Buildings Detection in VHR SAR Images Using Fully Convolution Neural Networks

Muhammad Shahzad, *Member, IEEE*, Michael Maurer, Friedrich Fraundorfer, Yuanyuan Wang, *Member, IEEE*, and Xiao Xiang Zhu, *Senior Member, IEEE*

***Abstract*— This paper addresses the highly challenging problem of automatically detecting man-made structures especially buildings in very high-resolution (VHR) synthetic aperture radar (SAR) images. In this context, this paper has two major contributions. First, it presents a novel and generic workflow that initially classifies the spaceborne SAR tomography (TomoSAR) point clouds—generated by processing VHR SAR image stacks using advanced interferometric techniques known as TomoSAR—into buildings and nonbuildings with the aid of auxiliary information (i.e., either using openly available 2-D building footprints or adopting an optical image classification scheme) and later back project the extracted building points onto the SAR imaging coordinates to produce automatic large-scale benchmark labeled (buildings/nonbuildings) SAR data sets. Second, these labeled data sets (i.e., building masks) have been utilized to construct and train the state-of-the-art deep fully convolution neural networks with an additional conditional random field represented as a recurrent neural network to detect building regions in a single VHR SAR image. Such a cascaded formation has been successfully employed in computer vision and remote sensing fields for optical image classification but, to our knowledge, has not been applied to SAR images. The results of the building detection are illustrated and validated over a TerraSAR-X VHR spotlight SAR image covering approximately 39 km^2—almost the whole city of Berlin— with the mean pixel accuracies of around 93.84%.**

***Index Terms*— Building detection, fully convolution neural networks (CNNs), OpenStreetMap (OSM), synthetic aperture radar (SAR), SAR tomography (TomoSAR), TerraSAR-X/TanDEM-X.**

Manuscript received October 13, 2017; revised February 12, 2018 and June 29, 2018; accepted July 9, 2018. Date of publication October 9, 2018; date of current version January 21, 2019. This work was supported in part by the European Research Council through the European Union's Horizon 2020 Research and Innovation Program (Acronym: *So2Sat*) under Grant ERC-2016-StG-714087 and in part by the Helmholtz Association through the Framework of the Young Investigators Group "SiPEO" under Grant VH-NG-1018. *(Muhammad Shahzad and Michael Maurer contributed equally to this work.) (Corresponding author: Xiao Xiang Zhu.)*

M. Shahzad is with the School of Electrical Engineering and Computer Science, National University of Sciences and Technology, Islamabad 44000, Pakistan, and also with the Signal Processing in Earth Observation, Technical University of Munich, 80333 Munich, Germany (e-mail: muhammad.shehzad@seecs.edu.pk).

M. Maurer and F. Fraundorfer are with the Institute of Computer Graphics and Vision, Graz University of Technology, 8010 Graz, Austria (e-mail: maurer@icg.tugraz.at; fraundorfer@icg.tugraz.at).

Y. Wang is with the Signal Processing in Earth Observation, Technical University of Munich, 80333 Munich, Germany (e-mail: yuanyuan.wang@dlr.de).

X. X. Zhu is with the Signal Processing in Earth Observation, Technical University of Munich, 80333 Munich, Germany, and also with the Remote Sensing Technology Institute, German Aerospace Center, 82234 Weßling, Germany (e-mail: xiao.zhu@dlr.de).

Color versions of one or more of the figures in this paper are available online at http://ieeexplore.ieee.org.

Digital Object Identifier 10.1109/TGRS.2018.2864716

I. Introduction

AUTOMATIC detection of man-made objects, in particular buildings from a single very high-resolution (VHR) synthetic aperture radar (SAR) image, is of great practical significance, particularly in applications having stringent temporal restrictions, e.g., emergency responses. However, owing to inherent complexity of SAR images caused by the so-called speckle effect together with radiometric distortions mainly originating due to side-looking geometry, scene interpretation using SAR images is highly challenging. Particularly in urban areas, such distortions render the data to be mainly characterized by multibounce, layover, and shadowing effects consequently giving rise to the need of automatic and robust algorithms for object detection from SAR images.

A variety of algorithms have been published in the literature that aims at the detection and reconstruction of buildings from SAR images. Typically, most of the developed approaches rely on auxiliary information, e.g., the multisensor data provided by the optical [1], [2] and light detection and ranging [3] sensors, geographic information system (GIS) data, e.g., 2-D building footprints [4], multidimensional data, e.g., polarimetric SAR (PolSAR) [5], or multiview/multiaspect data such as interferometric SAR (InSAR) [6]. These approaches have improved the feature extraction process by providing the complimentary information. To our knowledge, the literature using only a single SAR image in the context of building detection is rather sparse. Among few existing approaches, Quartulli and Datcu [7] employed an automatic stochastic algorithm to reconstruct buildings from a single SAR intensity image by modeling strong signals originated via dihedral scattering at the bottom and the layover at the roof edges of the building. Zhao *et al.* [8] proposed a building detection method based on a marker-controlled watershed algorithm. A similar approach that exploited layover and double bounce echoes to detect and determine the number of buildings from a single high-resolution image was provided in [9]. Ferro *et al.* [10] also developed a method that was primarily based on extracting a set of low-level bright (lines) and dark (shadows) primitives. Chen *et al.* [11] introduced a more recent 1-D range detector to determine the 2-D building footprints. The method could potentially reconstruct simple symmetrical building footprints but might fail for scenes containing more complex nonsymmetrical building shapes.

All the aforementioned approaches aim to extract buildings in an unsupervised (or data-driven) manner. Some researchers have also formulated the detection problem in a classification

framework to benefit from well-developed supervised learning methods typically used in computer vision [12], [13]. However, the effective utilization of such supervised learning methods has two practical limitations.

1) Extraction of distinctive features is necessary for reliable object detection.
2) A large annotated database is required, which is used for training and validation.

To address the first point, i.e., distinctive feature extraction, a number of approaches have been proposed. For example, raw pixels of images [13], magnitudes of 2-D Fourier coefficients [14], or discrete wavelet transform [15] have been used as features. Typically, feature extraction methods rely on heuristics in selecting appropriate features, and therefore, to cope with unaccounted situations (e.g., tolerance to incomplete views/poses in training data or randomness in speckle for different observations), expert knowledge is required to translate such discrepancies in the model for feature representation [16].

Recently, convolution neural networks (CNNs), a type of multilayered neural networks, have significantly outperformed previous methods and became the state of the art in image classification. Their power lies in the fact that they directly extract *high-level abstract* image features that allow replacing handcrafted features by the machine learning features fitting to the task at hand. They have special characteristics (i.e., shared weights architecture, local receptive fields, pooling, and spatial subsampling) that make them tolerant to high degree of image translations, skewing, scaling, rotation, and other forms of geometric distortions.

