Weakly Supervised 3-D Building Reconstruction From Monocular Remote Sensing Images

Weijia Li, Zhenghao Hu, Lingxuan Meng, Jinwang Wang, Juepeng Zheng, Runmin Dong, Conghui He, Gui-Song Xia, *Senior Member, IEEE*, Haohuan Fu, *Senior Member, IEEE*, and Dahua Lin, *Member, IEEE*

***Abstract*— 3-D building reconstruction from monocular remote sensing imagery is an important research problem that has been extensively studied for several decades. Although monocular remote sensing imagery is a more economic data source compared with the LiDAR data and multiview imagery, its limited information results in great challenges and restricts the performance of existing monocular reconstruction methods. Moreover, the expensive cost and the limited quantity of 3-D annotations also restrict the application scenes of existing methods, which are mostly based on fully supervised learning. In our previous work, we have proposed MTBR-Net, a monocular building reconstruction method that consists of a fully supervised multitask network and a postprocessing module for optimizing the reconstruction results. In this work, we further propose WS-MTBR-Net, a weakly supervised building reconstruction network that uses fewer 3-D annotations and achieves better performance in an end-to-end manner. Specifically, our WS-MTBR-Net fully leverages the relationship between different components of a 3-D building instance and the property of off-nadir images to improve the footprint segmentation boundary, based on six modified tasks and a new network structure with an improved feature warping module to support weakly supervised learning. We also design a new training strategy via a hybrid loss function that enables using the training samples with different annotation levels, i.e., complete 3-D annotations, 2-D footprint annotations, and image-level angle annotations. The results on the BONAI Shanghai and Xi'an test datasets demonstrate that our method achieves competitive performance when using 50% fewer 3-D-annotated samples, and improves the footprint segmentation $F1$-score by around 4% compared with current state-of-the-art.**

***Index Terms*— 3-D building reconstruction, high-resolution remote sensing images, multitask learning, weakly supervised learning.**

I. Introduction

As a FUNDAMENTAL task for large-scale city modeling, 3-D building reconstruction has been extensively studied for several decades. Although the aerial LiDAR data and multiview stereo imagery have become the primary data sources for many existing 3-D building reconstruction methods [1], [2], these data types are difficult to be used in large-scale building reconstruction due to the expensive cost, low frequency, limited coverage, and the requirement of multiple homologous images over the same area. The building reconstruction from monocular remote sensing imagery, on the contrary, provides a more economic solution for large-scale real-world applications. Meanwhile, the limited information of monocular remote sensing images and the diversity of building structures also result in great challenges for large-scale 3-D building reconstruction.

For 3-D building reconstruction from monocular remote sensing images, most recent methods are based on deep neural networks for building height regression [3], [4], [5], [6], [7], [8], which are inspired by the progress of supervised monocular depth estimation. Several studies aim at single-task height estimation from near-nadir images [3], [5], which take up a small proportion of the remote sensing images. For the building height estimation from monocular off-nadir images, some recent studies aim at learning the geocentric pose of buildings via using the additional information provided by the off-nadir images due to the parallax effect of roof and footprint [9], [10], or transfer deep learning models from a large-scale synthetic dataset to different real-world datasets in a few-shot cross-dataset setting [11]. However, these studies only focus on the single height estimation task instead of reconstructing the 3-D building model.

Manuscript received 30 September 2023; revised 15 January 2024; accepted 9 March 2024. Date of publication 18 March 2024; date of current version 29 March 2024. This work was supported in part by the National Natural Science Foundation of China under Grant T2125006, Grant 42201358, Grant 62325111, and Grant 42301390; and in part by the Jiangsu Innovation Capacity Building Program under Grant BM2022028. *(Corresponding authors: Juepeng Zheng; Haohuan Fu.)*

Weijia Li and Zhenghao Hu are with the School of Geospatial Engineering and Science and the Key Laboratory of Comprehensive Observation of Polar Environment, Ministry of Education, Sun Yat-sen University, Zhuhai 519082, China (e-mail: liweij29@mail.sysu.edu.cn; huzhh9@mail2.sysu.edu.cn).

Lingxuan Meng is with the School of Resources and Environment, Center for Information Geoscience, University of Electronic Science and Technology of China, Chengdu 611731, China (e-mail: xuanxuanling@std.uestc.edu.cn).

Jinwang Wang is with the School of Electronic Information, Wuhan University, Wuhan 430072, China (e-mail: jwwangchn@whu.edu.cn).

Juepeng Zheng is with the School of Artificial Intelligence, Sun Yat-sen University, Zhuhai 519082, China (e-mail: zhengjp8@mail.sysu.edu.cn).

Runmin Dong is with the Ministry of Education Key Laboratory for Earth System Modeling, Department of Earth System Science, Tsinghua University, Beijing 100084, China (e-mail: drm@mail.tsinghua.edu.cn).

Conghui He is with the Shanghai Artificial Intelligence Laboratory, Shanghai 200030, China, and also with SenseTime Research, Shenzhen 518038, China (e-mail: heconghui@pjlab.org.cn).

Gui-Song Xia is with the School of Computer Science, Wuhan University, Wuhan 430072, China (e-mail: guisong.xia@whu.edu.cn).

Haohuan Fu is with the Tsinghua Shenzhen International Graduate School, Tsinghua University, Shenzhen 518055, China, and also with the Ministry of Education Key Laboratory for Earth System Modeling and the Department of Earth System Science, Tsinghua University, Beijing 100084, China (e-mail: haohuan@tsinghua.edu.cn).

Dahua Lin is with the Department of Information Engineering, The Chinese University of Hong Kong, Hong Kong, China (e-mail: dhlin@ie.cuhk.edu.hk).

Digital Object Identifier 10.1109/TGRS.2024.3377694

Similar to the studies regarding joint depth estimation and semantic segmentation from monocular images, the building footprints or other types of semantic labels can also be used as extra useful information for height estimation, especially for the near-nadir images without the parallax effect. Consequently, the existing studies design multitask networks for joint footprint extraction and height estimation [6], [7], or exploit the semantic labels as the prior information for height estimation [4]. Although achieving better height estimation performance, these studies fail to explore the relationship between different components of a building instance (e.g., roof, footprint, and facade), and the relationship between the building heights and semantic types. Moreover, the existing monocular building reconstruction methods are designed for supervised learning, requiring a large number of expensive and fully annotated 3-D labels for network training.

However, due to the expensive annotation cost of the height information, the public available datasets for 3-D building reconstruction are still very insufficient, which also restricts the performance and application scenes of the existing 3-D building reconstruction methods. In Table I, we summarize the commonly used building datasets for footprint segmentation and 3-D reconstruction. ISPRS Potsdam and Vaihingen (denoted by ISPRS) [12] and Urban Semantic 3-D (denoted by US3D) [13] are the most popular datasets used in recent building reconstruction studies [4], [5], [6], [7], in which most of the images are near-nadir (with nearly overlapped roof and footprint). The DFC19 (extended from US3D) and ATL-SN4 are two datasets for monocular building height estimation proposed by [9], containing off-nadir images with a wider range of oblique viewing angles.

As shown in Table I, owing to the low annotation cost and the increase in free geographic data (e.g., Open Street Map), the public building footprint datasets have an extremely larger coverage and quantity compared with the building reconstruction datasets with 3-D annotations. These large-scale 2-D footprint datasets can provide new opportunities for improving the 3-D building reconstruction performance and have the potential of reducing the 3-D annotation requirement if they are effectively used by semi-supervised or weakly supervised methods.

In our previous work, we have proposed MTBR-Net [14], a multitask building reconstruction network that is trained in a fully supervised manner. In this work, we further propose WS-MTBR-Net, a weakly supervised building reconstruction network for monocular building reconstruction using fewer 3-D labels. Different from the existing methods that design a multitask network with a shared feature map and trained with fully annotated 3-D labels, the architecture design of our WS-MTBR-Net is based on the relationship between the main components of each 3-D building instance (roof, footprint, and the height) and the property of the off-nadir remote sensing images, with different head structures and feature maps designed for different prediction tasks. Moreover, it is a unified framework that is capable of using the training samples with different annotation levels (i.e., complete 3-D annotations, 2-D footprint annotations, and image-level angle annotations), owing to our proposed network architecture and a hybrid loss function. The results on the BONAI Shanghai and Xi'an test datasets [15] demonstrate that our method achieves competitive performance when only using a half 3-D labels of the state-of-the-art methods and improves the building segmentation $F1$-score of the current state-of-the-art by around 4%.

TABLE I

Summary of Several Popular Datasets for Building Footprint Segmentation and 3-D Reconstruction. The Public Building Footprint Datasets Have an Extremely Larger Coverage and Quantity Compared With the Public 3-D Building Reconstruction Datasets

Dataset	# Images	# Instances	Off-Nadir	Annotation
ARIS [16]	-	220,000	No	Roof
Microsoft Global [17]	-	1,240,000,000	No	Footprint
Open Buildings [18]	-	1,800,000,000	No	Footprint
CrowdAI [19]	340,000	2,915,000	No	Footprint
WHU [20]	8,189	120,000	No	Footprint
SpaceNet [21]	24,586	302,701	No	Footprint
MVOI [22]	60,000	126,747	Yes	Footprint
DFC19 [9]	3,200	500,000	Yes	3D
ATL-SN4 [9]	8,000	1,100,000	Yes	3D
BONAI [15]	3,300	268,958	Yes	3D
ISPRS [12]	33	-	No	3D

Our main contributions are summarized as follows.

1) We design WS-MTBR-Net, a weakly supervised building reconstruction network that fully explores the relationship between the main components of a 3-D building instance, based on six modified tasks and a new network structure with an improved feature warping module, achieving superior footprint segmentation and height estimation performance compared with the current state-of-the-art methods.
2) We propose a new training strategy via a hybrid loss function that enables the training of WS-MTBR-Net with different supervision levels, which further reduces the demand on large-scale training samples with expensive 3-D annotations compared with the existing supervised building reconstruction methods.
3) We conduct comprehensive experiments under different settings using: 1) entire 3-D-annotated samples; 2) partial 2-D-annotated samples + partial 3-D-annotated labels; and 3) entire 3-D-annotated samples + extra 2-D-annotated samples. The results show that our method achieves competitive performance using fewer 3-D labels, and significantly better performance using the same training set or extra 2-D labels compared with the current state-of-the-art methods.

II. Related Work

A. Building Footprint Extraction

Building footprint extraction from satellite or aerial images is a crucial prerequisite for 3-D building reconstruction, which has been broadly studied for decades. In recent years, deep neural networks have become the state-of-the-art methods for building extraction [23], [24], [25], which can be divided into three categories, i.e., the pixelwise segmentation methods, the corner-based methods, and the boundary-based methods.

For building segmentation tasks, instance and semantic segmentation networks have been broadly explored [26],

[27], [28] and achieved outstanding performances in many building extraction challenges, such as CrowdAI [19], DeepGlobe [21], and SpaceNet series [29]. Many studies use multitask segmentation network to improve the building segmentation performance. In [30], a multitask learning method was proposed to improve the building boundary prediction performance, which introduced an extra task to predict the distance to the border of buildings using an encoder–decoder network architecture. Yuan [31] proposed the signed distance representation for building footprint extraction which achieves better performance than the single-task fully connected network. Similarly, in [5], a modified signed distance function was introduced and jointly learned with other tasks for predicting both building footprint outlines and building heights. To reduce the demand on labeled datasets, several studies propose self-supervised or semi-supervised learning strategies for building footprint extraction [28], [32], [33]. In addition, some building footprint extraction methods are based on the building corners. These methods directly predict the vertices of a building polygon via a CNN-RNN [34], [35] or transformer [36] architecture, or combine the pixel-based multitask segmentation network with a graph-based polygon refinement network using a rule-based module [37], [38].

