CG-Net: Conditional GIS-Aware Network for Individual Building Segmentation in VHR SAR Images

Yao Sun, Yuansheng Hua, *Student Member, IEEE*, Lichao Mou, and Xiao Xiang Zhu, *Fellow, IEEE*

***Abstract*— Object retrieval and reconstruction from very-high-resolution (VHR) synthetic aperture radar (SAR) images are of great importance for urban SAR applications, yet highly challenging due to the complexity of SAR data. This article addresses the issue of individual building segmentation from a single VHR SAR image in large-scale urban areas. To achieve this, we introduce building footprints from geographic information system (GIS) data as a complementary information and propose a novel conditional GIS-aware network (CG-Net). The proposed model learns multilevel visual features and employs building footprints to normalize the features for predicting building masks in the SAR image. We validate our method using a high-resolution spotlight TerraSAR-X image collected over Berlin. Experimental results show that the proposed CG-Net effectively brings improvements with variant backbones. We further compare two representations of building footprints, namely, complete building footprints and sensor-visible footprint segments, for our task, and conclude that the use of the former leads to better segmentation results. Moreover, we investigate the impact of inaccurate GIS data on our CG-Net, and this study shows that CG-Net is robust against positioning errors in the GIS data. In addition, we propose an approach of ground truth generation of buildings from an accurate digital elevation model (DEM), which can be used to generate large-scale SAR image data sets. The segmentation results can be applied to reconstruct 3-D building models at level-of-detail (LoD) 1, which is demonstrated in our experiments.**

***Index Terms*— Deep convolutional neural network (CNN), geographic information system (GIS), individual building segmentation, large-scale urban areas, synthetic aperture radar (SAR).**

I. Introduction

VERY-HIGH-RESOLUTION (VHR) synthetic aperture radar (SAR) imagery has attracted many researchers in modeling and characterization of objects of interest in urban environments [2]–[8], as it is able to provide data being independent of sun illumination and insensitive to weather conditions. Such data source is particularly of interest to studies concerning areas frequently covered by clouds [9] and to applications of emergency response [10], [11]. However, because of side-looking imaging geometry and complex backscattering mechanism, SAR image interpretation is challenging, especially in urban areas where severe geometric distortions, such as layover and shadowing, further complicate SAR image understanding.

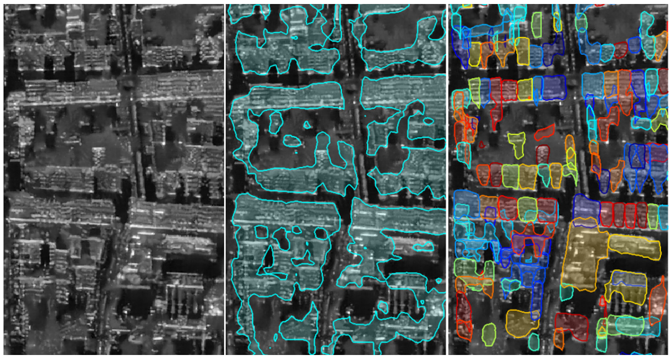


Fig. 1. Illustration of the difference between building semantic segmentation and individual building segmentation. From left to right: an SAR image, the result of building semantic segmentation [1], and the result of individual building segmentation (ours). In the middle image, all buildings are assigned the same label, while, in the right image, each individual building is identified as one class.

Buildings are the dominant structures in urban regions. The literature on retrieving information (e.g., footprint and height) from individual buildings on a large-scale VHR SAR image is still in its infancy. In [11] and [12], buildings are segmented from large-scale SAR images using deep networks. However, individual buildings cannot be recognized due to serious layover effects on high-rise buildings in urban areas. Fig. 1 shows the difference between building semantic segmentation results (middle) and our individual building segmentation results (right) in an SAR image (left). As can be seen, the latter is capable of not only providing pixelwise segmentation masks but also separating building instances. On the other hand, several works [4]–[6] develop tailored algorithms to perform accurate analyses for buildings in complex urban environments, but these methods are limited to be applied for large-scale areas. In this work, we are interested in individual

Manuscript received May 29, 2020; revised September 18, 2020; accepted November 16, 2020. Date of publication January 13, 2021; date of current version December 3, 2021. This work was supported in part by the European Research Council (ERC) through the European Union's Horizon 2020 Research and Innovation Program (Acronym: *So2Sat*) under Grant ERC-2016-StG-714087, in part by the Helmholtz Association through the Framework of Helmholtz Artificial Intelligence (HAICU)-Local Unit "Munich Unit @Aeronautics, Space and Transport (MASTr)," in part by the Helmholtz Excellent Professorship "Data Science in Earth Observation-Big Data Fusion for Urban Research," and in part by the German Federal Ministry of Education and Research (BMBF) in the framework of the International Future AI Laboratory "AI4EO–Artificial Intelligence for Earth Observation: Reasoning, Uncertainties, Ethics and Beyond". *(Corresponding author: Xiao Xiang Zhu.)*

The authors are with Remote Sensing Technology Institute, German Aerospace Center, 82234 Weßling, Germany, and also with the Data Science in Earth Observation, Technical University of Munich, 80333 Munich, Germany (e-mail: yao.sun@dlr.de; yuansheng.hua@dlr.de; lichao.mou@dlr.de; xiaoxiang.zhu@dlr.de).

Digital Object Identifier 10.1109/TGRS.2020.3043089

building segmentation from SAR images on a large scale. In what follows, we briefly explain the challenges of this task and review related work.

A. Challenges

Interpreting individual buildings in SAR images is highly challenging, mainly for two reasons. First, intensity values in SAR images are closely related to material types and structural shapes of objects. Therefore, consecutive buildings in the physical world are difficult to be separated from each other in an SAR image, unless in the presence of obvious material or structure changes at building boundaries. Second, even if buildings in the real world are not neighboring, they probably overlap with each other in the SAR image, which significantly increases the difficulty of image interpretation. Fig. 2 shows two typical urban areas in an optical image (the first column) and a VHR SAR image (the second column). Footprints and regions of buildings present in the SAR image are marked with different colors, as shown in the following two columns. It can be seen that some buildings severely overlap in the SAR image even if their corresponding footprints are not next to each other.

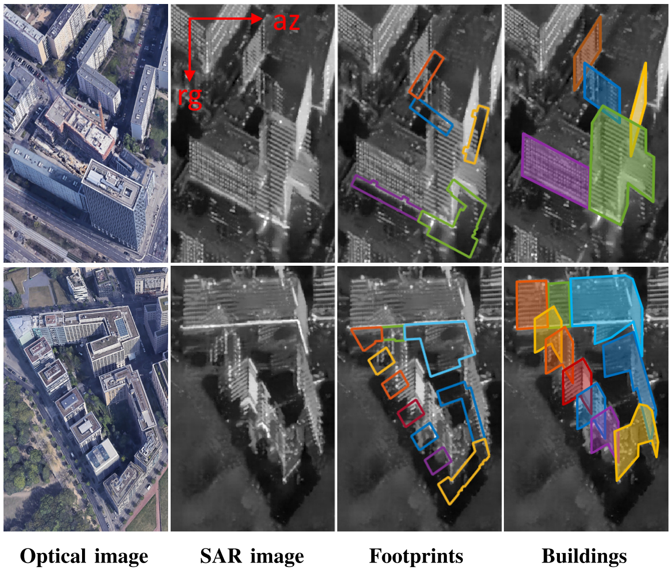


Fig. 2. Two typical urban areas shown in an optical image and an SAR image. In columns 3 and 4, footprints and the corresponding building regions in the SAR image are marked in different colors for reference. rg and az denote the range direction and the azimuth direction, respectively.

B. Related Work

Generally, building extraction approaches from SAR data can be grouped into the following two categories: data-driven methods and model-driven methods. The former extracts building features and then deduces building parameters. Two solutions based on this methodology have been developed. The first one makes an attempt at detecting line- or point-like features first and extracting building regions based on these features. For example, in [3], feature lines are identified using a line detector, and the layover areas are derived by extracting parallel edges; Xu and Jin [13] exploit a constant false alarm rate (CFAR) edge detector for line feature detection and apply a Hough transform for parallelogram-like wall area extraction; in [14] and [15], bright-line segments and regular spaced point-like features are detected and subsequently grouped to building footprints; and Ferro *et al.* [16] extract and combine a set of low-level features to create structured primitives. The second solution directly extracts building regions using segmentation techniques, such as active contour [17], rotating mask [18], mean-shift [19], and marker-controlled watershed transform [20]. In model-driven methods, an SAR image or InSAR phase is iteratively simulated using geometric and radiometric hypothesis [4], [5], [21]–[25]. The desired building parameters are progressively achieved by minimizing the difference between simulated and real data.

The majority of related studies are carried out on buildings with specific geometric shapes, e.g., rectangular- [26]–[28] or L-shaped footprints [20], [29], flat [30] or gable roofs [31], [32], and different heights [32]–[35]. Only a few studies address the problem of complex-shaped buildings [14], [15]. Furthermore, most studies investigate simple scenarios where a minimal distance between buildings is required to ensure that scattering effects of different buildings do not interfere with each other [4]–[6], [36]. In complex scenarios, possible overlapping areas between two buildings are usually assigned to one building [7], [37], which may cause incorrect estimations. By using an SAR tomography (TomoSAR) point cloud, Shahzad and Zhu [38] extract buildings without imposing constraints on building shapes and study scenarios. However, the TomoSAR technique [39] requires multiple SAR acquisitions that are generally unavailable for most areas and for applications with a stringent time limit, such as emergency response.

In addition to SAR data, some auxiliary data are introduced, e.g., building outlines extracted from optical images [6], [40] and footprint polygons obtained from geographic information system (GIS) data [7], [41], [42], for providing exact locations and geometric shapes of buildings in the real world. As illustrated in Fig. 2, in complex urban regions, the use of footprints is beneficial for tasks concerning individual buildings in SAR images. In exploiting the shape information, sensor-visible footprint segments, i.e., near-range segments in footprint polygons that correspond to sensor-visible walls, are desirable for extracting layover areas [7], [42]; contrarily, complete building footprints may provide additional information, especially for extracting roof areas of low-rise buildings [6]. Therefore, it leaves a question on how footprints can be effectively used. We demonstrate this issue in this work by comparing the results from both the footprint utilization.

