Remote Sensing of Environment 264 (2021) 112590

Contents lists available at ScienceDirect

Remote Sensing of Environment

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

A deep learning method for building height estimation using high-resolution multi-view imagery over urban areas: A case study of 42 Chinese cities

Yinxia Cao [a], Xin Huang [a,b,*]

[a] School of Remote Sensing and Information Engineering, Wuhan University, Wuhan 430079, PR China
[b] State Key Laboratory of Information Engineering in Surveying, Mapping and Remote Sensing, Wuhan University, Wuhan 430079, PR China

ARTICLE INFO

Keywords:
Building height
High-resolution
Multi-view
ZY-3
Multi-task
Deep learning

ABSTRACT

Knowledge of building height is critical for understanding the urban development process. High-resolution optical satellite images can provide fine spatial details within urban areas, while they have not been applied to building height estimation over multiple cities and the feasibility of mapping building height at a fine scale (< 5 m) remains understudied. Multi-view satellite images can describe vertical information of buildings, due to the inconsistent response of buildings (e.g., spectral and structural variations) to different viewing angles, but they have not been employed to deep learning-based building height estimation. In this context, we introduce high-resolution ZY-3 multi-view images to estimate building height at a spatial resolution of 2.5 m. We propose a multi-spectral, multi-view, and multi-task deep network (called M^{3}Net) for building height estimation, where ZY-3 multi-spectral and multi-view images are fused in a multi-task learning framework. A random forest (RF) method using multi-source features is also carried out for comparison. We select 42 Chinese cities with diverse building types to test the proposed method. Results show that the M^{3}Net obtains a lower root mean square error (RMSE) than the RF, and the inclusion of ZY-3 multi-view images can significantly lower the uncertainty of building height prediction. Comparison with two existing state-of-the-art studies further confirms the superiority of our method, especially the efficacy of the M^{3}Net in alleviating the saturation effect of high-rise building height estimation. Compared to the vanilla single/multi-task models, the M^{3}Net also achieves a lower RMSE. Moreover, the spatial-temporal transferability test indicates the robustness of the M^{3}Net to imaging conditions and building styles. The test of our method on a relatively large area (covering about 14,120 km^2) further validates the scalability of our method from the perspectives of both efficacy and quality. The source code will be made available at https://github.com/lauraset/BuildingHeightModel.

1. Introduction

Building height characterizes the vertical dimension of urban form and gives a basic insight into urban development. A recent study on urban growth typology shows that, upward and outward growth occurs in China and South Korea extensively, with a large increase of high-rise buildings (Mahtta et al., 2019). Particularly, building height provides essential knowledge for sustainable urban development, and plays a vital role in the fields of urban climate (Berger et al., 2017; Giridharan et al., 2004; Venter et al., 2020), pollution transmission (Hang et al., 2012), building energy consumption (Güneralp et al., 2017), population estimation (Leichtle et al., 2019; Tomás et al., 2016; Xie et al., 2015), and three-dimensional (3D) building reconstruction (Haala and Kada, 2010), among others. Thus, building height information is crucial for the comprehensive understanding of urban development.

Remote sensing techniques offer an effective tool for building height mapping. Nevertheless, few efforts have been made to estimate building height from high-spatial-resolution images (< 5 m) over multiple cities. Limited by 3D data availability, myriads of studies are concentrated on two-dimensional (2D) urban information extraction (Dell'Acqua and Gamba, 2003; Esch et al., 2017; Gamba et al., 2007; Gong et al., 2020; Li et al., 2018; Liu et al., 2019; Liu et al., 2018; Pesaresi et al., 2013; Schneider et al., 2010; Schug et al., 2020; Taubenböck et al., 2012), with only a few on 3D urban form analysis (Esch et al., 2020; Frolking et al.,

* Corresponding author at: School of Remote Sensing and Information Engineering, Wuhan University, Wuhan 430079, PR China.
E-mail address: xhuang@whu.edu.cn (X. Huang).

https://doi.org/10.1016/j.rse.2021.112590
Received 1 March 2021; Received in revised form 30 June 2021; Accepted 5 July 2021
Available online 12 July 2021

2013; Geiß et al., 2020; Geisß et al., 2019; Mahendra and Seto, 2019; Taubenböck et al., 2018). Recently, the Sentinel-1 satellite constellation offers free and globally available C-band Synthetic Aperture Radar (SAR) data at 10-m spatial resolution (Torres et al., 2012), and it is found that the recorded backscatter values are strongly related to building height (Koppel et al., 2017). In this context, Li et al. (2020b) proposed an indicator of Sentinel-1 backscatter intensity and then used it to develop a building height model based on seven cities of the United States (US). Through the model, they generated building height of all US cities with area $>$ 500 km^2 (e.g., New York, Chicago, and Los Angeles) at 500-m scale. Li et al. (2020a) trained random forest models (Breiman, 2001) and developed the first continental-scale (i.e., China, the US, and Europe) 3D building product, including building footprint, height, and volume, at a spatial resolution of 1 km. Further, Frantz et al. (2021) adopted support vector machine regression models (Cortes and Vapnik, 1995), and effectively retrieved 10-m building height for the whole Germany by synergistic use of Sentinel-1 and Sentinel-2 time series (Malenovský et al., 2012). These studies have well demonstrated the feasibility and the effectiveness of building height estimation at regional and global scales, but their spatial resolutions allow building heights only at aggregated spatial scales, and heterogeneities at individual building level can not be covered.

Generally, fine-scale building height can be estimated by three types of data: 1) Light Detection and Ranging (LiDAR), 2) radar, and 3) high-resolution optical imagery. LiDAR allows high accuracy measurements of building height (Baltsavias, 1999), and thus is widely applied to 3D building modelling (Rottensteiner, 2003; Sun and Salvaggio, 2013; Verma et al., 2006). However, the coverage of LiDAR is still limited due to its high acquisition cost. Alternatively, radar images hold great potentials for building height estimation, e.g., single/stereo SAR (Brunner et al., 2010; Soergel et al., 2009; Sun et al., 2017), interferometric SAR (Tison et al., 2007; Wegner et al., 2013), and tomographic SAR (Zhu and Bamler, 2010). Nevertheless, with the side-looking geometry, radar images usually record signals from a mixture of different microwave scattering mechanisms, leading to relatively high uncertainties of building height estimation (Sun et al., 2019). By contrast, high-resolution optical images can alleviate this issue, and provide fine spatial details and rich spectral information within urban areas. For single optical images, it is possible to retrieve building height from adjacent shadows (Liasis and Stavrou, 2016; Qi et al., 2016; Shao et al., 2011). Although this kind of technique is efficient for certain types of buildings, e.g., buildings with a height of $\leq$20 m (Qi et al., 2016), it largely relies on the accuracy of shadow detection. In addition, shadows in dense urban environments are often distorted, and thus incomplete. On the other hand, with the availability of stereo/multi-view images, building height can be readily estimated from digital surface models (DSMs) that are generated by stereo matching (Alobeid et al., 2009; Taubenböck et al., 2013; Tian et al., 2014). For instance, Liu et al. (2017) employed the morphological reconstruction method (Qin and Fang, 2014) to remove the terrain height from DSM, and thus obtained the off-terrain height, i.e., the normalized DSM (nDSM), to indicate the height of buildings. They found that DSM usually suffers from matching failures due to occlusions of buildings. Nevertheless, it should be noted that these high-resolution building height studies are limited to small or local regions, leaving the applicability of the methodology over multiple cities unknown.

As aforementioned, the existing methods are largely affected by the quality of shadow detection or image stereo matching. In this background, deep learning (DL) opens a new avenue for building height estimation. As a particular DL architecture, convolutional neural networks (CNNs) can automatically exploit multilevel features from raw images, and replace the conventional feature handcrafting. Consequently, CNNs have been increasingly applied to the remote sensing domain (Li et al., 2019; Yuan et al., 2020), and have achieved impressive results in urban-related studies (Chen et al., 2020; Gonzalez et al., 2020; Miura et al., 2020; Taubenböck et al., 2020; Zhou et al., 2020). Particularly, in Taubenböck et al. (2020), the local climate zones have been classified using CNN network and the height information has been contained on the structural types in an indirect way. Recently, a growing body of studies have explored the feasibility of predicting continuous height values (e.g., DSM) from single high-resolution optical images (Amirkolaee and Arefi, 2019; Carvalho et al., 2019; Ghamisi and Yokoya, 2018; Liebel et al., 2020; Mahmud et al., 2020; Mou and Zhu, 2018). For instance, Amirkolaee and Arefi (2019) developed a deep CNN to estimate DSM from single aerial images, and demonstrated its effectiveness on ISPRS datasets (Rottensteiner et al., 2012). Particularly, in order to fully exploit mutual information from different tasks, Carvalho et al. (2019) introduced the multi-task learning network (Naik and Rangwala, 2018) to simultaneously handle land cover mapping and the normalized DSM (nDSM) estimation, and they found that the multi-task learning performs better than the single-task approach. Although height estimation with single images has achieved success to some extent, it is still an ill-posed and challenging problem (Amirkolaee and Arefi, 2019). Therefore, it is natural to introduce multi-view images to lower the uncertainty of height estimation. Multi-view satellite images can provide vertical information of buildings, and have been applied to urban scene classification (Huang et al., 2018). However, to the knowledge of the authors, multi-view satellite images have not been considered in the building height regression studies, and their capability of height estimation remains unknown.

In summary, although the existing research has made progress on building height estimation, there still exist the following limitations:

1) Most high-resolution building height estimation studies are limited to local or small areas, and the investigation across multiple cities is lacking.
2) Multi-view satellite images are able to describing the vertical attribute of ground objects, but they have not been employed to deep learning-based building height estimation.

Given these issues and challenges, this study introduces the Chinese ZY-3 stereo satellite constellation that is composed of three satellites, i.e., ZY-3 01, 02, and 03, launched in 2012, 2016, and 2020, respectively. Each satellite can simultaneously acquire multi-spectral images (with a spatial resolution of 5.8 m) and multi-view images with nadir (2.1 m), $+22^\circ$ forward (2.5–3.5 m), and -22° backward (2.5–3.5 m) viewing angles (Huang et al., 2017; Tang et al., 2020). The combination of multi-spectral and multi-view images can offer fine spectral, spatial, and vertical information for building height estimation. In this context, we aim to estimate fine-scale (2.5 m) building height in 42 Chinese cities, in order to investigate the effectiveness of ZY-3 multi-view images for building height prediction and compare the proposed method as well as its result with the existing ones. Accordingly, we propose a multi-spectral, multi-view, and multi-task deep network (called M^3Net) that fuses ZY-3 multi-spectral and multi-view images with a multi-task learning framework. For the purpose of comparison, we also implement a random forest (RF) method using multi-source features for building height estimation. We aim to answer the following research questions:

1) Can ZY-3 multi-view imagery improve accuracy in building height estimation?
2) How is the capability and superiority of the proposed M^3Net for height estimation compared to the existing models and results?
3) Can the M^3Net generalize well across space and time over a large number of cities?

The remainder of this paper is arranged as follows. Section 2 introduces the study areas and data. The methodology is presented in Section 3. Next, the results are reported in Section 4 and discussed in Section 5. Finally, the conclusions are drawn in Section 6.

Table 1
Available samples (1 km × 1 km) in the 42 Chinese cities.

City	#Sample	City	#Sample	City	#Sample
Baotou	93	Hefei	78	Shanghai	253
Beijing	610	Hohhot	25	Shenyang	242
Changchun	81	Jinan	60	Shenzhen	271
Changsha	66	Kunming	12	Shijiazhuang	106
Chengdu	220	Lanzhou	23	Taiyuan	63
Chongqing	173	Lhasa	34	Tangshan	44
Dalian	83	Luoyang	29	Tianjin	203
Dongguan	145	Nanchang	16	Urumqi	41
Foshan	53	Nanjing	155	Wuhan	150
Fuzhou	35	Nanning	70	Wuxi	63
Guangzhou	363	Ningbo	40	Xi'an	183
Haikou	38	Ordos	4	Xining	29
Hangzhou	230	Qingdao	88	Yinchuan	17
Harbin	82	Quanzhou	63	Zhengzhou	89
Total	4723				

2. Study areas and data

A total of 42 Chinese cities were selected to assess the proposed method (Table 1), including 4 municipalities, 26 provincial capitals, and 12 large cities, and representing a variety of landscapes and urbanization levels. We focused on urban areas of these cities, since they are the main places of human activities and contain diverse buildings with different colors, shapes, sizes, and height, which is suitable for testing the generalization ability of the proposed method.

We collected ZY-3 optical images, reference building height, and other data, e.g., Sentinel-2 optical images, Sentinel-1 radar images, the Visible Infrared Imaging Radiometer Suite (VIIRS) nighttime light data, ALOS World 3D DSM, global impervious areas (GISA), and OpenStreetMap (OSM). Please notice that the proposed M^{3}Net only takes ZY-3 images as input, while as a comparison method, the random forest model uses both ZY-3 images and other data as input. The acquisition time of all the images should be close to the reference year 2015, considering the temporal consistency and data availability. Details of these data are presented below.

ZY-3 images with cloud cover <10% were collected from the Land Satellite Remote Sensing Application Center (LASAC) of China. The images acquired between 2014 and 2017 were considered, since a few cities were not fully covered in 2015. Note that each ZY-3 image scene contains four images with a ground swath of about 50 km, including the multi-spectral images (blue, green, red, and near-infrared bands with a

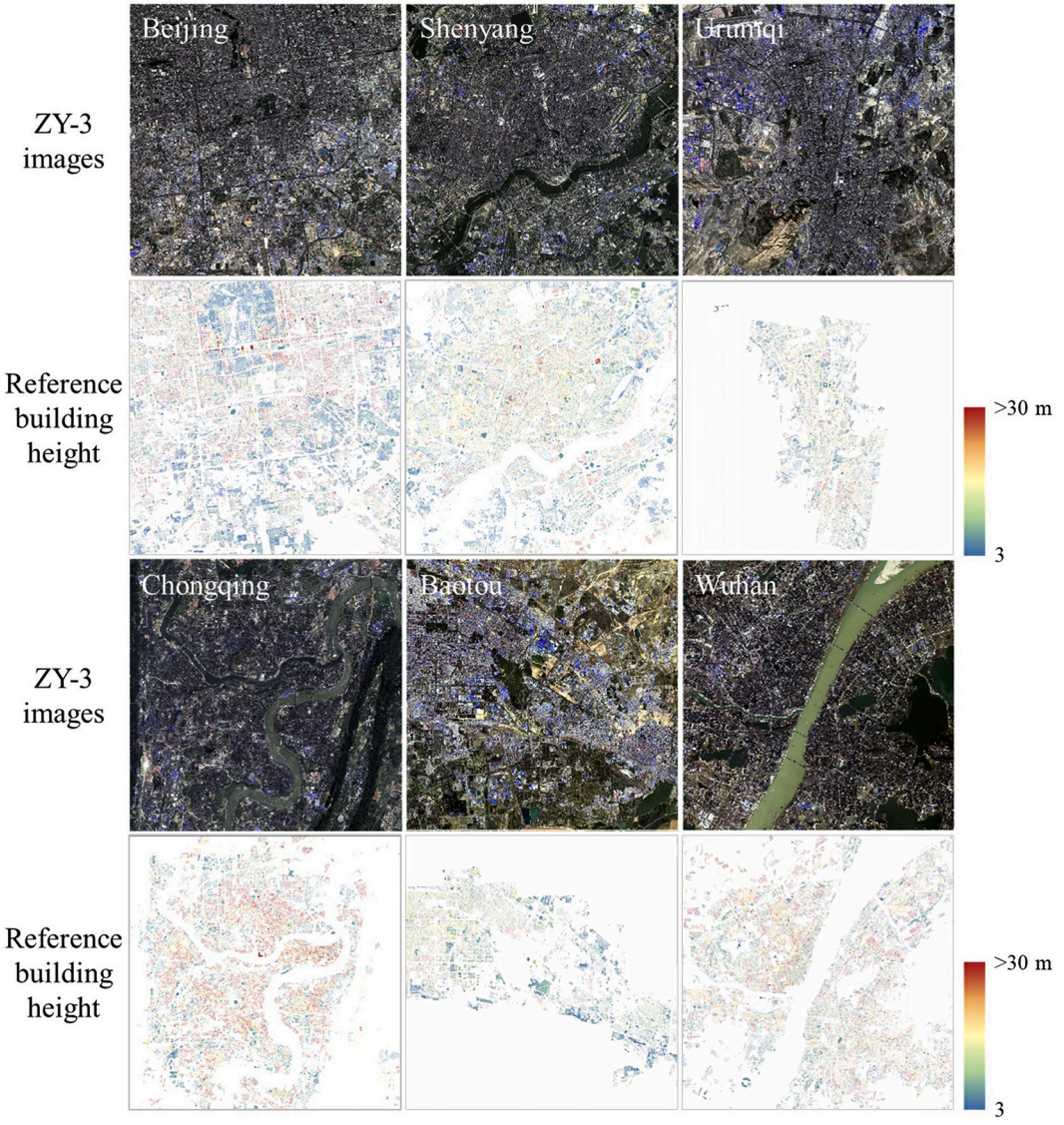

Fig. 1. ZY-3 images and the corresponding reference building height in six representative cities (i.e., Beijing, Shenyang, Urumqi, Chongqing, Wuhan, and Baotou). Each area has a spatial extent of 20 km × 20 km.

