3D Building Reconstruction from Monocular Remote Sensing Images

Weijia Li*[1,2], Lingxuan Meng*[2,3], Jinwang Wang[2,4], Conghui He[2], Gui-Song Xia[4], and Dahua Lin[1,5]

[1]CUHK-SenseTime Joint Lab, The Chinese University of Hong Kong
[2]SenseTime Research
[3]University of Electronic Science and Technology of China
[4]Wuhan University
[5]Shanghai AI Laboratory
{wjli,dhlin}@ie.cuhk.edu.hk, xuanxuanling@std.uestc.edu.cn, guisong.xia@whu.edu.cn

Abstract

3D building reconstruction from monocular remote sensing imagery is an important research problem and an economic solution to large-scale city modeling, compared with reconstruction from LiDAR data and multi-view imagery. However, several challenges such as the partial invisibility of building footprints and facades, the serious shadow effect, and the extreme variance of building height in large-scale areas, have restricted the existing monocular image based building reconstruction studies to certain application scenes, i.e., modeling simple low-rise buildings from near-nadir images. In this study, we propose a novel 3D building reconstruction method for monocular remote sensing images, which tackles the above difficulties, thus providing an appealing solution for more complicated scenarios. We design a multi-task building reconstruction network, named MTBR-Net, to learn the geometric property of oblique images, the key components of a 3D building model and their relations via four semantic-related and three offset-related tasks. The network outputs are further integrated by a prior knowledge based 3D model optimization method to produce the the final 3D building models. Results on a public 3D reconstruction dataset and a novel released dataset demonstrate that our method improves the height estimation performance by over 40% and the segmentation F1-score by 2% - 4% compared with current state-of-the-art.

1. Introduction

3D building reconstruction is an important and fundamental task for monitoring the human settlements and urban environment, assessing the disasters, maintaining the geographical information, etc. During the past few decades, existing methods were mostly based on aerial LiDAR data [31], which is difficult to be applied to large-scale areas due to the expensive cost, low frequency, and limited coverage. For large-scale applications, many approaches have been proposed towards building reconstruction from multi-view imagery [10]. Although the satellite images have a higher acquisition frequency and a larger coverage, the application scenarios of these methods are seriously restricted by the re-

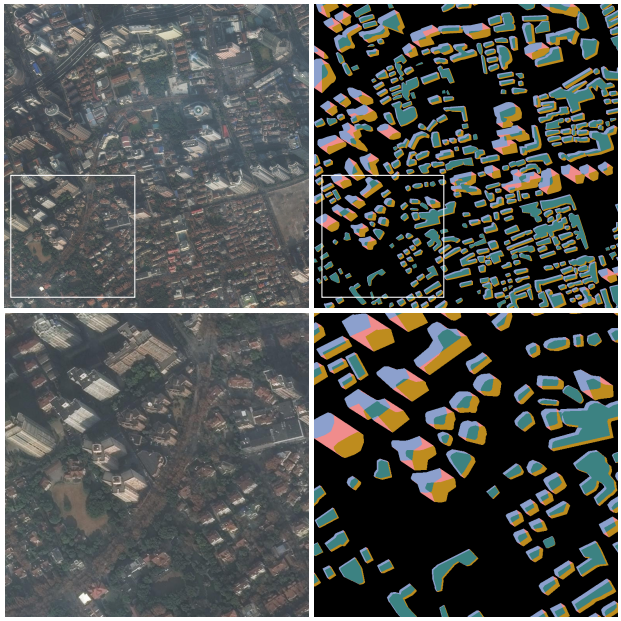


Figure 1. An example of the 3D building reconstruction result obtained from our approach. The purple, brown, pink and green colors denote the roof, footprint, facade and overlapped regions, respectively. Our method produces vector 3D model for buildings with complex shapes and an extreme variance of heights.

*Equal Contribution.

quirement of multiple homologous images over the same region [20]. The monocular image based building reconstruction, on the contrary, avoid such limitations and demonstrate great potential for large-scale applications, which has become an important research topic in recent years [22].

However, the limited information of a monocular image results in great challenges for 3D building reconstruction. As shown in Figure 1, some key components such as the footprints and facades are partially invisible on these images. The serious shadow effect also results in difficulties for accurate segmentation and reconstruction of different parts of a building. Moreover, in large-scale areas, the building heights vary across an extremely wide range. It is hard to directly learn a precise height value via a deep neural network. These challenges restrict the application scene of existing studies to reconstructing simple low-rise buildings from near-nadir images [16, 22, 29, 34].

As an important prerequisite for 3D building reconstruction, building footprint extraction has been extensively explored over a long period. Recent studies are mostly based on deep neural networks, such as semantic segmentation or instance segmentation models [1, 22, 33]. Several studies design polygonal building segmentation approaches to produce the vectorized outputs [17, 18, 19, 35]. Existing methods generally achieve satisfying results for low-rise buildings in near-nadir images, as the footprint contour is completely visible without the parallax effect. However, these methods often produce poor segmentation boundaries when extracting high-rise buildings from oblique images.

Motivated by the progress of monocular depth estimation, various methods have been proposed for building height estimation via deep neural networks [11, 16, 22, 25, 29, 34]. These methods focus on height estimation from near-nadir images, which only take up a small proportion of the remote sensing images. For oblique or off-nadir scenes, a recent study [7] proposed a monocular height estimation method via learning the geocentric pose of buildings, which is designed for single height estimation task instead of 3D building reconstruction. Besides these limitations, all the methods mentioned above produce raster outputs. Further post-processing is required for converting such outputs into the final vector 3D model for practical applications.