In recent years, deep neural networks have been becoming increasingly popular and shown success in remote sensing data analysis [43]–[52], including a wide range of applications using SAR data, such as classification [53]–[57], segmentation [58], [59], target recognition [60]–[63], and change detection [64]–[66]. Instead of relying on handcrafted features, deep networks can learn effective feature representations from raw data in an end-to-end fashion. However, one problem of applying deep networks to urban SAR analysis tasks is the

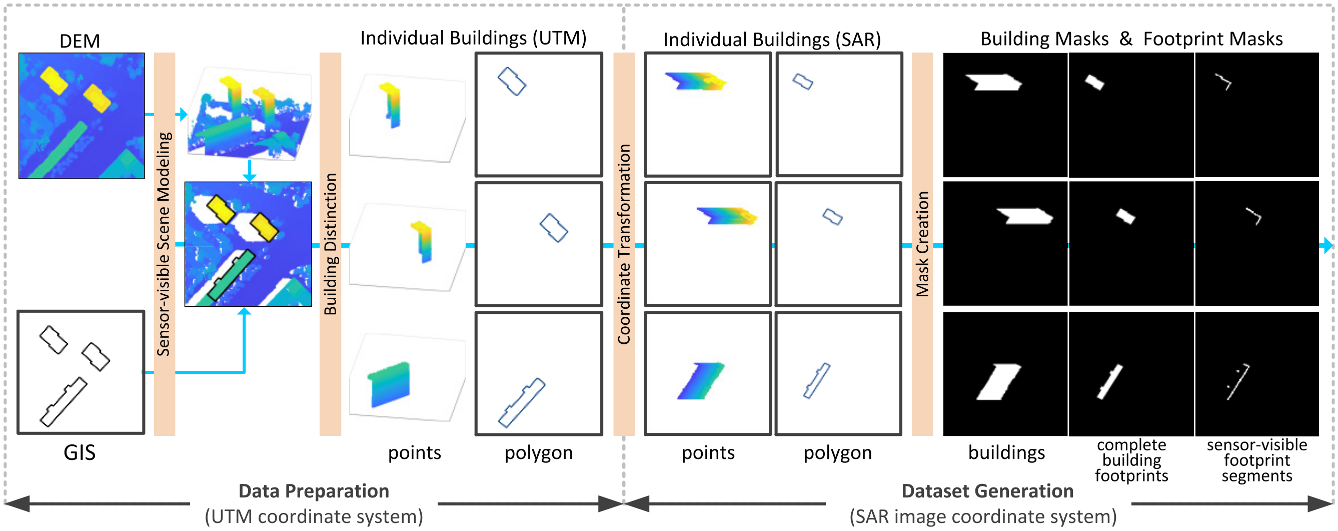

DEM
GIS
Sensor-visible Scene Modeling
Building Distinction
Individual Buildings (UTM)
points
polygon
Coordinate Transformation
Individual Buildings (SAR)
points
polygon
Mask Creation
Building Masks & Footprint Masks
buildings
complete building footprints
sensor-visible footprint segments
Data Preparation
(UTM coordinate system)
Dataset Generation
(SAR image coordinate system)

Fig. 3. Workflow for data set generation. We first collect DEM and GIS data in the UTM coordinate system and then project them to the SAR image coordinate system in order to generate building ground truth annotations and the corresponding footprints in our study area.

lack of annotation data. To address this issue, Wang *et al.* [67] take building polygons from the OpenStreetMap (OSM) data set and an official map as ground truth data and train a network to segment buildings in an urban scene. For building footprint extraction, Shermeyer *et al.* [68] presented a multisensor all-weather mapping (MSAW) data set containing airborne SAR images, optical images, and building footprint annotations, along with a deep network baseline model and benchmark. However, in these two works, building footprints, instead of building areas, are learning targets. By introducing a TomoSAR point cloud, Shahzad *et al.* [12] are able to acquire accurate building areas in an SAR image and take them as ground truth annotations to train a segmentation network for the purpose of building extraction. However, this work cannot differentiate individual buildings. As our survey of related work shows, there is a paucity of literature on using deep learning for VHR SAR image interpretation in complex urban areas, particularly aiming at segmenting individual and overlapping buildings.

C. Contributions

In this work, we intend to segment individual buildings in a large urban area by exploiting SAR images and building footprints. For the training of models, we generate pixelwise ground truth annotations from an accurate digital elevation model (DEM). The building footprints are acquired from GIS data. Afterward, a novel conditional GIS-aware network (CG-Net) has been proposed to first learn multilevel visual features and then employ GIS building footprint data to normalize these features for predicting final building masks. In addition, we compare two representations of building footprints, namely, complete building footprints and sensor-visible footprint segments, aiming to find out a more suitable representation way for this task.

The main contributions of this article are fourfold.

1) We propose a workflow for the segmentation of individual buildings in VHR SAR images with GIS data. To the best of our knowledge, this is the first time that individual buildings are studied on a large-scale SAR image, and deep networks are employed in the problem of individual building segmentation of SAR images.
2) We propose a network termed CG-Net, which is capable of significantly improving the performance of networks for our task by imposing constraints on the learning process.
3) We investigate the impact of inaccurate GIS data on CG-Net and find out that CG-Net is robust against positioning errors in GIS data. This study suggests that a large amount of open-sourced GIS data can be exploited for individual building segmentation in SAR images.
4) We propose a ground truth generation approach to produce building masks using an accurate DEM. We believe that our method can provide large potential in analyzing complex urban regions.

The remainder of this article is organized as follows. The detailed procedure of the data set generation is presented in Section II, and the proposed CG-Net is delineated in Section III. Section IV introduces the configuration of experiments and analyzes results. Section V demonstrates an application using the produced segmentation results. In Section VI, we conclude this article.

II. Data Set Generation

A. Overview

Building annotations (as ground truth data) and building footprints (as input data) in SAR images are necessary for training our network. For this reason, we propose a workflow that employs highly accurate DEM and GIS building footprints to automatically label building masks and their corresponding footprints in SAR images. Our data set is generated in two stages. First, sensor-visible 3-D building models (i.e., nonoccluded roofs and facades) and building footprints are prepared in the Universal Transverse Mercator (UTM) coordinate system. Second, they are projected to the SAR image coordinate system in order to generate building ground truth annotations and the corresponding footprints. Fig. 3 illustrates

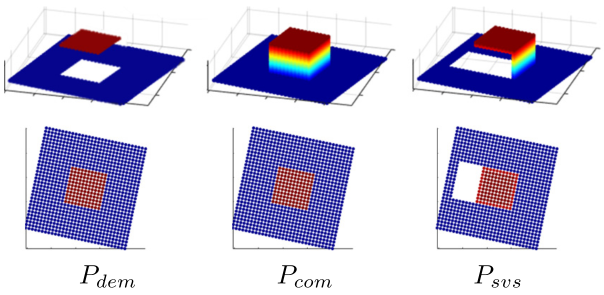


Fig. 4. Illustration of scene modeling steps with a simulated DEM in 3-D (first row) and 2-D (second row). (Left) DEM point cloud P_{dem}. (Middle) Complete point cloud P_{com} after adding vertical points. (Right) Sensor-visible point cloud P_{svs} after HPR.

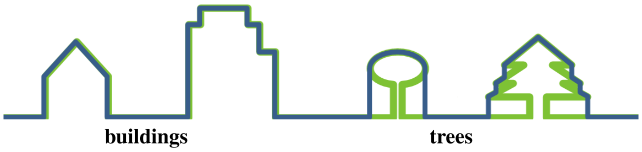


Fig. 5. Illustration of 2.5-D (dark blue) and 3-D (green) surface models. In 2.5-D representation, each 2-D point (x, y) is assigned to a unique height value z. Therefore, 2.5-D DEM can represent vertical walls of buildings, but not vertical surfaces of complex objects, such as trees.

the workflow, and for more details, refer to the following sections.

B. Data Preparation in the UTM Coordinate System

1) Sensor-Visible Scene Modeling: We first model a scene that can be viewed by a radar sensor in the UTM coordinate system. The procedure is conducted in three steps (see Fig. 4).

1) *DEM is Transformed to a Point Cloud P_{dem}:* Specifically, each pixel in the DEM with geolocation coordinates (x, y) and a height value h is represented as a point with coordinates (x, y, h), and hence, all pixels establish a nadir-looking 3-D point cloud P_{dem}.
2) *Complete 3-D Point Cloud P_{com} is Generated by Filling Vertical Data Gaps:* To be more specific, vertical structures, such as building walls, that are absent from P_{dem} are added through the following steps. We first detect building points that are located at height jumps. Afterward, at each detected point $g(x, y, h)$, a vertical point set $G = \{g_i(x_i, y_i, h_i) | i = 1, \ldots, m\}$ is added, where $x_i = x, y_i = y, h_i = h_0 + i \times h_{\text{step}}, h_i < h_e$. h_0 and h_e are the minimum and maximum heights in the neighborhood of g, h_{step} is a predefined height step, and the number of points is $m = (h_e - h_0)/h_{\text{step}}$. Eventually, a complete 3-D point cloud P_{com} is built by all vertical point sets and P_{dem}. Note that the DEM is 2.5-D instead of true 3-D, i.e., each 2-D point (x, y) is assigned to a unique height value z [69], that the vertical surfaces of complex objects are not represented, such as trees (see Fig. 5). Therefore, vertical points are only added to building areas in this step.
3) *Sensor-Visible Scene Point Cloud P_{svs} is Obtained Through a Visibility Test on the Point Cloud P_{com}:* Since a radar sensor only sees one side of a scene, points on the other side should be removed. To this end, the hidden point removal (HPR) algorithm [70] is applied.
In our process, the viewpoint in HPR is positioned on the line of sight of the radar sensor at a large distance away from the scene, in order to simulate an orthographic view in the azimuth of the radar sensor. In this way, sightlines from the viewpoint to objects in the scene are parallel to each other and orthogonal to the azimuth, enabling HPR to remove sensor-invisible points.

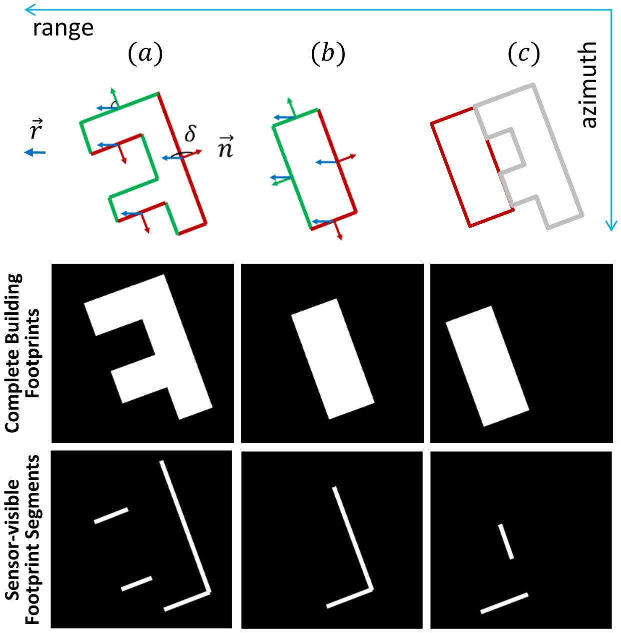

range
(a)
(b)
(c)
azimuth
$\vec{r}$
δ
$\vec{n}$
Complete Building Footprints
Sensor-visible Footprint Segments

Fig. 6. Examples of (top) visibility test of building footprints and (middle and bottom) two footprint representations. (a) and (b) Footprints of isolated buildings: red edges are sensor-visible, as the angle δ between the outward normal vector of an edge $\overrightarrow{n}$ and the range direction vector $\overrightarrow{r}$ is in the range of $(90°, 180°]$, while green ones are invisible. (c) Case that a footprint is touching another one; hence, common edges are sensor-invisible.

2) Building Distinction: In this step, we distinguish building points[1] for individual buildings. Given one building, its building points are selected from P_{svs} using its footprint. Note that there are two possible inconsistencies between the DEM and GIS data. First, if a building is contained in P_{svs} but not in GIS data, it is not selected from P_{svs}. Second, if a building is contained in GIS data but not in P_{svs}, i.e., points in the footprint region are not elevated than surrounding ground points, we exclude this building from our data set.

C. Data Set Generation in the SAR Image Coordinate System

1) Coordinate Transformation: The aforementioned procedures are carried out in the UTM coordinate system, and in our case, building points generated in the previous steps should be projected to the SAR image coordinate system; that is to say, coordinates (x, y, h) need to

[1] Building points refer to points in a point cloud that belong to the building class.