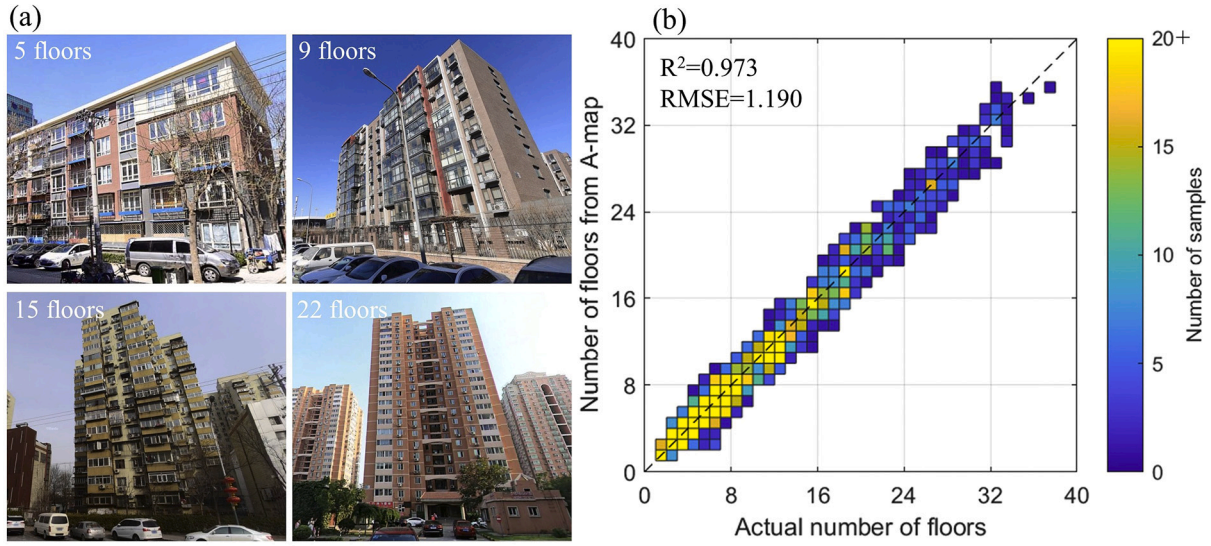

Fig. 2. (a) Examples of Baidu street view images used for obtaining the actual number of floors by visual interpretation. (b) Accuracy of the number of floors from A-map. RMSE: root mean square error.

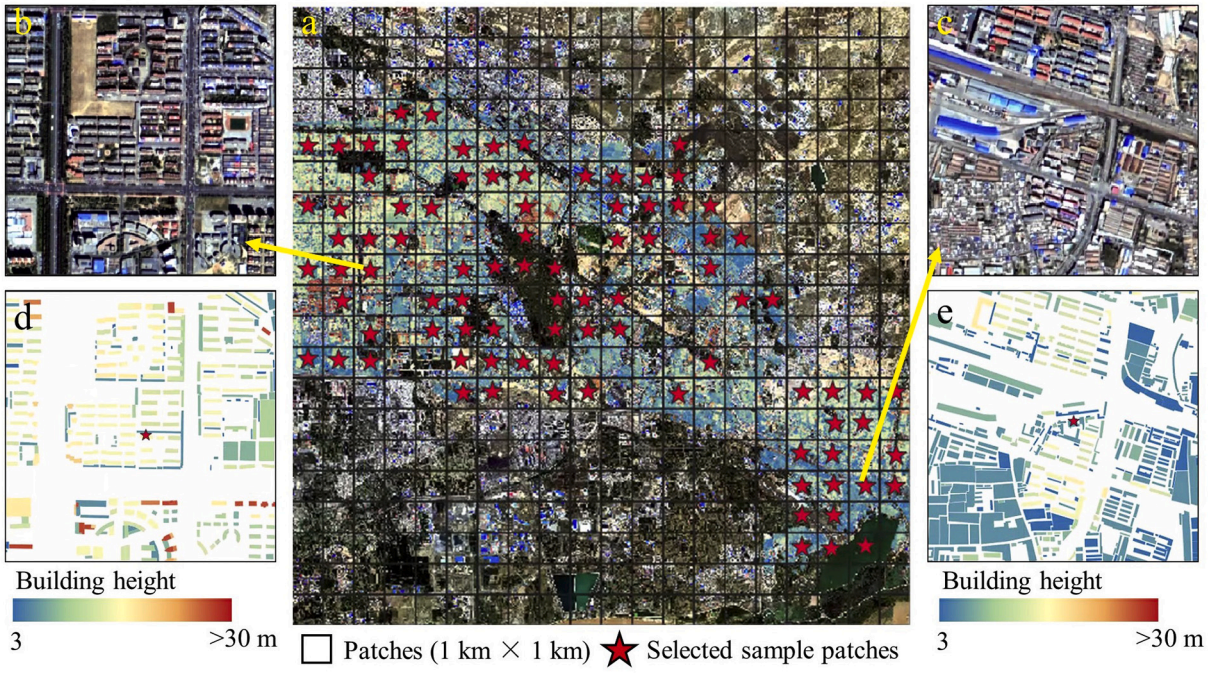

Fig. 3. Illustration of the sample selection in Baotou city (a). Graphs (b) and (c) show the ZY-3 images (1 km × 1 km) and graphs (d) and (e) are the corresponding reference building height maps.

spatial resolution of 5.8 m) and three panchromatic images (with nadir (2.1 m), +22° forward (2.5–3.5 m), and − 22° backward (2.5–3.5 m) viewing angles, respectively). The data pre-processing steps included radiometric correction, ortho-rectification, image-to-image registration, and pan-sharpening. For each ZY-3 image scene, the forward, backward, and multispectral images were geometrically registered to the nadir image by polynomial warping using automatically generated tie points (Kennedy and Cohen, 2003), with a registration error < 1 pixel. All images were resampled to 2.5 m. The multispectral images were then fused with nadir images by the Gram-Schmidt method (Laben and Brower, 2000) to increase spatial details of multispectral images. To reduce the radiometric difference between multi-view images, forward and backward images were normalized to the nadir image by the histogram matching algorithm (Gonzalez and Woods, 2002). Details of ZY-3 data pre-processing can be found in (Huang et al., 2020; Liu et al., 2019).

Reference building height data were acquired from a map service provider of China, A-map (https://amap.com) (see Fig. 1). A-map collected building footprints with number of floors in a lot of Chinese cities by field investigation, and publicly released the data in the vector form. To assess the accuracy of the number of floors from A-map, we manually interpreted the number of floors of 2324 buildings that are randomly distributed in 42 Chinese cities, with the aid of Baidu street view images (https://map.baidu.com, see Fig. 2(a) for example). As shown in Fig. 2(b), the RMSE is 1.190, verifying the high reliability of the number of floors from A-map. Then, we converted the number of

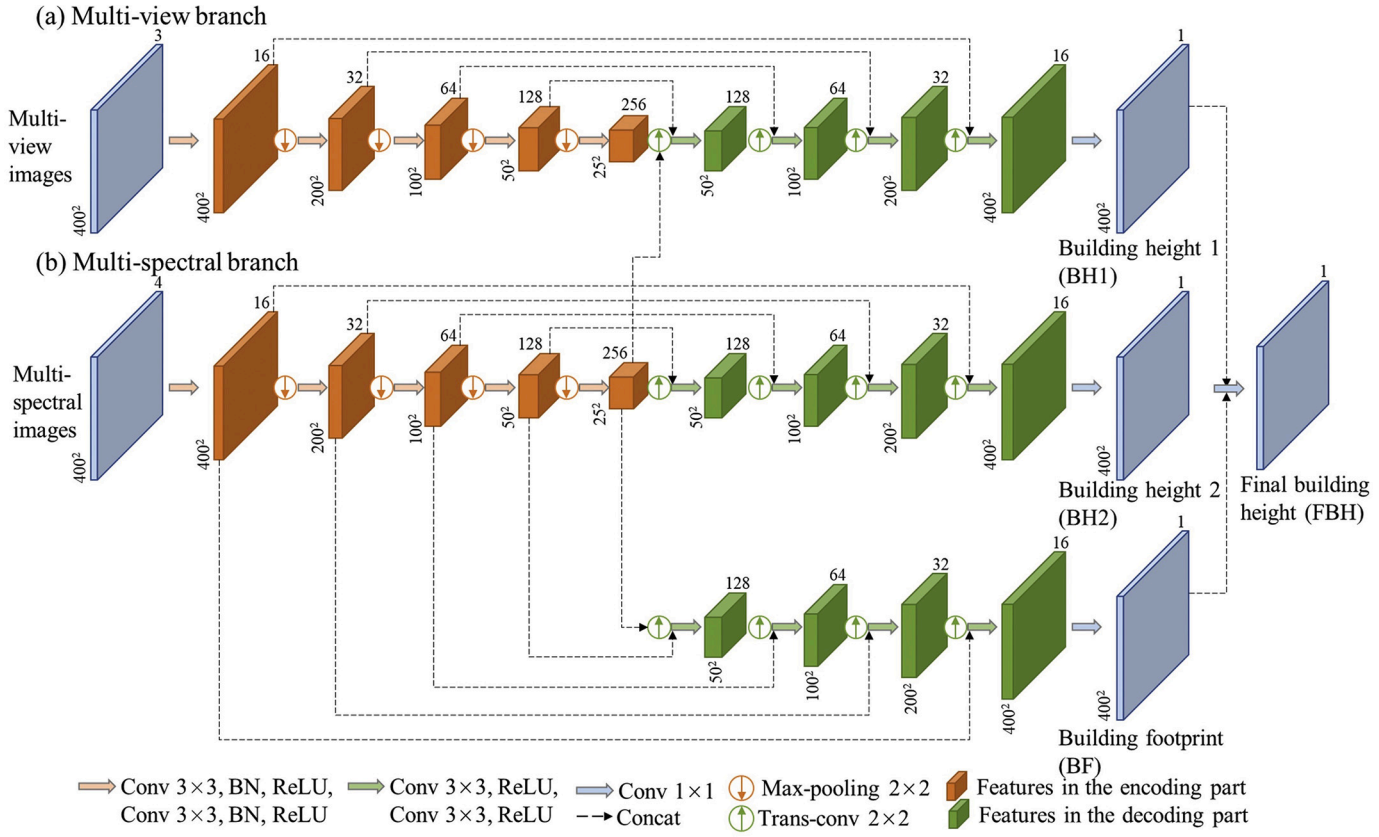

Fig. 4. Structure of the proposed M^{3}Net consisting of (a) the multi-view branch and (b) the multi-spectral branch. The dimension of features is described by the number of channels and resolution. For instance, 400^2 (pixels) refers to the width and the height of features, and 1 denotes the number of feature channels. Conv: convolution layer; BN: batch normalization; Trans-conv: transposed convolutional layer; ReLU: rectified linear unit activation function. 3 × 3 or 2 × 2 corresponds to the kernel size.

floors to building height under the assumption that each floor is 3 m (Li et al., 2020a; Zheng et al., 2017; Zhou et al., 2014). To the authors' knowledge, the building height provided by A-map is currently the most reliable data that can be publicly accessible, and it has been successfully used as reference data for continental-scale building height estimation in Li et al., 2020a. Note that affected by the temporal inconsistency, the original reference data may contain some false alarms and omissions in the reference year 2015. Therefore, we first clipped these data into 1 km × 1 km samples, and then retained the high-quality ones through careful visual interpretation. In this way, we obtained 4723 samples from the 42 cities (Table 1), and randomly selected 70%, 10%, and 20% of them for training, validation, and testing, respectively. An example of the sample selection is shown in Fig. 3, where graphs (d) and (e) show the reference building height within a 1 km × 1 km sample.

Other data, including Sentinel-2 top-of-atmosphere reflectance (TOA) images (Drusch et al., 2012), Sentinel-1 with VV and VH bands (Torres et al., 2012), VIIRS (Elvidge et al., 2017), and ALOS World 3D data (Takaku et al., 2020), were downloaded from the google earth engine (GEE) cloud computing platform (Gorelick et al., 2017). For Sentinel-2 images, we used the Sen2Cor algorithm (Main-Knorn et al., 2017) to generate bottom-of-atmosphere (BOA) reflectance product. Then, we calculated the median value of all cloud-free Sentinel-2 BOA images for each pixel to obtain the final composite image. For Sentinel-1 images, we converted them to the backscatter coefficients and obtained the composite Sentinel-1 image by calculating the mean values of all images for each pixel. Finally, the VIIRS nighttime images were aggregated into one composite image by taking the maximum value for each pixel. In addition, we acquired GISA that contains global annual impervious surface areas with a spatial resolution of 30 m from 1972 to 2019 from the website http://irsip.whu.edu.cn/resources/gisa.html (Huang et al., 2021). Finally, we obtained the road layer from OpenStreetMap (OSM) (Haklay and Weber, 2008), and ensured that it is close to the year 2015 by careful visual inspection. Note that these datasets were only used in the random forest method to estimate building height (see Section 3.3).

3. Methodology

3.1. Overview

ZY-3 high-resolution images have potential for predicting building height at a high spatial resolution, i.e., 2.5 m, which is able to reflect the individual building height. Accordingly, we proposed a multi-spectral, multi-view, and multi-task deep network (called M^{3}Net) that fuses ZY-3 multi-spectral and multi-view images in a multi-task learning framework to estimate building height. For the purposed of comparison, we also implemented a random forest (RF) method using multi-source features, which has been successfully applied to building height regression at relatively coarse scales (Geiß et al., 2020; Li et al., 2020a). Compared to the random forest using the hand-crafted features, the M^{3}Net can automatically and adaptively learn and extract features (e.g., texture and shape) from high-resolution images to predict building height. To assess the accuracy of the methods, we adopted the widely-used root mean square error (RMSE) metric ($RMSE = \sqrt{\frac{1}{N}\sum_{i=1}^{N}\left(\hat{y}_i - y_i\right)^2}$) to measure the difference between predicted ($\hat{y}$) and reference (y) building height values over all observations (with the number of N).