Moreover, several other studies propose boundary-based methods to combine the traditional active contour models with deep neural networks to improve the segmentation boundaries [39], [40], [41], which are mostly designed for single building extraction, i.e., the input images have been cropped by the ground-truth bounding boxes.

However, the existing methods of all three categories perform worse for extracting building footprints from off-nadir images, which constitute the main proportion in actual scenes. Especially for high-rise buildings, the existing methods produce poor footprint segmentation boundaries which are partially invisible on the off-nadir images. The existing methods directly predict the building footprint from the feature map extracted from the initial remote sensing images via a deep neural network. Our method, on the contrary, predicts the building footprint from a novel reconstructed feature map that is obtained by warping the feature map of building roof using the predicted offset vector (from roof to footprint), which not only significantly improves the footprint segmentation performance for off-nadir images but also enables the weakly supervised learning strategy for 3-D building reconstruction.

B. Monocular 3-D Building Reconstruction

There is an increasing number of studies for 3-D building reconstruction from monocular remote sensing images, owing to the inexpensive data acquisition costs and broad data coverage compared with reconstruction from LiDAR data [1] or multiview imagery [2], [42], [43], [44]. Traditional monocular 3-D building reconstruction methods are mostly based on the shadow information, lines or line intersections of the building outlines, and the meta information of satellites such as the sun–earth relative position [45], [46]. Complicated procedures with multiple steps are required for reconstructing the final 3-D building model.

Inspired by the progress of monocular depth estimation, the deep neural network has been used for monocular building height estimation in several recent studies [11], [47], [48]. Most of these studies are designed for height estimation from near-nadir images, in which the building roof and footprint are almost overlapped. Some methods use an encoder–decoder network to regress the height values [8], or use a generative adversarial network to simulate a height map [3]. Moreover, considering the limited information provided from the near-nadir images for height estimation, the semantic labels have been used as useful extra information in many existing studies. Some studies design a multitask network for joint footprint extraction and height estimation [6], [7], [47], while other studies exploit the semantic labels as prior information for height estimation [4].

In actual scenes, the off-nadir images constitute the most proportion of the remote sensing data, in which the parallax effect of roof and footprint results in more challenges for extracting the footprint outlines but provides additional useful information for building height estimation in the meantime. For off-nadir images, some recent studies [9], [10] proposed a monocular height estimation method via learning the geocentric pose of buildings (i.e., an imagewise flow angle and a pixelwise magnitude value) using a U-Net architecture [49]. However, these studies only focus on the height estimation task instead of reconstructing the 3-D building model.

In summary, the monocular building reconstruction methods in the existing studies require expensive and fully annotated 3-D labels for supervised learning, which are either designed for 3-D building reconstruction from near-nadir images or building height estimation from off-nadir images. In contrast, our proposed approach is a unified framework for weakly supervised 3-D building reconstruction with a new network architecture and six modified tasks, which leverages the relationship between building roofs and footprints on off-nadir images to enable 3-D building reconstruction with different supervision levels to reduce the demand for large-scale 3-D annotations.

C. Weakly Supervised and Semi-Supervised 3-D Reconstruction

Unlike the rapid progress of supervised 3-D reconstruction methods, the weakly supervised and semi-supervised 3-D reconstruction studies are still at an early stage [50], [51], [52], [53], [54], [55], [56], [57]. For weakly supervised 3-D reconstruction, Neverova et al. [52] introduced an intermediate representation that is defined as a segmentation of the hand into multiple parts, which contains important topological and structural information to enable the weakly supervised training for hand pose estimation. In Gwak et al. [55], a weak 2-D supervision type, i.e., the foreground mask, is effectively used as an alternative for the expensive 3-D CAD annotation via a raytrace pooling layer to enable the perspective projection and backpropagation. For semi-supervised methods, adversarial learning has been widely used in several semi-supervised 3-D reconstruction studies. For instance, Yang et al. [53] proposed a unified framework that combines two types of supervision,

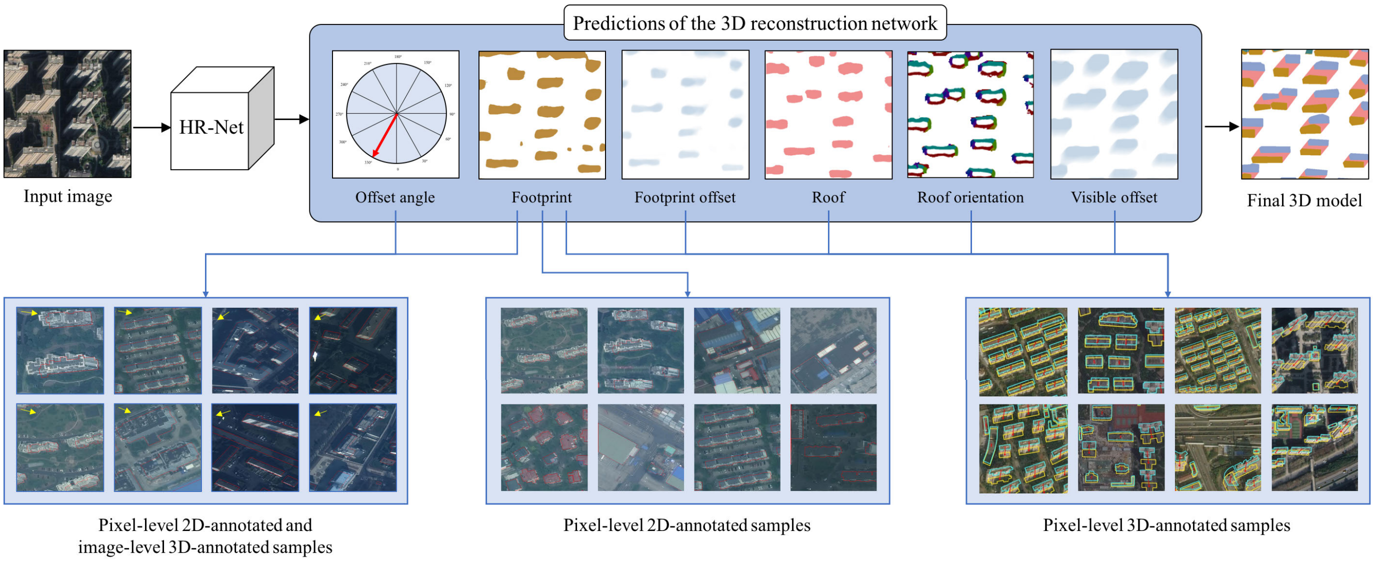
Predictions of the 3D reconstruction network
Input image
HR-Net
Offset angle
Footprint
Footprint offset
Roof
Roof orientation
Visible offset
Final 3D model
Pixel-level 2D-annotated and image-level 3D-annotated samples
Pixel-level 2D-annotated samples
Pixel-level 3D-annotated samples

Fig. 1. Overall framework of our WS-MTBR-Net. We convert the 3-D building reconstruction into six relevant tasks that are learned via an HR-Net-based multitask network, i.e., roof segmentation, roof boundary orientation prediction, visible part offset prediction, footprint offset prediction, footprint segmentation, and image-level offset angle prediction. To enable the training of WS-MTBR-Net with different supervision levels, we define a hybrid loss function that combines the losses calculated from three types of training samples. The network outputs are integrated into the final 3-D model based on a vectorization method.

i.e., a small number of cameras pose annotations to enforce the consistency of pose invariance and view point, and a large number of unlabeled images to enforce the realism of rendered 3-D shapes via an adversarial loss. In Ji et al. [54], a semi-supervised adversarial learning framework was proposed for monocular depth estimation, which only uses a small number of image–depth pairs and a large number of easily available monocular images to achieve high performance.

Different from the existing methods mentioned above, our proposed approach is designed based on the prior knowledge regarding the 3-D structure of a building instance on monocular remote sensing images. Specifically, our approach effectively uses the intrinsic relationship between the roof, footprint, and the height of each building instance, and the geometry property of the off-nadir remote sensing images, which significantly reduces the requirement for the expensive 3-D building annotations and makes full use of the large-scale and easily available 2-D footprint annotations. To the best of our knowledge, this is the first work for monocular 3-D building reconstruction with different supervision levels.

III. Methods

The overall framework of our WS-MTBR-Net is demonstrated in Fig. 1. Considering the 3-D structure of building instances on the monocular remote sensing image, we convert the 3-D building reconstruction into six relevant tasks that are learned via an HR-Net-based multitask network. Our WS-MTBR-Net includes three visible part prediction tasks (i.e., roof segmentation, roof boundary orientation prediction, and visible part offset prediction), two footprint-related prediction tasks (i.e., footprint offset prediction and footprint segmentation), and an image-level offset angle prediction task. To enable the training of WS-MTBR-Net with different supervision levels, we define a hybrid loss function that combines the losses calculated from three types of training samples. For the training samples with complete 3-D annotations, the loss function is calculated from all the visible part prediction tasks and footprint-related prediction tasks. For the samples with 2-D footprint and image-level offset angle annotations, the loss function is calculated from the footprint segmentation task and the image-level offset angle prediction task. For the samples with only 2-D footprint annotations, the loss function equals the footprint segmentation loss.

Compared with our previous work [14], WS-MTBR-Net makes improvements in terms of multitask definition, network architecture and the design of loss function. First, the roof/facade segmentation and skeleton orientation prediction tasks are replaced by roof segmentation and roof boundary orientation prediction tasks, which enables the effectiveness of our improved feature warping module under the premise that the roof and footprint of a building usually have the same contour shape. We also remove the skeleton segmentation task since WS-MTBR-Net is an end-to-end method that does not require the height vector optimization strategy used in [14]. Second, different network head structures and feature maps are designed for different prediction tasks, including an improved feature warping module for footprint segmentation under the weakly supervised conditions. Moreover, owing to our network architecture and hybrid loss function, the training of WS-MTBR-Net can be performed under different supervision levels.

In the following sections, we first make an analysis of the 3-D building structure on the off-nadir images and introduce the definitions of the six relevant tasks in our WS-MTBR-Net. Then we introduce the overall architecture and six task-specific heads of our WS-MTBR-Net. Next, we introduce the training of our WS-MTBR-Net, including the loss function of the six tasks under three levels of supervisions and the total loss. The implementation details and experimental settings will be introduced in Section IV.

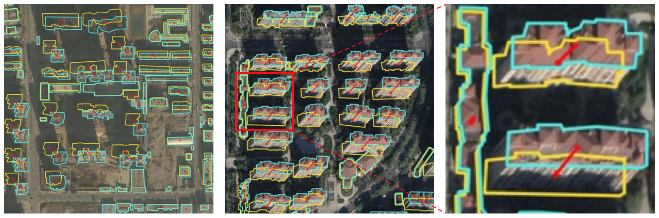

Fig. 2. Examples of the buildings on the off-nadir images. The contours of the building roofs are entirely visible on the monocular off-nadir images. The contours of the building footprints are partially invisible but have the same shape as the roof contour and can be obtained from moving the roof contour in the direction of an offset vector (denoted by the red arrows).

A. 3-D Structure Analysis of Building Instances

We first analyze the 3-D structure of building instances. Figure 2 shows some example of the buildings on the off-nadir images. The contours of the building roofs are entirely visible on the monocular off-nadir images. The contours of the building footprints, on the contrary, are partially invisible due to the parallax effects of the off-nadir images. Fortunately, for each building instance, the footprint contour often shares the same shape as the roof contour, which can be obtained from moving the roof contour in the direction of an offset vector (denoted by the red arrows in Fig. 2). Thus, our method is based on the premise that the building polygons of the roof and footprint are only under affine transformation, as the remote sensing images used in our method are rectified to have minimal shape deformation. Specifically, there are only translation transformations and little or no shape changes between the roof and footprint polygons. The length of offset vector (the relative height, denoted by H_r) can be converted into the actual building height (H_a) according to (1), in which α denotes the nadir angle ($\alpha \neq 0$) and R denotes the spatial resolution of the image

$$H_a = H_r \times R / \tan(\alpha). \tag{1}$$

In addition, the buildings on the same image often have the same offset angle (i.e., the red arrows of the same image are in parallel). Based on the above analysis, in our WS-MTBR-Net, we define six relevant tasks to enable the accurate reconstruction of buildings with fewer 3-D labels, including three visible part prediction tasks, two footprint-related prediction tasks, and an image-level angle prediction task.