In this work, we propose a novel method for 3D building reconstruction from monocular oblique remote sensing images. Our method solves the limitations of previous studies via: (1) a 3D building reconstruction network that converts the ill-posed problem into learning the visible parts of buildings and their relations via four semantic-related and three offset-related tasks; (2) a 3D model optimization method that further integrates the network outputs for improving the height estimation and polygonization, based on the prior knowledge of the building structure. Results demonstrate that our method improves the height estimation performance by over 40% and the segmentation F1-score by 2% - 4% compared with current state-of-the-art.

Our main contributions are summarized as follows:

- We design MTBR-Net, a multi-task building reconstruction network that effectively learns the geometric property of oblique images, the key elements of buildings and their relations, producing 3D models for buildings with various heights and complex shapes.
- We propose a 3D model optimization method that integrates the network outputs based on the prior knowledge of building structures, which further improves the height estimation accuracy and produces vector 3D models with valid shapes.
- We release a new dataset for monocular 3D building reconstruction, including oblique images of multiple views and over 200,000 annotated buildings of a wide range of heights.

2. Related work

2.1. Building extraction from remote sensing images

Building extraction methods have been extensively explored in both remote sensing and computer vision domains. The deep neural network based pixel-wise segmentation methods have become state-of-the-art for building extraction [24, 8, 22]. The multi-task learning strategy has been used in several studies via learning a distance transform [1] or a modified signed distance function from the building boundary [22]. Several other studies combine active contour models with deep neural network to improve the segmentation boundaries for single building segmentation [6, 23, 12]. In addition, some recent approaches produce polygonized footprints that are in a more desirable format for actual applications. Several polygonal segmentation methods are designed for simplifying the segmentation map [17, 18], while others predict the polygon vertices at each time step using a CNN-RNN architecture [19]. In general, existing building extraction methods regard different parts of a building instance as an unified entirety. These methods usually produce poor segmentation boundaries when extracting high-rise buildings from oblique images, which take up a substantial proportion in actual scenes. Our method, on the contrary, predicts the visible components (e.g. roof, facade, and skeleton) and their position relations (offset) via a multi-task network, and effectively integrates these predictions to produce accurate footprint polygons.

2.2. Building height estimation and reconstruction

Over a long period, a large number of building reconstruction methods are based on LiDAR data [31] and multi-

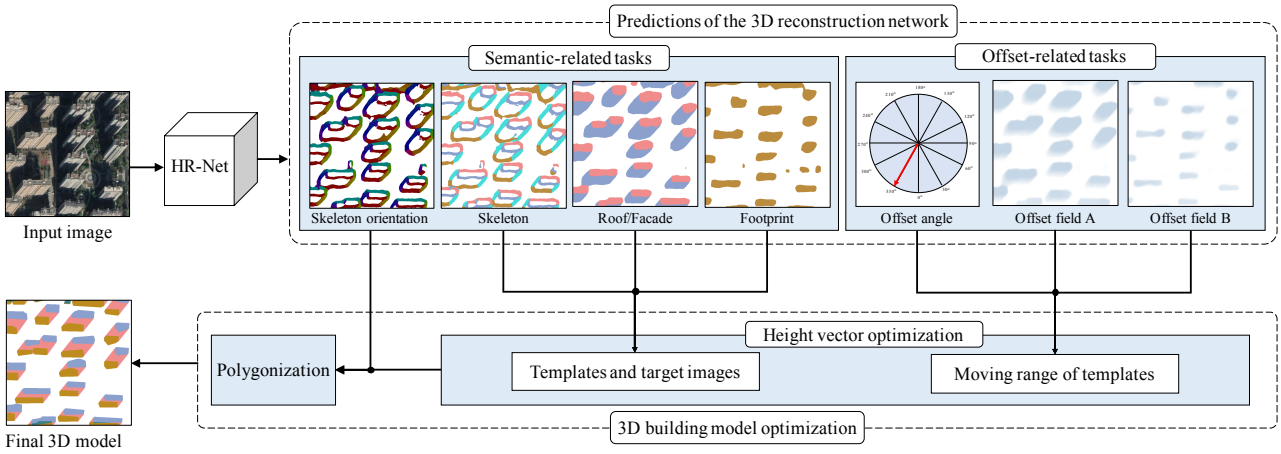


Figure 2. An overview of our proposed method. Taking a monocular remote sensing image as input, our MTBR-Net outputs a roof/facade segmentation map, a footprint segmentation map, a skeleton segmentation map, a skeleton orientation prediction map, a image-wise offset angle, and two pixel-wise offset field maps. The outputs of the four semantic-related tasks and three offset-related tasks are further integrated for height vector optimization and polygonization, producing the final vectorized 3D model.

view imagery [10, 3, 28, 20], which have limitations of expensive data acquisition costs and limited coverage, and require multiple homologous images over the same region. For monocular image based building reconstruction, traditional methods are based on the shadow information, lines and line intersections of the building outlines, etc., as well as the meta information of satellites such as the sun-earth relative position [15, 26]. These methods usually require a series of complex procedures for reconstructing the 3D building model from the above information.

Motivated by the progress of monocular depth estimation, several recent studies propose deep learning based methods for monocular building height estimation. Some studies propose a single-task network for building height estimation, via regressing height values using an encoder-decoder network [25] or simulating a height map using generative adversarial networks [11]. Several other studies are designed for both building footprint extraction and height estimation via a multi-task network [29, 34], or exploit the semantic labels as prior information for height estimation [16]. Different from our study, all these methods focus on height estimation from near-nadir images. Moreover, the raster reconstruction results require further post-processing process to generate the final 3D building model.

For building height estimation from oblique images, Christie et al. [7] proposed a monocular height estimation method via learning the geocentric pose of buildings, i.e, an image-wise flow angle and a pixel-wise magnitude value [27], under the prerequisite that buildings of the same image have the same offset angle. This study only focuses on single-task height estimation instead of 3D reconstruction, and the prerequisite is not always applicable. Our method, by contrast, includes an image-wise offset angle prediction task and two pixel-wise offset field prediction tasks, as well as several semantic-related tasks, which is applicable for images with different offset angles and produces vectorized 3D reconstruction results.