3.2. M^3Net for building height estimation

We proposed a multi-spectral, multi-view, and multi-task deep network (called M^3Net) for building height estimation (Section 3.2.1). A new loss function was proposed as the objective function of the network to weigh the losses of both building height estimation and building footprint extraction tasks based on the uncertainty of each task (Section 3.2.2). Under the experimental settings (Section 3.2.3), we trained the network and generated the final building height map at the spatial resolution of 2.5 m.

3.2.1. Structure of the M^3Net

Fig. 4 shows the structure of the proposed M^3Net, consisting of one branch for learning multi-spectral images (red/green/blue/near-infrared bands) and the other for multi-view images (nadir/forward/backward images). We adopted U-Net (Ronneberger et al., 2015), an efficient and widely-used encoder-decoder network with skip connections, as the basis of each branch. Encoder-decoder network is sequentially composed of two parts: 1) the encoder that compresses the input of arbitrary size into a feature representation and captures multi-level context information; 2) the decoder that recovers spatial details and predicts the output of the same size as the input from the feature representation. Compared to the classification network that yields a single label from an input of fixed size, the encoder-decoder network can obtain pixel-wise prediction from an input of arbitrary size. However, the feature representation usually suffers from low spatial resolution, leading to blurry boundaries in the output (Ma et al., 2019). Therefore, in order to retain fine-grained details for predicting the output, skip connection is introduced to reuse the features from the encoder by directly concatenating them to the decoder.

In the proposed network, the main components of the encoder-decoder structure include: 1) the convolution layer (written as Conv) that generates feature maps by convolving the input images with filter kernels (e.g., 3×3); 2) the max pooling layer that applies the max operation on a small neighborhood to down sample feature maps to improve local translation invariance; 3) the batch normalization layer (BN) that normalizes feature maps for each training mini-batch to lower internal covariate shift; 4) the transposed convolutional layer (Trans-conv, known as deconvolution) that can enlarge the feature maps through transposed convolution operations with filter kernels; and 5) the rectified linear unit (ReLU) activation function that enables the nonlinear modelling of the network by leaving positive values unchanged and setting negative values to zero.

For the multi-view branch (Fig. 4(a)), we used the combination of Conv, BN, and ReLU twice to map the input images to feature maps, and then a 2×2 max pooling operation was applied to down sample the feature maps in the encoding stage. The process was repeated four times and we doubled the number of feature channels after each max pooling operation. In the decoding stage, the size of feature maps was doubled each time by using the Trans-conv operation. Via the skip connection, we concatenated the feature maps from the Trans-conv operation and the same-scale feature maps from the encoder stage, and applied the combination of Conv and ReLU twice to yield new feature maps. We repeated the process four times and the number of feature channels was halved after each Trans-conv operation. Finally, a 1×1 Conv was used to generate the building height map. Through the multi-view branch, radiative and structural characteristics of buildings, e.g., their materials and sides, presented in multi-view images were automatically encoded to estimate building height.

The components of the multi-spectral branch (Fig. 4(b)) are similar to the multi-view branch. The main difference lies in the number of channels of the input images, i.e., four channels (red/green/blue/near-infrared bands) for the multi-spectral branch and three channels (nadir/forward/backward images) for the multi-view branch. Note that compared to putting all the images into one encoder, we adopted the network architecture of two encoders, for multi-spectral and multi-view images, respectively, which makes each encoder focus on a certain type of input. The two branches were then fused at intermediate and final stages of the whole model, in order to improve the reliability of height estimation. Particularly, for the multi-spectral branch, we reused the deepest feature map from multi-spectral images, i.e., the last layer in the encoding part, as the input of two decoders, which are designed to predict building height and footprints, respectively. Building height prediction and building footprint extraction tasks were learned simultaneously in order to boost the performance of single task. This technique is referred as the multi-task learning that aims to improve generalization by learning multiple related tasks in parallel (Caruana, 1997). In contrast to the single task learning, the multi-task learning can exploit the mutual information between multiple tasks and therefore is promising to increase the performance of each task. In this study, the second task, i.e., building footprint extraction, was regarded as additional supervision to support the optimization of the main task (i.e., building height estimation), by restricting the space of possible solutions. Finally, building height from the multi-view branch (written as BH1, see Fig. 4), and building height (BH2) and building footprints (BF) from the multi-spectral branch were concatenated to predict the final building height (FBH).

In summary, the contribution of the M^3Net is twofold. First, ZY-3 satellites can simultaneously provide multi-spectral images that contain rich spectral and textural information, and multi-view images that can describe vertical features of ground objects at a fine spatial resolution. The two kinds of images were learned by two encoders, respectively, and then the learned feature representations were fused to predict the final building height. To the best of the authors' knowledge, this is the first time that ZY-3 images were applied to the deep learning-based building height estimation. Second, we designed a multi-task learning framework, i.e., both building height estimation and building footprint extraction tasks were simultaneously learned by the proposed network, in order to boost the performance of the single one.

3.2.2. Weighted loss function

Instead of merely predicting FBH, we considered four maps, i.e., BH1, BH2, FBH, and BF as mentioned in the former section and optimized them using the loss function that aims to reduce the difference between the prediction and the reference by updating network parameters. However, it is challenging to appropriately weigh the loss of each map. Manual tuning is difficult and time-consuming to search for the optimal weights. To cope with this issue, we designed a weighted loss function based on task uncertainty (Kendall et al., 2017) to simultaneously minimize the prediction errors of four maps, i.e., BH1, BH2, FBH, and BF. This allows our network to learn the weight of each map dynamically and automatically, which, therefore, makes the weight selection more convenient and efficient. The weighted loss function is defined as:

$$\mathrm{Loss} = \sum_{i=1}^{4} (w_i L_i + r_i) \tag{1}$$

for the building height regression task (i.e., BH1, BH2, FBH):

$$w = \frac{exp(-log\sigma^2)}{2}, r = \frac{log\sigma^2}{2}$$

and for the building footprint extraction task (i.e., BF):

$$w = exp\left(-log\sigma^2\right), r = \frac{log\sigma^2}{2}$$

where L_i is the loss function for task i, w_i is the weight term, r_i represents the regularization term, and σ denotes the task uncertainty. The mean square error (MSE) is a widely-used regression loss function (Carvalho et al., 2018), and therefore was used for the building height regression task. The MSE is formulated as:

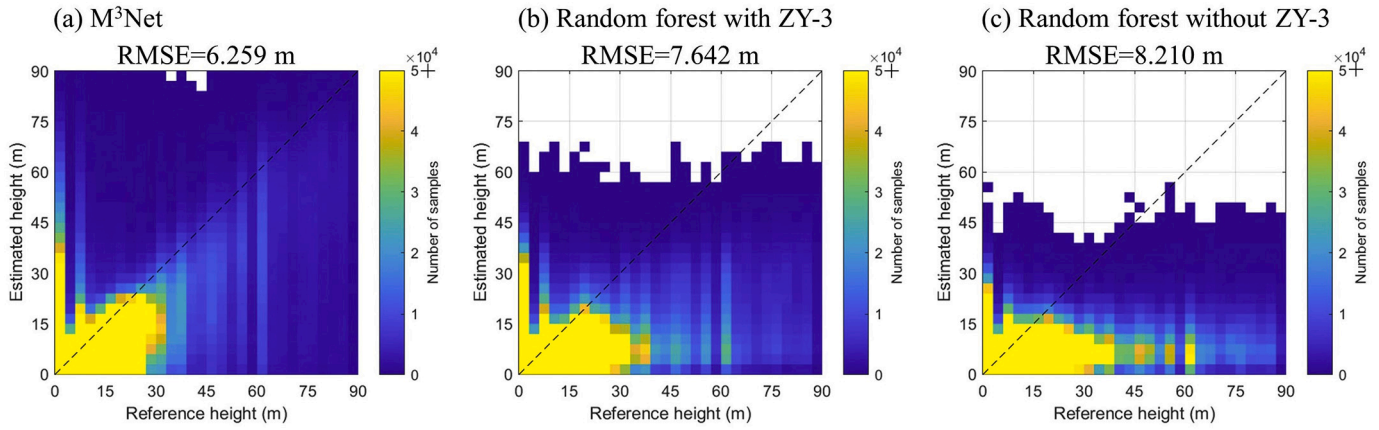

Fig. 5. Accuracies of building height prediction by (a) the M^{3}Net, (b) the random forest with ZY-3 images, and (c) the random forest without ZY-3 images. RMSE: root mean square error.

$$L_{MSE} = \frac{1}{N}\sum_{i=1}^{N}(\hat{y}_i - y_i)^2 \quad (2)$$

where $\hat{y}_i$ and y_i represent the predicted and the reference building height values for the observation i, and N is the number of all observations. In the building footprint extraction task, we used the popular binary cross entropy (BCE) loss function (De Boer et al., 2005):

$$L_{BCE} = -\frac{1}{N}\sum_{i=1}^{N}[y_i log(p_i) + (1-y_i) log(1-p_i)] \quad (3)$$

where y_i represents the reference building footprint label (1 for buildings and 0 for non-buildings) for the observation i, and p_i denotes the predicted probability of building footprints for the observation i. Note that the task uncertainty σ measures relative balancing of the tasks. Thus, one of the four uncertainty parameters (for BH1, BH2, FBH, and BF tasks) is fixed, and the other three ones are adaptively adjusted through the learning of the network. Specifically, in this study, we fixed the uncertainty parameter of the building footprint extraction (i.e., BF), in order to assess the effect of ZY-3 images on building height prediction (Section 5.1).

3.2.3. Network settings

We trained the M^{3}Net on the training set (Section 2) for 300 epochs with the initial learning rate of 0.001 decayed by factor 0.1 at epochs 200 and 250. The Adam optimizer (Kingma and Ba, 2015) was selected to optimize the network parameters, and the batch size was set to 16. The training procedures were implemented by the Pytorch framework and conducted on a Personal Computer with Intel Core CPU i9–7980XE at 2.60 GHz and a single NVIDIA GTX 1080 Ti GPU. In general, deep learning networks heavily rely on large amounts of training samples to avoid overfitting and improve the generalization ability (Shorten and Khoshgoftaar, 2019). However, high quality training samples are usually expensive and limited. To alleviate this issue, we enhanced the size of the training set by employing the data augmentation strategy (Buslaev et al., 2018) with a probability of 0.5, including image flipping horizontally and vertically, rotating by [0, 180°] with 15° interval, and colour space adjustment through stretching the input image to a range [0, λ] with a random value $\lambda \in$ [0.5,1].

3.3. Random forest building height estimation: A comparison method

As a comparison, we also carried out the random forest algorithm with two steps for building height estimation, which is briefly introduced as follows.

Step 1. Feature extraction. With ZY-3 images, we extracted the normalized difference vegetation index (NDVI) and the normalized difference water index (NDWI), since these two features can highlight non-urban areas (e.g., vegetation and water) (Szabo et al., 2016). MABI obtained from ZY-3 multi-view images, is able to denote the angular variations of buildings, and hence, is informative for the height estimation (Liu et al., 2019). The VV and VH bands of Sentinel-1 were used since they are strongly related to building height (Koppel et al., 2017). Sentinel-2, VIIRS, and ALOS World 3D data were also included given that they have been considered useful to indicate building areas and their height (Geiß et al., 2019; Levin and Zhang, 2017; Mushore et al., 2017; Takaku et al., 2020). In this study, we used ten bands of Sentinel-2 (i.e., four 10-m bands and six 20-m bands) and four indices extracted from Sentinel-2 including normalized difference built-up index (NDBI), urban index (UI), bare soil index (BI), and NDVI (Mushore et al., 2017). For the ALOS data, we extracted slope, aspect (Burrough and McDonnell, 1998), and normalized digital surface model. We took GISA as urban areas and removed roads within urban areas by using OSM, to alleviate the effect of non-building areas (e.g., vegetation and viaducts). Overall, we focused on 31 explanatory variables, including 21 images and 10 indices aforementioned, which were resampled to 2.5 m. All these explanatory variables were put together by layer stacking.

Step 2. Random forest regression. With all the explanatory variables and the reference building height, we trained a random forest regression model (Breiman, 2001) on the training set. The key parameters of random forest include the number of trees (denoted as *ntree*) and the number of features used for training each tree (denoted as *mtry*). In this study, the *mtry* was set to the default value, i.e., 1/3 of the total number of features. We searched for the *ntree* from 50 to 1000 with an interval of 50, and selected the optimal value with the lowest root mean square error based on the validation set.

4. Results

4.1. Performance of the M^{3}Net in 42 Chinese cites

Fig. 5 displays the accuracies of building height prediction results by the proposed M^{3}Net as well as the random forest (RF) method on the test set. In general, the M^{3}Net obtains lower root mean square errors (RMSE) than the RF, and the building height estimated by the M^{3}Net shows better agreement with the reference height compared to that predicted by the RF. Particularly, the RF model tends to underestimate the height

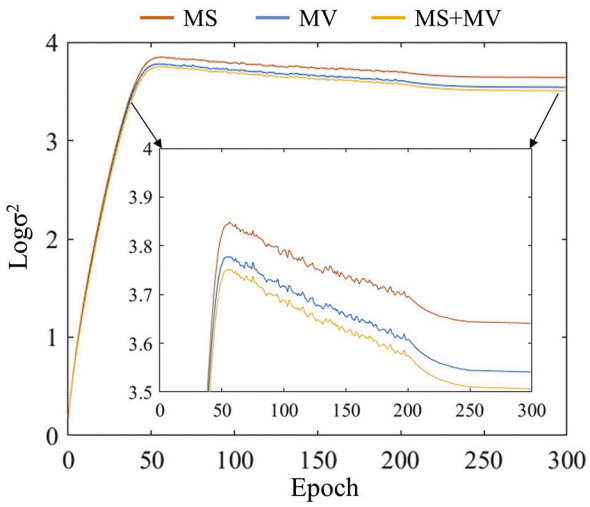

Fig. 6. Uncertainties of building height prediction with multi-spectral (MS), multi-view (MV), and all images (MS + MV).

of high-rise buildings (e.g., higher than 30 m), while the M^{3}Net can mitigate the saturation effect in estimating these high-rise buildings, indicating its better performance for building height estimation. The superiority of the M^{3}Net in estimating high-rise buildings can be also observed at coarse scales (see Figs. 12 and 14). In addition, the RF with ZY-3 images obtained a lower RMSE than that without ZY-3 images, verifying the effectiveness of ZY-3 images on the building height estimation (further analysis is given in Section 5.1).

Fig. 6 displays the curves of the task uncertainty during the training phase of the M^{3}Net. As mentioned in Section 3.2.2, we fixed the uncertainty parameter of the building footprint extraction task, and only adjusted the uncertainty parameters of the building height prediction tasks, including BH1 from the multi-view (MV) branch, BH2 from the multi-spectral (MS) branch, and FBH from the fusion of the two branches (MS + MV). To achieve better numerical stability, we optimized Logσ^2 instead of σ. It can be seen that the fused method (MS + MV) obtains the lowest task uncertainty, while the multi-spectral branch (MS) gives the largest task uncertainty. These results clearly demonstrate the effects of multi-view images on reducing the uncertainty of building height prediction.