We first design a roof segmentation task for the visible part prediction. To convert the raster results into vector 3-D model, we design an auxiliary task to predict the edge orientation of the roof boundary following [38]. Moreover, to estimate the height of each building, we design another task to predict the offset vector for the visible parts of each building, i.e., the complete roof regions and the visible facade regions. For each building instance, the offset vectors of the roof region are assigned as the same values, i.e., the vector from roof to footprint (denoted by $[O_x^r, O_y^r]$). For the pixels within the visible facade region, the offset vectors are assigned as $[\delta_x, \delta_y]$, i.e., the vector from the specific pixel to the corresponding pixel on the footprint contour, which has the same direction angle as vector $[O_x^r, O_y^r]$. The offset vectors of the background regions are assigned as $[0, 0]$.

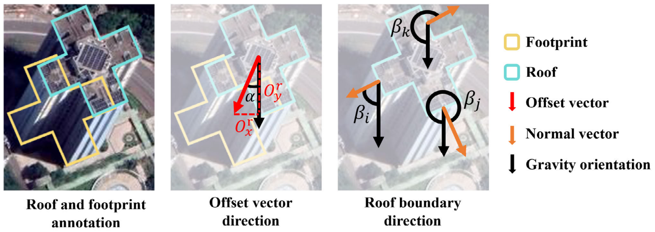

Fig. 3. Representation of two types of rotation directions. The left image shows the annotation of roof and footprint polygons. The middle and right images show the definitions of offset vector direction (denoted by α) and roof boundary direction (denoted by β), respectively.

Fig. 3 illustrates the definitions of the two types of rotation directions used in this work, i.e., offset vector direction and roof boundary direction. In terms of the offset vector direction (denoted by α), the accurate direction values were obtained from the accurate roof and footprint polygons, which were manually annotated during the dataset construction process [14], [15]. For each building, the offset vector is defined as the offset between the roof and footprint polygons in both the horizontal and vertical directions (i.e., $[O_x^r, O_y^r]$). For the roof boundary direction (denoted by β), the accurate direction value of each pixel on an edge is defined as the angle between its normal vector and the gravity orientation in a counterclockwise direction [38].

As mentioned above, the footprint contours are often partially invisible but have the same shape as the roof contour. Thus, we design a footprint offset vector prediction task for warping the feature map of the roof segmentation task, which enables our WS-MTBR-Net to reconstruct the 3-D building model with fewer 3-D labels and use the large-scale footprint annotations to improve the footprint segmentation performance. For each building instance, the offset vectors of the footprint regions are assigned as the same value, i.e., the offset vector of the corresponding roof regions $[O_x^r, O_y^r]$. The offset vectors of other regions are assigned as $[0, 0]$.

Similar to our previous work [14], we design an imagewise offset angle prediction task to provide additional supervisions for the samples without pixelwise 3-D annotations. For the monocular remote sensing images, the annotation of image-level offset angle requires much lower cost and efforts compared with the pixel-level height or offset vector annotations. The six tasks are jointly learned via a unified framework with samples of different annotation levels, which will be introduced in the following sections.

B. Network Architecture

Our WS-MTBR-Net is based on a modified HR-Net architecture [58], which is capable of maintaining high-resolution representations throughout the whole process and beneficial for remote sensing image analysis. After the four stages of the original HR-Net architecture, we obtained four feature maps of different resolutions. The number of channels of the four feature maps are C, $2C$, $4C$, and $8C$ (C is set as 12 in

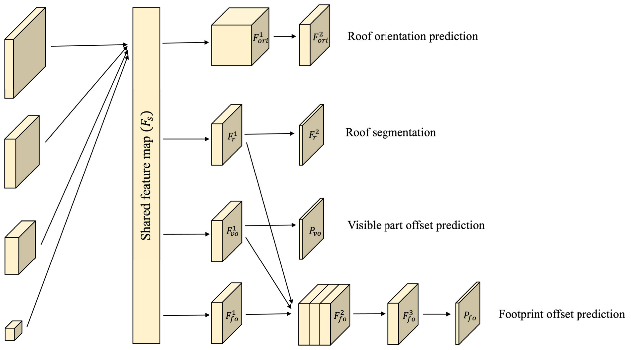
Shared feature map (F_s)
F^1_{ori}
F^2_{ori}
Roof orientation prediction
F^1_r
F^2_r
Roof segmentation
F^1_{vo}
P_{vo}
Visible part offset prediction
F^1_{fo}
F^2_{fo}
F^3_{fo}
P_{fo}
Footprint offset prediction

Fig. 4. Network structure of visible component prediction heads and footprint offset prediction head of our WS-MTBR-Net.

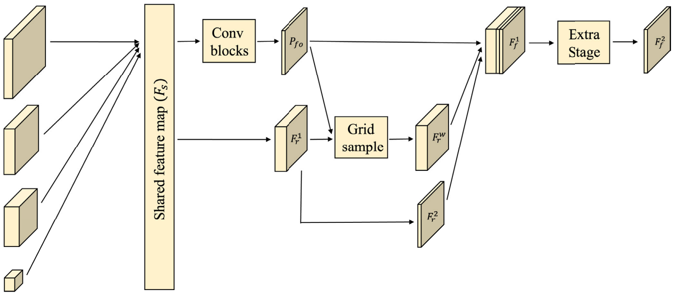

Fig. 5. Network structure of footprint segmentation head of our WS-MTBR-Net.

our experiments following [14]). The four feature maps are resampled to the same size and concatenated into a shared feature map of 15C channels (denoted by F_s). Then we design six task-specific heads with different structures according to the characteristic of each task. The details of each task-specific head of our WS-MTBR-Net are introduced as follows.

1) Visible Components Prediction Heads: For the three tasks regarding visible components prediction (i.e., roof segmentation, roof boundary orientation prediction, and visible part offset prediction), we design two segmentation heads and one regression head based on the original feature map. As shown in Figure 4, we first apply a 1 × 1 convolution to F_s for extracting the task-specific feature map for roof segmentation and roof boundary orientation prediction (denoted by F^1_r and F^1_{ori}). Then we apply another 1 × 1 convolution to F^1_r and F^1_{ori} and obtain the final feature map with the channel number equal to the class number (denoted by F^2_r and F^2_{ori}). Similarly, we first apply a 1 × 1 convolution to F_s for extracting the task-specific feature map for the visible part offset prediction task (denoted by F^1_{vo}). Then we apply another 1 × 1 convolution to F^1_{vo} and obtain the final prediction map with two channels (denoted by P_{vo}), indicating the predicted offset values in x and y directions of each pixel.

2) Footprint Offset Prediction Head: For footprint offset prediction, we design a regression head based on the concatenated original feature map of three relevant tasks. As shown in Fig. 4, we first apply a 1 × 1 convolution to F_s for extracting the task-specific feature map for footprint offset prediction (denoted by F^1_{fo}). Then we concatenate F^1_{fo} with the feature maps of roof segmentation and visible part offset prediction (F^1_r and F^1_{vo}). In our experiment, the numbers of channels for the above three feature maps (n_c) are all set as 1/4 of the channel number of F_s following [14], constituting the concatenated feature map (denoted by F^2_{fo}) of 135 channels. Finally, we apply two 1 × 1 convolutions to F^2_{fo} and obtain the final prediction map of two channels (denoted by P_{fo}).

3) Footprint Segmentation Head: Motivated by the intrinsic relationship between roof, offset, and footprint, we design a footprint segmentation head based on an improved feature warping module, of which the feature map of footprint segmentation is constructed from those of roof segmentation and footprint offset prediction. Our proposed footprint segmentation head not only improves the footprint segmentation performance compared with directly learning from the original feature map but also enables the weakly supervised 3-D building reconstruction from a large number of 2-D footprint annotation. Fig. 5 shows the structure of the footprint segmentation head based on the warped feature map. First, the feature map of the first 1 × 1 convolution of the roof segmentation task (F^1_r) is warped by the footprint offset prediction map P_{fo} (obtained via the process shown in Fig. 4, denoted by Conv blocks in Fig. 5), which is implemented using the grid_sample function provided by PyTorch [59]. The warped feature map (denoted by F^w_r) usually has clear roof contours at the footprint locations, which is beneficial for footprint segmentation due to the similarity of the roof and footprint contour. Then we concatenate F^w_r with the footprint offset prediction map (P_{fo}) and the final feature map of the roof segmentation task (F^2_r), constituting the warped feature map for footprint segmentation (denoted by F^1_f). Considering the difficulties in footprint segmentation, we design an extra stage to enable footprint segmentation from deeper feature maps. We resample F^1_f into four feature maps with different resolutions and apply the same operations as the former stage of HR-Net. Similarly, the four output feature maps are resampled to the same size and concatenated into the final feature map for footprint segmentation via 1 × 1 convolutions.

4) Offset Angle Prediction Head: To effectively use the image-level offset angle supervision, we design an offset angle prediction head based on the image-level feature vector, which is formulated as a classification problem to simplify the training process. The class definition and network structure of the offset angle prediction head is the same as those used in [14], i.e., based on the official structure and hyperparameter setting of the classification head introduced in [60]. Specifically, the four resolution feature maps are fed into a bottleneck and the output channels are increased from 12, 24, 48, and 96 to 128, 256, 512, and 1024, respectively. Then the high-resolution representation is downsampled by a two-strided 3 × 3 convolution with 256 output channels and added to the representation of the second high resolution. This process is repeated for two times to get a small resolution feature map of 1024 channels. Finally, the feature map is transformed from 1024 channels to 2048 channels via a 1 × 1 convolution followed by a global average pooling operation. The final 2048-D representation is fed into the classifier for offset angle prediction with 37 classes (36 angle classes and one unsure class).

C. Network Training

In this section, we first introduce the loss function of the six tasks in our WS-MTBR-Net. Then we introduce the loss function for our training samples with three levels of supervisions, i.e, samples with complete pixelwise 3-D supervision, samples with only pixel-wise 2-D footprint supervision, and samples with pixelwise 2-D footprint supervision and imagewise offset supervision. The total loss of our WS-MTBR-Net is introduced at the end of this section.

