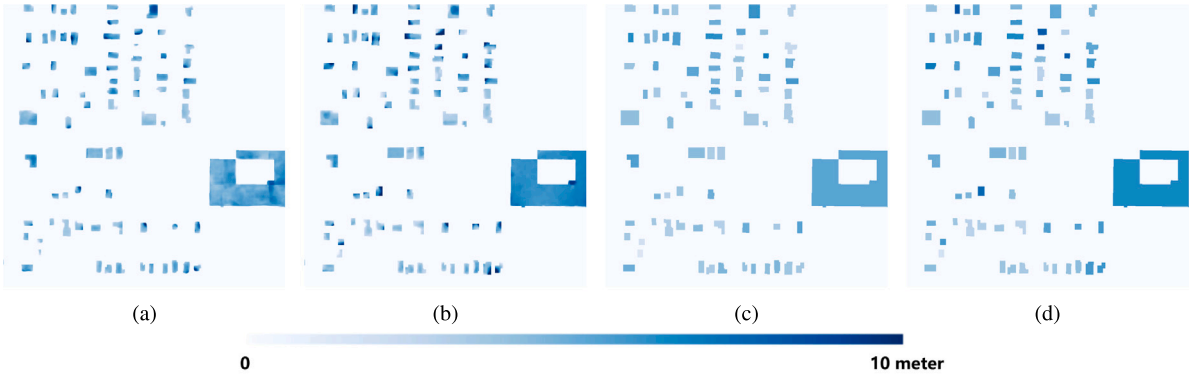


Fig. 13. (a) And (c) are the pixel-wise and instance-wise building height maps estimated by 3DCentripetalNet. (b) and (d) are the corresponding ground reference height maps. The example is from the Urban 3D dataset (spatial resolution: 50 cm/pixel).

from the method with the decoupling module show sharper building boundaries on both the ISPRS Vaihingen dataset and Urban 3D Dataset. This is because the use of a decoupling module is able to acquire the high spatial frequency component that indicates abrupt changes, thus, contributing to the network learning of building corners.

6. Discussion

In this section, we discuss the limitations of our 3DCentripetalNet and indicate future directions for improvements. The first limitation is that 3DCentripetalNet fails to perform well on off-nadir imagery