2.3. Building reconstruction datasets

Several public datasets provide both footprint annotations and pixel-wise height information that are generated from LiDAR data. ISPRS Potsdam and Vaihingen [14] and Urban Semantic 3D (US3D) [2] are two popular datasets used in many recent studies [11, 22, 29, 34, 16], in which most of the images are near-nadir and the roof and footprint are nearly overlapped. Recently, Christie et al. [7] proposed two new datasets, i.e, DFC19 and ATL-SN4, which extend the US3D [2] and SN4 [32] datasets to include additional images with a wider range of oblique viewing angles. Although a variety of annotations are provided, most annotation types (e.g. roof, facade, and height) are generated from point cloud data and have plenty of fragments and noises. These datasets are difficult to be used for vector 3D model reconstruction and performance evaluation at instance level. Unlike the existing 3D reconstruction datasets mentioned above, the dataset proposed in this study provides manually labeled roof, facade, footprint, and height of each building instance in complete shapes and vector format. The proposed dataset can be used for vector 3D model reconstruction from oblique remote sensing images and performance evaluation at both instance and pixel levels.

3. Methods

The overall framework of our proposed approach is demonstrated in Figure 2, which consists of two main components: (1) a multi-task deep neural network that produces the 3D building reconstruction model via four semantic-

related and three offset-related tasks. (2) a 3D model optimization module that integrates the network outputs to further improve the height estimation and produce vector 3D models with valid shapes. Taking a monocular remote sensing image as input, an HR-Net based multi-task network is designed for seven interrelated tasks, i.e. a roof/facade segmentation and a skeleton segmentation task for predicting the visible parts of a building; a skeleton orientation prediction task for polygonization; a footprint segmentation task based on our proposed feature warping module; an image-level offset angle prediction and two pixel-level offset field prediction tasks for predicting the relation between roof and footprint. The 3D reconstruction results obtained from the network is further optimized via a prior knowledge based method for improving the height estimation, and a skeleton orientation based polygonization method for producing vector 3D building models. In the following, we first introduce the definitions of the seven tasks. Then we introduce the training of our MTBR-Net and the 3D model optimization method. The implementation details are described at the end of this section.

3.1. Task definitions of MTBR-Net

3.1.1 Semantic-related tasks

Roof/Facade and Footprint: The semantic-related tasks are designed for producing the essential components of a 3D building model. We first design a task for roof and visible facade segmentation, which have complete contour on the monocular remote sensing images. The footprint contours, on the contrary, are often partially invisible but have the same shape as the roof contour. Under this prerequisite, our footprint segmentation task is based on warping the feature map of roof using the predicted offset field, which will be introduced in Section 3.2.

Building Skeleton: To learn the structure of a 3D building model, we define four types of semantic edges that are: (1) between roof and background (E_a), (2) between roof and facade (E_b), (3) between facade and background (E_c), (4) between facade and footprint (E_d). The four types of edges constitute the whole visible skeleton of a building on oblique images. For a building instance, E_b usually has the same shape as E_d, which serves as an important prior knowledge in the 3D building model optimization process.

Skeleton Orientation: Inspired by [18], we design a task to predict the edge orientation of the building skeleton, which will be utilized for converting the raster segmentation map into vector 3D model in the polygonization phase. For each pixel on the skeleton, its orientation is decided by the angle between the edge normal and the gravity direction in the clockwise direction, e.g., the α in Figure 3. The detailed definition can be found in [18].

Figure 3. Representation of different types of supervisions. The middle image shows the annotation of offset field A for the left remote sensing image. In the right image, the four types of semantic edge of the building skeleton are denoted by different colors. The angle between the black arrow and the gravity direction defines the skeleton orientation. The red arrow denotes the offset vector.

3.1.2 Offset-related tasks

The offset-related tasks are designed for estimating the height of each building. We encode the relative height of a building as an offset vector of two signed values (denoted by O_x and O_y), reflecting the direction from a roof to its corresponding footprint (denoted by the red arrows in Figure 3). The offset vector will be used for warping the roof to footprint in both network training and building model optimization phases, which can be further converted to the actual height based on the meta information (image resolution and nadir angle) [15, 26]. In general, the buildings of a single-source remote sensing image often have the same offset angle. However, some publicly available images, such as the Google Earth imagery, are mosaicked from different data sources with multiple offset angles. Considering both cases, we design an image-wise offset angle prediction task and two pixel-wise offset field prediction tasks.

Offset Field A: The first pixel-wise task is designed for predicting the offset vector for roof and facade regions, which will be used for warping the predicted roof segment to footprint after network training. For offset field A, the pixels of roof regions are assigned as the same values, i.e., the offset vector from roof to footprint, which is denoted by (O_x^r, O_y^r). The pixels within the facade region are assigned as (δ_x, δ_y) from the current pixel to the footprint contour, i.e., the values of E_b to E_d vary gradually from (O_x^r, O_y^r) to $(0, 0)$. The offset field values of background regions are set as $(0, 0)$. A visualization example of the offset field annotation can be found in the middle of Figure 3.

Offset Field B: The second pixel-wise task is designed for predicting the offset vector for footprint regions, which will be used for warping the feature map of roof/facade segmentation to footprint during training phase. For offset field B, the pixels of footprint regions are assigned as the offset field A values of the corresponding roof regions, i.e., (O_x^r, O_y^r), while the pixels in other regions are assigned as $(0, 0)$.

Offset Angle: Although pixel-wise tasks are good at handling images with multiple offset angles, it is hard to predict the offset angle accurately for low-rise buildings. This problem can be easily solved by image-wise offset prediction via

learning from nearby high-rise buildings.