For the purpose of visual inspection, Fig. 7 illustrates the estimated building height maps by the proposed M^{3}Net as well as the RF model in Beijing, Shenyang, and Urumqi. The building styles in the three areas are diverse due to their different urbanization process and urban planning policies. In the area of Beijing, building height values estimated by the RF show good agreement with those estimated by the M^{3}Net, especially in the low-rise building areas. However, for the RF, severe underestimations are clearly observed in the high-rise building areas, but this issue can be largely alleviated by the M^{3}Net. The same phenomenon is presented in the areas of Shenyang and Urumqi. Moreover, more spatial details, e.g., individual buildings (see zoomed-in areas in Fig. 7), are exhibited, due to the use of high-resolution ZY-3 images.

Moreover, the result of a typical area was visualized to show the change of building height from the center to the periphery (Fig. 8).

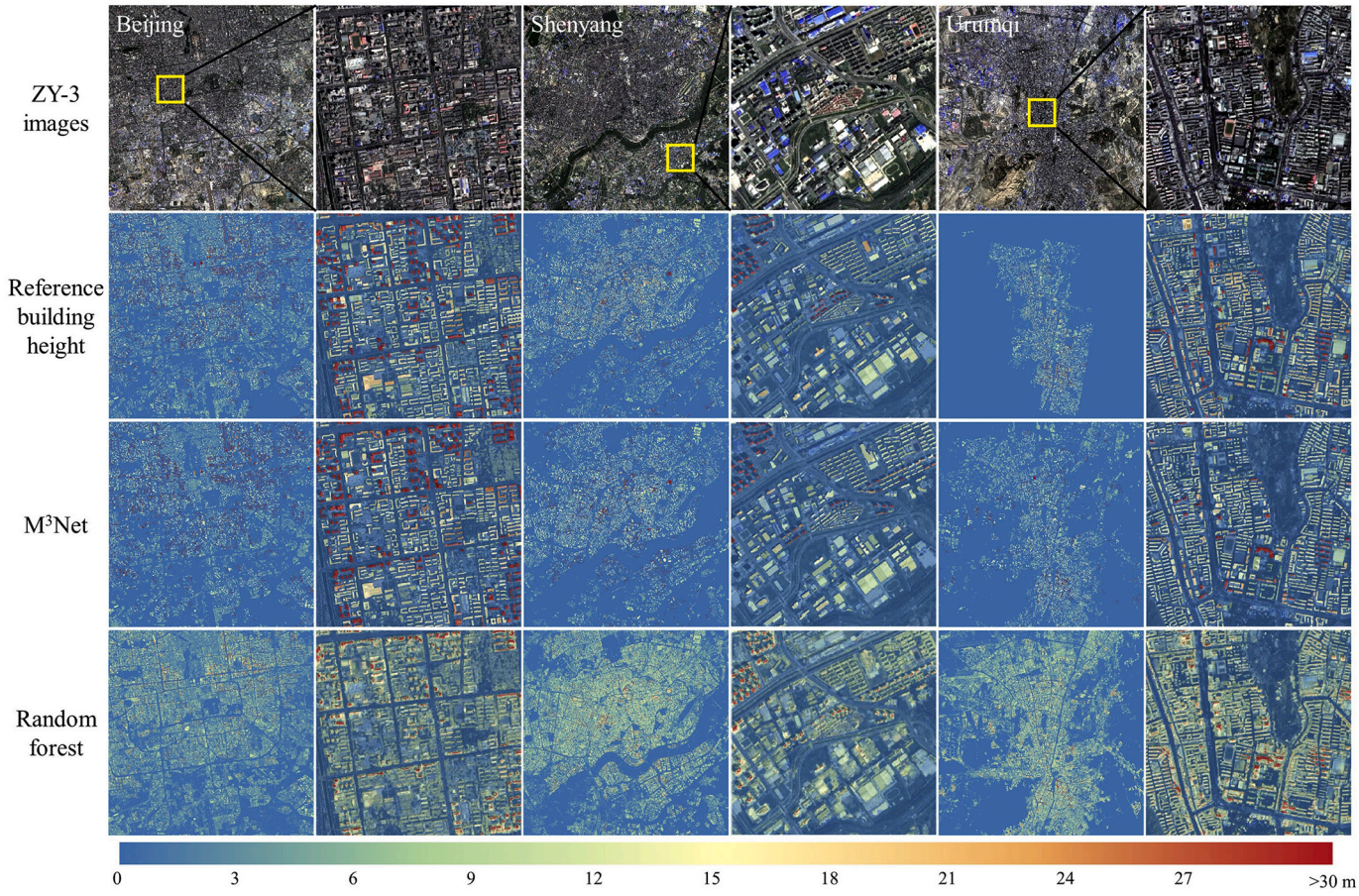

Fig. 7. Building height with a spatial extent of 20 km × 20 km predicted by the M^{3}Net and the random forest in Beijing, Shenyang, and Urumqi. Each zoomed-in area indicated by the yellow rectangle has a spatial extent of 2 km × 2 km. (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article.)

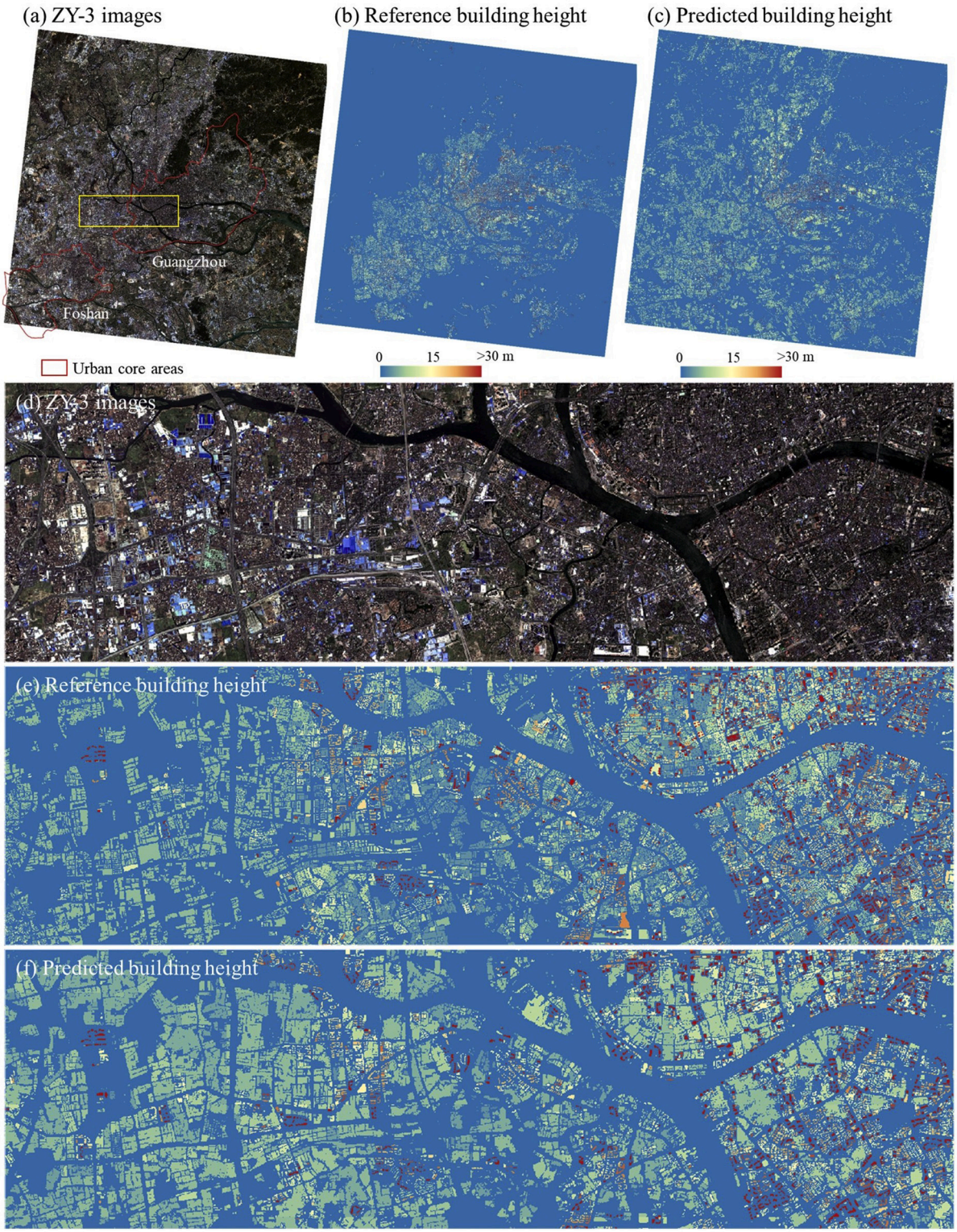

Fig. 8. Results of the Guangzhou city. (a) ZY-3 images covering about 2612 km^2. (b) Reference building height from A-map. (c) Predicted building height by the M^{3}Net. (d) The zoomed-in area (the yellow rectangle in graph (a)) with a spatial extent of 5 km × 17 km. (e) and (f) are the reference and the predicted building height of graph (d), respectively. (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article.)

Specifically, we collected a ZY-3 image scene acquired on April 14, 2015 covering approximately 2612 km^2. This image scene encompasses urban core areas (see the red polygons in Fig. 8(a)) as well as the peripheries of Guangzhou and Foshan cities. Building height reference data (Fig. 8(b)) was provided by A-map, and was mainly distributed in the urban core areas and their surroundings. The predicted result is shown in Fig. 8(c). Overall, the predicted building height is in good agreement with the reference, and we can observe a significant decrease trend of building height from the center areas to the peripheries, which is consistent with previous studies (Lin et al., 2020; Sun and Li, 2020). Furthermore, a zoomed-in area (the yellow rectangle in Fig. 8(a)) with a spatial extent of 5 km × 17 km is displayed in Fig. 8(d). It can be seen that the average

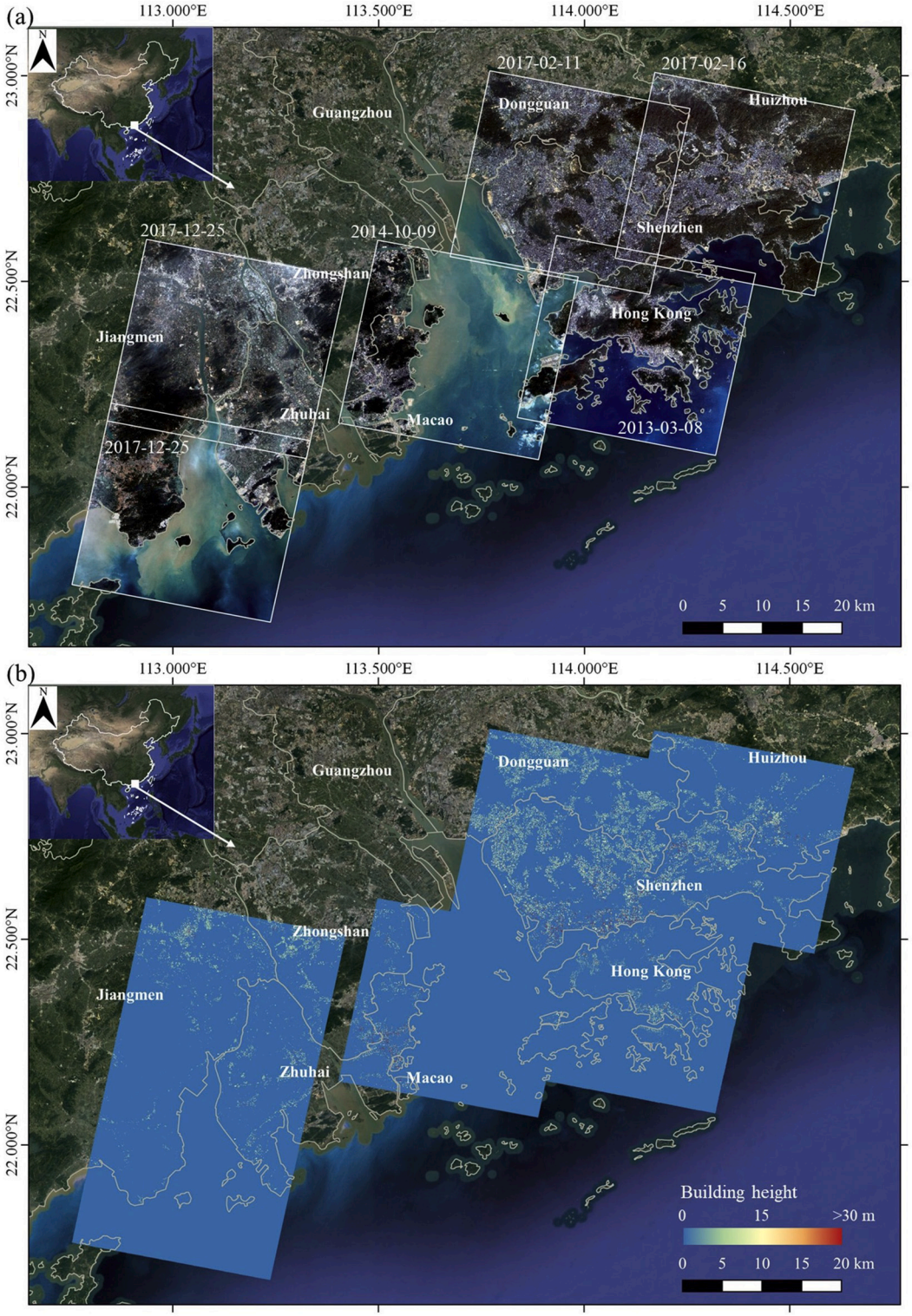

Fig. 9. Results in the southern coastline of China. (a) ZY-3 images. (b) Predicted building height by the M^{3}Net. Base map: Google Maps.

height of buildings gradually decreases from east (the urban core areas) to west (the peripheries). In addition, rich spatial details of buildings and non-buildings (e.g., water and roads) are exhibited in Fig. 8(f) by courtesy of high-resolution ZY-3 images. These results corroborate the effectiveness of the proposed method (i.e., the M^{3}Net) for high-resolution building height estimation.