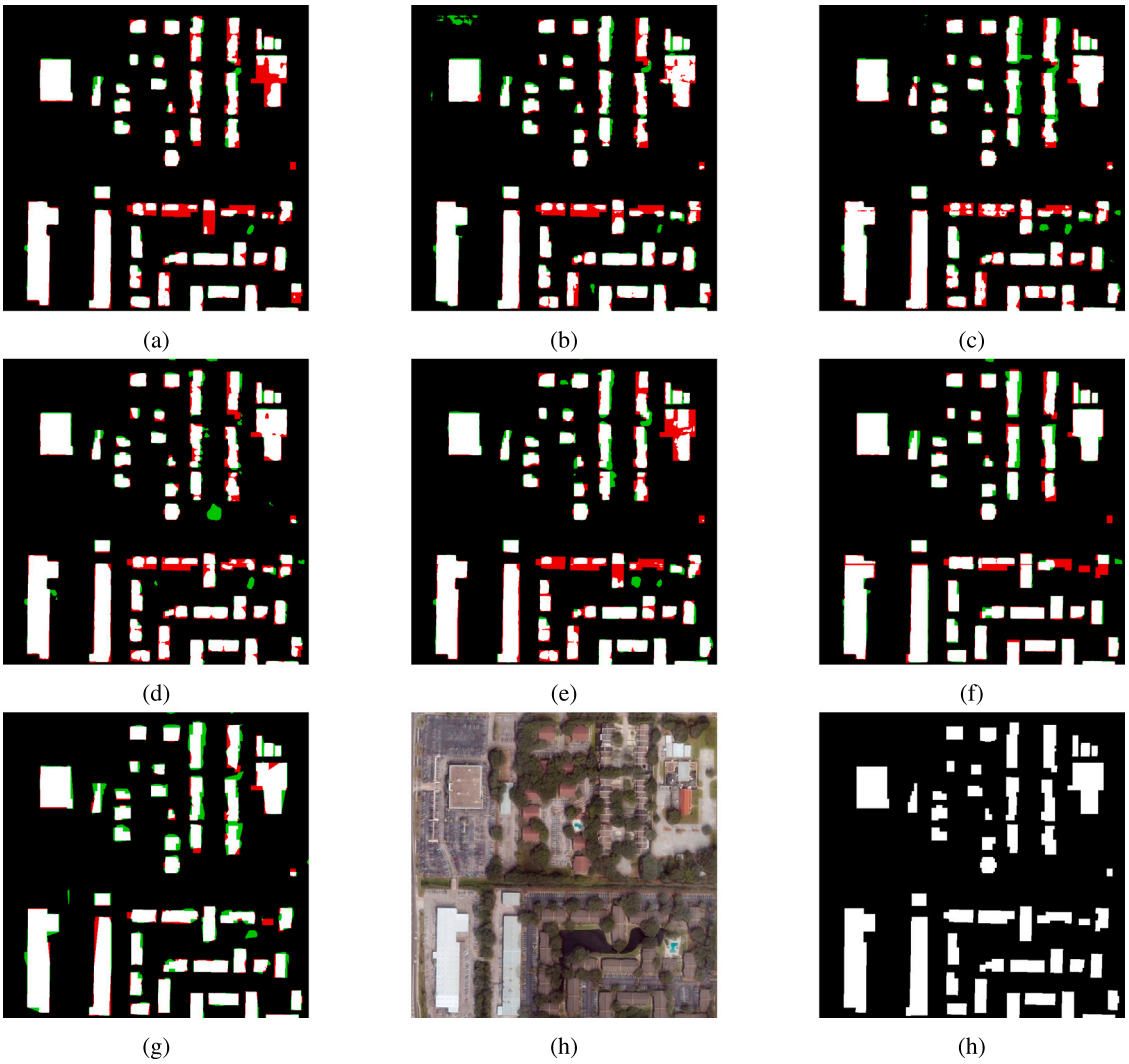


Fig. 14. Building footprints obtained by (a) Srivastava et al., (b) Elhousni et al., (c) Lin et al., (d) Wei et al., (e) Xu et al. (f) Chen et al., and (g) 3DCentripetalNet. (h) and (i) are the corresponding aerial imagery and ground truth from the Urban 3D dataset (spatial resolution: 50 cm/pixel). Pixel-based false negatives, false positives, and true positives are illustrated in red, green, and white, respectively. (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

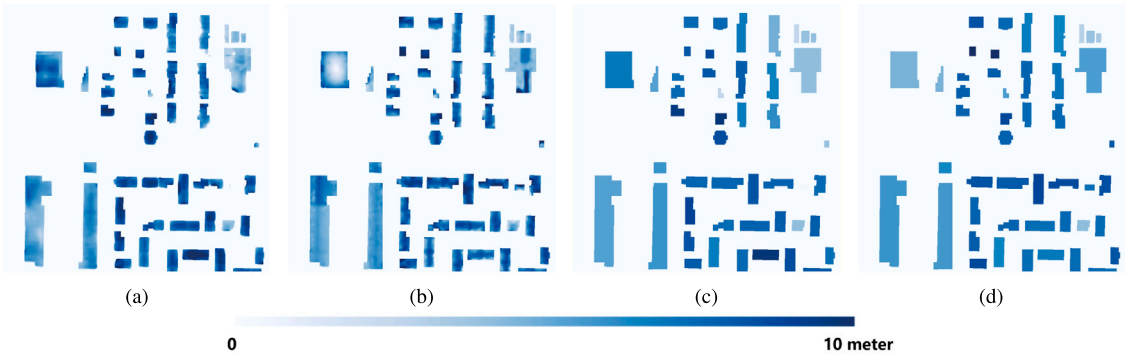


Fig. 15. (a) And (c) are the pixel-wise and instance-wise building height maps estimated by 3DCentripetalNet. (b) and (d) are the corresponding ground reference height maps. The example is from the Urban 3D dataset (spatial resolution: 50 cm/pixel).

where the building footprints and facades are partially invisible. As an example, the 3DCentripetalNet is trained with samples collected from the Urban 3D dataset and tested on off-nadir WorldView3 imagery. However, our method has difficulties in producing accurate building footprints from off-nadir imagery (c.f. Fig. 18). How to adjust 3DCentripetalNet for off-nadir imagery is worth delving into. In our future work, we would like to explore whether segmenting building facades could be beneficial to building height retrieval in such complex application scenarios. The second weakness lies in the inaccuracy of the buildings with complex structures. In Fig. 19, we show one failure case from the Urban 3D dataset. The reason is that the building object with complicated shapes has more corner points, leading to more