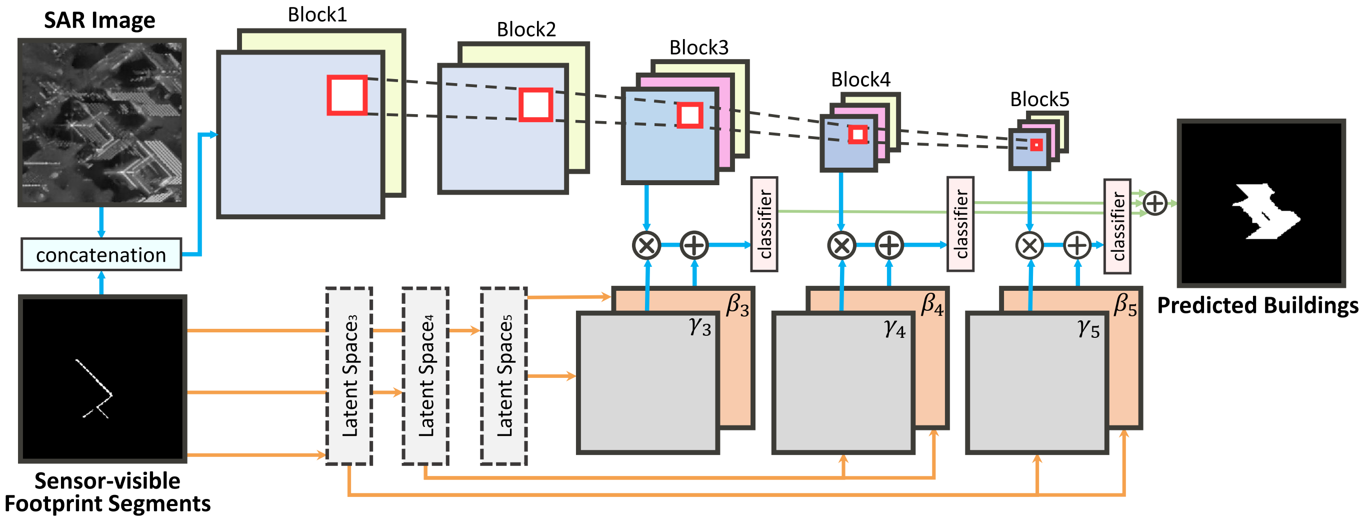


Fig. 7. Overview of the CG-Net architecture.

be transformed to *(range, azimuth)*. Moreover, building footprints are also projected to this coordinate system by using ground height values obtained from the DEM. Generally, the coordinate transformation of the point cloud from the UTM coordinate system to the SAR imaging coordinate system includes iterative solving Doppler-range-ellipsoid equations that can be implemented with different approaches [71]–[74]. In this work, radar coding was performed using German Aerospace Center's (DLR's) Integrated Wide Area Processor (IWAP) [75].

2) Mask Creation: Finally, according to *range-azimuth* coordinates of building points, we generate building ground truth masks, in which buildings are indicated by 1 and backgrounds are marked as 0. In addition, building footprint masks in the SAR image coordinate system are also created. Notably, in order to find out an effective way of using building footprints, we create two representations: complete building footprints and sensor-visible footprint segments. The latter is generated via a visibility test (see Fig. 6). Formally, let $\overrightarrow{n}$ be the outward normal vector of a polygon edge, $\overrightarrow{r}$ be the range direction vector, and $\delta \in [0°, 180°]$ be the angle between $\overrightarrow{n}$ and $\overrightarrow{r}$. A polygon edge is sensor-visible if $\delta \in (90°, 180°]$, and if a footprint is touching other footprints, common edges are invisible because they do not exist in the real world (e.g., Fig. 6(c)).

D. Postprocessing

Since the used SAR image and DEM are collected at different times, there might be inconsistencies resulted from urban changes, such as building construction and deconstruction. This leads to inaccurate ground truth data. We cope with the problem using intensity values of the given SAR image. In the SAR image, the intensity values are generally larger in building areas than ground areas. Therefore, a threshold is set to be the mode of the intensity values of the SAR image, to exclude buildings whose mean intensity values are smaller than the threshold.

III. Methodology

A. Overview

In this work, our goal is to train a network that takes a SAR image and building footprint as inputs and predicts the building area associated with the footprint in the SAR image. Since footprints and visible segments generated from GIS data can provide precise geometry and location information, we resort to exploiting such cues in our task and devise a network module that performs a conditional GIS-aware normalization. By utilizing the CG module, our network, termed CG-Net, can learn feature representations from not only SAR but also GIS data. Specifically, we employ VGG-16 [76] as the backbone of CG-Net to learn multilevel features from SAR images. Afterward, the outputs of the last three convolutional blocks are upsampled and fed into the CG module separately. Meanwhile, footprints or visible segments are imported into the CG module as complementary inputs in order to yield final predictions. In what follows, Section III-B illustrates the procedure of multilevel feature extraction. Section III-C introduces details of our CG module, and Section III-D details the configuration of our CG-Net.

B. Multilevel Feature Extraction Module

We make use of VGG-16 [76] as the backbone of our network to extract features from multiple layers, as these multilevel features help in recognizing buildings with variant scales. The backbone consists of five convolutional blocks, and each of them contains two or three convolutional layers. The size of their filters is 3×3. The outputs of all convolutional layers are activated by rectified linear unit (ReLU) [77], and 2×2 max-pooling layers with a pooling stride of 2 are interleaved among these blocks. Features learned from deep layers are considered to include high-level semantics, while those from shallow layers are low-level. Therefore, in this task, we utilize features learned from the last three blocks, i.e., Block3, Block4, and Block5 (see Fig. 7). Afterward, the extracted features are fed into the CG module separately.

C. Conditional GIS-Aware Normalization Module

An intuitive way to make use of GIS data is to simply concatenate them with SAR images and then feed them to a vanilla semantic segmentation network, such as fully convolutional networks (FCN). However, such a method might suffer from the inefficient use of GIS data and leads to unstructured predictions (see the third column in Fig. 13). To address this issue, in this article, we propose a conditional GIS-aware normalization module to distill the geometry information of individual buildings from GIS data and normalize final predictions with such information. Formally, let $\boldsymbol{m}_{\mathrm{gis}}$ be the mask of the complete building footprint or sensor-visible footprint segments with a spatial size of $W \times H$, and $\boldsymbol{x}_b$ denotes feature maps extracted from the bth convolutional block. The width and height of $\boldsymbol{x}_b$ are represented as W' and H', respectively. The number of channels is denoted as C'. We consider a naive conditional normalization procedure as follows:

$$\hat{\boldsymbol{x}}_b = \gamma_b \boldsymbol{x}_b + \beta_b \tag{1}$$

where γ_b and β_b represent a scale factor and a bias, respectively, and they indicate to what extent $\boldsymbol{x}_b$ should be scaled and shifted. The normalized $\boldsymbol{x}_b$ is denoted as $\hat{\boldsymbol{x}}_b$. A commonly used measure of γ and β is to calculate the standard deviation and mean of $\boldsymbol{x}_b$. Since $\boldsymbol{x}_b$ consists of more than one channel, γ and β are often computed in a channelwise manner, and thus, (1) can be rewritten as

$$\hat{\boldsymbol{x}}_{b,c} = \gamma_{b,c}(\boldsymbol{x}_{b,c}) \cdot \boldsymbol{x}_{b,c} + \beta_{b,c}(\boldsymbol{x}_{b,c}) \tag{2}$$

where c denotes the cth channel of $\boldsymbol{x}_b$ and ranges from 1 to C'. This equation can be easily extended to the batch normalization [78] by computing the standard deviation and mean of each $\boldsymbol{x}_{b,c}$ in a batch.

In our case, we want to normalize feature representations learned from SAR images, conditioned on GIS data. Our insight is that the GIS data imply coarse localization cues, and their use can guide the network to segment individual buildings accurately. Therefore, we reformulate (2) as follows:

$$\hat{\boldsymbol{x}}_{b,c,p,q} = \gamma_{b,c,p,q}(\boldsymbol{m}_{\mathrm{gis}}) \cdot \boldsymbol{x}_{b,c,p,q} + \beta_{b,c,p,q}(\boldsymbol{m}_{\mathrm{gis}}) \tag{3}$$

where $\gamma_{b,c,p,q}$ and $\beta_{b,c,p,q}$ indicate the scale factor and bias *learned* specifically for the pixel located at (p, q) in the c-th channel of $\boldsymbol{x}_b$. As a consequence, normalization parameters γ_b and β_b are formatted as matrices with a size of $W' \times H' \times C'$. Such a design enjoys an advantage that normalization parameters are learned in a data-driven manner, and thus these parameters are expected to be more adapted to $\boldsymbol{x}_b$. As to the implementation of (3), we first project $\boldsymbol{m}_{\mathrm{gis}}$ onto a latent space through 3 × 3 convolutions and then employ two convolutional layers to learn γ_b and β_b from the encoded $\boldsymbol{m}_{\mathrm{gis}}$. Subsequently, the elementwise multiplication of $\gamma_b(\boldsymbol{m}_{\mathrm{gis}})$ and $\boldsymbol{x}_b$ is performed, and the output is added to $\beta_b(\boldsymbol{m}_{\mathrm{gis}})$ pixel by pixel. Fig. 8 illustrates the architecture of our CG module.

D. Configuration of CG-Net

In order to fully exploit GIS data at multiple scales, we append three CG modules to the last three convolutional blocks of the backbone (see Fig. 7). However, a question is that spatial and channel dimensions of the extracted multilevel features are inconsistent with those of complete building footprints/sensor-visible footprint segments. To address this issue, we upsample these multilevel feature maps to match the spatial resolution of $\boldsymbol{m}_{\mathrm{gis}}$ via bilinear interpolation. Note that doing so would significantly increase the computation overhead of subsequent operations. Hence, we reduce the number of feature channels through 1 × 1 convolutions and modify the CG module (see Fig. 9) accordingly. The outputs of the CG module are squashed into the number of classes (two) and added via an elementwise addition operation to produce final segmentation results. Fig. 7 illustrates the architecture of the proposed CG-Net. Furthermore, we note that the proposed CG module is in a plug-and-play fashion and is flexible enough to enhance other semantic segmentation network architectures, e.g., DeepLabv3. For DeepLabv3, since it already fuses features from different layers in its architecture, we simply add our module right before the last layer.

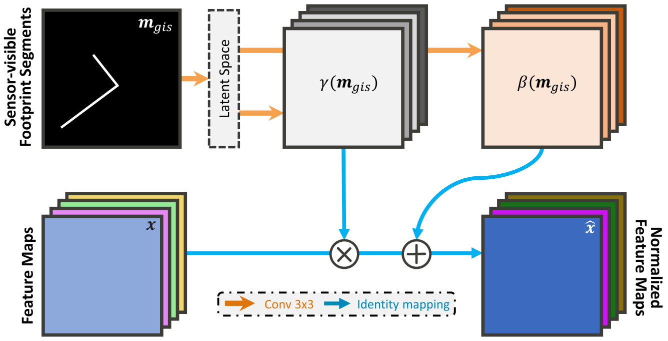


Fig. 8. Architecture of the proposed CG module. Here, we take the sensor-visible footprint segments as an example. γ and β are the normalization parameters learned from the sensor-visible footprint segments and used to normalize input feature maps with (3).

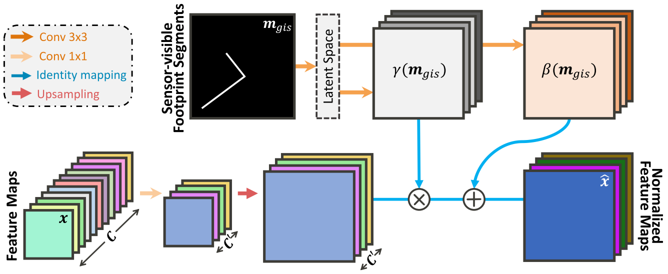


Fig. 9. Architecture of the final CG module. In advance of performing normalization, the channel of input feature maps is first reduced, and the spatial size is enlarged according to that of sensor-visible footprint segments.

IV. Experiments

A. Data Description

In our data set, a TerraSAR-X image was acquired in the high-resolution spotlight mode over Berlin with the pixel spacing[2] of 0.871 m in the azimuth direction and 0.455 m

[2] In SAR images, pixel spacing represents that the length one pixel corresponds to in the real world, while resolution indicates the minimum distance at which the radar can distinguish two close scatters.