4.2. Building height at a relatively large area

To evaluate the scalability of the proposed method to large areas, we applied our method to a relatively large area located in the Delta of the Pearl River, south of China (Fig. 9(a)). Six ZY-3 image scenes were collected from the Land Satellite Remote Sensing Application Center (LASAC) of China, covering about 14,120 km^2 and acquired between 2013 and 2017 with cloud cover <10%. Each ZY-3 image scene contains

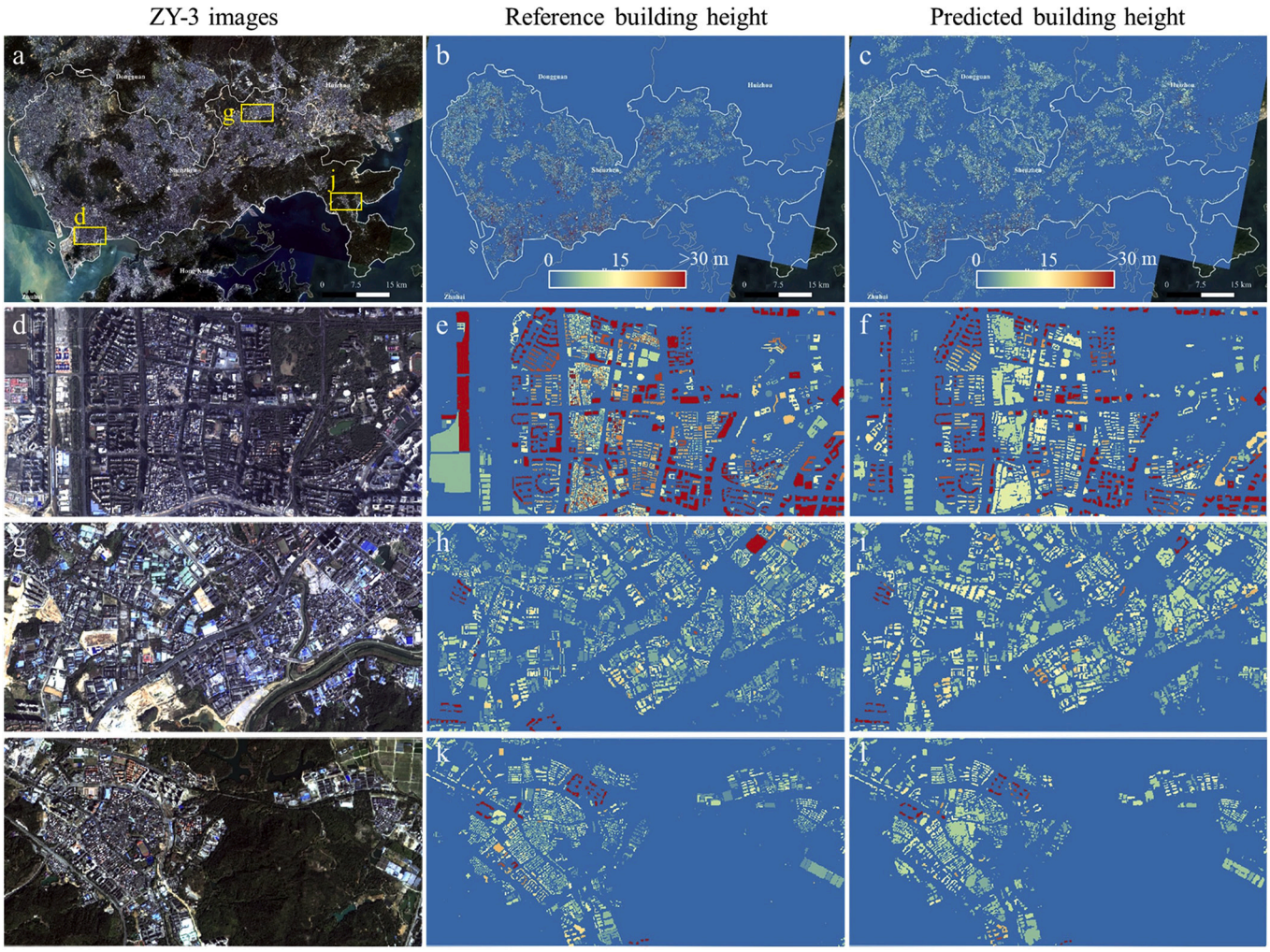

Fig. 10. Results of the Shenzhen city. (a) ZY-3 images covering the whole Shenzhen city. The white line in graph (a) denotes the city's boundary. (b) Reference building height from airborne LiDAR. (c) Predicted building height by the M^{3}Net. (d), (g) and (j) show the zoomed-in areas of the yellow rectangles in graph (a). Each zoomed-in area has a spatial extent of 2 km by 4 km. (e), (h), and (k) correspond to the reference building height of graphs (d), (g), and (j), respectively. (f), (i), and (l) correspond to the predicted building height of graphs (d), (g), and (j), respectively. (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article.)

four images with a ground swath of about 50 km, including the multi-spectral images (blue, green, red, and near-infrared bands with a spatial resolution of 5.8 m) and three panchromatic images (with nadir (2.1 m), +22° forward (2.5–3.5 m), and − 22° backward (2.5–3.5 m) viewing angles, respectively). All the images were preprocessed with radiometric correction, ortho-rectification, image-to-image registration, and pan-sharpening (Liu et al., 2019). Each image scene was then resampled to 2.5 m and has an average size of 24,029 × 23,884 pixels. We applied the proposed M^{3}Net model to this test area. The predicted result is shown in Fig. 9(b). It can be seen that building areas and non-building areas (e.g., water and forest) are well distinguished. Most of buildings are located in the north of Shenzhen, where the average height value is about 15 m (~5 floors). High-rise building areas are mainly distributed in the southwest of Shenzhen, and their height value is over 30 m on average. The runtime of the model is about 1.046 s for an area of 100 km^2 on a single GTX 1080 Ti GPU. The total inference time for the six images is approximately 229 s. Note that if more GPUs are available, the inference time can be further reduced in a parallel manner.

Moreover, the building height of the whole Shenzhen city acquired from airborne LiDAR data in 2017 was used as the reference to evaluate the quality of the produced building height. Note that airborne LiDAR can provide highly accurate elevation measurements and has been widely adopted as the building height reference data (Bonczak and

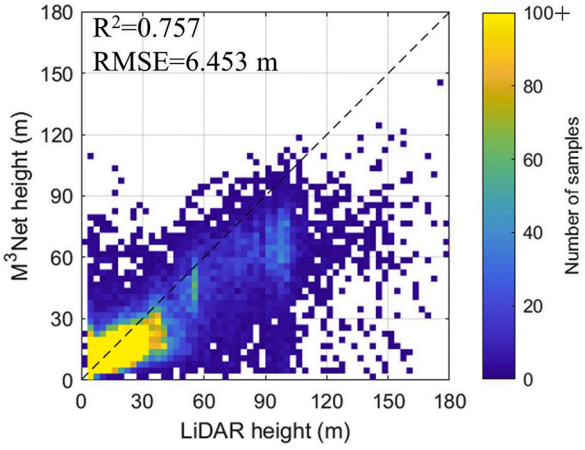

Fig. 11. Accuracy of building height predicted by the M^{3}Net in the Shenzhen city. RMSE: root mean square error.

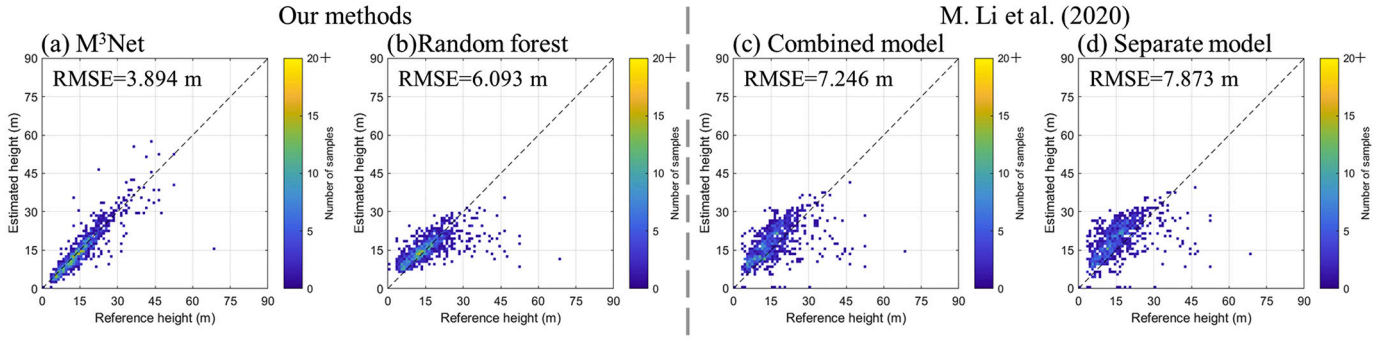

Fig. 12. Accuracies of building height predicted by (a-b) our methods and (c-d) those of Li et al., 2020a at 1-km scale. RMSE: root mean square error.

Kontokosta, 2019; Rottensteiner, 2003; Wang and Li, 2020; Yu et al., 2010). The acquired reference data was originally provided in the vector form, i.e., building footprints with height, and was then converted to its raster form with a spatial resolution of 2.5 m (Fig. 10(b)), for a comparison with the predicted height map. As depicted in Fig. 11, the RMSE of the proposed M^{3}Net is 6.453 m, showing its good agreement with the LiDAR height reference. Fig. 10(c) displays the predicted height by the M^{3}Net. We can observe that the M^{3}Net can effectively predict the height values for most of buildings in the high-rise, middle-rise, and low-rise building areas, corresponding to graphs (d), (g), and (j) in Fig. 10, respectively. Overall, from the perspectives of both efficiency (runtime) and quality, the aforementioned results show that the proposed M^{3}Net is promising for dealing with large-area building height estimation.

5. Discussions

5.1. Effects of ZY-3 images

ZY-3 satellites can simultaneously acquire high-resolution multi-spectral (5.8 m) and multi-view images with nadir (2.1 m), +22° forward (2.5–3.5 m), and − 22° backward (2.5–3.5 m) viewing angles over the same area. Multi-spectral images can provide rich spectral and textural features, while multi-view images can describe vertical information of the ground objects. Previous studies have successfully applied ZY-3 images to urban scene classification (Huang et al., 2018) for identifying high buildings and low-rise urban villages, and extracting built-up areas with the multi-angular features (Liu et al., 2019). However, multi-view satellite images have not been considered in the building height regression, and their capability of height estimation remains unknown.

Therefore, we jointly used multi-spectral and multi-view images by the proposed M^{3}Net for building height estimation, and designed a weighted loss function based on task uncertainty to automatically weigh the importance of building height estimated from the multi-view images, multi-spectral images, and their combination. The uncertainty curve (Fig. 6) indicated that in terms of building height estimation, multi-view images perform better than multi-spectral images, and it is necessary to synergistically use the two information sources. Moreover, we compared the random forest with and without ZY-3 images, and results (Fig. 5(b-c)) verified that the inclusion of ZY-3 images can significantly improve the accuracy of building height estimation. According to the visualized results (Fig. 7), the introduction of high-resolution ZY-3 images provided rich spatial details, e.g., buildings (with different colors, shapes, and sizes), shadow, roads and vegetation, which is helpful for discriminating between buildings and non-buildings, and more accurately estimating the height of buildings with different characteristics.

5.2. Comparison with existing studies

To further evaluate the performance of the proposed M^{3}Net, we compared it with three existing state-of-the-art methods for building height prediction:1) the random forest method (at the scale of 1-km) by using multi-source data, including Landsat-8 optical images, Sentinel-1 radar images, Enhanced Vegetation Index (EVI) from MODIS products, VIIRS, Global Urban Footprint (GUF) and OSM (Li et al., 2020a); 2) the indicator-based model integrating VH and VV information of Sentinel-1 radar data at the scale of 500-m (Li et al., 2020b); and 3) the vanilla single/multi-task deep learning models (Carvalho et al., 2019). When analyzing the first two methods, we also presented the result of our random forest approach (see Section 3.3) for a comparison. To unify the spatial resolution (or scale) among different methods and results, we adopted the spatial aggregation approach: 1) for the predicted building height (2.5 m) from the M^{3}Net, we aggregated it to a target scale (1-km or 500-m), according to the definition of building height (Li et al., 2020a; Li et al., 2020b); 2) as for our RF model (Section 3.3), we first aggregated all the explanatory variables to the target scale (e.g., 500-m or 1-km) by calculating the mean values in a pixel, and then trained a random forest model to predict the building height at the target scale. The building height reference data (Section 2) was also aggregated to the corresponding target scale for accuracy assessment.

In Li et al., 2020a, two models were developed: 1) a separate model trained for each region (e.g., China, the US, and Europe) and 2) a combined model developed for all the three regions. They found that the accuracies of the two models were comparable, but the combined model had a lower uncertainty than the separate model. They released the final building height maps (with a spatial resolution of 1-km) generated by both models in the three regions, where the building height was defined as the area-weighted height of all buildings in a pixel. We clipped the maps to our study areas, and compared our results with the maps produced by Li et al., 2020a at 1-km scale. It should be noted that the predictions are comparable, since we used the same building height reference as that adopted by Li et al., 2020a. As shown in Fig. 12, our methods obtain a lower RMSE than those developed by Li et al., 2020a. This is mostly due to the fact that our methods introduced high-resolution ZY-3 multi-spectral and multi-view images that can provide rich spectral, textural and vertical information of buildings. Moreover, we can observe that the M^{3}Net can alleviate the saturation effect of high-rise building estimation compared to the random forest method.

Fig. 13 compared the building height maps by our methods and those of Li et al., 2020a. In general, both results show similar trend in most areas. Differences mainly occur in areas with a mixture of high-rise buildings and bare land (see the second row in Fig. 13). The models of Li et al., 2020a tend to underestimate these areas, mostly owing to the relatively coarse resolution of images. In contrast, our methods can alleviate this issue, by courtesy of ZY-3 high-resolution images. Moreover, we can find that the M^{3}Net is more effective in both sparsely and densely populated areas compared to other ones. The main reason lies in

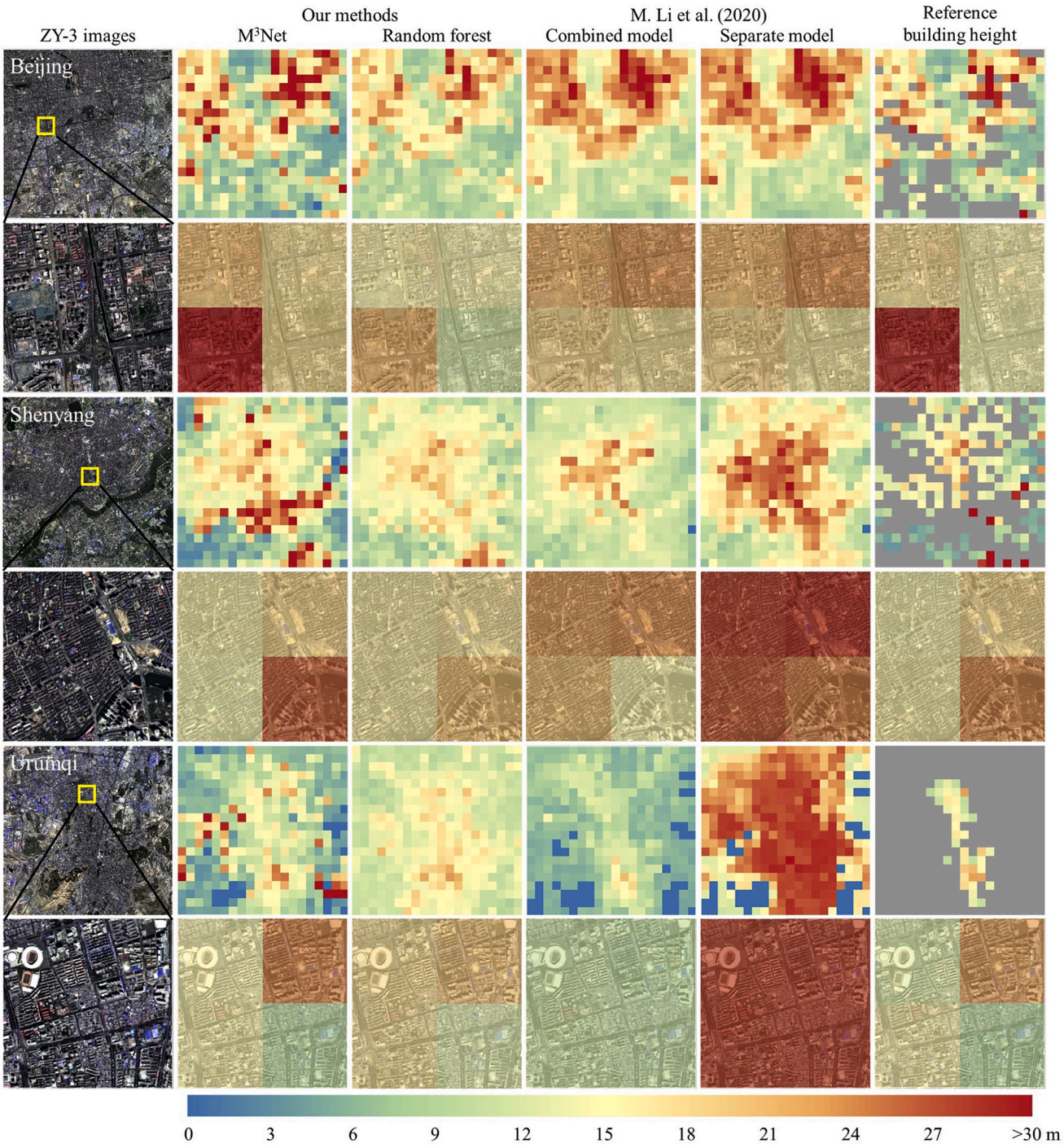

Fig. 13. Comparison of building height maps predicted by our methods and those of Li et al., 2020a in Beijing, Shenyang, and Urumqi. No data areas are colored as grey in the reference building height maps.

the multiscale and multilevel feature representation ability of the deep learning network (Yuan et al., 2020), in comparison with the hand-crafted feature extraction used by other methods. Moreover, for the M^{3}Net, the building height at 2.5-m scale was first generated and then aggregated to 1 km, while other methods predicted building height using images that were aggregated to 1 km. The "predict-then-aggregate" strategy adopted by the M^{3}Net can better exploit the spatial details and hence reduce the mixture effect within 1-km grid. This phenomenon was also mentioned by Frantz et al. (2021).