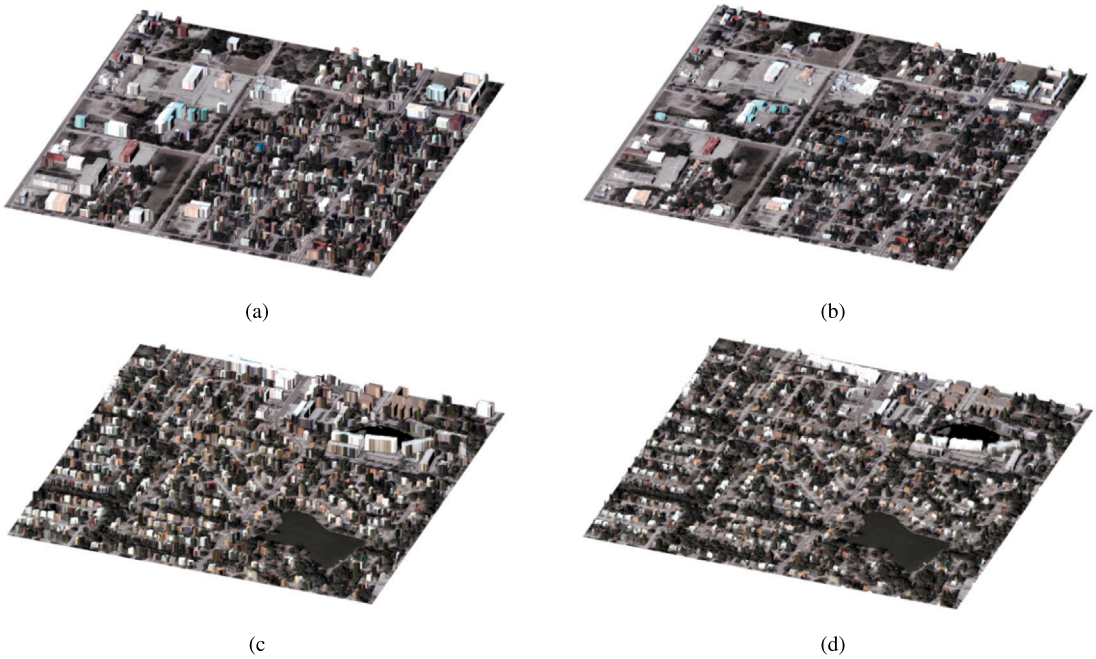


Fig. 16. (a) And (c) are the 3D building models with instance-wise heights estimated by 3DCentripetalNet. (b) and (d) are the 3D building models with pixel-wise heights estimated by 3DCentripetalNet. The example is from the Urban 3D dataset (spatial resolution: 50 cm/pixel).

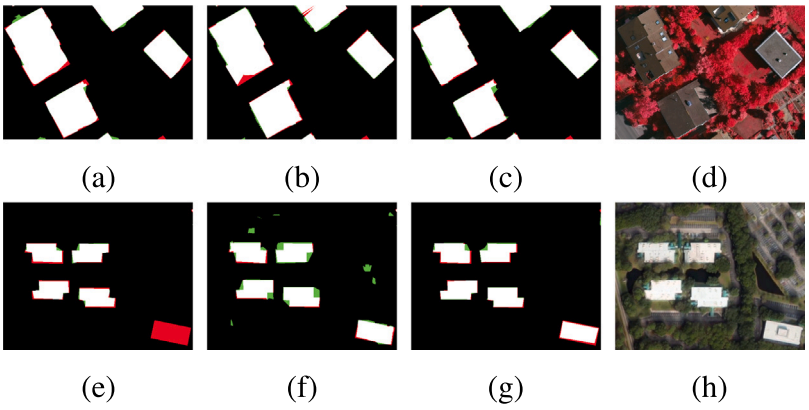


Fig. 17. Building footprints obtained by (a) without 3D centripetal shift representation, (b) without decoupling module, and (c) 3DCentripetalNet. (d) is the corresponding aerial imagery from the ISPRS Vaihingen dataset (spatial resolution: 9 cm/pixel). Building footprints obtained by (e) without 3D centripetal shift representation, (f) without decoupling module, and (g) 3DCentripetalNet. (h) is the corresponding aerial imagery from the Urban 3D dataset (spatial resolution: 50 cm/pixel). Pixel-based false negatives, false positives, and true positives are illustrated in red, green, and white, respectively. (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