3.2. Training of MTBR-Net

Our MTBR-Net is based on the HR-Net architecture [30]. The capacity of maintaining high-resolution representations throughout the whole process is beneficial for remote sensing images with a relatively low spatial resolution and a large image size. In our method, the footprint segmentation task is based on the warped feature map of roof/facade segmentation, while the other six tasks share the same feature representation. Each task owns a task-specific head that consists of two 1×1 convolution layers.

Offset-based Feature Warping Module: We design an offset-based feature warping module for footprint segmentation, which not only strengthen the relation constraints between semantic and offset-related tasks but improves the footprint segmentation performance. First, we warp the output feature map of the first 1×1 convolution of roof/facade segmentation task based on offset field B. Then the warped feature map is concatenated with the prediction map of offset field B and the feature map of roof/facade segmentation, constituting the feature map for footprint segmentation.

The roof/facade, footprint, skeleton, and orientation prediction tasks are formulated as pixel-wise segmentation problems. The loss function of the above tasks are denoted by $\mathcal{L}_{rf}$, $\mathcal{L}_{foot}$, $\mathcal{L}_{ske}$, and $\mathcal{L}_{ori}$ (uniformly denoted by $\mathcal{L}_{seg}$) and calculated according to formula 1, in which N denotes the number of pixels of an image; C denotes the number of classes; $y_{i,c}$ and $p(y_{i,c})$ denote the binary indicator and the predicted probability that pixel i belongs to class c.

$$\mathcal{L}_{seg} = -\frac{1}{N}\sum_{i=1}^{N}\sum_{c=1}^{C} y_{i,c} \times \log(p(y_{i,c})) \tag{1}$$

The total loss of four semantic-related tasks ($\mathcal{L}_{sem}$) is the weighted sum of each task-specific loss:

$$\mathcal{L}_{sem} = \mathcal{L}_{ori} + \alpha_1 \mathcal{L}_{rf} + \alpha_2 \mathcal{L}_{foot} + \alpha_3 \mathcal{L}_{ske} \tag{2}$$

For the offset-related tasks, we formulate the image-wise offset angle prediction as a classification problem to simplify the training process, and formulate the pixel-wise offset field prediction as a regression problem to obtain precise offset values. The loss of angle prediction task $\mathcal{L}_{ang}$ is calculated by formula 3, where K denotes the number of angle classes; y_k and $p(y_k)$ denote the binary indicator and the predicted probability for class k.

$$\mathcal{L}_{ang} = -\sum_{k=1}^{K} y_k \times \log(p(y_k)) \tag{3}$$

The loss of two offset field regression tasks ($\mathcal{L}_{field\_a}$ and $\mathcal{L}_{field\_b}$, uniformly denoted by $\mathcal{L}_{field}$) is calculated by the endpoint error according to formula 4,

$$\mathcal{L}_{field} = \frac{1}{N}\sum_{i=1}^{N} ||\vec{O}_i^{pred} - \vec{O}_i^{gt}||_2, \tag{4}$$

where the predicted offset $\vec{O}_i^{pred} = [O_{x,i}^{pred}, O_{y,i}^{pred}]$, the ground truth offset $\vec{O}_i^{gt} = [O_{x,i}^{gt}, O_{y,i}^{gt}]$. The total loss of three offset-related tasks ($\mathcal{L}_{off}$) can be calculated as:

$$\mathcal{L}_{off} = \mathcal{L}_{ang} + \mathcal{L}_{field\_a} + \mathcal{L}_{field\_b} \tag{5}$$

The total loss of our MTBR-Net can be summarized as:

$$\mathcal{L} = \mathcal{L}_{sem} + \mathcal{L}_{off} \tag{6}$$

3.3. Optimization of the 3D building model

We design a simple but effective method to further optimize the building reconstruction results via integrating the outputs of MTBR-Net, which consists of two major phases: (1) a prior knowledge based template matching method for optimizing the height estimation result for each building instance; (2) a skeleton orientation based polygonization method for converting the raster results into vector 3D building model with valid shapes.

The template matching method for height vector optimization is based on the prior knowledge that: (1) the edge between roof and facade (E_b) usually has the same shape as the edge between facade and footprint (E_d); (2) the roof often has the same contour shape as the footprint. For each building instance on the roof/facade segmentation map S_{rf}, we extract the template of E_b from the skeleton segmentation map S_{ske}, the template of E_b from S_{rf}, and the template of roof segment from S_{rf}. Accordingly, the target images for the above three templates are E_d on S_{ske}, E_d on and S_{rf}, and the footprint segments on S_{rf}.

For each building instance, let $\vec{V}$ denote the vector for moving the templates. Vector $\vec{V}$ has a fixed moving direction angle and a range of moving distance. The direction angle of $\vec{V}$ is determined either by the offset angle or the offset field A predictions. Specifically, for low-rise building instances (with the offset length smaller than a threshold T_{off}), the direction angle is assigned with the image-wise angle prediction result; otherwise, it is assigned with the angle of the average offset field of the roof region. The moving distance range of $\vec{V}$ is determined by the length of the average offset field of the roof region (len) and two used defined ratios (r_1 and r_2), i.e, $[r_1 \times len, r_2 \times len]$. We use the IoU between the templates and the corresponding target images to calculate the template matching score. The length of vector $\vec{V}$ is optimized via a grid search method with an interval of 1 pixel. For each building instance, the vector $\vec{V}$ that maximizes the matching score (IoU) is the final optimized height vector.