A. Related Work

There exist abundant works that employ CNNs to perform object detection in remote sensing images [17]–[19]. In this context, we refer the interested reader to an excellent recently published survey article containing a comprehensive review of deep learning techniques applied to optical remote sensing images [19]. In contrast, the use of CNNs over SAR images is up to now limited but consistently increasing. For instance, Profeta *et al.* [20] experimented with various CNN architectures on the moving and stationary target (MSTAR) SAR data set to achieve high classification accuracy. The MSTAR data set has also been utilized to perform SAR image segmentation in [21] and [22]. Ding *et al.* [16] investigated the capability of deep CNNs to address the issues in SAR target recognition, such as target translations, random speckle noise, and insufficient pose images in the training data. Utilization of CNNs in the PolSAR image classification has been demonstrated in [5]. Some researchers also explored CNNs to solve the change detection problem in SAR images [23]. Recently, the application of CNNs over TerraSAR-X spoltlight data stacks to classify built-up area has been demonstrated in [20]. The problem is particularly challenging, as the SAR images suffer from severe geometric distortions in urban areas, and therefore, they developed a robust multiscale CNN architecture to extract hierarchical features directly from SAR image patches. With the aim to develop benchmark SAR data set, Zhao *et al.* [24] also exploited CNNs over a TerraSAR-X spotlight data in image classification context and prepared a relatively large SAR image database containing five classes of object patches, including buildings, roads, vegetation, alongside, and water area. They demonstrated that the CNNs trained with fairly large training samples significantly improve the classification accuracy. Xu *et al.* [25] also demonstrated the use of CNNs over SAR images to extract buildings by manually preparing the training data set and later incorporating modern regularization techniques (e.g., data augmentation, dropout, and early stopping) to reduce testing errors.

As can be imagined, the precondition for the application of CNNs or any other supervised learning frameworks is the availability of annotated data sets. They are necessary not only to analyze and validate the performance of classification algorithms but are too required in the training phase where parts of annotated data are utilized to optimize prediction models. Lack of such annotated data sets is one of the major issues in the application of CNNs over SAR images. Manual (or somewhat interactive) annotation, as is done in the aforementioned approaches, is one potential solution. However, due to complex multiple scattering and different microwave scattering properties of the objects appearing in the scene possessing different geometrical and material features, the manual annotation often requires expert's knowledge (see Fig. 1) and easily becomes impractical when large scenes need to be processed. Apart from this, another possibility of generating such a reference SAR data set is by exploiting simulation-based methods as proposed, e.g., in [26]–[28]. However, such methods have their own limitations in a sense that they are either only capable of simulating simpler building shapes (see [28]) or typically require accurate models (3-D building models and/or accurate digital surface models) to precisely generate such ground-truth (GT) data which, in most cases, is not available. Thus, in view of the above, *automatic* annotation of SAR images, if possible, is essential.

B. Significant Contributions

The objective of this paper is twofold: first is to demonstrate the potential of automatic preparation of SAR training data sets for larger regions, and second, using the automatically prepared data set to train deep CNN architecture to detect buildings in a single VHR SAR image. This paper extends the initial idea [29] of automatic SAR annotation and performs a thorough analysis of the obtained SAR annotation and prediction results. The novel workflow presented in this paper involves the following.

1) Automatic generation of annotated SAR images using spaceborne SAR tomography (TomoSAR) point clouds generated by processing SAR image stacks via advanced interferometric technique known as TomoSAR [30], [31] together with auxiliary information to obtain subimage patches for training and validation.
2) Constructing a deep fully CNN with an additional conditional random field (CRF) represented as a recurrent neural network (RNN) to learn a classifier via transfer learning. Such a cascaded formation has been successfully employed in computer vision and remote sensing

(a) (b)

Fig. 1. Depicting the challenges of SAR image interpretation together with demonstrating the limitations of directly using the 2-D GIS building footprints onto the SAR image. (a) Optical image Google and (b) corresponding SAR image. *rg* and *az* refer to the range and azimuth coordinates, respectively. The three green polygons in (b) are the projections of available 2-D OSM building footprints depicted from top view in (a) onto the SAR image. It can be seen that when the illuminated scene contains elevated objects such as buildings, the so-called "layover" phenomenon (i.e., the superposition of multiple reflection sources in one pixel) occurs as a result of strong reflection of the façade in the SAR image which not only limits the direct usage of 2-D footprint projections for annotation/labeling but also makes the SAR image interpretation of urban areas highly challenging.

fields for optical image classification [32] but, to our knowledge, has not been applied to SAR images.

3) Utilizing the trained CNNs for the classification of pixels as belonging to building and nonbuilding for previously unseen input data.

The proposed workflow leads to the following contributions to the remote sensing community. We addressed the problem of automatic generation of annotated (labeled) data, which is always problematic to obtain in SAR images. In addition, we also addressed the usage of CNNs in SAR image classification, which is still a relatively new research area and has not been explored much. Last but not least, since the data sets used are widely available, the annotation approach is generic and may actually lead to new perspectives in producing benchmark data sets for SAR images.

II. Ground-Truth Generation (Annotation/Labeling of SAR Images)

Annotating an image is fundamental for the application of any supervised learning technique for segmentation/classification purposes. For this reason, we propose a novel workflow that utilizes the TomoSAR point clouds together with auxiliary information to automatically annotate (buildings/nonbuildings) SAR images of the area of interest. Before proceeding further, we briefly introduce these point clouds and later demonstrate their usage in such automatic annotation.

A. TomoSAR Point Cloud

TomoSAR is an advanced interferometric technique that actually aims at 3-D SAR imaging. It exploits the stacked SAR images acquired from slightly varying positions to build up a synthetic aperture in the third (i.e., elevation) axis, which consequently enables retrieving the precise 3-D localization of strong scatterers in a single azimuth–range SAR image pixel.