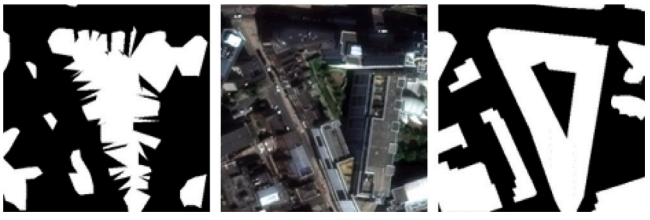


Fig. 18. (a) shows building footprints obtained by 3DCentripetalNet. (b) is the corresponding WorldView3 off-nadir (spatial resolution: 50 cm/pixel). (c) is the corresponding building footprints obtained from OpenStreetMap.

Fig. 19. (a) shows building footprints obtained by 3DCentripetalNet. (b) and (c) are the corresponding aerial imagery and ground truth from the Urban 3D dataset (spatial resolution: 50 cm/pixel).

difficulties and ambiguities in the connection of corners. In future work, we plan to improve corner detection and connection results. On the one hand, corner points can be defined as two types, i.e., convex corners and concave corners, in terms of their interior angle with respect to building polygon (Li et al., 2021b). On the other hand, a better corner connection method such as a graph neural network can be used (Zorzi et al., 2022). Moreover, we plan to incorporate frame field learning to improve the results of complex rooftop polygons (Girard et al., 2021).

7. Conclusion

An approach that retrieves building heights from monocular remote sensing imagery, namely 3DCentripetalNet, is proposed in this study. More specifically, our method first learns 3D centripetal shift representation, which is a novel representation of 3D buildings by incorporating both planar and vertical information of individual building instances. Afterward, a decoupling module is proposed to explicitly decouple the centripetal shift in z-direction into high and low frequency components, where high frequency part enables more accurate detection of building corners. Finally, a 3D modeling module is presented to retrieve building height information using the learned 3D centripetal shift representation and corner points. The performance of 3DCentripetalNet is investigated on two datasets: the ISPRS Vaihingen dataset (9 cm/pixel) and the Urban 3D dataset (50 cm/pixel). Results suggest that incorporating the 3D centripetal shift representation and decoupling module in our method is able to provide better results than other competitors. On the one hand, false alarms, where the background is misclassified as building, is able to be alleviated. On the other hand, complete structures and sharp boundaries of buildings are able to be well preserved. A subsequent study will be investigating the capability of 3DCentripetalNet in various tasks, e.g., 3D vehicle detection.

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Data availability

Public dataset.

Acknowledgments

The work is jointly supported by the Excellence Strategy of the Federal Government and the Länder through the TUM Innovation Network EarthCare, Germany, the Helmholtz Association through the Framework of the Helmholtz Excellent Professorship "Data Science in Earth Observation - Big Data Fusion for Urban Research"(grant number: W2-W3-100), by the German Federal Ministry of Education and Research (BMBF) in the framework of the international future AI lab "AI4EO – Artificial Intelligence for Earth Observation: Reasoning, Uncertainties, Ethics and Beyond" (grant number: 01DD20001) and by German Federal Ministry for Economic Affairs and Climate Action in the framework of the "national center of excellence ML4Earth" (grant number: 50EE2201C).

References

Amirkolaee, H.A., Arefi, H., 2019. Height estimation from single aerial images using a deep convolutional encoder-decoder network. ISPRS J. Photogramm. Remote Sens. 149, 50–66.

Biljecki, F., Ledoux, H., Stoter, J., 2016. An improved LOD specification for 3D building models. Comput. Environ. Urban Syst. 59, 25–37.

Cao, S., Du, M., Zhao, W., Hu, Y., Mo, Y., Chen, S., Cai, Y., Peng, Z., Zhang, C., 2020. Multi-level monitoring of three-dimensional building changes for megacities: trajectory, morphology, and landscape. ISPRS J. Photogramm. Remote Sens. 167, 54–70.

Cao, Y., Huang, X., 2021. A deep learning method for building height estimation using high-resolution multi-view imagery over urban areas: A case study of 42 Chinese cities. Remote Sens. Environ. 264, 112590.