Polygonization: Based on the skeleton orientation prediction, the raster roof segments can be simplified into polygons with valid shapes. For each instance, the pixels densely sampled from the roof contour constitute an initial

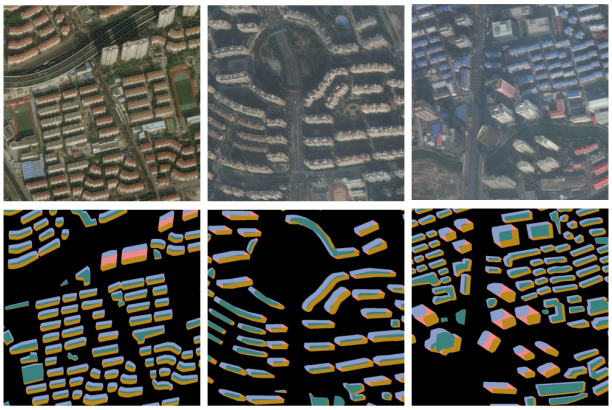


Figure 4. Examples of the 3D building reconstruction results of our method. The purple, brown, pink and green colors denote the roof, footprint, facade and overlapped regions, respectively.

vertex set. For each initial vertex, we calculate the absolute difference between its orientation class and its neighbour vertex. If it is greater than a given threshold (T_{ori}), the vertex will be selected as valid and remained; otherwise, it will be removed from the vertex set. The remaining valid vertices constitute the simplified roof polygon, which will be warped as the footprint based on the height vector. The simplified roof polygon, footprint polygon, and the height vector comprise the optimized 3D building model.

3.4. Implementation details

For the HR-Net architecture used in our MTBR-Net, the numbers of channels in the four stages are set as 12, 24, 48, and 96. The size of input image is 500×500 pixels. The weights for $\mathcal{L}_{sem}$ calculation (α_1, α_2, and α_3) are set as 3, 3 and 2. The number of classes for offset angle prediction and skeleton orientation prediction are both set as 36, indicating that the bin width of the angle is $10°$. For skeleton segmentation and skeleton orientation prediction tasks, the foreground types are assigned with larger loss weights than background (40:1 and 360:1 for two tasks) in order to predict thick edges that are more robust to the roof/facade segmentation results. For the height vector optimization, the threshold T_{off} used for determining the direction angle is set as 3 pixels. The ratios for determining the offset length range (r_1 and r_2) are set as 0.7 and 1.5. In the polygonization phase, the orientation difference threshold T_{ori} is set as 2, which indicates that the vertex with an interior angle of $160°$ to $200°$ will be regarded as an invalid vertex for simplifying the segmentation contour.

4. Experiments

4.1. Datasets

In this study, we propose a new dataset for 3D building reconstruction from monocular remote sensing images,

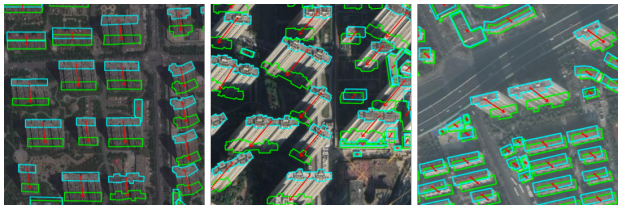


Figure 5. Examples of the GT buildings of our dataset. The annotators are required to manually annotate the roofs (cyan polygons) and the offset vectors (red arrows), producing the footprints (green polygons) with the same shape as the corresponding roofs.

which provides holistic annotations and solves the limitations of existing public datasets mentioned in Section 2.3. The dataset contains oblique remote sensing images of a diversity of view angles, which are collected from different data sources (e.g. Google Earth and Microsoft Virtual Earth). Over 200,000 buildings are annotated in our dataset, which are located in multiple cities of China (including Beijing, Shanghai, Harbin, Chengdu, Jinan and Xi'an). Figure 4 shows some examples of the 3D building reconstruction results obtained from our method. Figure 5 shows examples of the ground truth (GT) buildings of our dataset.

Our dataset contains 2,700 training images, 300 validation images and 300 test images, which are cropped into $1,024 \times 1,024$ pixels. To better evaluate the generalization capacity of the proposed method for large-scale applications, we divide our test dataset into an in-domain dataset containing 200 images located in the same city but different regions with the training dataset, and an out-domain dataset containing 100 images located in a new city that is not included in the training dataset. The whole dataset will be released on https://liweijia.github.io/projects/building_3d/.

The performance of the 3D building reconstruction results is evaluated from different perspectives in the following sections. In section 4.2, we evaluate the height estimation performance in terms of the offset vector, actual height and offset angle. In section 4.3, we evaluate the building roof segmentation and footprint extraction results. The effect of different components will be analyzed in Section 4.4.

4.2. Height estimation performance

The height estimation performance is evaluated on both our proposed dataset and DFC19 [7], a recent published 3D reconstruction dataset. For our proposed dataset, we compare the relative height (offset vector) estimation performance of our method with current state-of-the-art method [7], of which the network architecture is modified from U-Net to HR-Net for a fair comparison. Table 1 lists the EPE values obtained from two methods. For both methods, the pixel-wise offset prediction are converted into instance-wise results via calculating the average offset of each roof instance. We report the EPE of the roof instances within dif-

Table 1. Comparison of building height estimation on our proposed dataset. We report the EPE of the roof instances within different height range and the average EPE of all instances. Our method reduces the EPE of high-rise buildings by 5 to 24 pixels compared with [7].