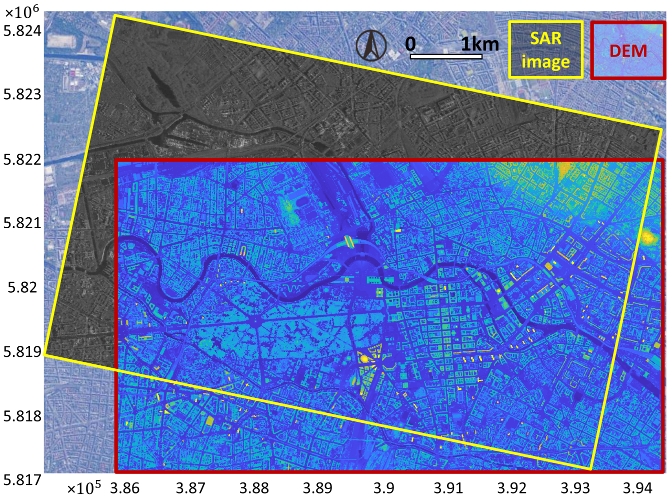


Fig. 10. Our study area in the UTM coordinate system that is the intersection between the SAR image and the DEM.

in the slant range direction. The incidence angle of this SAR image is 36°, and the heading angle is 194.34°. To reduce the speckle effect, the SAR image was filtered using a nonlocal InSAR algorithm [79]. Besides, building footprints in the study area were downloaded from Berlin 3D-Download Portal.[3] In order to yield ground truth annotations, we use a highly accurate DEM that was obtained via the stereo processing of aerial images with a resolution of 7 cm/pixel [80]. Fig. 10 illustrates our study region (the intersection area), the SAR image (yellow rectangle), and DEM (red rectangle). Notably, only data covering the study region are used for generating our data set.

By using the workflow described in Section II, building annotations and footprints are generated. Since we want to explore how GIS data can be effectively used for individual building segmentation, these two versions of footprint masks are produced: complete building footprints and sensor-visible footprint segments. Our data set, therefore, contains a 5736 × 10312 SAR image, two versions of footprint masks, and ground truths of individual buildings.

B. Training Details

In order to train an effective and robust segmentation network, we crop the SAR image into patches of 256×256 pixels with a stride of 150 pixels. Note that patches including incomplete footprints or ground truth annotations are discarded. Consequently, 30 056 buildings are remaining, and each of them has three patches: an SAR image patch, a footprint patch, and a ground truth mask. Among all buildings, 19 434 of them are utilized to train networks, and the others are test samples. Note that training and test regions do not overlap. The network takes one SAR patch and the corresponding GIS patch for one building as inputs. After predicting masks of all buildings, overlapping areas are obtained by overlaying all masks.

During the training phase, components of the proposed CG-Net are initialized with different strategies. Specifically, the multilevel feature extraction module is initialized with weights pretrained on ImageNet [81], and all convolutional layers in the CG modules are initialized with a Glorot uniform initializer. The network is implemented on TensorFlow and trained on one NVIDIA Tesla P100 16-GB GPU for 155k iterations. During the training procedure, all weights are updated through backpropagation, and we select Netrov Adam [82] as the optimizer. The parameters of this optimizer are set as recommended: $\epsilon = 1\mathrm{e}{-08}$, $\beta_1 = 0.9$, and $\beta_2 = 0.999$. The loss is defined as binary cross-entropy, as only two classes are considered in our data set, i.e., building segments and background. We initialize the learning rate as $2e{-3}$ and reduce it by a factor of $\sqrt{10}$ once the loss stops to decrease for two epochs. Moreover, we utilize a small batch size of 5 in our experiments.

TABLE I

NUMERICAL RESULTS USING SENSOR-VISIBLE FOOTPRINT SEGMENTS. THE HIGHEST VALUES OF DIFFERENT METRICS ARE HIGHLIGHTED IN **BOLD**

Model Name	P	F1 score	IoU	OA
FCN	0.6478	0.6808	0.5138	0.8340
FCN-CG	0.6553	0.6931	0.5303	0.9926
DeepLabv3	0.6635	0.6971	0.5351	0.9927
DeepLabv3-CG	**0.6852**	**0.7068**	**0.5465**	**0.9928**

C. Quantitative Evaluation

To evaluate the performance of networks, we calculate the F1 score as follows:

$$F1 = 2 \cdot \frac{P \cdot R}{P + R}, \quad P = \frac{tp}{tp + fp}, \quad R = \frac{tp}{tp + fn} \tag{4}$$

where P and R denote the precision and recall, respectively. In addition, the intersection over union (IoU) and overall accuracy (OA) are also calculated for a comprehensive comparison

$$\mathrm{IoU} = \frac{tp}{tp + fp + fn}, \quad \mathrm{OA} = \frac{tp + tn}{tp + tn + fp + fn} \tag{5}$$

where tp, fp, tn, and fn represent pixel-based true positives, false positives, true negatives, and false negatives for buildings, respectively.

In our experiments, we compare four models: FCN, FCN-CG, DeepLabv3, and DeepLabv3-CG. It is worth mentioning that FCN and DeepLabv3 are regarded as baselines, and their inputs are concatenations of SAR patches and their corresponding footprint patches. Both FCN-CG and DeepLabv3-CG are our proposed networks with different backbones.

Table I reports numerical results of different models on our data set, where sensor-visible footprint segments are used. Comparison of these results corroborates that the proposed CG module can improve the performance of individual building segmentation. Specifically, compared with FCN and DeepLabv3, FCN-CG and DeepLabv3-CG achieve improvements of 0.75% and 2.17% in the precision, respectively. Besides, increments of 1.23% and 1.65% in the mean F1 score and IoU can be observed by comparing FCN-CG and FCN, while improvements of 0.97% and 1.14% in the

[3] https://www.businesslocationcenter.de/downloadportal/

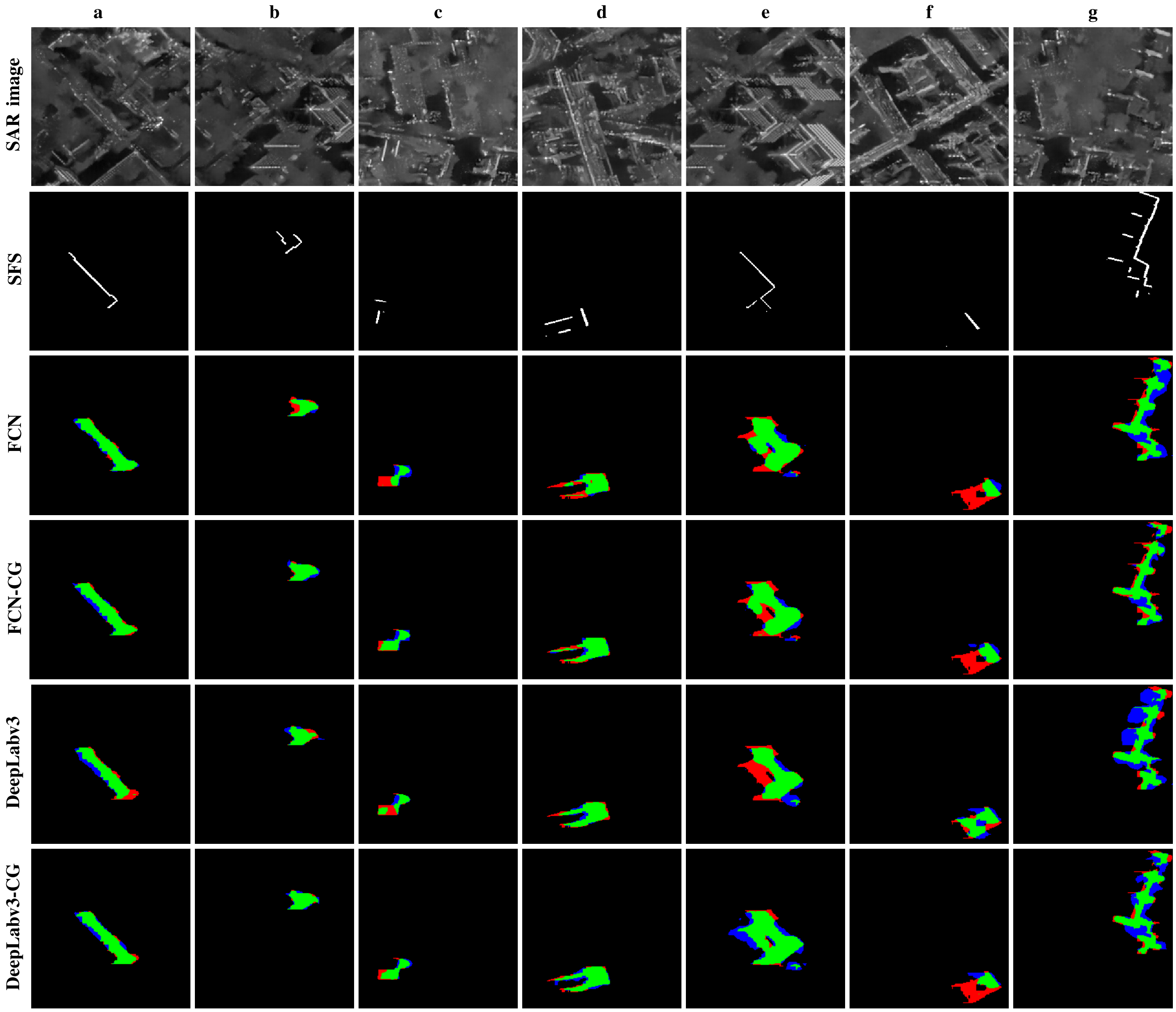


Fig. 11. Examples of segmentation results using sensor-visible footprint segments (SFS). Pixel-based true positives, false positives, and false negatives are marked in green, red, and blue, respectively.

TABLE II

NUMERICAL RESULTS USING COMPLETE BUILDING FOOTPRINTS. THE HIGHEST VALUES OF DIFFERENT METRICS ARE HIGHLIGHTED IN **BOLD**

Model Name	P	F1 score	IoU	OA
FCN	0.7045	0.7242	0.5676	0.9932
FCN-CG	0.7240	0.7362	0.5826	0.9935
DeepLabv3	0.7129	0.7337	0.5794	0.9935
DeepLabv3-CG	**0.7523**	**0.7508**	**0.6010**	**0.9937**

same metrics are achieved by introducing the CG module to DeepLabv3.

Table II presents results of variant models using complete building footprints. We can see that the results are consistent with those using sensor-visible footprint segments. For example, with the CG module, the precision improves 1.95% and 3.94% with the backbone, FCN and DeepLabv3, and the IoU increases 1.50% and 2.16%. To summarize, improvements achieved by FCN-CG and DeepLabv3-CG demonstrate the effectiveness of the proposed CG module, and DeepLabv3-CG can achieve the best performance in all four metrics on our data set. Moreover, we note that all models achieve relatively high OAs, and even the worst model can achieve an OA of 83.40%. This is because OA is computed by considering all pixels, while nonbuilding pixels, which are easily recognized, account for a large proportion.

D. Qualitative Evaluation

In addition to the quantitative evaluation, we visualize several segmentation results in Figs. 11 and 12. Pixel-based true positives, false positives, and false negatives are presented in green, red, and blue, respectively.