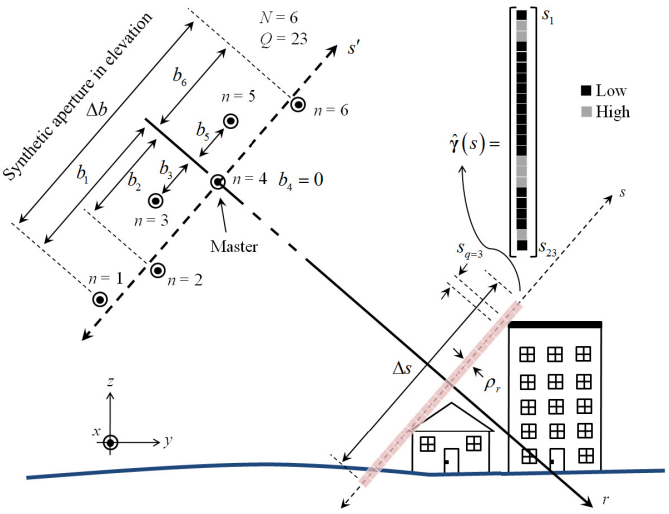


Fig. 2. Schematic of the TomoSAR imaging geometry. The elevation aperture is built by exploiting multipass/multibaselines (six in the depicted case) from slightly different viewing angles. It is shown that the backscattering contribution from the edge of two buildings and a small portion of ground is mapped onto single range–azimuth SAR image pixel. TomoSAR aims to estimate the depicted reflectivity profile $\hat{\gamma}(s)$ for discretized (pink region) elevation extent Δs. Typically, the discretization factor is much higher, i.e., $N \ll Q$ which renders (3) to be underdetermined (i.e., infinite solutions). s denotes the elevation axis, which is actually a curve but is usually approximated as a straight line due to large range distances.

The imaging geometry of SAR is shown in Fig. 2. In the following, the TomoSAR imaging model is briefly described.

Let N represent the number of observations, and the complex-valued SAR azimuth–range pixel value g_n of nth $(n = 1, \ldots, N)$ acquisition with the corresponding perpendicular baseline b_n (see Fig. 2) can be approximated as an integral of reflectivity function $\gamma(s)$ [30], [33]

$$g_n = \int_{\Delta s} \gamma(s) \exp(-j2\pi\xi_n s) ds$$

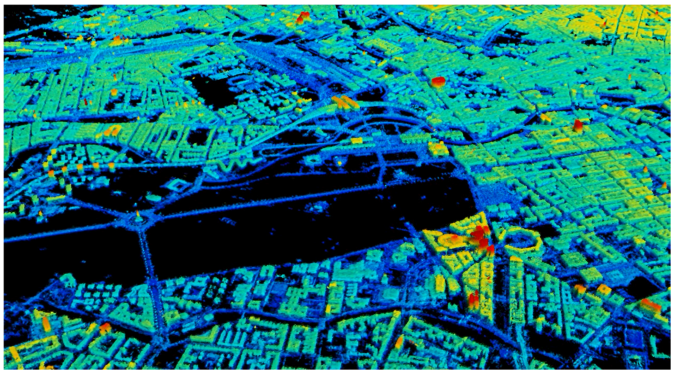


Fig. 3. TomoSAR point clouds generated from TerraSAR-X data stacks of ascending and descending orbits (Site: city of Berlin). The color represents height. Black areas are temporally decorrelated objects, e.g., vegetation or water.

with

$$\xi_n = -2b_n/\lambda r \tag{1}$$

where Δs denotes the span in elevation. Since it is well known that the far-field diffraction acts like a Fourier transform, the presented model is actually nothing, but Fourier transform of $\gamma(s)$ is sampled at discrete frequencies (in elevation) ξ_n.

The continuous model in (1) can be discretized along the elevation dimension into Q positions (i.e., $s_q \forall q = 1, \ldots, Q$) by replacing the integral with the sum as follows:

$$g_n = \sum_{q=1}^{Q} \exp(-j2\pi\xi_n s_q)\gamma(s_q) + \varepsilon_n \tag{2}$$

or alternatively in the matrix form as [30], [33]

$$\mathbf{g} = \mathbf{R}\boldsymbol{\gamma} + \boldsymbol{\varepsilon} \tag{3}$$

where $\mathbf{g} \in \mathbb{C}^{N\times 1}$ is the measurement vector with $g_n \forall n \in \{1, \ldots, N\}$, $\mathbf{R} \in \mathbb{C}^{N\times Q}$ is an irregularly sampled Fourier transform matrix with $R_{nq} = \exp(-j2\pi\xi_n s_q)$, $\boldsymbol{\gamma} \in \mathbb{C}^{Q\times 1}$ is the unknown discretized reflectivity vector with $\gamma(s_q)$, and $\boldsymbol{\varepsilon} \in \mathbb{C}^{N\times 1}$ is additive noise usually modeled as i.i.d complex circular Gaussian random variable.

TomoSAR aims to invert the imaging model presented in (3) to retrieve the unknown discrete reflectivity vector $\boldsymbol{\gamma}$. The reconstructed reflectivity profile along the elevation axis, thus, allows the separation of multiple layovered scatterers within single pixel [30], [31]. The retrieved scatterer information when geocoded into world coordinates generates TomoSAR point clouds. Fig. 3 shows the generated TomoSAR point cloud of the city of Berlin, Germany, using German Aerospace Center (DLR)'s tomographic precessing system—Tomo-GENESIS [34], [35].

In this paper, we utilized these TomoSAR point clouds in generating labeled SAR images. The basic idea is to classify each 3-D point as belonging to buildings and nonbuildings and later geocode them back into their corresponding SAR (i.e., in azimuth and range) coordinates. The classification of each point is obtained in two ways.

1) By exploiting information pertaining to already available 2-D building footprints.

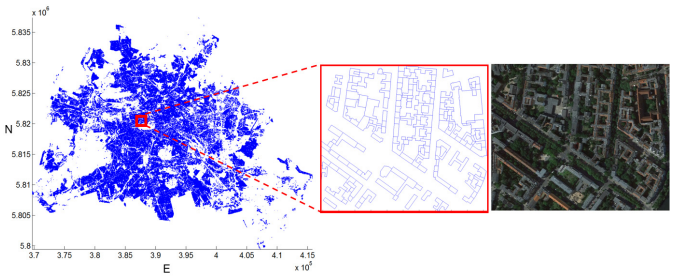


Fig. 4. GIS data of Berlin from OSM. (Left) 2-D building footprints. (Center) Zoomed-in region. (Right) Corresponding optical image of the zoomed-in region.

2) By classifying each TomoSAR point using an optical image classification scheme as proposed in [36]. This part is not the focus of this paper. Depending on the application, a different classification technique may be employed.

In the following, the two proposed methods to extract the building points in TomoSAR point cloud are described in detail.

B. Annotation Using TomoSAR Point Cloud and Openly Available OSM Data

To classify these point clouds, the 2-D building footprints from OpenStreetMap (OSM) are downloaded from Geofabrik's website,[1] which are subsequently utilized to automatically annotate the SAR image. The OSM is based on the crowdsourcing concept and has currently around 2 million registered users[2] [37]. It consists of a large number of available building footprints with positioning accuracies varying on the order of 4 m [37], [38]. The representation of building footprints is in the form of 2-D polygons having ordered list of vertices, i.e., pairs of latitude/longitude or Universal Transverse Mercator (UTM) coordinates as per WGS 84 coordinate system. The OSM data are openly available and have very high completeness percentage in many developed cities in Western Europe and USA. Fig. 4 shows an overview of the available 2-D building footprints in the Berlin city. The generated 3-D point cloud via TomoSAR inversion using SAR image stacks is already geocoded into UTM coordinates. Now, the idea is simple, and we extract all those TomoSAR points that lie within the OSM building polygons. For this purpose, we employed the classical ray casting algorithm [39], [40]. As a result, we are able to extract TomoSAR points that only belong to buildings. These building points are then projected back to SAR image coordinates (i.e., range and azimuth) to yield the building mask.