Dataset	Method	EPE of different height range (in pixels)										Average
		0-10	10-20	20-30	30-40	40-50	50-60	60-70	70-80	80-90	90-100	>100	EPE
In-domain	Christie et al. [7]	6.22	5.26	7.04	9.01	10.94	12.52	14.89	19.47	24.50	73.07	50.41	6.19
	Ours	**4.92**	**4.24**	**6.02**	**5.91**	**6.87**	**7.82**	**8.39**	**12.45**	**20.75**	**61.41**	**26.69**	**4.88**
Out-domain	Christie et al. [7]	7.99	**9.83**	9.81	10.41	13.31	16.11	19.41	24.13	21.27	26.17	75.21	12.31
	Ours	**6.63**	9.96	**8.33**	**8.56**	**9.32**	**9.45**	**12.55**	**15.75**	**10.76**	**11.82**	**52.52**	**9.59**

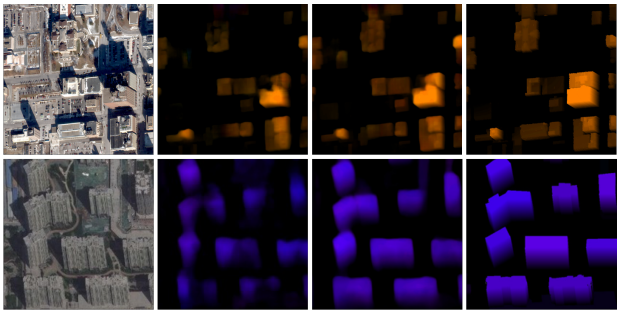


Figure 6. Examples of height estimation results on DFC19 (top) and our dataset (bottom) obtained from our method and current state-of-the-art [7]. Different colors represent different offset angles. The brightness of each color reflects the offset length.

ferent height range (the pixel length of offset vectors) and the average EPE of all instances. Our method reduces the EPE of high-rise buildings by 5 to 24 pixels compared with [7], indicating the superior of our offset encoding manner and the effectiveness of our multi-task learning strategies.

For DFC19, we report all metrics regarding the building height estimation following [7], including the pixel-level mean absolute error (MAE) and root mean square error (RMSE) of the actual height (in meters), the pixel-level endpoint error (EPE) in roof and facade regions (in pixels), and the image-level angle prediction error (in degree). As shown in Table 2, we compare the results of our method with the best results reported in [7] and two winning solutions [16, 34] in 2019 Data Fusion Contest [9]. For our method, we replace the offset field A prediction task with the flow vector prediction of [7] for evaluating the actual height, and calculate the average offset angle of roof regions for evaluating the image-level metric. Results show that our method significantly outperforms state-of-the-art in terms of all metrics, reducing the actual height RMSE and angle error by over 40%. Figure 6 provides a qualitative comparison of the results obtained from our method and the state-of-the-art results obtained from [7]. Results demonstrate that our method produces height estimation results with more accurate offset values and building boundaries.

4.3. Building segmentation performance

We further evaluate the roof and footprint segmentation results using our proposed dataset. To the best of our knowledge, this is the first dataset that provides manually labeled

Table 2. Comparison of building height estimation on DFC19 dataset, in terms of the MAE and RMSE of actual height (in meters), EPE of offset vector (in pixels), and angle error (in degrees).

Method	Actual Height		Offset EPE		Angle Error
	MAE	RMSE	Roof	Facade	
Kunwar [16]	8.33	19.65	-	-	-
Zheng et al. [34]	8.72	19.32	-	-	-
Christie et al. [7]	7.73	16.87	5.44	7.11	15.09
Ours	**4.75**	**9.57**	**4.67**	**5.35**	**8.40**

roof, offset, and footprint annotations, which are essential for our MTBR-Net. We compare the segmentation performance of our approach with the current state-of-the-art method for polygonal building segmentation proposed in Li et al. [18] and several other competitive segmentation methods [21, 13, 4, 5, 30]. We calculate the precision, recall, and F1-score (IoU ≥ 0.5) at instance-level following [18, 8].

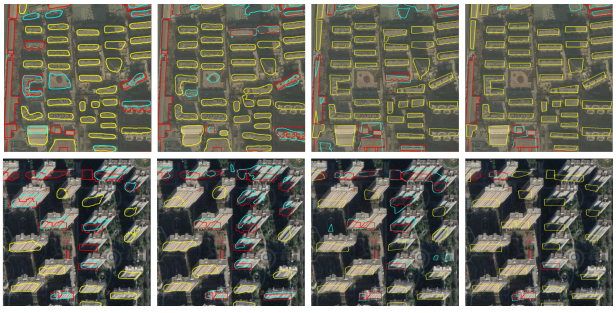
Table 3 lists the roof and footprint segmentation results on the in-domain and out-domain test datasets. Our method obtains the highest precision, recall, and F1-score for all cases. For the roof segmentation results, our method improves the F1-score of the single-task HR-Net by 4.6% and 10.6%, which indicates that the proposed interrelated tasks can effectively benefit the roof segmentation results via joint learning. Regarding the footprint extraction performance, our method improves the F1-score by 2.5% and 4.3% compared with current state-of-the-art [18], which indicates the effectiveness of warping the predicted roof instances to footprints using offset vectors. For all methods, the performance drop on the out-domain dataset is due to the change of test city as well as the increasing ratio of very high-rise buildings compared with the in-domain dataset. The runtime of our method is about 2.8 seconds per test image on a Titan Xp GPU. Figure 7 provides a qualitative comparison of the footprint extraction results. Results show that our method produces polygonal footprints with the most accurate boundaries even for high-rise buildings. On the other hand, our method has difficulties in accurately reconstructing the extremely adjacent building instances, buildings without a clear boundary between the roof and the facade, and buildings with non-flat roofs (such as the family houses), which should be improved in our future work.

4.4. Ablation study

In this section, we analyze the effect of the main novel modules of our proposed approach, including: (1) the

Table 3. Building roof and footprint segmentation results of different methods, in terms of precision, recall and F1-score (%). Our method improves the roof segmentation F1-score by 1.6% and 3.0%, and improves the footprint segmentation F1-score by 2.5% and 4.3% compared with current state-of-the-art [18].