We formulate the roof segmentation, footprint segmentation, and roof contour orientation prediction tasks as pixelwise semantic segmentation problems. The loss of the three tasks ($\mathcal{L}_r$, $\mathcal{L}_f$, and $\mathcal{L}_{\text{ori}}$; uniformly denoted by $\mathcal{L}_{\text{seg}}$) is calculated according to (2), in which N denotes the pixel number of an image; C denotes the class number; and $y_{i,c}$ and $p(y_{i,c})$ denote the binary indicator and the predicted probability that pixel i belongs to class c, respectively,

$$\mathcal{L}_{\text{seg}} = -\frac{1}{N}\sum_{i=1}^{N}\sum_{c=1}^{C} y_{i,c} \times \log\left(p\left(y_{i,c}\right)\right). \tag{2}$$

We formulate the visible part offset prediction and the footprint offset prediction as pixelwise regression problems. The loss of the two offset regression tasks ($\mathcal{L}_{vo}$ and $\mathcal{L}_{fo}$; uniformly denoted by $\mathcal{L}_{\text{off}}$) is calculated by the endpoint error according to (3), in which $\vec{O}_i^{\text{pred}}$ denotes the predicted offset vector, i.e., $[O_{x,i}^{\text{pred}}, O_{y,i}^{\text{pred}}]$; $\vec{O}_i^{gt}$ denotes the corresponding ground-truth offset vector, i.e., $[O_{x,i}^{gt}, O_{y,i}^{gt}]$

$$\mathcal{L}_{\text{off}} = \frac{1}{N}\sum_{i=1}^{N}\left|\left|\vec{O}_i^{\text{pred}} - \vec{O}_i^{gt}\right|\right|_2. \tag{3}$$

We formulate the imagewise offset angle prediction as a classification problem for simplifying the training process. The loss of angle prediction task $\mathcal{L}_{\text{ang}}$ is calculated by (4), where K denotes the class number of angle prediction; y_k and $p(y_k)$ denote the binary indicator and the predicted probability for class k, respectively,

$$\mathcal{L}_{\text{ang}} = -\sum_{k=1}^{K} y_k \times \log(p(y_k)). \tag{4}$$

In our proposed unified framework, the training samples can be divided into three categories according to the supervision level: 1) samples with full pixelwise 3-D supervision, which are denoted by $\mathcal{X}^F = \{x_1^F, x_2^F, \ldots, x_{n1}^F\}$; 2) samples with partial 3-D supervision (i.e., only imagewise offset supervision), which are denoted by $\mathcal{X}^P = \{x_1^P, x_2^P, \ldots, x_{n2}^P\}$; and 3) samples with no 3-D supervision, which are denoted by $\mathcal{X}^N = \{x_1^N, x_2^N, \ldots, x_{n3}^N\}$. All the training samples of the three categories are provided with the pixelwise 2-D footprint annotations.

The loss function for the samples in $\mathcal{X}^F$ (denoted by $\mathcal{L}_{\mathcal{X}^F}$) is defined as the sum of all the segmentation losses and pixelwise offset prediction losses, which can be calculated according to the following equation:

$$\mathcal{L}_{\mathcal{X}^F} = \alpha_1\mathcal{L}_r + \alpha_2\mathcal{L}_f + \mathcal{L}_{\text{ori}} + \mathcal{L}_{vo} + \mathcal{L}_{fo}. \tag{5}$$

For the samples in $\mathcal{X}^P$, as only the 2-D footprint and imagewise offset angle annotations are available, the loss function (denoted by $\mathcal{L}_{\mathcal{X}^P}$) is defined as the sum of footprint segmentation loss and offset angle classification loss according to the following equation:

$$\mathcal{L}_{\mathcal{X}^P} = \mathcal{L}_f + \mathcal{L}_{\text{ang}}. \tag{6}$$

Similarly, as the samples in $\mathcal{X}^N$ only have the 2-D footprint annotations, the loss function (denoted by $\mathcal{L}_{\mathcal{X}^N}$) is defined as the footprint segmentation loss according to the following equation:

$$\mathcal{L}_{\mathcal{X}^N} = \mathcal{L}_f. \tag{7}$$

The final hybrid loss is defined as the total loss of the three categories of training samples according to the following equation:

$$\mathcal{L} = \mathcal{L}_{\mathcal{X}^F} + \mathcal{L}_{\mathcal{X}^P} + \mathcal{L}_{\mathcal{X}^N}. \tag{8}$$

D. Vectorization of the 3-D Models

After obtaining the outputs of the WS-MTBR-Net, we apply a simple but effective vectorization method to integrate the network outputs into the final 3-D building model, which is constituted by a roof polygon, a footprint polygon, and an instancewise height vector. Different from our previous work [14] that uses a height vector optimization strategy based on multiple types of network outputs, we simply calculate the average offset field values in the roof region as the instancewise offset vector for each building, which can be converted to the actual height according to Section III-A.

The process of constructing roof and footprint polygons is similar to the one used in our previous work [14]. Specifically, based on the prediction of the roof boundary orientation, the raster roof segments can be vectorized into polygons with valid shapes. For each densely sampled vertex of the roof segment contour, if the absolute difference in orientation class between the vertex and its neighbor vertex is greater than three, the vertex will be selected as valid and remained. The remaining valid vertices constitute the simplified roof polygon, which will be warped as the footprint polygon based on the offset vector. The simplified roof polygon, footprint polygon, and height vector comprise the final vector 3-D building model.

IV. Experimental Results Analysis

A. Datasets

In our experiments, we evaluate the 3-D building reconstruction results in terms of the height estimation performance and the footprint segmentation performance on the BONAI dataset proposed in our previous work [15], which provides holistic 3-D building annotations for both footprint segmentation and height estimation. We also analyze the effect of using additional footprint segmentation datasets as the extra training dataset on improving the footprint segmentation performance. The details are introduced as follows.

1) Our Previously Proposed BONAI Dataset [15]: We evaluate the proposed method on the BONAI dataset [15] for 3-D building reconstruction from monocular remote sensing images. The dataset includes over 200 000 manually annotated building footprints and the corresponding height vectors, which solves the limitations of the existing datasets and can be used for the evaluation of both the aspects. The images are collected from different data sources (e.g., Google Earth and Microsoft Virtual Earth) with a diversity of view angles. Our BONAI dataset [15] contains 2700 training images, 300 validation images and 300 test images, which are in a size of 1024 × 1024 pixels. The training and validation images of BONAI are collected from five cities of China, i.e., Shanghai, Beijing, Harbin, Jinan, and Chengdu. Moreover, BONAI contains two test datasets: 1) BONAI Shanghai dataset (the in-domain dataset in [14]), which contains 200 images located in the same city but different regions with the training images; and 2) BONAI Xi'an dataset (the out-domain dataset in [14]), which contains 100 images located in a new city that is not included in the training images. The whole dataset can be downloaded from https://github.com/jwwangchn/BONAI.

2) Extra Building Footprint Datasets: Besides the BONAI dataset [15], we use some additional public building footprint datasets to analyze its effect on improving the footprint segmentation performance, including the SpaceNet DeepGlobe dataset [21], the Microsoft Global building dataset [17], and WHU building dataset [20]. As the Microsoft Global building dataset only provides the footprint annotations without the original imagery, we download the Google Earth imagery corresponding to the Salt Lake City area following [38] and crop the image to obtain the training samples of this dataset. The above three datasets constitute the extra 2-D-annotated dataset (denoted by $\mathcal{X}^E$).

B. Evaluation Metrics

We evaluate the 3-D building reconstruction results from two aspects, i.e., the offset vector prediction performance and the footprint segmentation performance. The offset vector prediction performance is evaluated in terms of the endpoint error (denoted by EPE, in pixels) [10], [14], [15], i.e., the Euclidean distance between the endpoints of the predicted offset vector and ground-truth offset vector. We also report the EPE of the building instances within different offset length ranges and the average EPE of all the instances (in pixels). For the footprint segmentation performance, we use the instance-level evaluation metrics to evaluate the footprint segmentation results. Specifically, we calculate the precision, recall, and $F1$-score under the IoU threshold of 0.5, which have been widely used in previous building segmentation studies and challenges [19], [21], [38]. Furthermore, the size and computation complexity of our model are, respectively, evaluated in terms of the number of parameters and the FLoating point OPerations (FLOPs).

C. Experimental Settings and Comparison Methods

In the training process of our WS-MTBR-Net, the original images of all the datasets are randomly cropped into 500 × 500 pixels due to memory limitation. For $\mathcal{L}_{\mathcal{X}^F}$ calculation, we set higher weights for the roof segmentation and footprint segmentation tasks following [14], i.e., α_1 and α_2 are set as 3. We use random rotation, cropping, scaling, flipping, and Gaussian blur for data augmentation. We train the WS-MTBR-Net on 16 NVIDIA Titan Xp GPUs using stochastic gradient descent (SGD) as the optimizer, with a batch size of 16 for 2000 epochs, an initial learning rate of 0.01, a momentum of 0.9, and a weight decay of 10^{-4}. Under the above settings, the FLOPs of our model is 192.12 G and the number of parameters is 26.62 M.

For evaluating the building height estimation results, we provide a thorough comparison between our method and the state-of-the-art methods designed for pixelwise offset estimation from monocular off-nadir images, including Christie et al. (CVPR 2020) [9] and MTBR-Net (ICCV 2021) [14]. For Christie et al. [9], we replace the flow vector prediction task with the visible part offset prediction task and replace the U-Net with the HR-Net architecture for a fair comparison with our method following [14]. For MTBR-Net [14], we remove the height vector optimization strategy (w/o optimization in [14]) to guarantee a fair comparison of the MTBR-Net and WS-MTBR-Net architectures in terms of the offset prediction performance.

For evaluating the building footprint segmentation results, we compare our results with those of three instance segmentation methods (Mask R-CNN [61], Cascade Mask R-CNN [62], and PANet [63]) and the HR-Net-based semantic segmentation method [58]. We use ResNet-50 [64] pretrained on the ImageNet [65] with FPN [66] as the backbone of the three instance segmentation methods. All three methods are trained with a batch size of 32 on 16 NVIDIA Titan Xp GPUs for 24 epochs and a learning rate starting from 0.02 and decreasing by a factor of 0.1 at the 16th and 22nd epoch, using SGD with a weight decay of 0.0001 and a momentum of 0.9 as the optimizer, which are implemented based on mmdetection [67] with the default data augmentation and recommended hyperparameter settings. We use the same data augmentation, hyperparameter setting, and backbone architecture as our approach for the HR-Net-based semantic segmentation method [58]. We also compare our method with the two state-of-the-art methods for building footprint extraction on the BONAI datasets, i.e., MTBR-Net (ICCV 2021) [14] and LOFT (TPAMI 2022) [15]. For MTBR-Net [14], we use the same experimental settings as those used for height estimation evaluation. For LOFT [15], we use the default experimental settings and models provided by the authors to evaluate the footprint segmentation performance.

D. Building Height Estimation

In this section, we compare the height estimation performance of our method with two state-of-the-art methods for pixelwise offset estimation on the BONAI Shanghai and Xi'an test datasets. Table II lists the EPE of the roof instances in different height ranges (the offset vector length, in pixels) and the average EPE of all the instances obtained from different methods. We report the height estimation metrics of our method