Here, one may argue that if the auxiliary information, e.g., 2-D building footprints, is being taken into account why not directly use them instead of projecting the building points in the TomoSAR point clouds back to the SAR coordinates. The rationale against this is clearly shown in Fig. 1. The fact is that the inevitable side-looking SAR imaging geometry

[1]GEOFABRIK downloads, http://download.geofabrik.de/europe/germany/berlin.html.

[2]Stats—OSM Wiki, http://wiki.openstreetmap.org/wiki/Stats.

results in undesired occlusion and geometric distortions (such as layover and multibounce) that renders elevated objects (e.g., buildings in urban areas) to appear bright and as being projected toward the sensor consequently limiting the application potential of directly projecting such auxiliary information.

C. Annotation Using TomoSAR Point Cloud and Optical Image Classification

Since the OSM is a crowdsourcing project, the lack of interest and unavailability of suitably qualified personnel, especially in the underdeveloped countries, may give rise to low completeness issues. Consequently, the use of OSM data for generating such reference (labeled) building masks may not be suitable. In such a case, an alternative way to extract building points in TomoSAR point cloud might be to perform the semantic classification of TomoSAR point clouds. To this end, we adopt an approach [36] that performs optical image classification, generates an optical 3-D point cloud, and subsequently coregisters (fuses or matches) them with TomoSAR point clouds to achieve such labeling. Since this part is not the focus of this paper, therefore the readers are kindly referred to the original literature [36] for more details. In the following, we briefly describe the main working steps of the algorithm.

1) Optical Image Classification: The optical images are classified patchwisely using the bag of words (BoW) method [41], which is a well-known technique in the computer vision community. Training patches are manually selected in the original image. The classification is done patchwisely in the large aerial image. The local feature used in BoW is simply the RGB value in a 3×3 sliding window in the patch. The classifier is a linear support vector machine [42] implemented in an open source library VLFeat [43].

2) Coregistration of Optical and TomoSAR Point Clouds: An optical 3-D optical point cloud is generated from a set of nine high-resolution aerial images using commercial Pix4D software [44]. Because of the different imaging geometry of SAR and optical images, TomoSAR and optical point clouds are different in point density on façade and flat areas. This drives the coregistration algorithm to be developed in the following way.

1) *Edge Extraction:*
   a) The optical point cloud is rasterized onto a 2-D height image by computing the mean heights of points inside each 3×3 grid cell.
   b) Similarly, the point density of TomoSAR point cloud is estimated on the rasterized 2-D grid by counting the number of points also inside each 3×3 grid cell.
   c) The edges in the optical height image and in the TomoSAR point density image are detected using the Sobel filter [45]. These edges correspond to the façade locations in the two point clouds.
2) *Initial Alignment:*
   a) The coarse horizontal alignment is performed by cross-correlating the two edge images, while the coarse vertical alignment is achieved by cross-correlating the height histogram of the two point clouds.
   b) These coarse alignments are fed as an initial solution to a robust iterative closest point (ICP) algorithm in the next step, which provides the final coregistration solution.
3) *Refined Solution:*
   a) The façade points in the TomoSAR point clouds are then removed, because the optical point cloud contains nearly no façade point.
   b) To refine the coregistration of the two-point clouds, an anisotropic ICP with robustly estimated covariance matrices using an M-estimator is applied. Considering the large quantity of points compared with the few coregistration parameters to be estimated, the resulting coregistration accuracy is quite high [36].

3) Projection of Label From Optical Image to SAR Image: Upon successful coregistration, the 2-D classification labels from the optical images are projected to the 3-D TomoSAR point cloud using the estimated camera parameters. Each TomoSAR point classified as belonging to building is then projected to SAR coordinates. After some image morphology, a binary mask of the buildings is generated.

III. Architecture for SAR Building Detection Network

A. Brief Introduction to CNNs

Extracting buildings in an SAR image represents a pixelwise classification task. In computer vision, this has been done using texton boost [46], texton forests [47], or general random forests [48]. All these methods rely on features that are handcrafted and thus prone to not always fit to the problem to classify or at least takes a lot of manual interaction to select suitable features for the specific task. Nowadays, these classification problems are tackled using CNNs. One benefit of CNNs is the fact that just the structure of the network is manually designed and all the parameters, which describe how the features are calculated, are automatically learned using training data. Furthermore, it is well known that CNNs are suitable for transfer learning. This means that a network trained for a specific task can be reused for another task. Therefore, parts of the network can be redesigned and the unchanged part of the new network can be initialized using the parameters of the original network and fine-tuned using task-specific training data. This ability of neural networks motivated us to use a semantic segmentation network from computer vision as a base for our SAR image classification network. Another not negligible feature is that neural networks are highly parallelizable and thus suited for efficient processing using GPUs. The well-known frameworks for CNNs are Caffe or Theano. In our experiments, the Caffe framework has been employed.

B. Proposed Architecture

The network architecture of the fully convolutional network (FCN) is based on the FCN structure of Long *et al.* [49].

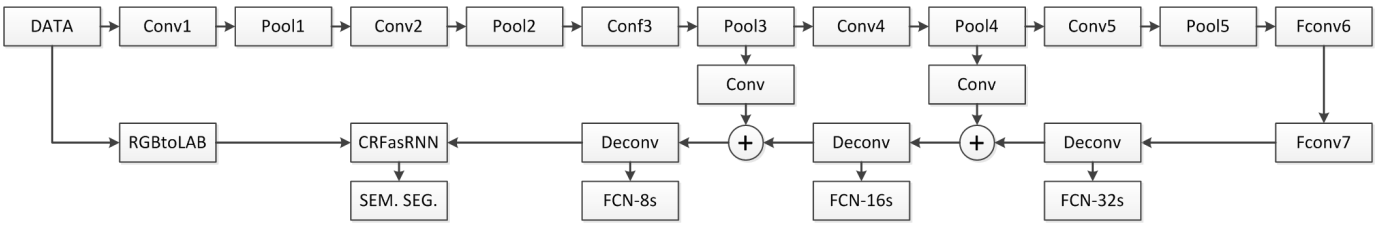


Fig. 5. Overview of the semantic segmentation network. The first part of our network calculates a feature for each input pixel by exploiting an FCN with in-network upsampling and skip-and-fuse architecture to fuse coarse, semantic, and local, appearance information. The second part of the network adds binary potentials (i.e., adding constraints to give neighboring pixels with a similar intensity the same label) by using the dense CRF-RNN as proposed in [32].