Method	In-domain dataset (Roof)			In-domain dataset (Footprint)			Out-domain dataset (Roof)			Out-domain dataset (Footprint)		
	Precision	Recall	F1-score	Precision	Recall	F1-score	Precision	Recall	F1-score	Precision	Recall	F1-score
Cascade Mask R-CNN [4]	66.68	67.06	66.87	61.27	61.48	61.37	48.39	48.74	48.56	40.73	39.31	40.00
Mask R-CNN [13]	67.98	69.35	68.66	63.43	63.85	63.64	59.65	52.09	55.62	50.30	41.29	45.35
PANet [21]	68.38	67.98	68.18	64.03	61.91	62.95	62.11	50.46	55.68	52.54	41.03	46.08
HR-Net [30]	68.78	66.09	67.41	64.19	64.29	64.24	55.76	46.62	50.78	41.95	35.06	38.20
Li et al. [18]	71.76	69.25	70.48	65.71	66.37	66.04	60.44	56.40	58.35	49.69	45.77	47.65
Ours (w/o optimization)	**72.72**	**71.37**	**72.04**	66.85	**68.05**	67.44	**65.20**	**57.97**	**61.37**	54.34	46.37	50.04
Ours (w/ optimization)	**72.72**	**71.37**	**72.04**	**69.47**	67.71	**68.58**	**65.20**	**57.97**	**61.37**	**56.45**	**48.17**	**51.98**



Mask R-CNN [13] HR-Net [30] Li et al. [18] Ours

Figure 7. Building footprint extraction results of different methods. The yellow, cyan, and red polygons denote the TP, FP, and FN. Our method produces much more accurate footprint boundaries than other three methods.

offset-related prediction tasks; (2) the offset-based feature warping module; (3) the prior knowledge based 3D model optimization module. Table 4 lists the results on two test datasets obtained by successively applying the above modules, in terms of both footprint extraction F1-score and height estimation EPE. The results of the Baseline method are obtained from [18]. The second row (+ Offset Field) shows the height estimation results of offset field A prediction and the footprint extraction results obtained from warping the predicted roof instances to footprints based on the average offset value of each roof instance. The third row (+ Feature Warp) shows the results obtained from applying the offset-based feature warping module, which are calculated from the footprint segmentation task. The final row (+ Optimization) shows the results obtained from applying the prior knowledge based 3D model optimization method.

Results show that the footprint extraction score can be improved by 0.8% via warping the roofs to footprints using the offset field A prediction. Moreover, compared with the baseline that directly predicting the footprint without feature warping, the F1-score can be improved by 1.4% and 2.4% via applying the feature warping module, indicating the effectiveness of using offset field B prediction to warp the feature map of roof/facade prediction for footprint segmentation. The prior knowledge based model optimization method further improves the building height estimation and

Table 4. Results of ablation study on two test datasets, in terms of the footprint segmentation F1-score (%) and the height estimation EPE (in pixels).

Method	Segmentation F1-score (↑)		Height estimation EPE (↓)	
	In-domain	Out-domain	In-domain	Out-domain
Baseline	66.04	47.65	-	-
+ Offset Field	66.79	48.49	5.26	10.45
+ Feature Warp	67.44	50.04	5.17	10.21
+ Optimization	**68.58**	**51.98**	**4.88**	**9.59**

footprint extraction results via effectively using all types of predictions, producing the footprints with the highest F1-score while maintaining same contour shape as the roofs.

5. Conclusion

In this paper, we have presented a novel 3D building model reconstruction method that produces vector 3D building model with accurate roof, facade, footprint, and height from monocular remote sensing images. Qualitative and quantitative evaluations demonstrate the significant advantages of our approach over state-of-the-art methods. The effect of different components of our approach is also verified in the ablation study. To the best of our knowledge, this is the first work that produces vectorized 3D building model reconstruction results from monocular remote sensing images using deep neural networks. We believe that this paper provides effective solutions for 3D building reconstruction in large-scale and complex application scenes. In our future work, we would like to explore more effective strategies for improving the 3D reconstruction results, such as utilizing more prior knowledges regarding the building structure, and improving the multi-task learning process via adding more constraints based on the relation of different components of a building instance.

Acknowledgements. This work has been supported by the Centre of Perceptual and Interactive Intelligence, CUHK Agreement TS1712093, the General Research Fund (GRF) of Hong Kong (No. 14205719), and Theme-based Research Scheme 2020/21 (No. T41-603/20-R). Gui-Song Xia was supported by the National Natural Science Foundation of China under Grant 61922065, Grant 61771350, Grant 41820104006.

References

[1] Benjamin Bischke, Patrick Helber, Joachim Folz, Damian Borth, and Andreas Dengel. Multi-task learning for segmentation of building footprints with deep neural networks. In *IEEE International Conference on Image Processing (ICIP)*, pages 1480–1484. IEEE, 2019.

[2] Marc Bosch, Kevin Foster, Gordon Christie, Sean Wang, and Myron Brown. Semantic stereo for incidental satellite images. In *IEEE Winter Conference on Applications of Computer Vision (WACV)*, 2019.

[3] Randi Cabezas, Julian Straub, and John W. Fisher. Semantically-aware aerial reconstruction from multi-modal data. In *IEEE International Conference on Computer Vision (ICCV)*, 2016.

[4] Zhaowei Cai and Nuno Vasconcelos. Cascade r-cnn: Delving into high quality object detection. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 6154–6162, 2018.

[5] Kai Chen, Jiangmiao Pang, Jiaqi Wang, Yu Xiong, Xiaoxiao Li, Shuyang Sun, Wansen Feng, Ziwei Liu, Jianping Shi, Wanli Ouyang, et al. Hybrid task cascade for instance segmentation. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 4974–4983, 2019.

[6] Dominic Cheng, Renjie Liao, Sanja Fidler, and Raquel Urtasun. Darnet: Deep active ray network for building segmentation. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 7431–7439, 2019.