Li et al. (2020b) developed an indicator-based model by relating the log-transformed building height (written as ln H) to the VV and VH backscatter coefficients of Sentinel-1 radar data at 500-m scale (see the developed model in Fig. 14(d)). For the purpose of comparison, here the indicator model proposed by Li et al. (2020b) was used to predict the building height of 42 China's cities, based on the China's training samples, at the resolution of 500 m. Specifically, based on the China's samples, the parameters of the indicator model, i.e., a, b, c, and γ, were determined as 21.183, 0.055, −19.785, and 5, respectively. Note that, according to Li et al. (2020b), building height was defined as the mean height of building and non-buildings within the 500-m grid. Therefore, to ensure a fair comparison, we also adopted the same definition in this experiment. As illustrated in Fig. 14, our methods achieve a lower RMSE

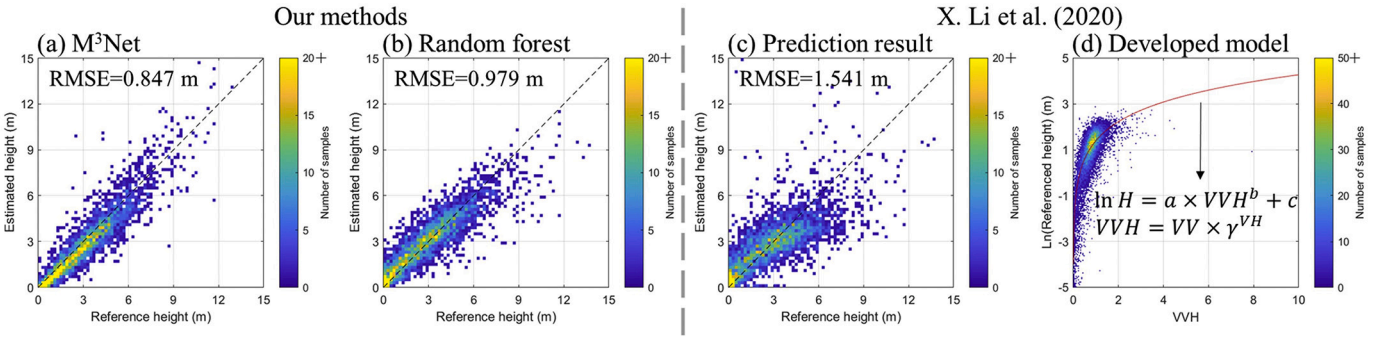

Fig. 14. Accuracies of building height predicted by (a-b) our methods and (c-d) the method of Li et al. (2020b) at 500-m scale. RMSE: root mean square error.

than the indicator model, and the majority of estimated building height values are located along the one-to-one line, suggesting the superiority of our methods.

Fig. 15 displays the building height maps in Beijing, Shenyang, and Urumqi, for a visual inspection. Overall, building height maps generated by Li et al. (2020b) and our methods are congruent in most areas. The inconsistency mainly exists in the high-rise building areas (see the close-ups in Fig. 15). For the indicator model of Li et al. (2020b), although the backscatter coefficients from Sentinel-1 radar data can capture variations of height, they could be different for buildings with similar heights but at the same time similar for buildings with different heights, which is caused by complex and diverse building materials and layouts, and limited information from the radar data (Koppel et al., 2017). In this regard, multi-source data (e.g., optical, multi-view, radar, and nighttime light) were considered in our RF method, in order to alleviate the uncertainty of individual feature. However, as aforementioned, we can still observe that, the RF model has the tendency to underestimate the height of high-rise buildings (see the close-ups in Fig. 15). By contrast, the M^{3}Net can well capture the variations of height from high-rise to low-rise building areas, owing to its powerful feature representation ability and the adopted "predict-then-aggregate" strategy, as mentioned in the comparison with the methods of Li et al. (2020a).

Finally, we compared the proposed M^{3}Net with the vanilla single/multi-task models (Carvalho et al., 2019) (Fig. S1). The vanilla single-task model used one encoder to process all images, and one decoder to predict building height, while the vanilla multi-task model adopted two decoders for building height estimation and building footprint extraction, respectively. Note that the main difference between the vanilla multi-task model and our method (the M^{3}Net) is that the former only adopted one encoder to process all images, while the latter fused two encoders to extract features from multi-spectral and multi-view images, respectively. Results show that our network outperforms the vanilla single/multi-task models, and the multi-task structure is superior to the single-task one in terms of building height prediction (Fig. S2). Details are provided in the Appendix A.

5.3. Spatial-temporal transferability

The spatial-temporal transferability refers to the ability of models trained on a certain time and a certain region to generalize to a new time and a new region, and this ability is vital in the automated large-scale mapping task. It is challenging to achieve an effective spatial-temporal transferability due to the differences in the complex imaging conditions (e.g., illumination and atmospheric effects) and diverse building types. In this research, to test the temporal transferability, we collected 1131 new reference building height samples (a spatial extent of 1 km × 1 km for each one) in five cities of China (Beijing, Shanghai, Shenzhen, Wuhan, and Xi'an) (Table 2). These samples collected on new dates covered the same areas as those used to train the M^{3}Net, in order to ensure that the difference between new and old images is merely the acquisition time. On the other hand, to test the spatial transferability, we acquired 1931 reference building height samples (a spatial extent of 1 km × 1 km for each one) in three cities of the United States (US) (Chicago, Des Moines, and Los Angeles). The reference building height data are publicly accessible from the local governments, and data sources are listed in Table 2. As depicted in Fig. 16, we can observe that buildings of the US are different from those of China in their colour, size, height, and spatial distribution, which, therefore, is suitable for assessing the robustness of the network to diverse building styles.

According to the temporal transferability result shown in Fig. 17, RMSE of the network increases less than 1.5 m for all the five cities when the network was applied to images on new dates, which shows the robustness of the proposed M^{3}Net to different imaging conditions to some extent.

With respect to the spatial transferability, we compared the M^{3}Net under four settings: 1) the China model—the M^{3}Net trained only with the training set of China; 2) the US model—the M^{3}Net trained only with the training set of the US; 3) the combined model—the M^{3}Net trained with both training sets of the US and China from scratch; and 4) the fine-tuned model—the M^{3}Net trained with the training set of China but fine-tuned with the training set of the US. The difference between the combined model (setting 3) and the fine-tuned (setting 4) model lies in that the former was initialized with random weights, but the latter was initialized with weights trained on the training set of China. Therefore, compared to the combined model, the fine-tuned model already has the ability of building height estimation on the source region (China), but it needs to adapt to the target region (the US). In this way, the demand of training data and time can be significantly reduced. Note that the samples of the US were randomly split into spatially disjoint training (70%), validation (10%) and test sets (20%), which is the same as the sample selection of China (Section 2). Fig. 18 shows the RMSE of the four models on the test set of the US. It is interesting to see that even if the China model was directly applied to the three US cities (without any sample from the US), the RMSE values for all the cities are satisfactory (3.3 m at average), indicating the robustness of the model in the case of spatial transferability. We can also find that when using both training sets of China and the US (i.e., settings 3 and 4), the fine-tuned model (setting 4) obtains a lower RMSE compared to the combined model (setting 3), suggesting that the fine-tuned approach can adapt better to a new region. The fine-tuned approach is more efficient since it is built on a well-trained model (e.g., the China model).

5.4. Scalability and limitations

With regard to the scalability of our method (the M^{3}Net), we have tested it at a relatively large area at the southern coastline of China (Section 4.2). We used six ZY-3 images covering about 14,120 km^2, and the total inference time for the six images was approximately 229 s with a single GTX 1080 Ti GPU. The predicted building height in the whole Shenzhen city achieved RMSE of 6.453 m and showed good agreement

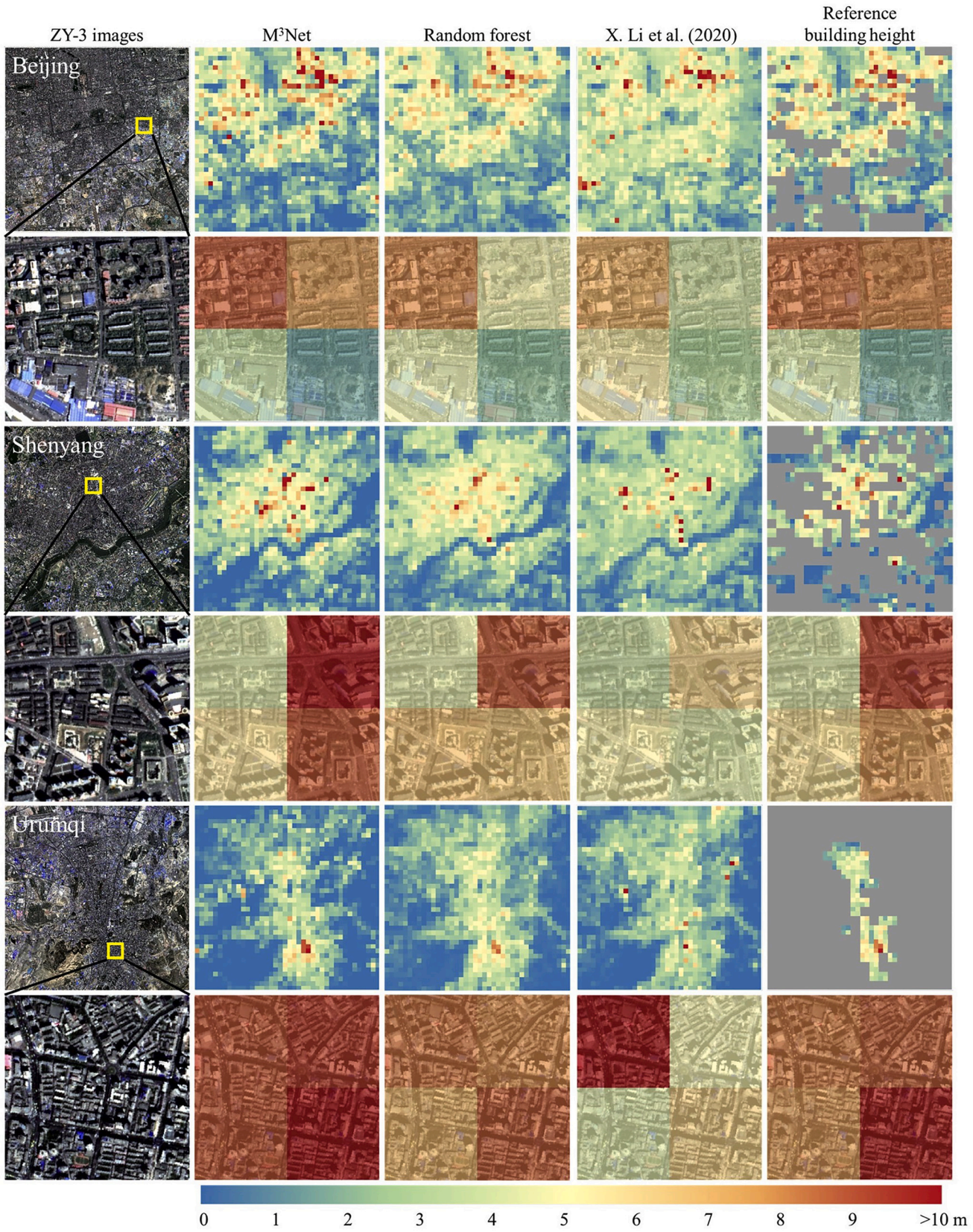

Fig. 15. Comparison of building height maps predicted by our methods and the model of Li et al. (2020b) in Beijing, Shenyang, and Urumqi. No data areas are colored as grey in the reference building height maps.

Table 2
Cities in China and the US for spatial-temporal transferability test.

	City	Old date	New date	#Sample	Sample source
Temporal test	Beijing	20,170,515	20,180,407	544	https://amap.com
	Shanghai	20,160,903	20,150,505	241	
	Shenzhen	20,170,211	20,131,223	131	
	Wuhan	20,160,124	20,170,112	123	
	Xi'an	20,150,512	20,170,627	92	
	Total			1131	
Spatial test	Chicago		20,200,307	474	https://clearinghouse.isgs.illinois.edu
	Des Moines		20,200,419	302	https://www.dsm.city/city_of_des_moines_gis_data
	Los Angeles		20,130,129	1155	https://egis-lacounty.hub.arcgis.com
	Total			1931	

with the LiDAR height reference (Figs. 10-11). Therefore, from the perspectives of both efficacy (runtime) and quality, the predicted results confirmed that it is potential to scale the M^{3}Net to large-area building height estimation. In future studies, we will attempt to apply our method to larger areas for high-resolution building height estimation.

There still exist some limitations in this study. Our test on rural landscape is limited. In this research, we focused on the urban areas, since they are the main places of human activities and contain diverse buildings with different colors, shapes, sizes, and height, and, hence, they are very suitable for testing the generalization ability of the proposed method. As for the rural areas, it is unfortunate that their reference data for building height is not available currently, and thus the test of our method on rural landscape is limited and difficult. In future, when the samples in rural areas are available, our method can be conveniently applied or evaluated. The second issue is the transferability of the M^{3}Net. Although the M^{3}Net showed good generalization ability across

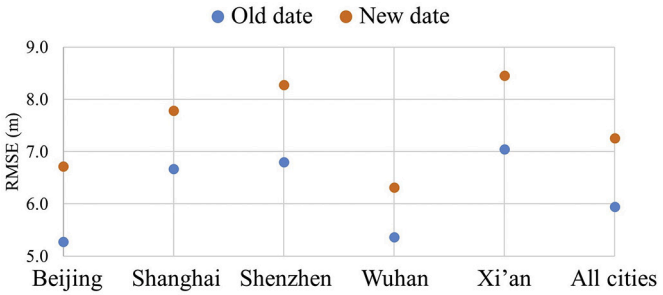

Fig. 17. Root mean square errors (RMSE) of the M^{3}Net in the five cities of China (Beijing, Shanghai, Shenzhen, Wuhan, and Xi'an).