Fig. 11 shows the results of models using sensor-visible footprint segments. We can observe a general improvement in quality from FCN/DeepLabv3 to FCN-CG/DeepLabv3-CG, especially for buildings in column *b*, *c*, and *g*. For buildings with simple structures (e.g., the building in column *a*), all models are able to offer satisfactory segmentation results,

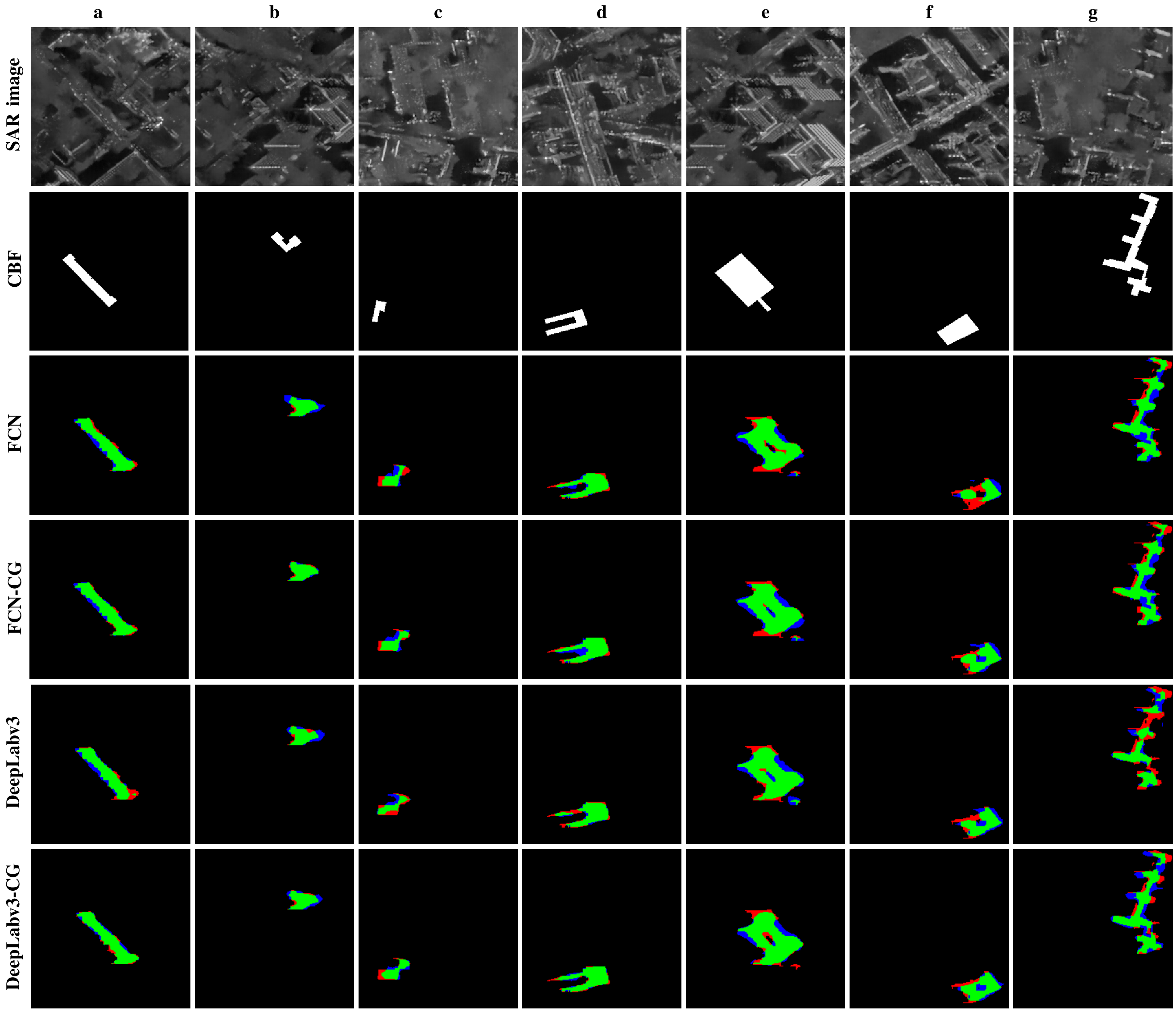


Fig. 12. Examples of segmentation results using complete building footprints (CBF). Pixel-based true positives, false positives, and false negatives are marked in green, red, and blue, respectively.

while for those with complicated shapes (see column *e*), large undersegmentation areas (see red pixels) can be seen in predicted building masks. Besides, the utilization of the proposed CG module can effectively reduce oversegmentation in final predictions.

Fig. 12 presents the results of models using complete footprints. They indicate that our CG module can ease both oversegmentation (see blue pixels in column *b*) and undersegmentation (see red pixels in column *e*) problems to a considerable extent. Moreover, examples in the third row, column *f* and the fifth row, column *f* show that the connectivity of segmentation results are disrupted (see green pixels), while the integration of the CG module can alleviate such a problem. A similar phenomenon can also be seen in column *d* and *g* that exploiting the CG module can enhance the connectivity of predictions. In summary, the proposed CG module effectively improves segmentation results.

E. Comparison of Complete Building Footprints and Sensor-visible Footprint Segments

From Tables I and II, we can see that models trained with complete building footprints surpass those trained with sensor-visible footprint segments. For instance, DeepLabv3-CG trained on complete footprints improves the F1 score and IoU by 4.40% and 5.45%, respectively, compared with that learned with sensor-visible segments.

Fig. 13 provides segmentation results of two patches using two versions of footprint masks, and different buildings are marked in different colors (50% transparency). Note that individual building masks are predicted separately, and then the masks of buildings in the same patch are plotted together to visualize the overlapping areas. Here, patch 1 presents a simple scenario, in which buildings are isolated and show clear signatures in the SAR image. In this case, all models can obtain good segmentation results. Patch 2 shows a fairly complicated scene, where two consecutive buildings exist in

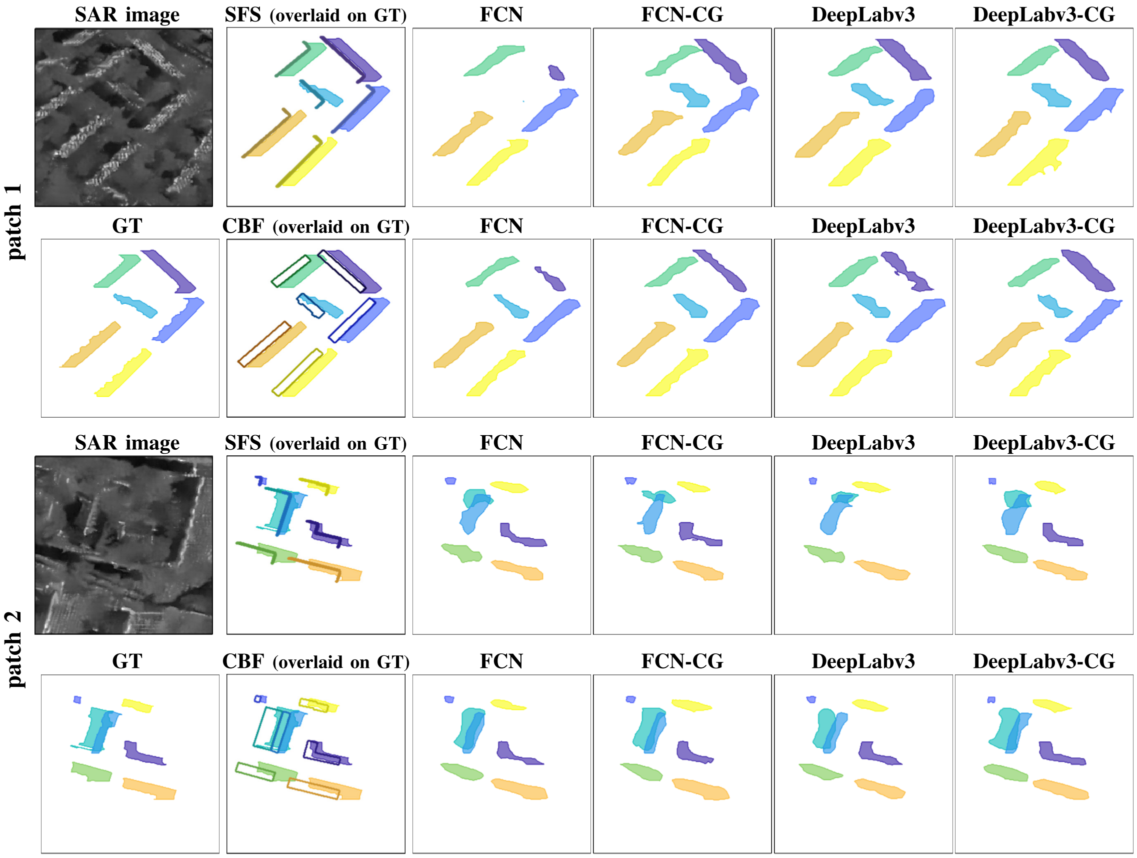


Fig. 13. Examples of segmentation results from different models on two patches, using complete building footprints (CBF) and sensor-visible footprint segments (SFS). CBF and SFS are overlaid on the ground truth (GT) to visualize the difference between building footprints and buildings. Different buildings are plotted in different colors (50% transparency).

the center (see buildings in cyan and blue), and SAR signatures are unclear. Although all networks can still successfully segment isolated buildings, the two overlapped buildings are not correctly segmented by models trained with sensor-visible footprint segments (see the third row of Fig. 13). This is because the mask of sensor-visible footprint segments for the building on the left contains only one edge, which does not provide adequate information. Moreover, we notice that the overlapping region between these two buildings can only be well identified by models trained with complete building footprints.

Overall, these results suggest that complete building footprints are more befitting for the segmentation of individual buildings than sensor-visible footprint segments. This may be because the former delivers more information, especially for low-rise buildings.

F. Can CG-Net Work With Inaccurate GIS Data?

So far, building footprints used in our experiments are highly accurate as they are acquired from official GIS data. However, most openly available GIS data, such as OpenStreetMap (OSM), often contain positioning errors. To test the performance of CG-Net in such cases, we conduct supplementary experiments on training our CG-Net with inaccurate building footprints and discuss the impact of positioning errors in GIS data.

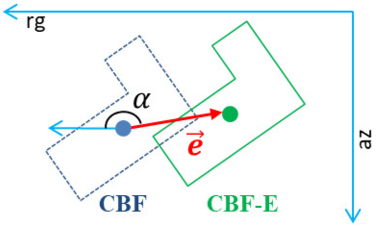


Fig. 14. Illustration of generating building footprints with positioning errors. Positioning error $\vec{e}$ is added to building footprint CBF, resulting in that CBF-E. rg and az denote the range direction and the azimuth direction, respectively. α is the angle between $\vec{e}$ and rg.

First, we generate inaccurate CBF, termed CBF-E, by injecting positioning errors. As illustrated in Fig. 14, $\vec{e}$ is the added positioning error, and α is the angle between $\vec{e}$ and the range direction. According to the quality assessment study of OSM in [83], the average offset of building footprints is 4.13 m with a standard deviation of 1.71 m. Therefore, we consider the positioning error as a variable whose magnitude is Gaussian distributed, i.e., $|\vec{e}| \sim \mathcal{N}(\mu = 4.13, \sigma^2 = 1.71^2)$. Since the offset may point to different directions, we assume that the direction of $\vec{e}$ is uniformly distributed, i.e., α is uniformly distributed in the range of $[0°, 360°)$. For simplicity, let α be discrete: $\alpha \sim \text{DiscreteUniform}(0°, 359°)$. Note that this is the most difficult case that all footprints contain positioning errors.

TABLE III
NUMERICAL RESULTS OF DEEPLABV3-CG TRAINED USING CBF AND CBF-E

GIS data used for training	P	F1 score	IoU	OA
CBF	0.7523	0.7508	0.6010	0.9937
CBF-E	0.7221	0.7146	0.5560	0.9927

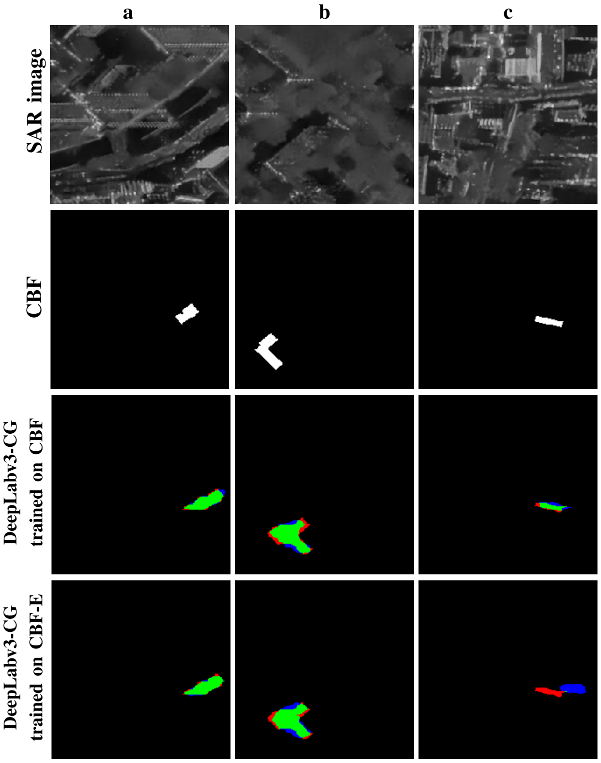


Fig. 15. Examples of segmentation results of networks trained using CBFs and networks trained using building footprints with positioning errors (abbreviated as CBF-E). Pixel-based true positives, false positives, and false negatives are marked in green, red, and blue, respectively.