[7] Gordon Christie, Rodrigo Rene Rai Munoz Abujder, Kevin Foster, Shea Hagstrom, Gregory D Hager, and Myron Z Brown. Learning geocentric object pose in oblique monocular images. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 14512–14520, 2020.

[8] Ilke Demir, Krzysztof Koperski, David Lindenbaum, Guan Pang, Jing Huang, Saikat Basu, Forest Hughes, Devis Tuia, and Ramesh Raskar. Deepglobe 2018: A challenge to parse the earth through satellite images. In *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*, 2018.

[9] DFC19. Data Fusion Contest 2019, https://github.com/pubgeo/dfc2019.

[10] Liuyun Duan and Florent Lafarge. Towards large-scale city reconstruction from satellites. In *European Conference on Computer Vision (ECCV)*, 2016.

[11] Pedram Ghamisi and Naoto Yokoya. Img2dsm: Height simulation from single imagery using conditional generative adversarial net. *IEEE Geoence Remote Sensing Letters*, pages 1–5, 2018.

[12] Shir Gur, Tal Shaharabany, and Lior Wolf. End to end trainable active contours via differentiable rendering. *arXiv*, 2019.

[13] Kaiming He, Georgia Gkioxari, Piotr Dollár, and Ross Girshick. Mask r-cnn. In *Proceedings of the IEEE international conference on computer vision (CVPR)*, pages 2961–2969, 2017.

[14] ISPRS. ISPRS 2D semantic labeling challenge, http://www2.isprs.org/commissions/comm3/wg4/semantic-labeling.html.

[15] M. Izadi. Three-dimensional polygonal building model estimation from single satellite images. *IEEE Transactions on Geoence Remote Sensing*, 50(6):2254–2272, 2012.

[16] Saket Kunwar. U-net ensemble for semantic and height estimation using coarse-map initialization. In *IGARSS 2019-2019 IEEE International Geoscience and Remote Sensing Symposium*, pages 4959–4962. IEEE, 2019.

[17] Muxingzi Li, Florent Lafarge, and Renaud Marlet. Approximating shapes in images with low-complexity polygons. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2020.

[18] Weijia Li, Wenqian Zhao, Huaping Zhong, Conghui He, and Dahua Lin. Joint semantic–geometric learning for polygonal building segmentation. In *AAAI*, 2021.

[19] Zuoyue Li, Jan Dirk Wegner, and Aurélien Lucchi. Topological map extraction from overhead images. In *Proceedings of the IEEE International Conference on Computer Vision (CVPR)*, pages 1715–1724, 2019.

[20] Jin Liu and Shunping Ji. 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 (CVPR)*, pages 6050–6059, 2020.

[21] Shu Liu, Lu Qi, Haifang Qin, Jianping Shi, and Jiaya Jia. Path aggregation network for instance segmentation. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 8759–8768, 2018.

[22] Jisan Mahmud, True Price, Akash Bapat, and Jan Michael Frahm. Boundary-aware 3d building reconstruction from a single overhead image. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2020.

[23] Diego Marcos, Devis Tuia, Benjamin Kellenberger, Lisa Zhang, Min Bai, Renjie Liao, and Raquel Urtasun. Learning deep structured active contours end-to-end. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 8877–8885, 2018.

[24] Sharada Prasanna Mohanty. Crowdai dataset: the mapping challenge. https://www.aicrowd.com/challenges/. 2018.

[25] Lichao Mou and Xiao Xiang Zhu. Im2height: Height estimation from single monocular imagery via fully residual convolutional-deconvolutional network. 2018.

[26] Ali Ozgun Ok, Caglar Senaras, and Baris Yuksel. Automated detection of arbitrarily shaped buildings in complex environments from monocular vhr optical satellite imagery. *IEEE Transactions on Geoence and Remote Sensing*, 51(3):1701–1717, 2013.

[27] Olaf Ronneberger, Philipp Fischer, and Thomas Brox. U-net: Convolutional networks for biomedical image segmentation. In *International Conference on Medical image computing and computer-assisted intervention*, pages 234–241. Springer, 2015.

[28] Ewelina Rupnik, Marc Pierrot-Deseilligny, and Arthur Delorme. 3d reconstruction from multi-view vhr-satellite images in micmac. *Isprs Journal of Photogrammetry Remote Sensing*, 139(MAY):201–211, 2018.

[29] Shivangi Srivastava, Michele Volpi, and Devis Tuia. Joint height estimation and semantic labeling of monocular aerial images with cnns. In *Igarss IEEE International Geoscience Remote Sensing Symposium*, 2017.

[30] Ke Sun, Bin Xiao, Dong Liu, and Jingdong Wang. Deep high-resolution representation learning for human pose estimation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2019.

[31] Vivek Verma, Rakesh Kumar, and Stephen Hsu. 3d building detection and modeling from aerial lidar data. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2006.

[32] Nicholas Weir, David Lindenbaum, Alexei Bastidas, Adam Van Etten, Sean McPherson, Jacob Shermeyer, Varun Kumar, and Hanlin Tang. Spacenet mvoi: a multi-view overhead imagery dataset. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 992–1001, 2019.

[33] Kang Zhao, Jungwon Kang, Jaewook Jung, and Gunho Sohn. Building extraction from satellite images using mask r-cnn with building boundary regularization. In *CVPR Workshops*, pages 247–251, 2018.

[34] Zhuo Zheng, Yanfei Zhong, and Junjue Wang. Pop-net: Encoder-dual decoder for semantic segmentation and single-view height estimation. In *IGARSS 2019-2019 IEEE International Geoscience and Remote Sensing Symposium*, pages 4963–4966. IEEE, 2019.

[35] Stefano Zorzi, Ksenia Bittner, and Friedrich Fraundorfer. Machine-learned regularization and polygonization of building segmentation masks. In *2020 IEEE International Conference of Pattern Recognition*, 2020.