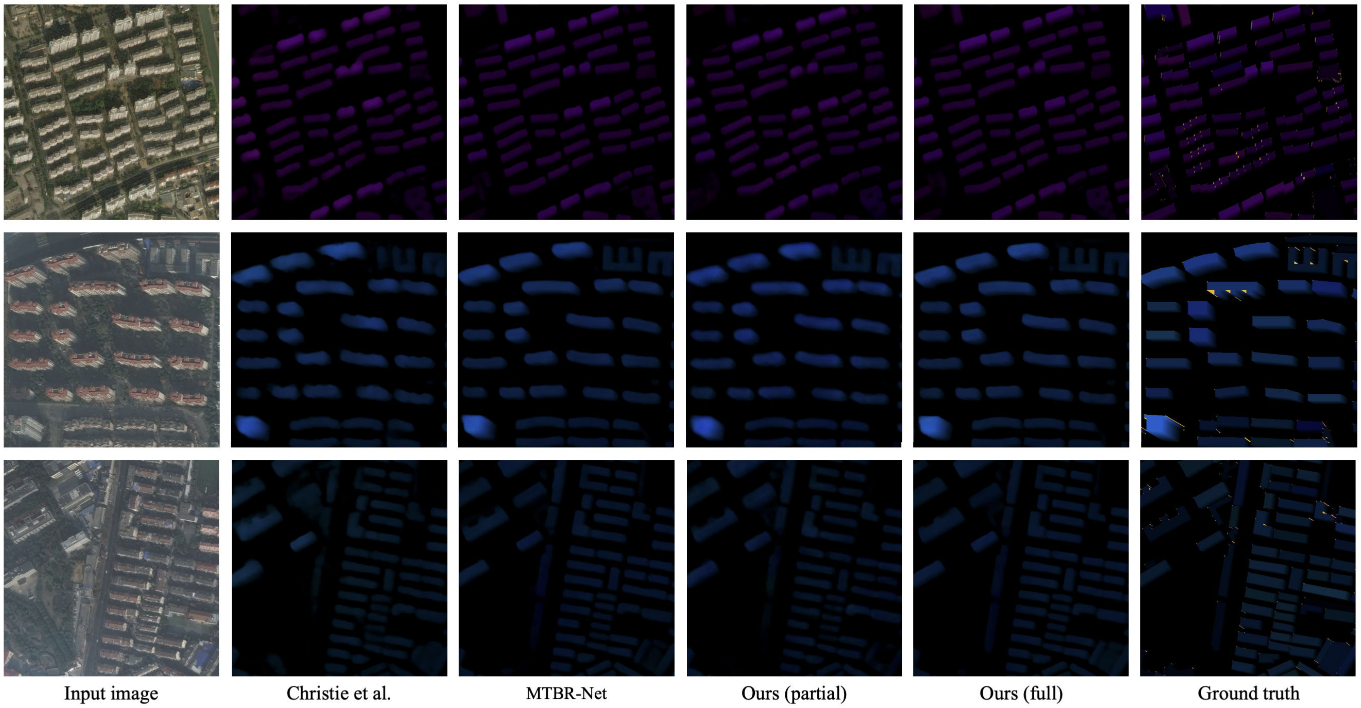

Fig. 6. Examples of height estimation results of different methods on the BONAI Shanghai test dataset. Different colors represent different offset angles. The brightness of each color reflects the offset length.

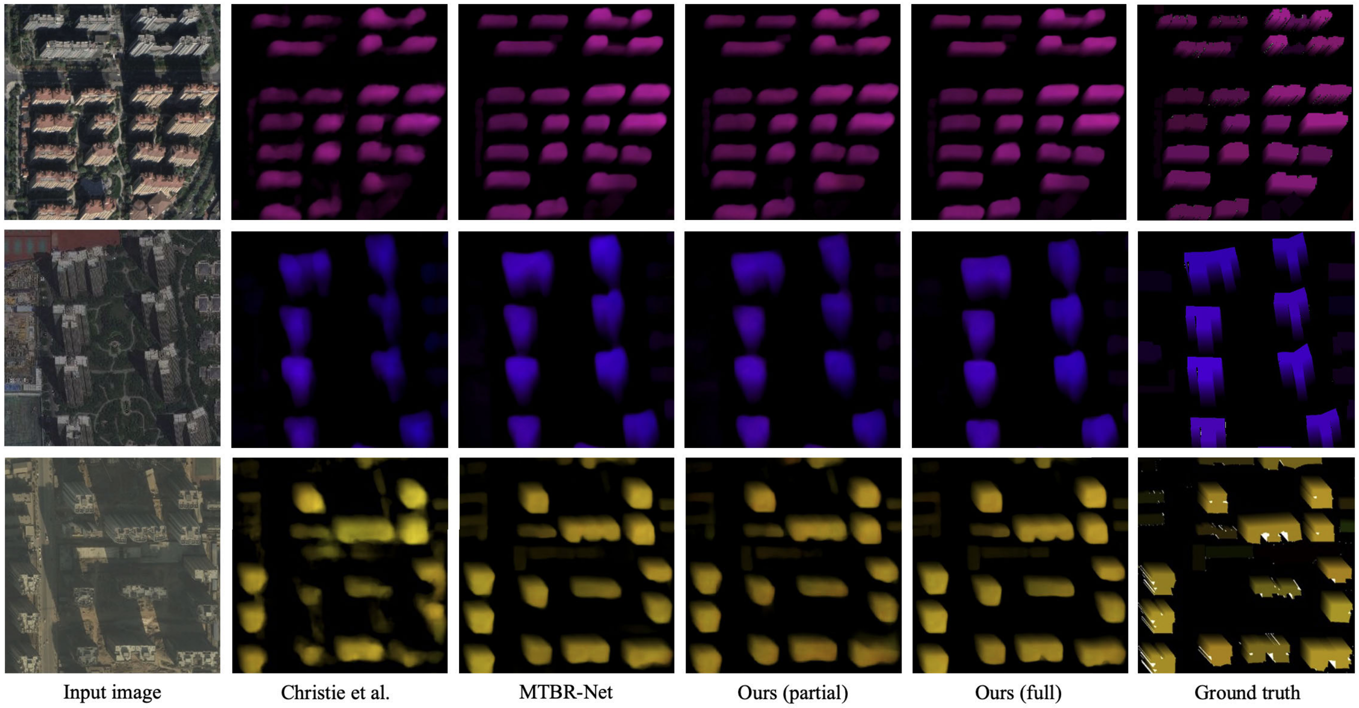

Fig. 7. Examples of height estimation results of different methods on the BONAI Xi'an test dataset. Different colors represent different offset angles. The brightness of each color reflects the offset length.

under two settings, i.e., the fully supervised training with the whole 3-D-annotated samples (denoted as full) and the weakly supervised training with 50% 3-D-annotated samples + 50% 2-D-annotated samples (denoted as partial). The results show that our method can reduce the EPE compared with the other three methods when using the same training set, achieving a significant performance gain for high-rise buildings. Moreover, our method achieves better performance when only using 50% 3-D-annotated samples compared with the results of [9] when using all 3-D-annotated samples, indicating the effectiveness of our method for weakly supervised building reconstruction with fewer 3-D labels. Figs. 6 and 7 provide a qualitative

TABLE II

Building Height Estimation Results Obtained From Different Methods in Terms of EPE. We Report the EPE of the Roof Instances Within Different Height Ranges and the Average EPE of All Instances. Our Method Significantly Reduces the EPE of High-Rise Buildings Compared With Other Three Methods When Using the Same Training Set (Full), and Achieves Better Performance When Only Using 50% 3-D-Annotated Samples (Partial) Compared With [9]

Dataset	Method	EPE of different height range (in pixels)											Average EPE
		0-10	10-20	20-30	30-40	40-50	50-60	60-70	70-80	80-90	90-100	>100	
BONAI Shanghai	Christie et al. [9]	6.22	5.26	7.04	9.01	10.94	12.52	14.89	19.47	24.50	73.07	50.41	6.19
	MTBR-Net [14]	5.24	4.58	**6.05**	6.23	7.20	**8.00**	**8.63**	13.35	**19.66**	51.14	31.34	5.17
	Ours (partial)	6.03	5.54	6.48	7.74	9.14	11.15	11.87	16.56	21.53	**49.55**	41.20	6.07
	Ours (full)	**5.04**	**4.51**	6.24	**6.09**	**7.20**	8.32	8.67	**12.62**	20.05	50.39	**25.27**	**5.08**
BONAI Xi’an	Christie et al. [9]	7.99	9.83	9.81	10.41	13.31	16.11	19.41	24.13	21.27	26.17	75.21	12.31
	MTBR-Net [14]	7.21	**9.67**	**7.92**	**8.71**	**10.02**	10.43	13.40	18.00	12.63	15.12	58.45	10.21
	Ours (partial)	**6.38**	9.97	9.52	8.91	11.62	16.80	22.99	31.01	18.19	21.85	67.93	11.35
	Ours (full)	6.68	9.72	8.16	9.30	10.04	**9.78**	**12.29**	**15.61**	**10.94**	**11.87**	**53.76**	**9.64**

TABLE III

Building Footprint Segmentation Results of Different Methods, in Terms of Precision, Recall, and $F1$-Score (%). Our Method Improves the $F1$-Score by 4.8%–16.0% Compared With the Other Methods When Using the Same 3-D-Annotated Training Samples

Method	BONAI Shanghai dataset			BONAI Xi’an dataset		
	Precision	Recall	F1-score	Precision	Recall	F1-score
Cascade-RCNN [62]	61.27	61.48	61.37	40.73	39.31	40.00
Mask-RCNN [61]	63.43	63.85	63.64	50.30	41.29	45.35
PANet [63]	64.03	61.91	62.95	52.54	41.03	46.08
HR-Net [58]	64.19	64.29	64.24	41.95	35.06	38.20
MTBR-Net [14]	66.85	68.05	67.44	54.34	46.37	50.04
LOFT [15]	65.55	**70.99**	68.16	54.03	47.49	50.72
Ours	**69.29**	68.81	**69.05**	**55.39**	**53.06**	**54.20**

comparison between our method and the two state-of-the-art methods on the BONAI Shanghai and Xi’an test datasets, respectively. The results demonstrate that the height estimation results obtained from our method have more accurate offset values and clearer building boundaries.

E. Building Footprint Segmentation

In this section, we compare the footprint segmentation performance of our method with several competitive instance and semantic segmentation methods [58], [61], [62], [63], [68], and the two state-of-the-art methods for extracting building footprints from off-nadir images, i.e., MTBR-Net [14] and LOFT [15]. Table III lists the segmentation performance of different methods on the BONAI Shanghai and Xi’an test datasets, in terms of precision, recall, and $F1$-score (with the IoU threshold of 0.5) at instance level following [21], [37]. A qualitative comparison of footprint segmentation results on the BONAI Shanghai and Xi’an test datasets is provided in Figs. 8 and 9, respectively. Note that all the experimental results in this section are derived using the same 3-D-annotated training samples, and the experimental results using the extra 2-D-annotated datasets will be analyzed in Section IV-F. The experimental results demonstrate that our method significantly improves the $F1$-score by 4.8%–16.0% compared with the instance and semantic segmentation methods that directly extract the building footprints. Moreover, our method improves the $F1$-score by 0.9%–4.2% compared with MTBR-Net and LOFT that are specifically designed for extracting off-nadir building footprints based on roof segmentation results and offset prediction results. The performance gain also indicates the effectiveness of using the predicted footprint offset to warp the roof segmentation feature map for footprint segmentation.

TABLE IV

Influence of Feature Warping Module on the Performance of Our Method, in Terms of the Footprint Segmentation $F1$-Score (%)and the Height Estimation EPE (in Pixels)

Strategy	Segmentation F1-score (↑)		Height estimation EPE (↓)	
	Shanghai	Xi’an	Shanghai	Xi’an
w/o warping	66.79	48.49	5.26	10.45
w/ original warping	67.44	50.04	5.17	10.21
w/ improved warping	**69.05**	**54.20**	**5.08**	**9.64**

F. Ablation Study

1) Influence of the Feature Warping Module: To evaluate the effect of the improved feature warping module proposed in our WS-MTBR-Net, we analyze the performance obtained from different feature warping strategies using the same 3-D-annotated training samples. Table IV lists the results on our test datasets in terms of the footprint segmentation $F1$-score and the height estimation EPE. The first row shows the results obtained from the multitask HR-Net based on the shared feature map (without feature warping module). The second row shows the results when using the original feature warping module following MTBR-Net [14]. The third row shows the results obtained using the improved feature warping module proposed in our WS-MTBR-Net. Compared with using the shared feature map and the original feature warping module in [14], the performance of both footprint segmentation and height estimation could be improved when using the improved feature warping module.

2) Influence of the Hybrid Loss: To further evaluate the effect of using different combinations of the hybrid loss in our WS-MTBR-Net, we analyze the performance of our method obtained from the training sets with different supervision levels. As shown in Table V, the first row lists the baseline results obtained from the original HR-Net (without feature map warping) trained by the 2-D-annotated footprint dataset. The second row ($\mathcal{X}^F + \mathcal{X}^N$) shows the results of our WS-MTBR-Net when using 50% 3-D-annotated training samples and 50% 2-D-annotated training samples. The third row ($\mathcal{X}^F + \mathcal{X}^P$) shows the results obtained from using 50% 3-D-annotated training samples and 50% training samples with pixel-level 2-D annotations and image-level 3-D annotations.