Then, we train DeepLabv3-CG using CBF-E and SAR patches and test the trained network with a clean test set. DeepLabv3-CG is chosen because it performs best among all the networks. The parameter settings of the network remain the same as previous experiments, as described in Section IV-B.

The results are listed in Table III. As can be seen, comparing to results using CBF, the precision of the network trained on CBF-E is decreased by 3.02%, the F1 score is reduced by 3.62%, and the IoU is decreased by 4.5%. However, it still gives competent segmentation results. For visual comparison, Fig. 15 shows the results of DeepLabv3-CG trained with CBF-E and CBF. For the building in column c, DeepLabv3-CG trained with CBF performs much better than that with CBF-E. However, the predictions for buildings in columns a and b are visually very similar. Moreover, we observed that predictions from DeepLabv3-CG trained on CBF-E are satisfactory for most buildings.

The experiments show that, although weakened by positioning errors in GIS data, the proposed CG-Net is robust even in the most difficult case. This finding suggests that a large amount of existing open-sourced GIS data, such as OSM, can be exploited for segmenting individual buildings in SAR images.

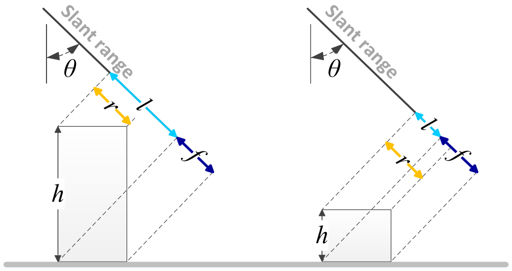


Fig. 16. Projection geometry of two flat-roof buildings in a slant-range SAR image. θ is the incidence angle. h is the building height. l, r, and f denote the length of layover, roof, and footprint areas in a slant-range SAR image, respectively.

V. FURTHER APPLICATION: RECONSTRUCTION OF LOD1 BUILDING MODELS FROM AN SAR IMAGE

Building models can be created at different levels-of-detail (LoD). According to the terminology of CityGML [84], LoD1 models represent buildings as blocks with flat roof structures and can be reconstructed by extruding footprints with building heights. Here, we regard the average roof height as the building height.[4] In this section, we demonstrate the process of reconstructing LoD1 models using our predicted individual building masks.

Fig. 16 illustrates the projection geometry of two flat-roof buildings in a constant azimuth profile of an SAR image. θ is the incidence angle. l, r, and f denote the length of layover, roof, and footprint areas in the slant-range SAR image, respectively. Notably, the building region in the SAR image contains both the layover and the roof areas. The layover area coincides with the building region when the building height h is large, e.g., the case in Fig. 16 (left), and it is covered by the building region when h is small, e.g., the case in Fig. 16 (right). In both cases, the layover area can be calculated by subtracting the footprint from the building region. Therefore, l is estimated to be the length of the layover area in the slant-range direction, and h can be computed with the following equation:

$$h = l / \cos\theta. \tag{6}$$

From the predicted individual building masks (see Fig. 17), we calculate building heights with (6). Afterward, LoD1 building models are created by extruding building footprints with obtained heights. Fig. 18 presents example LoD1 models superimposed on the SAR image in the study area. It can be observed that buildings with large l (pointed by yellow arrows) are predicted as high-rise, while those with small l (pointed by red arrows) are reconstructed as low-rise buildings. This is in line with reality. We further evaluate the estimated height against the mean height from the accurate DEM for each building. The mean height error that we achieve in the

[4] http://en.wiki.quality.sig3d.org/index.php/Modeling_Guide_for_3D_Objects_-_Part_2:_Modeling_of_Buildings_(LoD1,_LoD2,_LoD3)

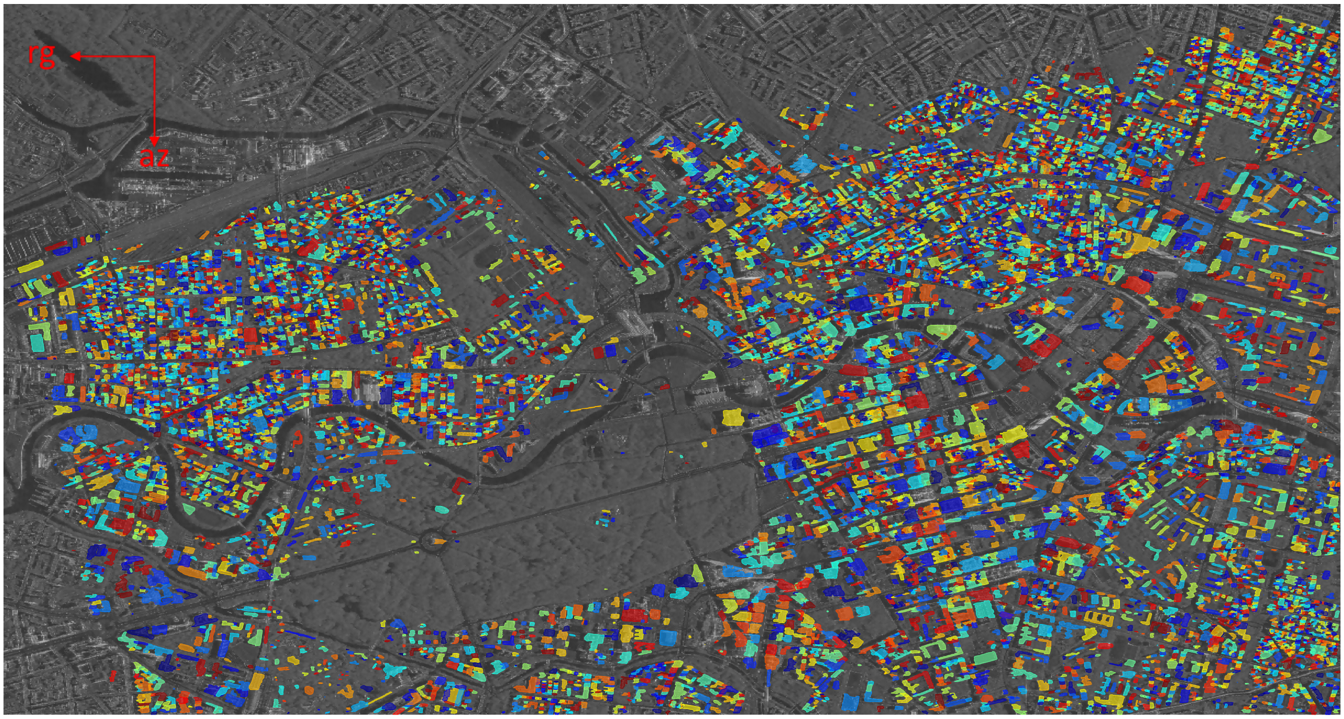


Fig. 17. Segmentation results in the study area obtained by DeepLabv3-CG. The building segments are plotted with different colors translucently for visualizing the layover areas between buildings. rg and az denote the range direction and the azimuth direction, respectively.

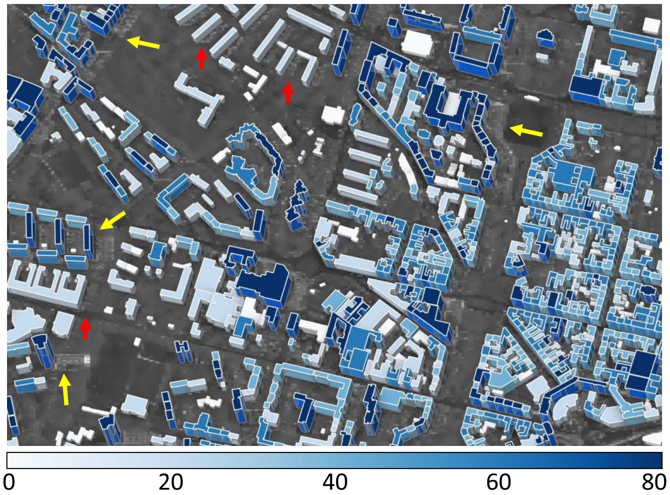


Fig. 18. Example LoD1 building models in the study area superimposed on the SAR image. Layover areas of some buildings are visible, as pointed by the yellow and red arrows. Building heights are color-coded.

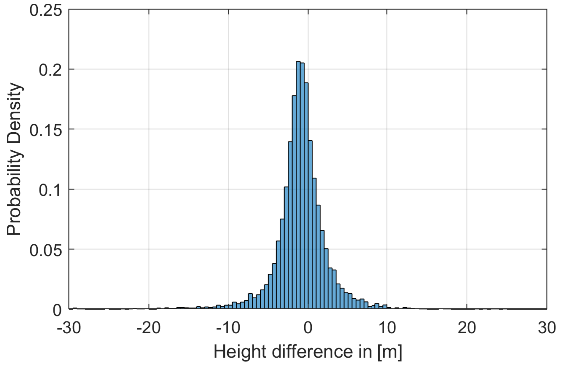


Fig. 19. Histogram of building height errors in the study area.

study site is 2.39 m. The histogram of height errors is shown in Fig. 19.

VI. Conclusion

In this article, we propose a CG-Net to segment individual buildings from a large-scale VHR SAR image. We also propose an approach for generating ground truth annotations of buildings using a high-resolution DEM. The proposed method is evaluated in the Berlin area, using a high-resolution spotlight TerraSAR-X image and building footprints obtained from GIS data. Both qualitative and quantitative results demonstrate the effectiveness of the proposed CG-module. Compared with competitors, DeepLabv3-CG achieves the best F1 score of 75.08%. In addition, we compare two building footprint representations, namely complete building footprints and sensor-visible footprint segments. Experimental results suggest that the use of complete building footprints leads to better results. Further experiments of training the networks using inaccurate GIS data suggest that CG-Net is robust in presence of positioning errors in GIS data. In addition, we demonstrate an application of our results, i.e., LoD1 building model reconstruction. In the future, we are interested in applying the proposed data generation workflow to areas of various urban morphologies and using our CG-Net to reconstruct LoD1 building models from TerraSAR-X and TanDEM-X stripmap images.

Acknowledgment

The authors would like to thank Dr. H. Hirschmüller of DLR-RM for providing the optical DEM.

References

[1] Y. Sun, Y. Hua, L. Mou, and X. X. Zhu, “Large-scale building height estimation from single VHR SAR image using fully convolutional network and GIS building footprints,” in *Proc. Joint Urban Remote Sens. Event (JURSE)*, May 2019, pp. 1–4.

[2] G. Franceschetti, A. Iodice, and D. Riccio, "A canonical problem in electromagnetic backscattering from buildings," *IEEE Trans. Geosci. Remote Sens.*, vol. 40, no. 8, pp. 1787–1801, Aug. 2002.

[3] F. Tupin and M. Roux, "Detection of building outlines based on the fusion of SAR and optical features," *ISPRS J. Photogramm. Remote Sens.*, vol. 58, nos. 1–2, pp. 71–82, Jun. 2003.

[4] R. Guida, A. Iodice, and D. Riccio, "Height retrieval of isolated buildings from single high-resolution SAR images," *IEEE Trans. Geosci. Remote Sens.*, vol. 48, no. 7, pp. 2967–2979, Jul. 2010.

[5] D. Brunner, G. Lemoine, L. Bruzzone, and H. Greidanus, "Building height retrieval from VHR SAR imagery based on an iterative simulation and matching technique," *IEEE Trans. Geosci. Remote Sens.*, vol. 48, no. 3, pp. 1487–1504, Mar. 2010.

[6] H. Sportouche, F. Tupin, and L. Denise, "Extraction and three-dimensional reconstruction of isolated buildings in urban scenes from high-resolution optical and SAR spaceborne images," *IEEE Trans. Geosci. Remote Sens.*, vol. 49, no. 10, pp. 3932–3946, Oct. 2011.

[7] W. Liu and F. Yamazaki, "Building height detection from high-resolution TerraSAR-X imagery and GIS data," in *Proc. Joint Urban Remote Sens. Event*, Apr. 2013, pp. 33–36.

[8] X. X. Zhu and M. Shahzad, "Facade reconstruction using multiview spaceborne TomoSAR point clouds," *IEEE Trans. Geosci. Remote Sens.*, vol. 52, no. 6, pp. 3541–3552, Jun. 2014.