Fig. 16. Data sources in the five cities of China (Beijing, Shanghai, Shenzhen, Wuhan, and Xi'an) and the three cities of the US (Chicago, Des Moines, and Los Angeles).

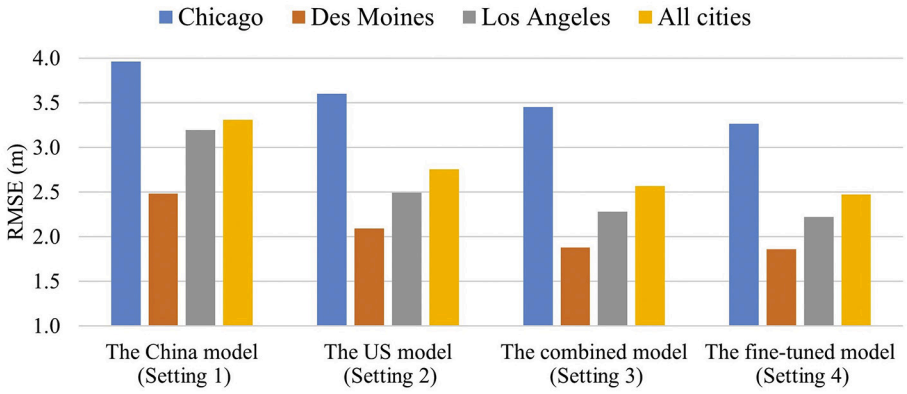

Fig. 18. Root mean square errors (RMSE) of four models in the three cities of the US (Chicago, Des Moines, and Los Angeles).

space and time in this study, it may not generalize well to new regions that are totally different from the study areas. In future research, we plan to consider the transfer learning technique that can adapt the knowledge from the source domain (the existing task) to the target domain (the new task) and alleviate the problem of the insufficient training data in the target domain (Pan and Yang, 2009; Tan et al., 2018). For instance, in the spatial transferability test (Section 5.3), we presented a fine-tuned approach to reuse the network learned from the training set of China (i.e., the source domain) and then fine-tuned the network on the training set of the US (i.e., the target domain). But more sophisticated transfer learning methods, e.g., adaptation of data distributions (Tuia et al., 2016), can be investigated to fully exploit the samples from the source domain to improve the network performance on the target domain in future work. The third issue is the data access limitation. ZY-3 imagery used in this research was obtained from the Land Satellite Remote Sensing Application Center (LASAC) of China, and this data source is not available publicly. The data access limitation makes it difficult to map high-resolution building height over large areas. However, we show the effects of ZY-3 images on estimating building height of 42 Chinese cities, which provides a valuable reference for the relevant research.

6. Conclusions

In this study, we aimed to estimate building height with high-resolution multi-view imagery in 42 Chinese cities. With respect to high-resolution building height estimation, most of the existing studies are limited to local or small areas, while the investigation across multiple cities is still lacking. High-resolution images hold potentials for estimating building height at a fine scale, while most studies focus on a coarse resolution. Multi-view satellite images can well characterize the vertical dimension of buildings; however, they have not been employed to deep learning-based building height estimation. Given these issues, we introduced high-resolution ZY-3 satellites that can simultaneously acquire multi-spectral and multi-view images. We proposed the multi-spectral, multi-view, and multi-task deep network (M^3Net) for building height estimation. In the M^3Net, we mapped the ZY-3 multi-spectral and multi-view images to deep feature representations by two encoders, respectively, and then fused the learned feature representations for building height estimation. Furthermore, we incorporated building height estimation and footprint extraction in a multi-task learning network, in order to boost the performance of the single task. As a comparison, we implemented a random forest (RF) method using multi-source features.

A total of 42 Chinese cities were used to test the performance of the proposed method. The results showed that the M^3Net outperformed the RF model, and particularly the former can mitigate the saturation effect in estimating the height of high-rise buildings (over 30 m) to some extent. Furthermore, we found that the inclusion of ZY-3 multi-view images can significantly lower the uncertainty of building height prediction. We compared our method with two existing state-of-the-art studies on building height estimation, and found that the M^3Net can well capture the variations of height in both densely and sparsely populated areas, due to the powerful feature representation ability of deep learning network. In addition, by comparing with the vanilla single/multi-task models, the proposed M^3Net obtained a lower RMSE, suggesting that it can make better use of ZY-3 multi-view and multi-spectral information. Notably, in the experiments of the spatial-temporal transferability, the M^3Net exhibited satisfactory robustness to imaging conditions and distinct building styles. We tested our method at a relatively large area covering about 14,120 km^2, and the result validated the good scalability of our method from the perspectives of both efficacy (runtime) and quality. These findings confirmed that high-resolution ZY-3 images are valuable to building height estimation and the M^3Net holds great potentials for automated large-scale building height mapping.

Declaration of Competing Interest

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

Acknowledgments

The research was supported by the National Natural Science Foundation of China under Grants 41771360 and 41971295.

Appendix A. Supplementary data

Supplementary data to this article can be found online at https://doi.org/10.1016/j.rse.2021.112590.

References

Alobeid, A., Jacobsen, K., Heipke, C., 2009. Building height estimation in urban areas from very high resolution satellite stereo images. In: IntArchPhRS, 38 (p. Part 1–4-7/WS).

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

Baltsavias, E.P., 1999. A comparison between photogrammetry and laser scanning. ISPRS J. Photogramm. Remote Sens. 54, 83–94. https://doi.org/10.1016/S0924-2716(99)00014-3.

Berger, C., Rosentreter, J., Voltersen, M., Baumgart, C., Schmullius, C., Hese, S., 2017. Spatio-temporal analysis of the relationship between 2D/3D urban site characteristics and land surface temperature. Remote Sens. Environ. 193, 225–243.

Bonczak, B., Kontokosta, C.E., 2019. Large-scale parameterization of 3D building morphology in complex urban landscapes using aerial LiDAR and city administrative data. Comput. Environ. Urban. Syst. 73, 126–142. https://doi.org/10.1016/j.compenvurbsys.2018.09.004.

Breiman, L., 2001. Random forests. Mach. Learn. 45, 5–32. https://doi.org/10.1023/A:1010933404324.
Brunner, D., Lemoine, G., Bruzzone, L., Greidanus, H., 2010. Building height retrieval from VHR SAR imagery based on an iterative simulation and matching technique. IEEE Trans. Geosci. Remote Sens. 48, 1487–1504. https://doi.org/10.1109/TGRS.2009.2031910.
Burrough, P.A., McDonnell, R.A., 1998. Principles of Geographical Information Systems. Oxford university press.
Buslaev, A., Parinov, A., Khvedchenya, E., Iglovikov, V.I., Kalinin, A.A., 2018. Albumentations: Fast and flexible image augmentations. arXiv 11, 125.
Caruana, R., 1997. Multitask learning. Mach. Learn. 28, 41–75.
Carvalho, M., Le Saux, B., Trouve-Peloux, P., Almansa, A., Champagnat, F., 2018. On regression losses for deep depth estimation. Proc. Int. Conf. Image Process. 2915–2919. https://doi.org/10.1109/ICIP.2018.8451312.
Carvalho, M., Le Saux, B., Trouve-Peloux, P., Champagnat, F., Almansa, A., 2019. Multitask learning of height and semantics from aerial images. IEEE Geosci. Remote Sens. Lett. 1–5. https://doi.org/10.1109/lgrs.2019.2947783.
Chen, T.-H.K., Qiu, C., Schmitt, M., Zhu, X.X., Sabel, C.E., Prishchepov, A.V., 2020. Mapping horizontal and vertical urban densification in Denmark with Landsat time-series from 1985 to 2018: a semantic segmentation solution. Remote Sens. Environ. 251, 112096.
Cortes, C., Vapnik, V., 1995. Support-vector networks. Mach. Learn. 20, 273–297.
De Boer, P.-T., Kroese, D.P., Mannor, S., Rubinstein, R.Y., 2005. A tutorial on the cross-entropy method. Ann. Oper. Res. 134, 19–67.
Dell'Acqua, F., Gamba, P., 2003. Texture-based characterization of urban environments on satellite SAR images. IEEE Trans. Geosci. Remote Sens. 41, 153–159.
Drusch, M., Del Bello, U., Carlier, S., Colin, O., Fernandez, V., Gascon, F., Hoersch, B., Isola, C., Laberinti, P., Martimort, P., Meygret, A., Spoto, F., Sy, O., Marchese, F., Bargellini, P., 2012. Sentinel-2: ESA's optical high-resolution mission for GMES operational services. Remote Sens. Environ. 120, 25–36. https://doi.org/10.1016/j.rse.2011.11.026.
Elvidge, C.D., Baugh, K., Zhizhin, M., Hsu, F.C., Ghosh, T., 2017. VIIRS night-time lights. Int. J. Remote Sens. 38, 5860–5879.
Esch, T., Heldens, W., Hirner, A., Keil, M., Marconcini, M., Roth, A., Zeidler, J., Dech, S., Strano, E., 2017. Breaking new ground in mapping human settlements from space–the global urban footprint. ISPRS J. Photogramm. Remote Sens. 134, 30–42.
Esch, T., Zeidler, J., Palacios-Lopez, D., Marconcini, M., Roth, A., Mönks, M., Leutner, B., Brzoska, E., Metz-Marconcini, A., Bachofer, F., Loekken, S., Dech, S., 2020. Towards a large-scale 3D modeling of the built environment-joint analysis of tanDEM-X, sentinel-2 and open street map data. Remote Sens. 12 https://doi.org/10.3390/RS12152391.
Frantz, D., Schug, F., Okujeni, A., Navacchi, C., Wagner, W., van der Linden, S., Hostert, P., 2021. National-scale mapping of building height using Sentinel-1 and Sentinel-2 time series. Remote Sens. Environ. 252, 112128. https://doi.org/10.1016/j.rse.2020.112128.
Frolking, S., Milliman, T., Seto, K.C., Friedl, M.A., 2013. A global fingerprint of macro-scale changes in urban structure from 1999 to 2009. Environ. Res. Lett. 8, 24004. https://doi.org/10.1088/1748-9326/8/2/024004.
Gamba, P., Dell'Acqua, F., Lisini, G., Trianni, G., 2007. Improved VHR urban area mapping exploiting object boundaries. IEEE Trans. Geosci. Remote Sens. 45, 2676–2682. https://doi.org/10.1109/TGRS.2007.899811.
Geiß, C., Schrade, H., Aravena Pelizari, P., Taubenböck, H., 2020. Multistrategy ensemble regression for mapping of built-up density and height with Sentinel-2 data. ISPRS J. Photogramm. Remote Sens. 170, 57–71. https://doi.org/10.1016/j.isprsjprs.2020.10.004.
Geisß, C., Leichtle, T., Wurm, M., Pelizari, P.A., Standfus, I., Zhu, X.X., So, E., Siedentop, S., Esch, T., Taubenbock, H., 2019. Large-area characterization of urban morphology - mapping of built-up height and density using TanDEM-X and Sentinel-2 data. IEEE J. Sel. Top. Appl. Earth Obs. Remote Sens. 12, 2912–2927. https://doi.org/10.1109/JSTARS.2019.2917755.
Ghamisi, P., Yokoya, N., 2018. IMG2DSM: height simulation from single imagery using conditional generative adversarial net. IEEE Geosci. Remote Sens. Lett. 15, 794–798. https://doi.org/10.1109/LGRS.2018.2806945.
Giridharan, R., Ganesan, S., Lau, S.S.Y., 2004. Daytime urban heat island effect in high-rise and high-density residential developments in Hong Kong. Energy Build. 36, 525–534. https://doi.org/10.1016/j.enbuild.2003.12.016.
Gong, P., Li, X., Wang, J., Bai, Y., Chen, B., Hu, T., Liu, X., Xu, B., Yang, J., Zhang, W., Zhou, Y., 2020. Annual maps of global artificial impervious area (GAIA) between 1985 and 2018. Remote Sens. Environ. 236, 111510. https://doi.org/10.1016/j.rse.2019.111510.
Gonzalez, R.C., Woods, R.E., 2002. Digital Image Processing.
Gonzalez, D., Rueda-Plata, D., Acevedo, A.B., Duque, J.C., Ramos-Pollán, R., Betancourt, A., García, S., 2020. Automatic detection of building typology using deep learning methods on street level images. Build. Environ. 177, 106805. https://doi.org/10.1016/j.buildenv.2020.106805.
Gorelick, N., Hancher, M., Dixon, M., Ilyushchenko, S., Thau, D., Moore, R., 2017. Google earth engine: planetary-scale geospatial analysis for everyone. Remote Sens. Environ. 202, 18–27.
Güneralp, B., Zhou, Y., Ürge-Vorsatz, D., Gupta, M., Yu, S., Patel, P.L., Fragkias, M., Li, X., Seto, K.C., 2017. Global scenarios of urban density and its impacts on building energy use through 2050. Proc. Natl. Acad. Sci. U. S. A. 114, 8945–8950. https://doi.org/10.1073/pnas.1606035114.
Haala, N., Kada, M., 2010. An update on automatic 3D building reconstruction. ISPRS J. Photogramm. Remote Sens. 65, 570–580. https://doi.org/10.1016/j.isprsjprs.2010.09.006.
Haklay, M., Weber, P., 2008. OpenStreetMap: user-generated street maps. IEEE Pervasive Comput. 7, 12–18. https://doi.org/10.1109/MPRV.2008.80.
Hang, J., Li, Y., Sandberg, M., Buccolieri, R., Di Sabatino, S., 2012. The influence of building height variability on pollutant dispersion and pedestrian ventilation in idealized high-rise urban areas. Build. Environ. 56, 346–360. https://doi.org/10.1016/j.buildenv.2012.03.023.
Huang, X., Wen, D., Li, J., Qin, R., 2017. Multi-level monitoring of subtle urban changes for the megacities of China using high-resolution multi-view satellite imagery. Remote Sens. Environ. 196, 56–75. https://doi.org/10.1016/j.rse.2017.05.001.
Huang, X., Chen, H., Gong, J., 2018. Angular difference feature extraction for urban scene classification using ZY-3 multi-angle high-resolution satellite imagery. ISPRS J. Photogramm. Remote Sens. 135, 127–141. https://doi.org/10.1016/j.isprsjprs.2017.11.017.
Huang, X., Cao, Y., Li, J., 2020. An automatic change detection method for monitoring newly constructed building areas using time-series multi-view high-resolution optical satellite images. Remote Sens. Environ. 244. https://doi.org/10.1016/j.rse.2020.111802.
Huang, X., Li, J., Yang, J., Zhang, Z., Li, D., Liu, X., 2021. 30-m global impervious surface area dynamics and urban expansion pattern observed by Landsat satellites: from 1972 to 2019, (in press). Sci. China Earth Sci. https://doi.org/10.1007/s11430-020-9797-9.
Kendall, A., Gal, Y., Cipolla, R., 2017. Multi-task learning using uncertainty to weigh losses for scene geometry and semantics. ArXiv 7482–7491.
Kennedy, R.E., Cohen, W.B., 2003. Automated designation of tie-points for image-to-image coregistration. Int. J. Remote Sens. 24, 3467–3490. https://doi.org/10.1080/0143116021000024249.
Kingma, D.P., Ba, J., 2015. Adam: a method for stochastic optimization. Proceedings of International Conference on Learning Representations, pp. 1–13.
Koppel, K., Zalite, K., Voormansik, K., Jagdhuber, T., 2017. Sensitivity of Sentinel-1 backscatter to characteristics of buildings. Int. J. Remote Sens. 38, 6298–6318. https://doi.org/10.1080/01431161.2017.1353160.
Laben, C.A., Brower, B.V., 2000. Process for Enhancing the Spatial Resolution of Multispectral Imagery Using Pan-Sharpening.
Leichtle, T., Lakes, T., Zhu, X.X., Taubenböck, H., 2019. Has Dongying developed to a ghost city? - evidence from multi-temporal population estimation based on VHR remote sensing and census counts. Comput. Environ. Urban. Syst. 78, 101372. https://doi.org/10.1016/j.compenvurbsys.2019.101372.
Levin, N., Zhang, Q., 2017. A global analysis of factors controlling VIIRS nighttime light levels from densely populated areas. Remote Sens. Environ. 190, 366–382.
Li, X., Zhou, Y., Zhu, Z., Liang, L., Yu, B., Cao, W., 2018. Mapping annual urban dynamics (1985–2015) using time series of Landsat data. Remote Sens. Environ. 216, 674–683.
Li, J., Huang, X., Gong, J., 2019. Deep neural network for remote-sensing image interpretation: status and perspectives. Natl. Sci. Rev. 6, 1082–1086.
Li, M., Koks, E., Taubenböck, H., van Vliet, J., 2020a. Continental-scale mapping and analysis of 3D building structure. Remote Sens. Environ. 245, 111859.
Li, X., Zhou, Y., Gong, P., Seto, K.C., Clinton, N., 2020b. Developing a method to estimate building height from Sentinel-1 data. Remote Sens. Environ. 240, 111705. https://doi.org/10.1016/j.rse.2020.111705.
Liasis, G., Stavrou, S., 2016. Satellite images analysis for shadow detection and building height estimation. ISPRS J. Photogramm. Remote Sens. 119, 437–450. https://doi.org/10.1016/j.isprsjprs.2016.07.006.
Liebel, L., Bittner, K., Krner, M., 2020. A generalized multi-task learning approach to stereo DSM filtering in urban areas. ISPRS J. Photogramm. Remote Sens. 166, 213–227.
Lin, J., Wan, H., Cui, Y., 2020. Analyzing the spatial factors related to the distributions of building heights in urban areas: a comparative case study in Guangzhou and Shenzhen. Sustain. Cities Soc. 52, 101854.
Liu, C., Huang, X., Wen, D., Chen, H., Gong, J., 2017. Assessing the quality of building height extraction from ZiYuan-3 multi-view imagery. Remote Sens. Lett. 8, 907–916. https://doi.org/10.1080/2150704X.2017.1335904.
Liu, X., Hu, G., Chen, Y., Li, X., Xu, X., Li, S., Pei, F., Wang, S., 2018. High-resolution multi-temporal mapping of global urban land using Landsat images based on the Google earth engine platform. Remote Sens. Environ. 209, 227–239.
Liu, C., Huang, X., Zhu, Z., Chen, H., Tang, X., Gong, J., 2019. Automatic extraction of built-up area from ZY3 multi-view satellite imagery: analysis of 45 global cities. Remote Sens. Environ. 226, 51–73. https://doi.org/10.1016/j.rse.2019.03.033.
Ma, L., Liu, Y., Zhang, X., Ye, Y., Yin, G., Johnson, B.A., 2019. Deep learning in remote sensing applications: a meta-analysis and review. ISPRS J. Photogramm. Remote Sens. 152, 166–177. https://doi.org/10.1016/j.isprsjprs.2019.04.015.
Mahendra, A., Seto, K.C., 2019. Upward and Outward Growth: Managing Urban Expansion for More Equitable Cities in the Global South.
Mahmud, J., Price, T., Bapat, A., Frahm, J.M., P.R. (CVPR), 2020. Boundary-Aware 3D Building Reconstruction From a Single Overhead Image.
Mahtta, R., Mahendra, A., Seto, K.C., 2019. Building up or spreading out? Typologies of urban growth across 478 cities of 1 million+. Environ. Res. Lett. 14 https://doi.org/10.1088/1748-9326/ab59bf.
Main-Knorn, M., Pflug, B., Louis, J., Debaecker, V., Müller-Wilm, U., Gascon, F., 2017. Sen2Cor for sentinel-2. In: Image and Signal Processing for Remote Sensing XXIII. International Society for Optics and Photonics, p. 1042704.
Malenovský, Z., Rott, H., Cihlar, J., Schaepman, M.E., García-Santos, G., Fernandes, R., Berger, M., 2012. Sentinels for science: potential of Sentinel-1,-2, and-3 missions for scientific observations of ocean, cryosphere, and land. Remote Sens. Environ. 120, 91–101.
Miura, H., Aridome, T., Matsuoka, M., 2020. Deep learning-based identification of collapsed, non-collapsed and blue tarp-covered buildings from post-disaster aerial images. Remote Sens. https://doi.org/10.3390/rs12121924.