To additionally integrate binary potentials, we add a CRF represented as RNN [32]. This gives us an end-to-end trainable network, as shown in Fig. 5.

In detail, the first part of our network calculates a feature for each input pixel. Therefore, we exploit an FCN with in-network upsampling and skip-and-fuse architecture to fuse coarse, semantic and local, and appearance information [50]. As we are using an FCN, we exploit the ability to not only classify a single pixel as proposed in [5], [21], and [24] but also perform image segmentation for input images of arbitrary size at once. Thus, we eliminate overhead calculations resulting from the sliding window approach.

The second part of the network adds binary potentials. This means that it adds constraints to give neighboring pixels with a similar intensity the same label. This is typically done using a Markov random field or to be more precise the special case of a fully connected CRF as presented by Krähenbühl and Koltun [51] whose overall energy function can be characterized as follows [32], [51]:

$$E(\mathbf{x}) = \sum_i \psi_u(x_i) + \sum_{i<j} \psi_p(x_i, x_j). \tag{4}$$

Inference of the CRF involves finding a configuration (or labeling) $\mathbf{x}$, such that the total unary $\psi_u(x_i)$ and pairwise $\psi_p(x_i, x_j)$ energy components (or potentials/costs) are together minimized. Unary potentials measure the inverse likelihood (and thus, the cost) of the pixel i being assigned a label x_i, while the pairwise energy components measure the simultaneous cost of assigning labels x_i, x_j to pixels i, j. It typically provides an image-dependent smoothing term that favors assigning similar labels to neighboring pixels having similar properties. Specifically, in our model, the unary energies are obtained from a CNN (FCN-8s architecture of [49] as mentioned earlier). This network is primarily based on the VGG-16 network but has been modified to perform semantic segmentation instead of image classification. The pairwise energies, on the other hand, have been modeled as weighted Gaussians as follows [32], [51]:

$$\psi_p(x_i, x_j) = \mu(x_i, x_j) \sum_{m=1}^{M} w^m G^m(\mathbf{f}_i, \mathbf{f}_j) \tag{5}$$

where each G^m for $m = 1, 2, \ldots, M$ is Gaussian kernel applied on feature vectors. The feature vector $\mathbf{f}_i$ of pixel i is derived from image features, such as RGB values and 2-D spatial location. w^m are linear combination weights, while μ is the label compatibility function that is a simple Potts model $\mu(x_i, x_j) = [x_i \neq x_j]$ in our case.

As an end-to-end trainable network is preferable, we added the dense CRF represented as a RNN further called CRF-RNN as proposed in [32].

This network was then modified to get a pixelwise two-class classification representing building and nonbuilding.

IV. Implementation of Training Algorithm

We performed staged training as mentioned in [50], because it is less prone to divergence. First, the single-stream FCN-32s is trained, and then, the training is continued with the two-stream FCN-16s and the three-stream FCN-8s. Next, the CRF-RNN is added and trained by keeping the FCN-8s part constant. Finally, a fine-tuning of the complete network has been performed. Each stage was trained for 400 000 iterations with constant learning rate ($1e^{-10}$, $1e^{-12}$, $1e^{-14}$, $1e^{-12}$, and $1e^{-12}$ for each stage, respectively) a momentum of 0.99, a weight decay of 0.0005, and a pixelwise softmax loss (that has been averaged over 100 images each epoch).

As the network contains convolutional layers as well as pooling layers, the resulting segmented image is reduced in dimension. This is compensated by in-network upsampling layers whose parameters are initialized as bilinear filtering and further refined while training. Moreover, as suggested in [32], in all our experiments, during training, we fixed the number of mean-field iterations in the CRF-RNN to 5 to avoid vanishing/exploding gradient problems and to reduce the training time. However, the number of iterations was raised to 10 for deploying/inference (when evaluating the test images). Moreover, the compatibility transform parameters of the CRF-RNN were initialized using the Potts model.

V. Experimental Results and Validation

A. Data Set Description

To validate our approach, we employed SAR data sets consisting of a TerraSAR-X high-resolution spotlight image and a 3-D TomoSAR point cloud of Berlin. The SAR image has a spatial resolution of about 0.588 and 1.1 m in range and azimuth directions, respectively. The image was acquired from ascending orbit with an incidence angle of 36°, which almost provides a full coverage of the whole city. The 3-D TomoSAR point clouds have been generated from stacks of 102 TerraSAR-X high spotlight images from ascending and descending orbits covering almost the whole city of

Fig. 6. SAR intensity image covering almost the whole city of Berlin.

Fig. 7. Automatically generated mask of building regions using OSM + TomoSAR point clouds for the SAR intensity image shown in Fig. 6.

Berlin using the Tomo-GENESIS software developed at the DLR [34], [35]. The number of points in the Berlin data set is approximately 30 million.

The optical images used for annotation were attained in March 2014 and include nine UltraCam aerial images of Berlin having an altitude of around 4000 m. The ground spacing is roughly 20 cm/pixel. The camera positions and orientations were measured by an onboard GPS and inertial measurement unit with the standard deviations of about 5 cm and $5 \times 10^{-4^\circ}$, respectively.

B. Results of Automatic Annotation

Fig. 6 shows the SAR intensity image covering almost the whole city of Berlin (around 39 km^2), while Fig. 7 shows the resulting mask of building regions obtained automatically using the OSM building footprints. Similarly, Figure 8 presents the SAR intensity image together with resulting building masks obtained using the optical image classification scheme. Since the completeness percentage of OSM data is quite high for many cities in Europe and USA, it can be seen that automatic annotation/labeling using this data is quite generic and has the potential of producing benchmark SAR data sets, which is still missing within the relevant community. However, although quite a lot of buildings are present, it is also worth to mention that since it is a crowdsourcing project, there are still a few missing buildings and inner yards. Fig. 9 shows a couple of such examples. In addition to this, there

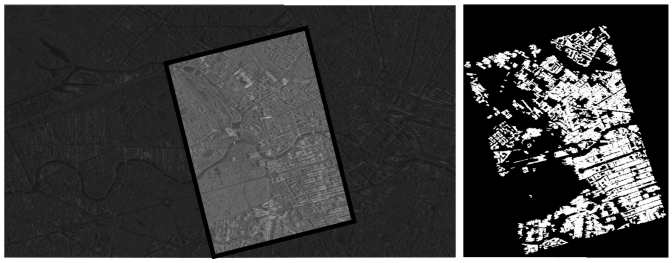


Fig. 8. (Left) SAR intensity image partly covering the city of Berlin—highlighted region and (Right) corresponding generated mask of the highlighted region using optical image classification + TomoSAR point clouds.

are also false annotations such as some parts of the railway tracks originating from the Berlin central station which have also been labeled as building structure in the OSM data (see Fig. 10). As a consequent, when OSM data are utilized to extract building points in the TomoSAR point cloud, points belonging to such railway tracks are misclassified as buildings and when projected back to SAR image coordinates yields false annotation/labeling. Although limited but on the other hand, the use of optical image classification and TomoSAR point cloud avoids this false labeling as depicted in Fig. 10(d) and produces better annotation results but may be restrictive in a sense to generate large-scale data sets.