[9] B. Huang, Y. Li, X. Han, Y. Cui, W. Li, and R. Li, "Cloud removal from optical satellite imagery with SAR imagery using sparse representation," *IEEE Geosci. Remote Sens. Lett.*, vol. 12, no. 5, pp. 1046–1050, May 2015.

[10] D. Brunner, G. Lemoine, and L. Bruzzone, "Earthquake damage assessment of buildings using VHR optical and SAR imagery," *IEEE Trans. Geosci. Remote Sens.*, vol. 48, no. 5, pp. 2403–2420, May 2010.

[11] T.-L. Wang and Y.-Q. Jin, "Postearthquake building damage assessment using multi-mutual information from pre-event optical image and postevent SAR image," *IEEE Geosci. Remote Sens. Lett.*, vol. 9, no. 3, pp. 452–456, May 2012.

[12] M. Shahzad, M. Maurer, F. Fraundorfer, Y. Wang, and X. X. Zhu, "Buildings detection in VHR SAR images using fully convolution neural networks," *IEEE Trans. Geosci. Remote Sens.*, vol. 57, no. 2, pp. 1100–1116, Feb. 2019.

[13] F. Xu and Y.-Q. Jin, "Automatic reconstruction of building objects from multiaspect meter-resolution SAR images," *IEEE Trans. Geosci. Remote Sens.*, vol. 45, no. 7, pp. 2336–2353, Jul. 2007.

[14] E. Michaelsen, U. Soergel, and U. Thoennessen, "Perceptual grouping for automatic detection of man-made structures in high-resolution SAR data," *Pattern Recognit. Lett.*, vol. 27, no. 4, pp. 218–225, Mar. 2006.

[15] U. Soergel, E. Michaelsen, A. Thiele, E. Cadario, and U. Thoennessen, "Stereo analysis of high-resolution SAR images for building height estimation in cases of orthogonal aspect directions," *ISPRS J. Photogramm. Remote Sens.*, vol. 64, no. 5, pp. 490–500, Sep. 2009.

[16] A. Ferro, D. Brunner, and L. Bruzzone, "Automatic detection and reconstruction of building radar footprints from single VHR SAR images," *IEEE Trans. Geosci. Remote Sens.*, vol. 51, no. 2, pp. 935–952, Feb. 2013.

[17] R. D. Hill, C. P. Moate, and D. Blacknell, "Estimating building dimensions from synthetic aperture radar image sequences," *IET Radar, Sonar Navigat.*, vol. 2, no. 3, p. 189, 2008.

[18] R. Bolter and F. Leberl, "Shape-from-shadow building reconstruction from multiple view SAR images," in *Proc. 24th Workshop Austrian Assoc. Pattern Recognit. (ÖAGM/AAPR)*, 2000, pp. 199–206.

[19] F. Cellier, H. Oriot, and J.-M. Nicolas, "Introduction of the mean shift algorithm in SAR imagery: Application to shadow extraction for building reconstruction," in *Proc. EARSeL Workshop 3D-Remote Sens.*, 2005, pp. 10–11.

[20] L. Zhao, X. Zhou, and G. Kuang, "Building detection from urban SAR image using building characteristics and contextual information," *EURASIP J. Adv. Signal Process.*, vol. 2013, no. 1, p. 56, Dec. 2013.

[21] H. Sportouche, F. Tupin, and L. Denise, "Building extraction and 3D reconstruction in urban areas from high-resolution optical and SAR imagery," in *Proc. Joint Urban Remote Sens. Event*, May 2009, pp. 1–11.

[22] A. Thiele, C. Dubois, E. Cadario, and S. Hinz, "GIS-supported iterative filtering approach for building height estimation from InSAR data," in *Proc. Eur. Conf. Synth. Aperture Radar (EUSAR)*, 2012, pp. 19–22.

[23] Y. Zhang, X. Sun, A. Thiele, and S. Hinz, "Stochastic geometrical model and Monte Carlo optimization methods for building reconstruction from InSAR data," *ISPRS J. Photogramm. Remote Sens.*, vol. 108, pp. 49–61, Oct. 2015.

[24] M. Quartulli and M. Datcu, "Stochastic geometrical modeling for built-up area understanding from a single SAR intensity image with meter resolution," *IEEE Trans. Geosci. Remote Sens.*, vol. 42, no. 9, pp. 1996–2003, Sep. 2004.

[25] R. Guida, A. Iodice, D. Riccio, and U. Stilla, "Model-based interpretation of high-resolution SAR images of buildings," *IEEE J. Sel. Topics Appl. Earth Observ. Remote Sens.*, vol. 1, no. 2, pp. 107–119, Jun. 2008.

[26] E. Simonetto, H. Oriot, and R. Garello, "Rectangular building extraction from stereoscopic airborne radar images," *IEEE Trans. Geosci. Remote Sens.*, vol. 43, no. 10, pp. 2386–2395, Oct. 2005.

[27] Y. Wang, F. Tupin, C. Han, and J.-M. Nicolas, "Building detection from high resolution PolSAR data by combining region and edge information," in *Proc. IEEE Int. Geosci. Remote Sens. Symp. (IGARSS)*, 2008, pp. IV-153–IV-156.

[28] B. Liu, K. Tang, and J. Liang, "A bottom-up/top-down hybrid algorithm for model-based building detection in single very high resolution SAR image," *IEEE Geosci. Remote Sens. Lett.*, vol. 14, no. 6, pp. 926–930, Jun. 2017.

[29] F. Zhang, Y. Shao, X. Zhang, and T. Balz, "Building L-shape footprint extraction from high resolution SAR image," in *Proc. Joint Urban Remote Sens. Event*, Apr. 2011, pp. 273–276.

[30] J. D. Wegner, J. R. Ziehn, and U. Soergel, "Combining high-resolution optical and InSAR features for height estimation of buildings with flat roofs," *IEEE Trans. Geosci. Remote Sens.*, vol. 52, no. 9, pp. 5840–5854, Sep. 2014.

[31] A. Thiele, E. Cadario, K. Schulz, and U. Soergel, "Analysis of gable-roofed building signature in multiaspect InSAR data," *IEEE Geosci. Remote Sens. Lett.*, vol. 7, no. 1, pp. 83–87, Jan. 2010.

[32] J. Chen, C. Wang, H. Zhang, F. Wu, B. Zhang, and W. Lei, "Automatic detection of low-rise gable-roof building from single submeter SAR images based on local multilevel segmentation," *Remote Sens.*, vol. 9, no. 3, p. 263, Mar. 2017.

[33] R. Guo and X. X. Zhu, "High-rise building feature extraction using high resolution spotlight TanDEM-X data," in *Proc. Eur. Conf. Synth. Aperture Radar (EUSAR)*, 2014, pp. 1–4.

[34] W. Liu, K. Suzuki, and F. Yamazaki, "Height estimation for high-rise buildings based on InSAR analysis," in *Proc. Joint Urban Remote Sens. Event (JURSE)*, 2015, pp. 1–4.

[35] K. Tang, B. Liu, and B. Zou, "High-rise building detection in dense urban area based on high resolution SAR images," in *Proc. IEEE Int. Geosci. Remote Sens. Symp. (IGARSS)*, Jul. 2016, pp. 1568–1571.

[36] S. Chen, H. Wang, F. Xu, and Y.-Q. Jin, "Automatic recognition of isolated buildings on single-aspect SAR image using range detector," *IEEE Geosci. Remote Sens. Lett.*, vol. 12, no. 2, pp. 219–223, Feb. 2015.

[37] P.-P. Lu, K.-N. Du, W.-D. Yu, and H. Feng, "New building signature extraction method from single very high-resolution synthetic aperture radar images based on symmetric analysis," *J. Appl. Remote Sens.*, vol. 9, no. 1, Jul. 2015, Art. no. 095072.

[38] M. Shahzad and X. X. Zhu, "Automatic detection and reconstruction of 2-D/3-D building shapes from spaceborne TomoSAR point clouds," *IEEE Trans. Geosci. Remote Sens.*, vol. 54, no. 3, pp. 1292–1310, Mar. 2016.

[39] X. X. Zhu and R. Bamler, "Very high resolution spaceborne SAR tomography in urban environment," *IEEE Trans. Geosci. Remote Sens.*, vol. 48, no. 12, pp. 4296–4308, Dec. 2010.

[40] J. D. Wegner, U. Soergel, and A. Thiele, "Building extraction in urban scenes from high-resolution InSAR data and optical imagery," in *Proc. Joint Urban Remote Sens. Event (JURSE)*, 2009, pp. 1–6.

[41] A. Thiele, S. Hinz, and E. Cadario, "Combining GIS and InSAR data for 3D building reconstruction," in *Proc. IEEE Int. Geosci. Remote Sens. Symp.*, Jul. 2010, pp. 2418–2421.

[42] Y. Sun, M. Shahzad, and X. X. Zhu, "Building height estimation in single SAR image using OSM building footprints," in *Proc. Joint Urban Remote Sens. Event (JURSE)*, Mar. 2017, pp. 1–7.

[43] L. Zhang, L. Zhang, and B. Du, "Deep learning for remote sensing data: A technical tutorial on the state of the art," *IEEE Geosci. Remote Sens. Mag.*, vol. 4, no. 2, pp. 22–40, Jun. 2016.

[44] X. X. Zhu *et al.*, "Deep learning in remote sensing: A comprehensive review and list of resources," *IEEE Geosci. Remote Sens. Mag.*, vol. 5, no. 4, pp. 8–36, Dec. 2017.

[45] L. Mou, Y. Hua, and X. X. Zhu, "A relation-augmented fully convolutional network for semantic segmentation in aerial scenes," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2019, pp. 12416–12425.

[46] N. Kussul, M. Lavreniuk, S. Skakun, and A. Shelestov, "Deep learning classification of land cover and crop types using remote sensing data," *IEEE Geosci. Remote Sens. Lett.*, vol. 14, no. 5, pp. 778–782, May 2017.

[47] G. Cheng, C. Yang, X. Yao, L. Guo, and J. Han, "When deep learning meets metric learning: Remote sensing image scene classification via learning discriminative CNNs," *IEEE Trans. Geosci. Remote Sens.*, vol. 56, no. 5, pp. 2811–2821, May 2018.

[48] L. Mou and X. Xiang Zhu, "IM2HEIGHT: Height estimation from single monocular imagery via fully residual convolutional-deconvolutional network," 2018, *arXiv:1802.10249*. [Online]. Available: http://arxiv.org/abs/1802.10249

[49] R. Kemker, C. Salvaggio, and C. Kanan, "Algorithms for semantic segmentation of multispectral remote sensing imagery using deep learning," *ISPRS J. Photogramm. Remote Sens.*, vol. 145, pp. 60–77, Nov. 2018.

[50] N. Audebert, B. Le Saux, and S. Lefèvre, "Beyond RGB: Very high resolution urban remote sensing with multimodal deep networks," *ISPRS J. Photogramm. Remote Sens.*, vol. 140, pp. 20–32, Jun. 2018.

[51] Y. Hua, L. Mou, and X. X. Zhu, "Relation network for multilabel aerial image classification," *IEEE Trans. Geosci. Remote Sens.*, vol. 58, no. 7, pp. 4558–4572, Jul. 2020, doi: 10.1109/TGRS.2019.2963364.

[52] L. Mou, Y. Hua, and X. X. Zhu, "Relation matters: Relational context-aware fully convolutional network for semantic segmentation of high-resolution aerial images," *IEEE Trans. Geosci. Remote Sens.*, vol. 58, no. 11, pp. 7557–7569, Nov. 2020, doi: 10.1109/TGRS.2020.2979552.

[53] Q. Lv, Y. Dou, X. Niu, J. Xu, J. Xu, and F. Xia, "Urban land use and land cover classification using remotely sensed SAR data through deep belief networks," *J. Sensors*, vol. 2015, pp. 1–10, Jul. 2015.