Mou, L., Zhu, X.X., 2018. IM2HEIGHT: HEIGHT Estimation from Single Monocular Imagery Via Fully Residual Convolutional-Deconvolutional Network, pp. 1–13.
Mushore, T.D., Odindi, J., Dube, T., Mutanga, O., 2017. Prediction of future urban surface temperatures using medium resolution satellite data in Harare metropolitan city, Zimbabwe. Build. Environ. 122, 397–410. https://doi.org/10.1016/j.buildenv.2017.06.033.
Naik, A., Rangwala, H., 2018. Multi-task learning. SpringerBriefs Comput. Sci. 28, 75–88. https://doi.org/10.1007/978-3-030-01620-3_5.
Pan, S.J., Yang, Q., 2009. A survey on transfer learning. IEEE Trans. Knowl. Data Eng. 22, 1345–1359.
Pesaresi, M., Huadong, G., Blaes, X., Ehrlich, D., Ferri, S., Gueguen, L., Halkia, M., Kauffmann, M., Kemper, T., Lu, L., 2013. A global human settlement layer from optical HR/VHR RS data: concept and first results. IEEE J. Sel. Top. Appl. Earth Obs. Remote Sens. 6, 2102–2131.
Qi, F., Zhai, J.Z., Dang, G., 2016. Building height estimation using Google earth. Energy Build. 118, 123–132. https://doi.org/10.1016/j.enbuild.2016.02.044.
Qin, R., Fang, W., 2014. A hierarchical building detection method for very high resolution remotely sensed images combined with DSM using graph cut optimization. Photogramm. Eng. Remote. Sens. 80, 873–883. https://doi.org/10.14358/PERS.80.9.873.
Ronneberger, O., Fischer, P., Brox, T., 2015. U-net: convolutional networks for biomedical image segmentation. In: International Conference on Medical Image Computing and Computer-Assisted Intervention. Springer, pp. 234–241.
Rottensteiner, F., 2003. Automatic generation of high-quality building models from lidar data. IEEE Comput. Graph. Appl. 23, 42–50. https://doi.org/10.1109/MCG.2003.1242381.
Rottensteiner, F., Sohn, G., Jung, J., Gerke, M., Baillard, C., Benitez, S., Breitkopf, U., 2012. The ISPRS benchmark on urban object classification and 3D building reconstruction. ISPRS Ann. Photogramm. Remote Sens. Spat. Inf. Sci. 293–298. I-3 (2012), Nr. 1 1.
Schneider, A., Friedl, M.A., Potere, D., 2010. Mapping global urban areas using MODIS 500-m data: New methods and datasets based on 'urban ecoregions'. Remote Sens. Environ. 114, 1733–1746. https://doi.org/10.1016/j.rse.2010.03.003.
Schug, F., Frantz, D., Okujeni, A., van der Linden, S., Hostert, P., 2020. Mapping urban-rural gradients of settlements and vegetation at national scale using Sentinel-2 spectral-temporal metrics and regression-based unmixing with synthetic training data. Remote Sens. Environ. 246, 111810. https://doi.org/10.1016/j.rse.2020.111810.
Shao, Y., Taff, G.N., Walsh, S.J., 2011. Shadow detection and building-height estimation using IKONOS data. Int. J. Remote Sens. 32, 6929–6944.
Shorten, C., Khoshgoftaar, T.M., 2019. A survey on image data augmentation for deep learning. J. Big Data 6, 1–48.
Soergel, U., Michaelsen, E., Thiele, A., Cadario, E., Thoennessen, U., 2009. Stereo analysis of high-resolution SAR images for building height estimation in cases of orthogonal aspect directions. ISPRS J. Photogramm. Remote Sens. 64, 490–500. https://doi.org/10.1016/j.isprsjprs.2008.10.007.
Sun, W., Li, T., 2020. Building height trends and their influencing factors under China's rapid urbanization: a case study of Guangzhou, 1960–2017. Chin. Geogr. Sci. 30, 993–1004.
Sun, S., Salvaggio, C., 2013. Aerial 3D building detection and modeling from airborne LiDAR point clouds. IEEE J. Sel. Top. Appl. Earth Obs. Remote Sens. 6, 1440–1449.
Sun, Y., Shahzad, M., Zhu, X.X., 2017. Building height estimation in single SAR image using OSM building footprints. In: 2017 Joint Urban Remote Sensing Event (JURSE), pp. 1–4. https://doi.org/10.1109/JURSE.2017.7924549.
Sun, Y., Hua, Y., Mou, L., Zhu, X.X., 2019. Large-scale building height estimation from single VHR SAR image using fully convolutional network and GIS building footprints. In: 2019 Joint Urban Remote Sensing Event (JURSE). IEEE, pp. 1–4.
Szabo, S., Gácsi, Z., Balazs, B., 2016. Specific features of NDVI, NDWI and MNDWI as reflected in land cover categories. Landsc. Environ. 10, 194–202.
Takaku, J., Tadono, T., Doutsu, M., Ohgushi, F., Kai, H., 2020. Updates of 'AW3D30' ALOS global digital surface model with other open access datasets. Int. Arch. Photogramm. Remote. Sens. Spat. Inf. Sci. 43, 183–189.
Tan, C., Sun, F., Kong, T., Zhang, W., Yang, C., Liu, C., 2018. A Survey on Deep Transfer Learning, arXiv. Springer International Publishing. https://doi.org/10.1007/978-3-030-01424-7.
Tang, X., Gao, X., Cao, H., Mo, F., Wang, Z., Xu, W., Zhu, G., Yue, Q., Hu, F., Zhu, H., Lu, J., 2020. The China ZY3-03 Mission: surveying and mapping Technology for High-Resolution Remote Sensing Satellites. IEEE Geosci. Remote Sens. Mag. 8, 8–17. https://doi.org/10.1109/MGRS.2019.2929770.
Taubenböck, H., Esch, T., Felbier, A., Wiesner, M., Roth, A., Dech, S., 2012. Monitoring urbanization in mega cities from space. Remote Sens. Environ. 117, 162–176. https://doi.org/10.1016/j.rse.2011.09.015.
Taubenböck, H., Klotz, M., Wurm, M., Schmieder, J., Wagner, B., Wooster, M., Esch, T., Dech, S., 2013. Delineation of central business districts in mega city regions using remotely sensed data. Remote Sens. Environ. 136, 386–401. https://doi.org/10.1016/j.rse.2013.05.019.
Taubenböck, H., Kraff, N.J., Wurm, M., 2018. The morphology of the Arrival City - a global categorization based on literature surveys and remotely sensed data. Appl. Geogr. 92, 150–167. https://doi.org/10.1016/j.apgeog.2018.02.002.
Taubenböck, H., Debray, H., Qiu, C., Schmitt, M., Wang, Y., Zhu, X.X., 2020. Seven city types representing morphologic configurations of cities across the globe. Cities 105, 102814. https://doi.org/10.1016/j.cities.2020.102814.
Tian, J., Cui, S., Reinartz, P., 2014. Building change detection based on satellite stereo imagery and digital surface models. IEEE Trans. Geosci. Remote Sens. 52, 406–417. https://doi.org/10.1109/TGRS.2013.2240692.
Tison, C., Tupin, F., Maitre, H., 2007. A fusion scheme for joint retrieval of urban height map and classification from high-resolution interferometric SAR images. IEEE Trans. Geosci. Remote Sens. 45, 496–505. https://doi.org/10.1109/TGRS.2006.887006.
Tomás, L., Fonseca, L., Almeida, C., Leonardi, F., Pereira, M., 2016. Urban population estimation based on residential buildings volume using IKONOS-2 images and lidar data. Int. J. Remote Sens. 37, 1–28.
Torres, R., Snoeij, P., Geudtner, D., Bibby, D., Davidson, M., Attema, E., Potin, P., Rommen, B., Floury, N., Brown, M., 2012. GMES Sentinel-1 mission. Remote Sens. Environ. 120, 9–24.
Tuia, D., Persello, C., Bruzzone, L., 2016. Domain adaptation for the classification of remote sensing data: an overview of recent advances. IEEE Geosci. Remote Sens. Mag. 4, 41–57. https://doi.org/10.1109/MGRS.2016.2548504.
Venter, Z.S., Brousse, O., Esau, I., Meier, F., 2020. Hyperlocal mapping of urban air temperature using remote sensing and crowdsourced weather data. Remote Sens. Environ. 242, 111791. https://doi.org/10.1016/j.rse.2020.111791.
Verma, V., Kumar, R., Hsu, S., 2006. 3D building detection and modeling from aerial LIDAR data. In: Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition, pp. 2213–2220. https://doi.org/10.1109/CVPR.2006.12.
Wang, X., Li, P., 2020. Extraction of urban building damage using spectral, height and corner information from VHR satellite images and airborne LiDAR data. ISPRS J. Photogramm. Remote Sens. 159, 322–336. https://doi.org/10.1016/j.isprsjprs.2019.11.028.
Wegner, J.D., Ziehn, J.R., Soergel, U., 2013. Combining high-resolution optical and InSAR features for height estimation of buildings with flat roofs. IEEE Trans. Geosci. Remote Sens. 52, 5840–5854.
Xie, Y., Weng, A., Weng, Q., 2015. Population estimation of urban residential communities using remotely sensed morphologic data. IEEE Geosci. Remote Sens. Lett. 12, 1111–1115. https://doi.org/10.1109/LGRS.2014.2385597.
Yu, B., Liu, H., Wu, J., Hu, Y., Zhang, L., 2010. Automated derivation of urban building density information using airborne LiDAR data and object-based method. Landsc. Urban Plan. 98, 210–219. https://doi.org/10.1016/j.landurbplan.2010.08.004.
Yuan, Q., Shen, H., Li, T., Li, Z., Li, S., Jiang, Y., Xu, H., Tan, W., Yang, Q., Wang, J., 2020. Deep learning in environmental remote sensing: achievements and challenges. Remote Sens. Environ. 241, 111716.
Zheng, Z., Zhou, W., Wang, J., Hu, X., Qian, Y., 2017. Sixty-year changes in residential landscapes in Beijing: a perspective from both the horizontal (2D) and vertical (3D) dimensions. Remote Sens. 9, 992.
Zhou, C., Wang, Z., Chen, Q., Jiang, Y., Pei, J., 2014. Design optimization and field demonstration of natural ventilation for high-rise residential buildings. Energy Build. 82, 457–465. https://doi.org/10.1016/j.enbuild.2014.06.036.
Zhou, W., Ming, D., Lv, X., Zhou, K., Bao, H., Hong, Z., 2020. SO–CNN based urban functional zone fine division with VHR remote sensing image. Remote Sens. Environ. 236, 111458. https://doi.org/10.1016/j.rse.2019.111458.
Zhu, X.X., Bamler, R., 2010. Very high resolution Spaceborne SAR tomography in urban environment. IEEE Trans. Geosci. Remote Sens. 48, 4296–4308. https://doi.org/10.1109/TGRS.2010.2050487.