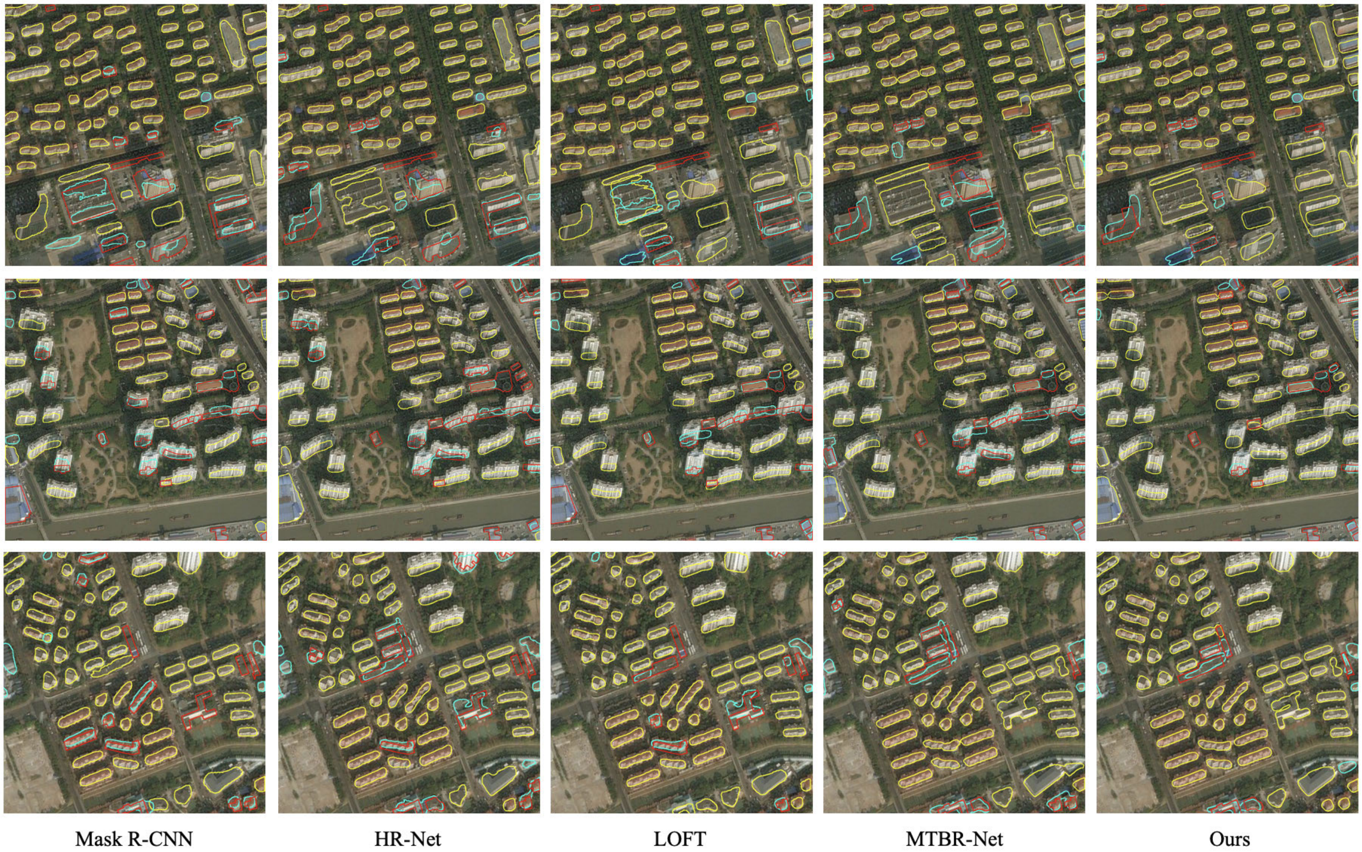

Fig. 8. Examples of building footprint segmentation results of different methods on the BONAI Shanghai test dataset. The yellow, cyan, and red polygons denote the TP, FP, and FN, respectively. Our method produces much more accurate footprint boundaries compared with other methods.

TABLE V

PERFORMANCE OF OUR METHOD OBTAINED FROM DIFFERENT SUPERVISION LEVELS, IN TERMS OF THE FOOTPRINT SEGMENTATION $F1$-SCORE (%) AND THE HEIGHT ESTIMATION EPE (IN PIXELS)

Training set	Segmentation F1-score (↑)		Height estimation EPE (↓)	
	Shanghai	Xi'an	Shanghai	Xi'an
Baseline	64.24	38.20	-	-
$\mathcal{X}^F + \mathcal{X}^N$	68.83	53.54	6.07	11.35
$\mathcal{X}^F + \mathcal{X}^P$	68.91	53.80	5.50	10.39
$\mathcal{X}^F + \mathcal{X}^F$	69.05	54.20	5.08	9.64
$\mathcal{X}^F + \mathcal{X}^F + \mathcal{X}^E$	**70.33**	**56.26**	**4.93**	**9.62**

TABLE VI

PERFORMANCE OF OUR METHOD USING DIFFERENT PERCENTAGES OF FULLY ANNOTATED TRAINING SAMPLES ($\mathcal{X}^F$), IN TERMS OF THE FOOTPRINT SEGMENTATION $F1$-SCORE (%) AND THE HEIGHT ESTIMATION EPE (IN PIXELS)

Percentage	Segmentation F1-score (↑)		Height estimation EPE (↓)	
	Shanghai	Xi'an	Shanghai	Xi'an
10%	57.84	40.39	10.69	15.79
20%	60.35	44.41	9.68	14.34
30%	63.89	47.86	8.41	13.51
40%	66.39	51.18	7.35	12.66
50%	68.83	53.54	6.07	11.35

The fourth row ($\mathcal{X}^F + \mathcal{X}^F$) shows the results obtained from the whole 3-D-annotated training samples. The final row ($\mathcal{X}^F + \mathcal{X}^F + \mathcal{X}^E$) shows the results obtained from using the whole 3-D-annotated training samples and the extra 2-D-annotated training samples. The results demonstrate that the footprint segmentation $F1$-score can be improved by 4.6% and 15.3% when using our WS-MTBR-Net with 50% 3-D-annotated training samples. Moreover, the footprint segmentation and height estimation performance can be successively improved via adding additional supervision types or extra training samples, with the best $F1$-score and EPE obtained from training with the whole 3-D-annotated samples and extra 2-D-annotated samples.

3) Influence of the Training Sample Percentage: To evaluate the effect of using different training set sizes starting from very low percentages, we analyze the performance of our WS-MTBR-Net using different percentages of fully annotated training samples ($\mathcal{X}^F$) ranging from 10% to 50%. As shown in Table VI, with the percentage of fully annotated training samples increases, the offset vector prediction performance and the footprint segmentation performance improve. In addition, when the percentage fully annotated training samples is very low, the offset vector prediction results become poor, and incorrect offset vector prediction will further lead to poor footprint segmentation results.

G. Limitation Analysis

There are several limitations of our method in terms of footprint segmentation and height estimation performance, especially in difficult scenarios. Fig. 10 demonstrates some

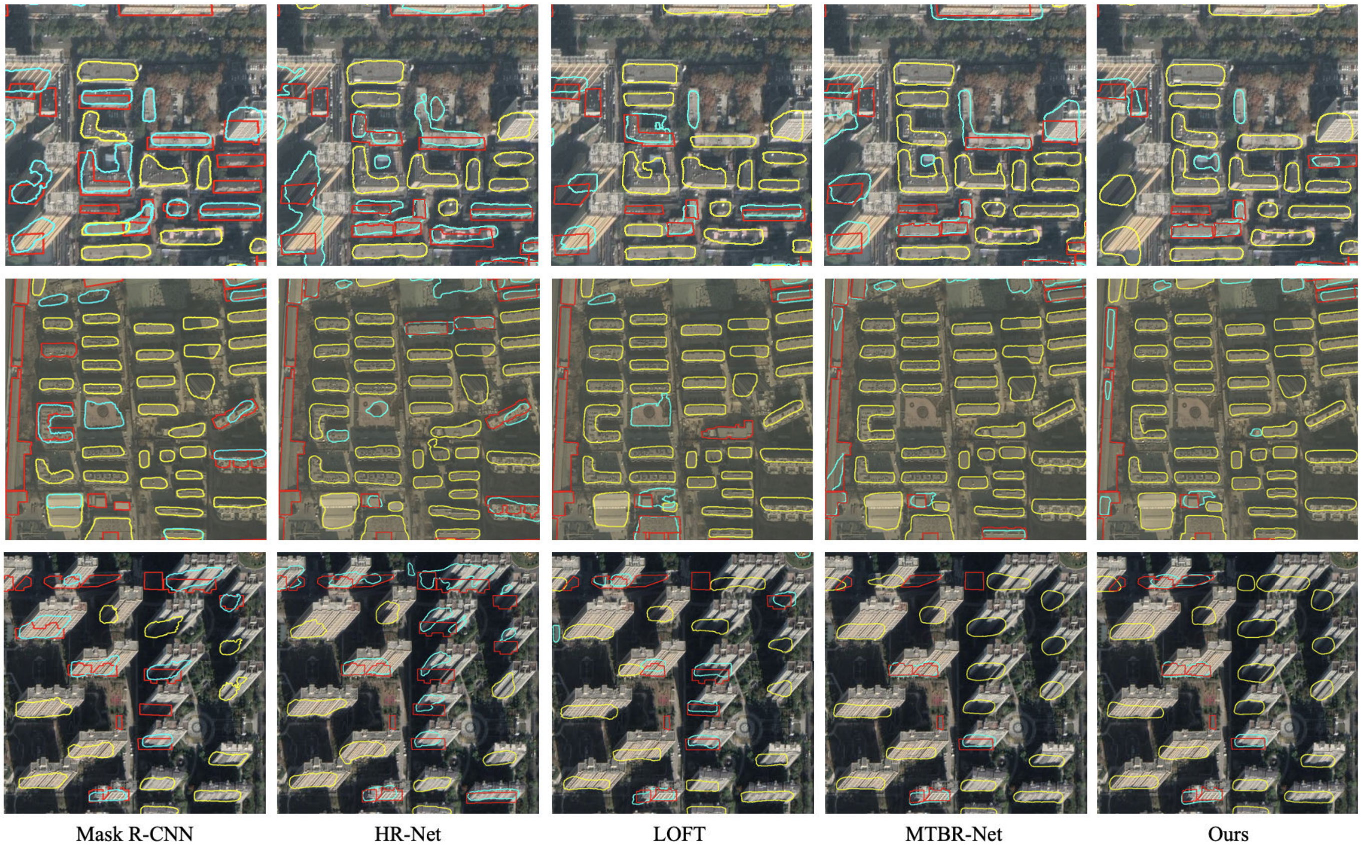

Fig. 9. Examples of building footprint segmentation results of different methods on the BONAI Xi'an test dataset. The yellow, cyan, and red polygons denote the TP, FP, and FN, respectively. Our method produces much more accurate footprint boundaries compared with other methods.