C. Accuracy Analysis of Automatic SAR Annotation

To perform the precise accuracy analysis of the produced annotations, we have manually labeled the building pixels in the SAR image covering an area of around 3.3 km^2 in the Berlin city. Fig. 11 shows the selected SAR image, while Fig. 12 shows its corresponding GT annotation obtained by manual labeling of building pixels/regions in the selected SAR region. For qualitative evaluation, Figs. 13 and 14 show the common and difference maps for visual comparison. The difference maps are obtained by subtracting the produced annotation masks OSM-Ref and Opt-Ref from the GT annotated mask, respectively. The green pixels in Fig. 14 indicate no change, while the red pixels denote the missing buildings and the blue pixels show the regions labeled as buildings in the generated building masks using the two proposed annotation schemes but not present in the GT reference mask. For quantitative evaluation, Table I shows the performance of the proposed annotation schemes using the common and difference maps by employing the standard precision/recall evaluation metrics computed as Precision $(\%) = 100 \times (t_p/(t_p + f_p))$ and Recall $(\%) = 100 \times (t_p/(t_p + f_n))$, where t_p are the number of white pixels (true positives) in the common image, while f_n and f_p are the number of red (false negatives) and blue pixels (false positives) in the difference image, respectively.

The evaluation statistics in Table I depicts that both the proposed annotation methods correctly label building pixels with good accuracy. However, in terms of completeness, OSM-Ref shows less relative accuracy owing to the already mentioned fact that a few buildings are missing in the crowdsourced OSM building footprint data. In this context, the use of

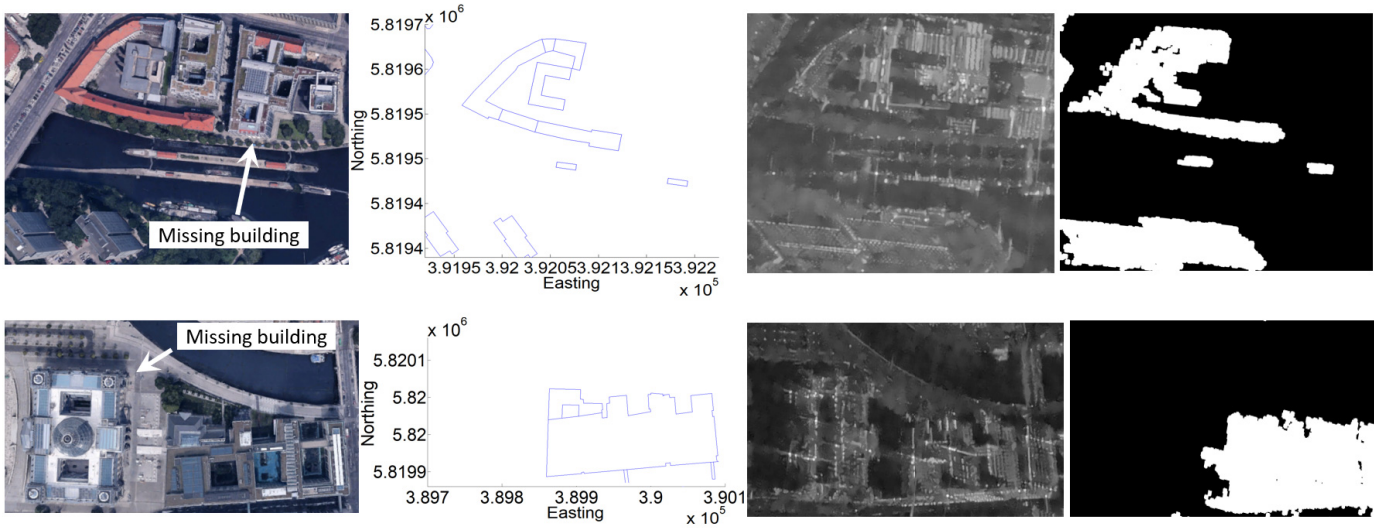
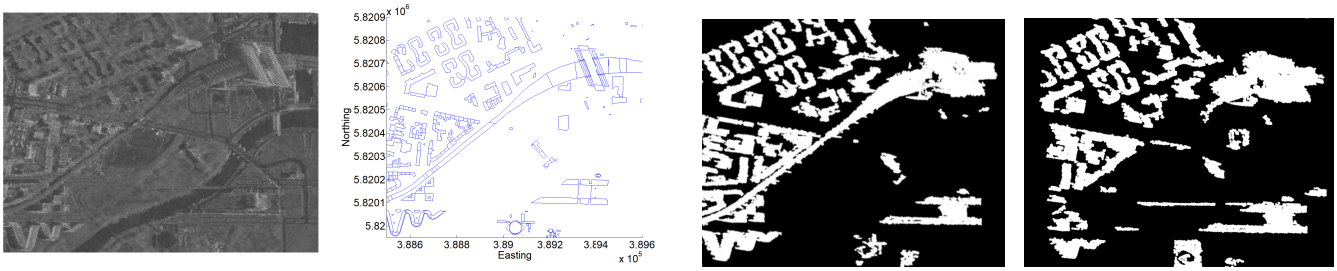


Fig. 9. Missing buildings in the 2-D OSM GIS data. The first column shows the optical images of the buildings which are missing in the OSM polygonal data as shown in UTM coordinates in the second column. The third column presents the corresponding SAR images, while the fourth column shows the GT generated by projecting the building points—extracted using auxiliary OSM GIS data—in the TomoSAR point clouds to the SAR image coordinates



(a) (b) (c) (d)

Fig. 10. Close-up views of Berlin central station to show the false labeling in the GT generated using TomoSAR point cloud and the auxiliary OSM data. (a) SAR image of the area of interest (Berlin central station). (b) 2-D OSM building polygons. It can be seen that the railway track originating from the Berlin central station is falsely characterized as building structure in the OSM data. (c) TomoSAR points belonging to this track is misclassified as building points and when projected back to SAR image coordinates yields false labeling. (d) Close-up view of the GT (labeled SAR image) of the same area generated by projecting the building points—extracted using the optical image classification scheme [36]—in the TomoSAR point clouds to the SAR image coordinates. In contrast, the railway track is now correctly labeled as nonbuilding in the generated GT.