[54] Y. Zhou, H. Wang, F. Xu, and Y.-Q. Jin, "Polarimetric SAR image classification using deep convolutional neural networks," *IEEE Geosci. Remote Sens. Lett.*, vol. 13, no. 12, pp. 1935–1939, Dec. 2016.

[55] Z. Zhang, H. Wang, F. Xu, and Y.-Q. Jin, "Complex-valued convolutional neural network and its application in polarimetric SAR image classification," *IEEE Trans. Geosci. Remote Sens.*, vol. 55, no. 12, pp. 7177–7188, Dec. 2017.

[56] Z. Zhao, L. Jiao, J. Zhao, J. Gu, and J. Zhao, "Discriminant deep belief network for high-resolution SAR image classification," *Pattern Recognit.*, vol. 61, pp. 686–701, Jan. 2017.

[57] J. Geng, H. Wang, J. Fan, and X. Ma, "Deep supervised and contractive neural network for SAR image classification," *IEEE Trans. Geosci. Remote Sens.*, vol. 55, no. 4, pp. 2442–2459, Apr. 2017.

[58] Y. Duan, F. Liu, L. Jiao, P. Zhao, and L. Zhang, "SAR image segmentation based on convolutional-wavelet neural network and Markov random field," *Pattern Recognit.*, vol. 64, pp. 255–267, Apr. 2017.

[59] F. Mohammadimanesh, B. Salehi, M. Mahdianpari, E. Gill, and M. Molinier, "A new fully convolutional neural network for semantic segmentation of polarimetric SAR imagery in complex land cover ecosystem," *ISPRS J. Photogramm. Remote Sens.*, vol. 151, pp. 223–236, May 2019.

[60] S. Chen and H. Wang, "SAR target recognition based on deep learning," in *Proc. Int. Conf. Data Sci. Adv. Analytics (DSAA)*, Oct. 2014, pp. 541–547.

[61] J. Ding, B. Chen, H. Liu, and M. Huang, "Convolutional neural network with data augmentation for SAR target recognition," *IEEE Geosci. Remote Sens. Lett.*, vol. 13, no. 3, pp. 364–368, Mar. 2016.

[62] M. Kang, K. Ji, X. Leng, and Z. Lin, "Contextual region-based convolutional neural network with multilayer fusion for SAR ship detection," *Remote Sens.*, vol. 9, no. 8, p. 860, Aug. 2017.

[63] F. Gao, T. Huang, J. Sun, J. Wang, A. Hussain, and E. Yang, "A new algorithm for SAR image target recognition based on an improved deep convolutional neural network," *Cognit. Comput.*, vol. 11, no. 6, pp. 809–824, Dec. 2019.

[64] M. Gong, J. Zhao, J. Liu, Q. Miao, and L. Jiao, "Change detection in synthetic aperture radar images based on deep neural networks," *IEEE Trans. Neural Netw. Learn. Syst.*, vol. 27, no. 1, pp. 125–138, Jan. 2016.

[65] F. Gao, J. Dong, B. Li, and Q. Xu, "Automatic change detection in synthetic aperture radar images based on PCANet," *IEEE Geosci. Remote Sens. Lett.*, vol. 13, no. 12, pp. 1792–1796, Dec. 2016.

[66] J. Geng, X. Ma, X. Zhou, and H. Wang, "Saliency-guided deep neural networks for SAR image change detection," *IEEE Trans. Geosci. Remote Sens.*, vol. 57, no. 10, pp. 7365–7377, Oct. 2019.

[67] X. Wang, L. Cavigelli, M. Eggimann, M. Magno, and L. Benini, "HR-SAR-net: A deep neural network for urban scene segmentation from high-resolution SAR data," 2019, *arXiv:1912.04441*. [Online]. Available: http://arxiv.org/abs/1912.04441

[68] J. Shermeyer *et al.*, "SpaceNet 6: Multi-sensor all weather mapping dataset," 2020, *arXiv:2004.06500*. [Online]. Available: http://arxiv.org/abs/2004.06500

[69] R. Weibel and M. Heller, "Digital terrain modeling," in *Geographical Information Systems: Principles and Applications*, D. J. Maguire, M. F. Goodchild, and D. W. Rhind, Eds. London, U.K.: Longman, 1991, pp. 269–297.

[70] S. Katz, A. Tal, and R. Basri, "Direct visibility of point sets," *ACM Trans. Graph.*, vol. 26, no. 3, p. 24, Jul. 2007.

[71] J. C. Curlander, "Location of spaceborne SAR imagery," *IEEE Trans. Geosci. Remote Sens.*, vols. GE–20, no. 3, pp. 359–364, Jul. 1982.

[72] M. Schwabisch, "A fast and efficient technique for SAR interferogram geocoding," in *Proc. Sens. Manag. Environment. IEEE Int. Geosci. Remote Sensing. Symp. (IGARSS)*, Jul. 1998, pp. 1100–1102.

[73] T. Toutin, "Review article: Geometric processing of remote sensing images: Models, algorithms and methods," *Int. J. Remote Sens.*, vol. 25, no. 10, pp. 1893–1924, May 2004.

[74] A. Roth, M. Huber, and D. Kosmann, "Geocoding of TerraSAR-X data," in *Proc. Int. Congr. ISPRS*, 2004, pp. 840–844.

[75] F. R. Gonzalez, N. Adam, A. Parizzi, and R. Brcic, "The integrated wide area processor (IWAP): A processor for wide area persistent scatterer interferometry," in *Proc. ESA Living Planet Symp.*, 2013, p. 353.

[76] K. Simonyan and A. Zisserman, "Very deep convolutional networks for large-scale image recognition," in *Proc. IEEE Int. Conf. Learn. Represent. (ICLR)*, Jun. 2015, pp. 1–14.

[77] A. Krizhevsky, I. Sutskever, and G. E. Hinton, "ImageNet classification with deep convolutional neural networks," in *Proc. Adv. Neural Inf. Process. Syst. (NIPS)*, 2012, pp. 84–90.

[78] S. Ioffe and C. Szegedy, "Batch normalization: Accelerating deep network training by reducing internal covariate shift," in *Proc. Int. Conf. Mach. Learn. (ICML)*, 2015, pp. 1–11.

[79] G. Baier, X. X. Zhu, M. Lachaise, H. Breit, and R. Bamler, "Nonlocal InSAR filtering for DEM generation and addressing the staircasing effect," in *Proc. Eur. Conf. Synth. Aperture Radar (EUSAR)*, Jun. 2016, pp. 1–6.

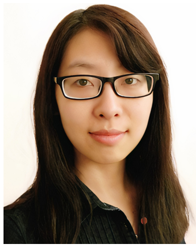
[80] H. Hirschmuller, "Stereo processing by semiglobal matching and mutual information," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 30, no. 2, pp. 328–341, Feb. 2008.

[81] J. Deng, W. Dong, R. Socher, L.-J. Li, K. Li, and L. Fei-Fei, "ImageNet: A large-scale hierarchical image database," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2009, pp. 248–255.

[82] T. Dozat. *Incorporating Nesterov Momentum Into Adam*. Accessed: Oct. 30, 2019. [Online]. Available: http://cs229.stanford.edu/proj2015/054_report.pdf

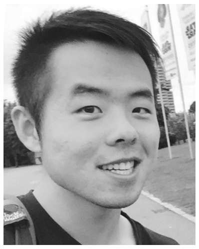
[83] H. Fan, A. Zipf, Q. Fu, and P. Neis, "Quality assessment for building footprints data on OpenStreetMap," *Int. J. Geographical Inf. Sci.*, vol. 28, no. 4, pp. 700–719, Apr. 2014.

[84] T. H. Kolbe, G. Gröger, and L. Plümer, "CityGML: Interoperable access to 3D city models," in *Proc. Int. Symp. Geo-Inf. Disaster Manage. (GiDM)*, 2005, pp. 883–899.



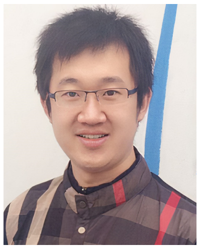
Yao Sun received the bachelor's degree in cartography and geo-information system from Wuhan University, Wuhan, China, in 2012, and the master's degree in Earth oriented space science and technology (ESPACE) from the Technical University of Munich (TUM), Munich, Germany, in 2016. She is pursuing the Ph.D. degree with the German Aerospace Center (DLR), Weßling, Germany and TUM.

Her main research interests are remote sensing, computer vision, and deep learning, especially object reconstruction from SAR data.



Yuansheng Hua (Student Member, IEEE) received the bachelor's degree in remote sensing science and technology from Wuhan University, Wuhan, China, in 2014, and the master's degree in Earth oriented space science and technology (ESPACE) from the Technical University of Munich (TUM), Munich, Germany, in 2018. He is pursuing the Ph.D. degree with the German Aerospace Center (DLR), Weßling, Germany and TUM.

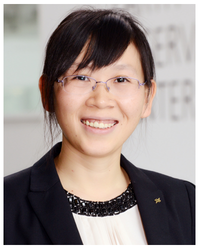
In 2019, he was a Visiting Researcher with the Wageningen University and Research, Wageningen, The Netherlands. His research interests include remote sensing, computer vision, and deep learning, especially their applications in remote sensing.



Lichao Mou received the bachelor's degree in automation from the Xi'an University of Posts and Telecommunications, Xi'an, China, in 2012, the master's degree in signal and information processing from the University of Chinese Academy of Sciences (UCAS), Beijing, China, in 2015, and the Dr.-Ing. degree from the Technical University of Munich (TUM), Munich, Germany, in 2020.

He is a Guest Professor with the Munich AI Future Laboratory AI4EO, TUM and the Head of Visual Learning and Reasoning Team with the Department "EO Data Science", Remote Sensing Technology Institute (IMF), German Aerospace Center (DLR), Weßling, Germany. Since 2019, he has been an AI Consultant for the Helmholtz Artificial Intelligence Cooperation Unit (HAICU). In 2015, he spent six months with Computer Vision Group, University of Freiburg, Freiburg, Germany. In 2019, he was a Visiting Researcher with Cambridge Image Analysis Group (CIA), University of Cambridge, Cambridge, U.K. From 2019 to 2020, he was a Research Scientist with DLR-IMF.

Dr. Mou was a recipient of the first place in the 2016 IEEE GRSS Data Fusion Contest and the Finalists for the Best Student Paper Award at the 2017 Joint Urban Remote Sensing Event and 2019 Joint Urban Remote Sensing Event.



Xiao Xiang Zhu (Fellow, IEEE) received the M.Sc., Dr.-Ing., and Habilitation degrees in signal processing from the Technical University of Munich (TUM), Munich, Germany, in 2008, 2011, and 2013, respectively.

She is the Professor for Data Science in Earth Observation (former: Signal Processing in Earth Observation) with TUM and the Head of the Department "EO Data Science," Remote Sensing Technology Institute, German Aerospace Center (DLR), Weßling, Germany. Since 2019, she has been a Co-Coordinator of the Munich Data Science Research School. Since 2019, she has been the Head of the Helmholtz Artificial Intelligence–Research Field "Aeronautics, Space and Transport." Since May 2020, she has been the Director of the International Future AI Laboratory "AI4EO–Artificial Intelligence for Earth Observation: Reasoning, Uncertainties, Ethics and Beyond," Munich. Since October 2020, she has also been serving on the Board of Directors of the Munich Data Science Institute (MDSI), TUM. She was a Guest Scientist or a Visiting Professor with the Italian National Research Council (CNR-IREA), Naples, Italy, Fudan University, Shanghai, China, The University of Tokyo, Tokyo, Japan, and the University of California at Los Angeles, Los Angeles, CA, USA, in 2009, 2014, 2015, and 2016, respectively. Her main research interests are remote sensing and Earth observation, signal processing, machine learning, and data science, with a special application focus on global urban mapping.

Dr. Zhu is a member of Young Academy (Junge Akademie/Junges Kolleg) at the Berlin-Brandenburg Academy of Sciences and Humanities and the German National Academy of Sciences Leopoldina and the Bavarian Academy of Sciences and Humanities. She is an Associate Editor of the IEEE TRANSACTIONS ON GEOSCIENCE AND REMOTE SENSING.