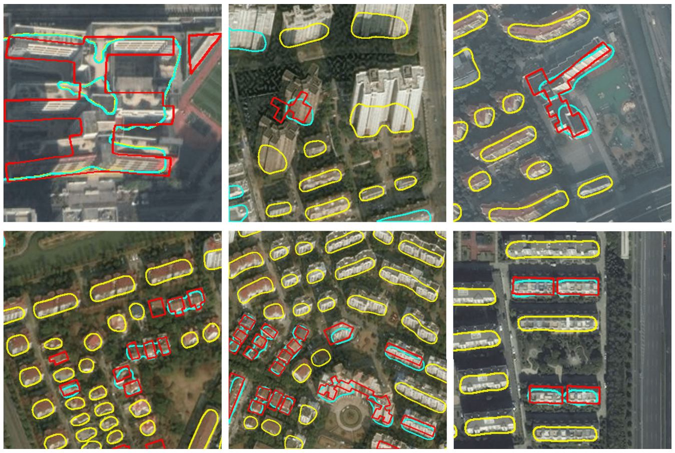

Fig. 10. Some typical failures of footprint segmentation results. The yellow, cyan, and red polygons denote the TP, FP, and FN, respectively.

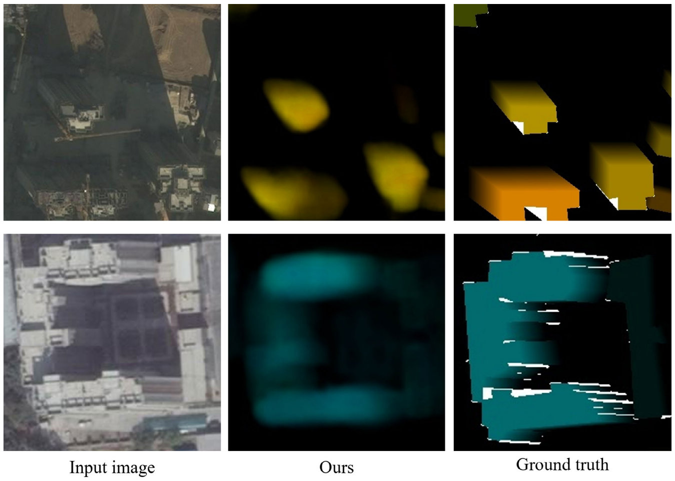

Fig. 11. Some typical failures of height estimation results. Different colors represent different offset angles. The brightness of each color reflects the offset length.

typical failure cases of footprint segmentation results. For buildings with complex shapes (the first row), our model may produce footprint segmentation results with inaccurate boundaries. For densely distributed buildings (the second row), our model may incorrectly extract multiple buildings as one building. Fig. 11 demonstrates some typical failure cases of height prediction results. Due to the shadow effect in high-rise buildings or small facades in low-rise buildings, our model may predict inaccurate offset orientation angles (the first row) and offset lengths (the second row).

V. Conclusion

In this article, we have presented a new method for weakly supervised building reconstruction from monocular remote sensing images, which is capable of reconstructing the accurate 3-D building models with fewer 3-D labels. Qualitative and quantitative evaluations demonstrate that our approach achieves significant performance gain compared with the state-of-the-art under different experiment settings. The effect of the improved feature warping module and using different combinations of the hybrid loss to train our WS-MTBR-Net is also analyzed in the ablation study. To the best of our knowledge, this is the first weakly supervised approach for 3-D building reconstruction from monocular remote sensing images. We believe that our method provides effective and economic solutions for 3-D building reconstruction in complex real-world scenes and significantly reduces the demand for large-scale expensive 3-D annotations. In our future work, we would like to explore more effective strategies for improving the 3-D reconstruction performance and explore weakly supervised approaches for 3-D building reconstruction and city modeling from multiview imagery or multimodal data.

References

[1] V. Verma, R. Kumar, and S. Hsu, "3D building detection and modeling from aerial LiDAR data," in *Proc. IEEE Comput. Soc. Conf. Comput. Vis. Pattern Recognit.*, Jun. 2006, pp. 2213–2220.

[2] L. Duan and F. Lafarge, "Towards large-scale city reconstruction from satellites," in *Proc. Eur. Conf. Comput. Vis. (ECCV)*, 2016, pp. 89–104.

[3] P. Ghamisi and N. Yokoya, "IMG2DSM: Height simulation from single imagery using conditional generative adversarial net," *IEEE Geosci. Remote Sens. Lett.*, vol. 15, no. 5, pp. 794–798, May 2018.

[4] S. Kunwar, "U-Net ensemble for semantic and height estimation using coarse-map initialization," in *Proc. IEEE Int. Geosci. Remote Sens. Symp.*, Jul. 2019, pp. 4959–4962.

[5] J. Mahmud, T. Price, A. Bapat, and J. M. Frahm, "Boundary-aware 3D building reconstruction from a single overhead image," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2020, pp. 441–451.

[6] S. Srivastava, M. Volpi, and D. Tuia, "Joint height estimation and semantic labeling of monocular aerial images with CNNS," in *Proc. IEEE Int. Geosci. Remote Sens. Symp. (IGARSS)*, Jul. 2017, pp. 5173–5176.

[7] Z. Zheng, Y. Zhong, and J. Wang, "Pop-Net: Encoder-dual decoder for semantic segmentation and single-view height estimation," in *Proc. IEEE Int. Geosci. Remote Sens. Symp.*, Jul. 2019, pp. 4963–4966.

[8] Y. Mao et al., "Elevation estimation-driven building 3-D reconstruction from single-view remote sensing imagery," *IEEE Trans. Geosci. Remote Sens.*, vol. 61, 2023, Art. no. 5608718.

[9] G. Christie, R. R. R. M. Abujder, K. Foster, S. Hagstrom, G. D. Hager, and M. Z. Brown, "Learning geocentric object pose in oblique monocular images," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2020, pp. 14512–14520.

[10] G. Christie, K. Foster, S. Hagstrom, G. D. Hager, and M. Z. Brown, "Single view geocentric pose in the wild," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. Workshops (CVPRW)*, Jun. 2021, pp. 1162–1171.

[11] Z. Xiong, W. Huang, J. Hu, and X. X. Zhu, "THE benchmark: Transferable representation learning for monocular height estimation," *IEEE Trans. Geosci. Remote Sens.*, vol. 61, 2023, Art. no. 5620514.

[12] ISPRS. *ISPRS 3D Semantic Labeling Challenge*. Accessed: Sep. 20, 2023. [Online]. Available: https://www2.isprs.org/commissions/comm2/wg4/benchmark/3d-semantic-labeling/

[13] M. Bosch, K. Foster, G. Christie, S. Wang, G. D. Hager, and M. Brown, "Semantic stereo for incidental satellite images," in *Proc. IEEE Winter Conf. Appl. Comput. Vis. (WACV)*, Jan. 2019, pp. 1524–1532.

[14] W. Li, L. Meng, J. Wang, C. He, G.-S. Xia, and D. Lin, "3D building reconstruction from monocular remote sensing images," in *Proc. IEEE/CVF Int. Conf. Comput. Vis.*, Oct. 2021, pp. 12548–12557.

[15] J. Wang, L. Meng, W. Li, W. Yang, L. Yu, and G.-S. Xia, "Learning to extract building footprints from off-nadir aerial images," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 45, no. 1, pp. 1294–1301, Jan. 2023.

[16] Q. Chen, L. Wang, Y. Wu, G. Wu, Z. Guo, and S. L. Waslander, "TEMPORARY REMOVAL: Aerial imagery for roof segmentation: A large-scale dataset towards automatic mapping of buildings," *ISPRS J. Photogramm. Remote Sens.*, vol. 147, pp. 42–55, Jan. 2019.

[17] Microsoft. *Microsoft Global Building Footprints*. Accessed: Sep. 20, 2023. [Online]. Available: https://github.com/microsoft/GlobalMLBuildingFootprints

[18] W. Sirko et al., "Continental-scale building detection from high resolution satellite imagery," 2021, *arXiv:2107.12283*.

[19] S. P. Mohanty. (2018). *Crowdai Dataset: The Mapping Challenge*. [Online]. Available: https://www.aicrowd.com/challenges/

[20] S. Ji, Y. Shen, M. Lu, and Y. Zhang, "Building instance change detection from large-scale aerial images using convolutional neural networks and simulated samples," *Remote Sens.*, vol. 11, no. 11, p. 1343, Jun. 2019.

[21] I. Demir et al., "DeepGlobe 2018: A challenge to parse the earth through satellite images," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. Workshops (CVPRW)*, Jun. 2018, pp. 172–181.

[22] N. Weir et al., "SpaceNet MVOI: A multi-view overhead imagery dataset," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV)*, Oct. 2019, pp. 992–1001.

[23] E. Maltezos, A. Doulamis, N. Doulamis, and C. Ioannidis, "Building extraction from LiDAR data applying deep convolutional neural networks," *IEEE Geosci. Remote Sens. Lett.*, vol. 16, no. 1, pp. 155–159, Jan. 2019.

[24] Y. Wang, L. Gu, X. Li, and R. Ren, "Building extraction in multitemporal high-resolution remote sensing imagery using a multifeature LSTM network," *IEEE Geosci. Remote Sens. Lett.*, vol. 18, no. 9, pp. 1645–1649, Sep. 2021.

[25] L. Wang, S. Fang, X. Meng, and R. Li, "Building extraction with vision transformer," *IEEE Trans. Geosci. Remote Sens.*, vol. 60, 2022, Art. no. 5625711.

[26] W. Li, C. He, J. Fang, J. Zheng, H. Fu, and L. Yu, "Semantic segmentation-based building footprint extraction using very high-resolution satellite images and multi-source GIS data," *Remote Sens.*, vol. 11, no. 4, p. 403, Feb. 2019.

[27] X. Liu et al., "Building instance extraction method based on improved hybrid task cascade," *IEEE Geosci. Remote Sens. Lett.*, vol. 19, pp. 1–5, 2022.

[28] A. Yu, B. Liu, X. Cao, C. Qiu, W. Guo, and Y. Quan, "Pixel-level self-supervised learning for semi-supervised building extraction from remote sensing images," *IEEE Geosci. Remote Sens. Lett.*, vol. 19, pp. 1–5, 2022.

[29] A. Van Etten, D. Lindenbaum, and T. M. Bacastow, "SpaceNet: A remote sensing dataset and challenge series," 2018, *arXiv:1807.01232*.

[30] B. Bischke, P. Helber, J. Folz, D. Borth, and A. Dengel, "Multi-task learning for segmentation of building footprints with deep neural networks," in *Proc. IEEE Int. Conf. Image Process. (ICIP)*, Oct. 2019, pp. 1480–1484.

[31] J. Yuan, "Learning building extraction in aerial scenes with convolutional networks," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 40, no. 11, pp. 2793–2798, Nov. 2018.

[32] E. Protopapadakis, A. Doulamis, N. Doulamis, and E. Maltezos, "Stacked autoencoders driven by semi-supervised learning for building extraction from near infrared remote sensing imagery," *Remote Sens.*, vol. 13, no. 3, p. 371, Jan. 2021.

[33] Q. Li, Y. Shi, and X. X. Zhu, "Semi-supervised building footprint generation with feature and output consistency training," *IEEE Trans. Geosci. Remote Sens.*, vol. 60, 2022, Art. no. 5623217.

[34] Z. Li, J. D. Wegner, and A. Lucchi, "Topological map extraction from overhead images," in *Proc. IEEE Int. Conf. Comput. Vis. (CVPR)*, Oct. 2019, pp. 1715–1724.

[35] W. Zhao, C. Persello, and A. Stein, "Building outline delineation: From aerial images to polygons with an improved end-to-end learning framework," *ISPRS J. Photogramm. Remote Sens.*, vol. 175, pp. 119–131, May 2021.

[36] Y. Hu, Z. Wang, Z. Huang, and Y. Liu, "PolyBuilding: Polygon transformer for building extraction," *ISPRS J. Photogramm. Remote Sens.*, vol. 199, pp. 15–27, May 2023.

[37] W. Li, W. Zhao, H. Zhong, C. He, and D. Lin, "Joint semantic-geometric learning for polygonal building segmentation," in *Proc. AAAI*, 2021, pp. 1958–1965.