TABLE I

QUANTITATIVE EVALUATION STATISTICS OF AUTOMATICALLY PRODUCED SAR ANNOTATED MASKS

Evaluation Metrics	OSM-Ref	Opt-Ref
t_p	5614059	6580131
f_p	1191211	1290182
f_n	1573086	607014
t_n	12408130	11343087
Precision (Correctness %)	82.49	83.61
Recall (Completeness %)	78.11	91.55

accurate cadastral maps may help in achieving a high degree of recall/completeness.

In the following, we present the experimental results and its analysis obtained by employing the deep learning-based staged network architecture exploiting both these automatic annotations.

Fig. 11. SAR image of the selected 3.3-km^2 area with the following UTM coordinates. 33U (Top left) (389072 E, 5822399 N), (Bottom left) (388741 E, 5820939 N), (Top right) (391201 E, 5821922 N), and (Bottom right) (390900 E, 5820460 N).

D. Preparation of Training Data

We prepared the data set for training by taking 11 out of 16 of the labeled input images covering almost the whole city of Berlin (using OSM + TomoSAR point cloud) and created patches of 256×256 pixels with an overlap of 32 pixels. Furthermore, these patches are augmented by flipping and rotation. Finally, we got 26 312 image patches for training

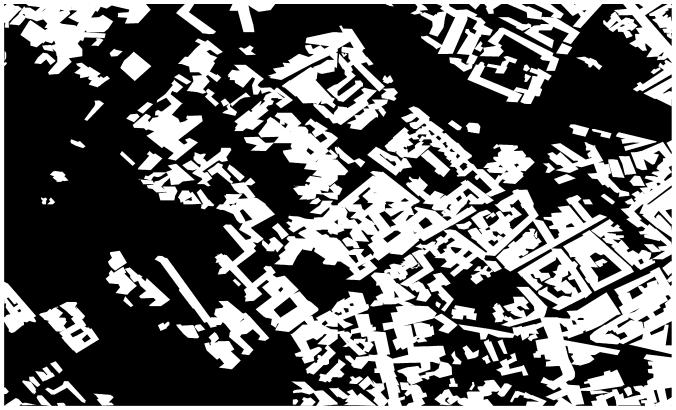
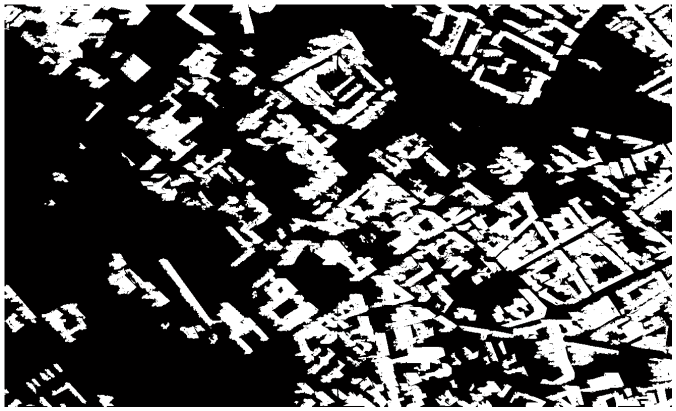
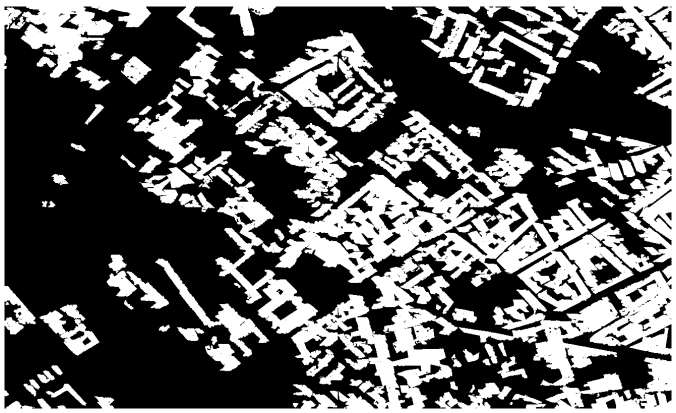


Fig. 12. GT mask obtained after manual labeling of building pixels/regions of the SAR image depicted in Fig. 11. The mask is used for accuracy analysis of the generated SAR annotations using the two proposed schemes.



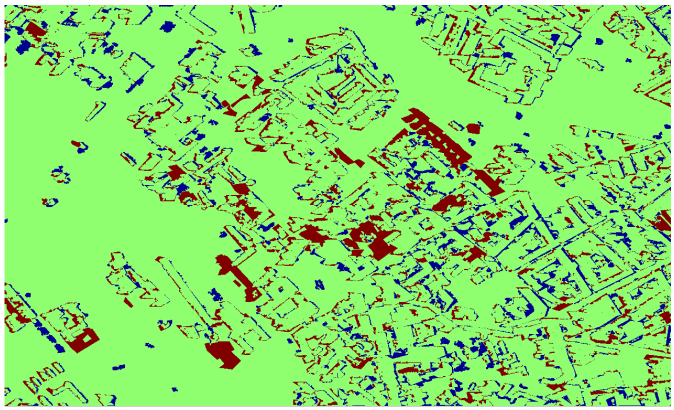
(a)



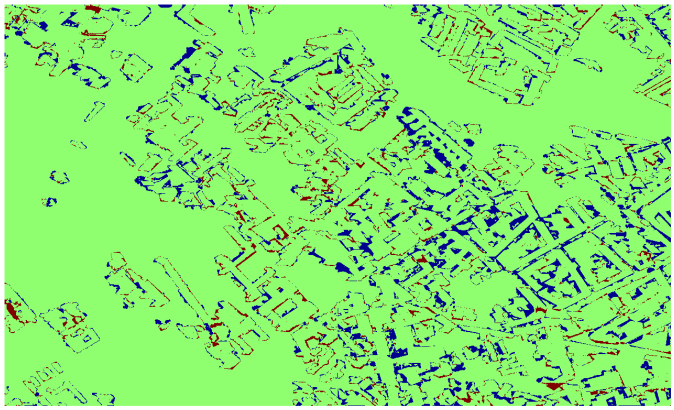
(b)

Fig. 13. Common map of the produced building masks using the proposed annotation schemes and the reference GT map. (a) OSM-Ref ∩ GT. (b) Opt-Ref ∩ GT.

and used the remaining 5 out of 16 of the labeled input images for testing. In case of Optical + TomoSAR point cloud, we vertically sliced the highlighted SAR image region shown in Fig. 8 in four equal parts and took the first and last for testing/validation and the two in the center for training. It is also important to mention that to reduce speckle effect, we first performed nonlocal filtering of the SAR images prior to training using the algorithm as proposed in [52].