Chen, S., Mou, L., Li, Q., Sun, Y., Zhu, X.X., 2021. Mask-height R-CNN: An end-to-end network for 3D building reconstruction from monocular remote sensing imagery. In: IGARSS 2021-2021 IEEE International Geoscience and Remote Sensing Symposium. IEEE, pp. 1202–1205.

Chen, Z., Qin, Q., Lin, L., Liu, Q., Zhan, W., 2012. DEM densification using perspective shape from shading through multispectral imagery. IEEE Geosci. Remote Sens. Lett. 10 (1), 145–149.

Cote, M., Saeedi, P., 2012. Automatic rooftop extraction in nadir aerial imagery of suburban regions using corners and variational level set evolution. IEEE Trans. Geosci. Remote Sens. 51 (1), 313–328.

Elhousni, M., Zhang, Z., Huang, X., 2021. Height prediction and refinement from aerial images with semantic and geometric guidance. IEEE Access 9, 145638–145647.

Frantz, D., Schug, F., Okujeni, A., Navacchi, C., Wagner, W., van der Linden, S., Hostert, P., 2021. National-scale mapping of building height using sentinel-1 and sentinel-2 time series. Remote Sens. Environ. 252, 112128.

Ghamisi, P., Yokoya, N., 2018. IMG2DSM: Height simulation from single imagery using conditional generative adversarial net. IEEE Geosci. Remote Sens. Lett. 15 (5), 794–798.

Girard, N., Smirnov, D., Solomon, J., Tarabalka, Y., 2021. Polygonal building extraction by frame field learning. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 5891–5900.

Goldberg, H., Brown, M., Wang, S., 2017. A benchmark for building footprint classification using orthorectified RGB imagery and digital surface models from commercial satellites. In: 2017 IEEE Applied Imagery Pattern Recognition Workshop. AIPR, IEEE, pp. 1–7.

Goodfellow, I., Pouget-Abadie, J., Mirza, M., Xu, B., Warde-Farley, D., Ozair, S., Courville, A., Bengio, Y., 2014. Generative adversarial nets. Adv. Neural Inf. Process. Syst. 27.

Hepp, B., Nieß ner, M., Hilliges, O., 2018. Plan3d: Viewpoint and trajectory optimization for aerial multi-view stereo reconstruction. ACM Trans. Graph. 38 (1), 1–17.

Horn, B.K., 1990. Height and gradient from shading. Int. J. Comput. Vis. 5 (1), 37–75.

Huang, X., Zhang, L., 2011. A multidirectional and multiscale morphological index for automatic building extraction from multispectral GeoEye-1 imagery. Photogramm. Eng. Remote Sens. 77 (7), 721–732.

ISPRS, 0000. 2D Semantic Labeling - Vaihingen data http://www2.isprs.org/commissions/comm3/wg4/semantic-labeling.html.

Izadi, M., Saeedi, P., 2011. Three-dimensional polygonal building model estimation from single satellite images. IEEE Trans. Geosci. Remote Sens. 50 (6), 2254–2272.

Karantzalos, K., Argialas, D., 2009. A region-based level set segmentation for automatic detection of man-made objects from aerial and satellite images. Photogramm. Eng. Remote Sens. 75 (6), 667–677.

Li, M., Koks, E., Taubenböck, H., van Vliet, J., 2020a. Continental-scale mapping and analysis of 3D building structure. Remote Sens. Environ. 245, 111859.

Li, X., Li, X., Zhang, L., Cheng, G., Shi, J., Lin, Z., Tan, S., Tong, Y., 2020b. Improving semantic segmentation via decoupled body and edge supervision. In: European Conference on Computer Vision. ECCV, Springer, pp. 435–452.

Li, W., Meng, L., Wang, J., He, C., Xia, G.-S., Lin, D., 2021a. 3D building reconstruction from monocular remote sensing images. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. ICCV, pp. 12548–12557.

Li, Q., Mou, L., Hua, Y., Shi, Y., Zhu, X.X., 2022a. CrossGeoNet: A framework for building footprint generation of label-Scarce Geographical Regions. Int. J. Appl. Earth Obs. Geoinf. (ISSN: 1569-8432) 111, 102824. http://dx.doi.org/10.1016/j.jag.2022.102824, URL https://www.sciencedirect.com/science/article/pii/S1569843222000267.

Li, Q., Shi, Y., Zhu, X.X., 2022b. Semi-supervised building footprint generation with feature and output consistency training. IEEE Trans. Geosci. Remote Sens..