[38] W. Li et al., "Joint semantic–geometric learning for polygonal building segmentation from high-resolution remote sensing images," *ISPRS J. Photogramm. Remote Sens.*, vol. 201, pp. 26–37, Jul. 2023.

[39] D. Cheng, R. Liao, S. Fidler, and R. Urtasun, "DARNet: Deep active ray network for building segmentation," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2019, pp. 7431–7439.

[40] L. Zhang et al., "Learning deep structured active contours end-to-end," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, Jun. 2018, pp. 8877–8885.

[41] S. Gur, T. Shaharabany, and L. Wolf, "End to end trainable active contours via differentiable rendering," 2019, *arXiv:1912.00367*.

[42] R. Cabezas, J. Straub, and J. W. Fisher, "Semantically-aware aerial reconstruction from multi-modal data," in *Proc. IEEE Int. Conf. Comput. Vis. (ICCV)*, Dec. 2016, pp. 2156–2164.

[43] E. Rupnik, M. Pierrot-Deseilligny, and A. Delorme, "3D reconstruction from multi-view VHR-satellite images in MicMac," *ISPRS J. Photogramm. Remote Sens.*, vol. 139, pp. 201–211, May 2018.

[44] J. Liu and S. Ji, "A novel recurrent encoder–decoder structure for large-scale multi-view stereo reconstruction from an open aerial dataset," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, May 2020, pp. 6050–6059.

[45] M. Izadi and P. Saeedi, "Three-dimensional polygonal building model estimation from single satellite images," *IEEE Trans. Geosci. Remote Sens.*, vol. 50, no. 6, pp. 2254–2272, Jun. 2012.

[46] A. O. Ok, C. Senaras, and B. Yuksel, "Automated detection of arbitrarily shaped buildings in complex environments from monocular VHR optical satellite imagery," *IEEE Trans. Geosci. Remote Sens.*, vol. 51, no. 3, pp. 1701–1717, Mar. 2013.

[47] Z. Gao et al., "Joint learning of semantic segmentation and height estimation for remote sensing image leveraging contrastive learning," *IEEE Trans. Geosci. Remote Sens.*, vol. 61, 2023, Art. no. 5614015.

[48] Q. Li et al., "3DCentripetalNet: Building height retrieval from monocular remote sensing imagery," *Int. J. Appl. Earth Observ. Geoinf.*, vol. 120, Jun. 2023, Art. no. 103311.

[49] O. Ronneberger, P. Fischer, and T. Brox, "U-Net: Convolutional networks for biomedical image segmentation," in *Proc. Int. Conf. Med. Image Comput. Comput.-Assist. Intervent.* Munich, Germany: Springer, Oct. 2015, pp. 234–241.

[50] Y. Chen, Z. Tu, L. Ge, D. Zhang, R. Chen, and J. Yuan, "SO-HandNet: Self-organizing network for 3D hand pose estimation with semi-supervised learning," in *Proc. IEEE/CVF Int. Conf. Comput. Vis.*, 2019, pp. 6961–6970.

[51] R. Mitra, N. B. Gundavarapu, A. Sharma, and A. Jain, "Multiview-consistent semi-supervised learning for 3D human pose estimation," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2020, pp. 6906–6915.

[52] N. Neverova, C. Wolf, F. Nebout, and G. W. Taylor, "Hand pose estimation through semi-supervised and weakly-supervised learning," *Comput. Vis. Image Understand.*, vol. 164, pp. 56–67, Nov. 2017.

[53] G. Yang, Y. Cui, S. Belongie, and B. Hariharan, "Learning single-view 3D reconstruction with limited pose supervision," in *Proc. Eur. Conf. Comput. Vis. (ECCV)*, 2018, pp. 86–101.

[54] R. Ji et al., "Semi-supervised adversarial monocular depth estimation," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 42, no. 10, pp. 2410–2422, Oct. 2020.

[55] J. Gwak, C. B. Choy, M. Chandraker, A. Garg, and S. Savarese, "Weakly supervised 3D reconstruction with adversarial constraint," in *Proc. Int. Conf. 3D Vis. (3DV)*, 2017, pp. 263–272.

[56] J. Han, Y. Yang, D. Zhang, D. Huang, D. Xu, and F. De La Torre, "Weakly-supervised learning of category-specific 3D object shapes," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 43, no. 4, pp. 1423–1437, Apr. 2021.

[57] C. Li, A. Morel-Forster, T. Vetter, B. Egger, and A. Kortylewski, "Robust model-based face reconstruction through weakly-supervised outlier segmentation," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2023, pp. 372–381.

[58] K. Sun, B. Xiao, D. Liu, and J. Wang, "Deep high-resolution representation learning for human pose estimation," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2019, pp. 5693–570.

[59] A. Paszke et al., "PyTorch: An imperative style, high-performance deep learning library," 2019, *arXiv:1912.01703*.

[60] K. Sun et al., "High-resolution representations for labeling pixels and regions," 2019, *arXiv:1904.04514*.

[61] K. He, G. Gkioxari, P. Dollár, and R. Girshick, "Mask R-CNN," in *Proc. IEEE Int. Conf. Comput. Vis. (ICCV)*, Oct. 2017, pp. 2961–2969.

[62] Z. Cai and N. Vasconcelos, "Cascade R-CNN: Delving into high quality object detection," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, Jun. 2018, pp. 6154–6162.

[63] S. Liu, L. Qi, H. Qin, J. Shi, and J. Jia, "Path aggregation network for instance segmentation," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, 2018, pp. 8759–8768.

[64] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proc. CVPR*, vol. 16, 2016, pp. 770–778.

[65] O. Russakovsky et al., "ImageNet large scale visual recognition challenge," *Int. J. Comput. Vis.*, vol. 115, no. 3, pp. 211–252, Dec. 2015.

[66] T. Y. Lin, P. Dollàr, R. Girshick, K. He, B. Hariharan, and S. Belongie, "Feature pyramid networks for object detection," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jul. 2017, pp. 2117–2125.

[67] K. Chen et al., "MMDetection: Open MMLab detection toolbox and benchmark," 2019, *arXiv:1906.07155*.

[68] K. Chen et al., "Hybrid task cascade for instance segmentation," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, Jul. 2019, pp. 4974–4983.

Weijia Li received the bachelor's degree from the Department of Computer Science, Sun Yat-Sen University, Guangzhou, China, in 2014, and the Ph.D. degree from the Department of Earth System Science, Tsinghua University, Beijing, China, in 2019.

From 2019 to 2021, she was a Post-Doctoral Researcher with the CUHK-Sensetime Joint Laboratory (MMLab), Department of Information Engineering, The Chinese University of Hong Kong (CUHK), Hong Kong. She is currently an Associate Professor with the School of Geospatial Engineering and Science, Sun Yat-Sen University. Her research interests include remote sensing image understanding, computer vision, and deep learning.

Zhenghao Hu received the bachelor's degree from the School of Geospatial Engineering and Science, Sun Yat-Sen University, Zhuhai, China, in 2023, where he is currently pursuing the M.S. degree with the School of Geospatial Engineering and Science.

His research interests include remote sensing image understanding, computer vision, and deep learning.

Lingxuan Meng received the B.S. degree from the University of Electronic Science and Technology of China (UESTC), Chengdu, China, in 2016, and the Ph.D. degree from the School of Resources and Environment, UESTC, in 2022.

His research interests include computer vision and deep learning.

Jinwang Wang received the B.Eng. degree in communication engineering from Lanzhou University, Lanzhou, China, in 2016, and the Ph.D. degree in communication and information system from Wuhan University, Wuhan, China, in 2021.

His research interests include computer vision and remote sensing, particularly focusing on object detection in aerial images.

Juepeng Zheng received the bachelor's degree from the College of Surveying and Geoinformatics, Tongji University, Shanghai, China, in 2019, and the Ph.D. degree from the Department of Earth System Science, Tsinghua University, Beijing, China, in 2023.

He is currently an Assistant Professor with the School of Artificial Intelligence, Sun Yat-Sen University, Zhuhai, China. He is also a Researcher with the National Supercomputing Center, Shenzhen. His research interests include remote sensing image understanding, high-performance computing, and deep learning.

Runmin Dong received the Ph.D. degree in ecology from the Department of Earth System Science, Tsinghua University, Beijing, China, in 2022.

She is a Post-Doctoral Researcher with the Department of Earth System Science, Tsinghua University. Her research interests include remote sensing, artificial intelligence, land cover mapping, super-resolution, image fusion, image synthesis, and self-supervised learning.

Conghui He received the bachelor's degree from Sun Yat-sen University, Guangzhou, China, in 2013, and the Ph.D. degree from Tsinghua University, Beijing, China, in 2018.

He is a Research Director at SenseTime, Shenzhen, China, as well as a Research Scientist and a PI at the Shanghai AI Laboratory, Shanghai, China. His research interests include high-performance computing, reconfigurable computing, graph computing, and computer vision.

Gui-Song Xia (Senior Member, IEEE) received the Ph.D. degree in image processing and computer vision from the CNRS Information Processing and Communications Laboratory, Télécom ParisTech, Paris, France, in 2011.

From 2011 to 2012, he was a Post-Doctoral Researcher with the Centre de Recherche en Mathématiques de la Décision, CNRS, Paris Dauphine University, Paris, for one and a half years. He was a Visiting Scholar with DMA, École Normale Supérieure (ENS-Paris), Paris, for two months, in 2018. He was also a Guest Professor with the Future Laboratory AI4EO, Technical University of Munich, Munich, Germany. He is currently a Yongyi Distinguished Professor of computer vision and photogrammetry with Wuhan University, Wuhan, China. His research interests include mathematical modeling of images and videos, structure from motion, perceptual grouping, and remote sensing image understanding.

Dr. Xia served on the editorial boards of several journals, including *ISPRS Journal of Photogrammetry and Remote Sensing*, *Pattern Recognition*, *Signal Processing: Image Communication*, *EURASIP Journal on Image and Video Processing*, *Journal of Remote Sensing*, and *Frontiers in Computer Science: Computer Vision*.

Haohuan Fu (Senior Member, IEEE) received the Ph.D. degree in computing from Imperial College, London, U.K., in 2009.

He is currently a Professor with the Tsinghua Shenzhen International Graduate School, the Ministry of Education Key Laboratory for Earth System Modeling, and the Department of Earth System Science, Tsinghua University, Beijing, China. He is also the Deputy Director with the National Supercomputing Center, Wuxi, China. His research interests include high-performance computing in Earth and environmental sciences, computer architectures, performance optimizations, and programming tools in parallel computing.

Dr. Fu was a recipient of the ACM Gordon Bell Prize in 2016 and 2017, and the Most Significant Paper Award by FPL in 2015.

Dahua Lin (Member, IEEE) received the B.Eng. degree from the University of Science and Technology of China (USTC), Hefei, China, in 2004, the M.Phil. degree from The Chinese University of Hong Kong (CUHK), Hong Kong, in 2006, and the Ph.D. degree from the Massachusetts Institute of Technology (MIT), Cambridge, MA, USA, in 2012.

He is currently an Associate Professor with the Department of Information Engineering, The Chinese University of Hong Kong (CUHK), Hong Kong, and the Director of CUHK-SenseTime Joint Laboratory. Prior to joining CUHK, he served as a Research Assistant Professor at the Toyota Technological Institute at Chicago, Chicago, IL, USA, from 2012 to 2014. His research interest covers computer vision and machine learning.

Dr. Lin serves on the editorial board of the *International Journal of Computer Vision (IJCV)*. He also serves as an Area Chair for multiple conferences, including ECCV 2018, ACM Multimedia 2018, BMVC 2018, CVPR 2019, BMVC 2019, AAAI 2020, and CVPR 2021.