Li, Q., Taubenböck, H., Shi, Y., Auer, S., Roschlaub, R., Glock, C., Kruspe, A., Zhu, X.X., 2022c. Identification of undocumented buildings in cadastral data using remote sensing: Construction period, morphology, and landscape. Int. J. Appl. Earth Obs. Geoinf. 112, 102909.

Li, S., Yang, B., 2008. Multifocus image fusion using region segmentation and spatial frequency. Image Vis. Comput. 26 (7), 971–979.

Li, W., Zhao, W., Zhong, H., He, C., Lin, D., 2021b. Joint semantic-geometric learning for polygonal building segmentation. In: Proceedings of the AAAI Conference on Artificial Intelligence, vol. 35 no. 3. pp. 1958–1965.

Lin, J., Jing, W., Song, H., Chen, G., 2019. ESFNet: Efficient network for building extraction from high-resolution aerial images. IEEE Access 7, 54285–54294.

Liu, J., Ji, S., 2020. A novel recurrent encoder-decoder structure for large-scale multi-view stereo reconstruction from an open aerial dataset. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 6050–6059.

Long, J., Shelhamer, E., Darrell, T., 2015. Fully convolutional networks for semantic segmentation. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. CVPR, pp. 3431–3440.

Mahmud, J., Price, T., Bapat, A., Frahm, J.-M., 2020. Boundary-aware 3D building reconstruction from a single overhead image. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. CVPR, pp. 441–451.

Mou, L., Hua, Y., Zhu, X.X., 2019. A relation-augmented fully convolutional network for semantic segmentation in aerial scenes. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. CVPR, pp. 12416–12425.

Mou, L., Zhu, X.X., 2018. IM2HEIGHT: Height estimation from single monocular imagery via fully residual convolutional-deconvolutional network. arXiv preprint arXiv:1802.10249.

Ok, A.O., Senaras, C., Yuksel, B., 2012. Automated detection of arbitrarily shaped buildings in complex environments from monocular VHR optical satellite imagery. IEEE Trans. Geosci. Remote Sens. 51 (3), 1701–1717.

Paoletti, M., Haut, J., Ghamisi, P., Yokoya, N., Plaza, J., Plaza, A., 2020. U-IMG2DSM: Unpaired simulation of digital surface models with generative adversarial networks. IEEE Geosci. Remote Sens. Lett..

Pentland, A., 1988. Shape information from shading: a theory about human perception. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. ICCV, IEEE, pp. 404–413.

Rajabi, M.A., Blais, J.R., 2004. Optimization of DTM interpolation using SFS with single satellite imagery. J. Supercomput. 28 (2), 193–213.

Ronneberger, O., Fischer, P., Brox, T., 2015. U-Net: Convolutional networks for biomedical image segmentation. In: International Conference on Medical Image Computing and Computer-Assisted Intervention. Springer, pp. 234–241.

Senaras, C., Ozay, M., Vural, F.T.Y., 2013. Building detection with decision fusion. IEEE J. Sel. Top. Appl. Earth Obs. Remote Sens. 6 (3), 1295–1304.

Srivastava, S., Volpi, M., Tuia, D., 2017. Joint height estimation and semantic labeling of monocular aerial images with CNNs. In: IGARSS 2017-2017 IEEE International Geoscience and Remote Sensing Symposium. IEEE, pp. 5173–5176.

Wei, S., Ji, S., Lu, M., 2019. Toward automatic building footprint delineation from aerial images using CNN and regularization. IEEE Trans. Geosci. Remote Sens. 58 (3), 2178–2189.

Xu, L., Liu, Y., Yang, P., Chen, H., Zhang, H., Wang, D., Zhang, X., 2021. HA U-Net: Improved model for building extraction from high resolution remote sensing imagery. IEEE Access 9, 101972–101984.

Zheng, Z., Zhong, Y., Wang, J., 2019. Pop-Net: Encoder-dual decoder for semantic segmentation and single-view height estimation. In: IGARSS 2019-2019 IEEE International Geoscience and Remote Sensing Symposium. IEEE, pp. 4963–4966.

Zhou, X., Zhuo, J., Krahenbuhl, P., 2019. Bottom-up object detection by grouping extreme and center points. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. CVPR, pp. 850–859.

Zorzi, S., Bazrafkan, S., Habenschuss, S., Fraundorfer, F., 2022. Polyworld: Polygonal building extraction with graph neural networks in satellite images. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 1848–1857.