(a)



(b)

Fig. 14. Difference map generated by subtracting the results of generated SAR annotations or training samples from the manually annotated GT mask. (a) OSM-Ref–GT. (b) Opt-Ref–GT. Note that the green pixels indicate no difference between the generated and GT masks and the red pixels indicate missing buildings, while the blue indicates the pixels labeled as belonging to buildings using the proposed annotated schemes but not present in the reference GT mask.

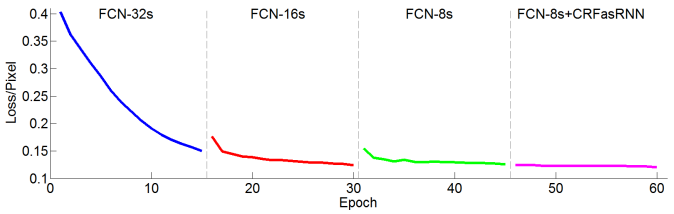


Fig. 15. Learning curves across different stages of the network. The loss is normalized by dividing with the number of pixels of the training image. One epoch represents all training images being passed through the network once.

E. Performance Evaluation of the Trained Network

1) Evaluation Metrics: To evaluate the performance of different networks, we use the metrics that are variations on pixel accuracy (PA) and region intersection over union (IU) and are commonly used for evaluating semantic segmentation and scene parsing algorithms [49], [50]. For each class, the IU score is computed as $(t_p/(t_p + f_p + f_n))$, where t_p (true positives) are the number of correctly classified pixels, f_p (false positives) are the number of wrongly classified pixels, and f_n (false negatives) are the number of pixels

TABLE II

ACCURACY ANALYSIS OF OBTAINED RESULTS USING DIFFERENT STAGES OF THE TRAINED NETWORK WITH THE FOLLOWING DETAILS: TRAINING AND TESTING/VALIDATION USING OSM-REF DATA

Network Architecture	PA	MA	MIU	FWIU	FAR	QR
FCN-32s	83.28	82.40	69.47	71.98	20.39	71.58
FCN-16s	83.46	82.43	69.70	72.22	20.12	71.85
FCN-8s	83.49	82.45	69.75	72.27	20.06	71.90
FCN-8s (CRF-RNN)	83.54	82.60	69.86	72.35	20.00	71.97

TABLE III

ACCURACY ANALYSIS OF OBTAINED RESULTS USING DIFFERENT STAGES OF THE TRAINED NETWORK WITH THE FOLLOWING DETAILS: TRAINING USING OSM-REF DATA SET AND TESTING/VALIDATION USING OSM-GT

Network Architecture	PA	MA	MIU	FWIU	FAR	QR
FCN-32s	89.87	91.32	79.89	82.16	11.36	81.72
FCN-16s	91.35	92.78	82.52	84.54	9.54	84.17
FCN-8s	91.52	92.97	82.81	84.81	9.34	84.45
FCN-8s (CRF-RNN)	92.13	93.84	83.97	85.82	8.61	85.48

wrongly not classified as belonging to a particular class. If we denote n_{ij} as the number of pixels of class i predicted to belong to class j, n_N as the number of classes, and t_i as the total number of pixels belonging to class i, then the following evaluation metrics have been computed [49], [50].

1) *PA:*

$$\frac{\sum_i n_{ii}}{\sum_i t_i}.$$

2) *Mean Accuracy (MA):*

$$\left(\frac{1}{n_N}\right)\sum_i \frac{n_{ii}}{t_i}.$$

3) *Mean IU:*

$$\left(\frac{1}{n_N}\right)\sum_i \frac{n_{ii}}{t_i + \sum_j n_{ji} - n_{ii}}.$$

4) *Frequency-Weighted IU:*

$$\left(\sum_k t_k\right)^{-1}\sum_i \frac{t_i n_{ii}}{t_i + \sum_j n_{ji} - n_{ii}}.$$

In addition to the above-mentioned four metrics, the following two metrics have also been computed.

1) *False Alarm Rate:*

$$\frac{f_p}{t_p} \equiv \frac{\sum_i \sum_j n_{ij}}{\sum_i n_{ii}}.$$

2) *Quality Rate:*

$$\frac{t_p}{t_p + f_p + f_n} \equiv \frac{\sum_i n_{ii}}{\sum_i \left(t_i + \sum_j n_{ji} - n_{ii}\right)}.$$

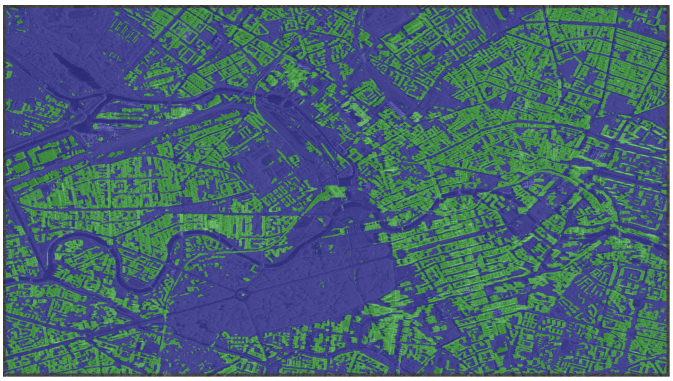


Fig. 16. Input SAR image of Berlin city as depicted in Fig. 6 with an overlay of the semantic segmentation. Results computed using the OSM-Ref annotated data set with FCN-8s with CRF-RNN network.

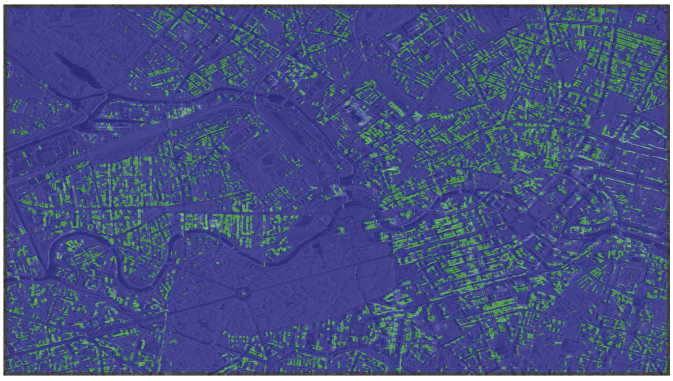


Fig. 17. Input SAR image of Berlin city as depicted in Fig. 6 with an overlay of the semantic segmentation. Results computed using the Opt-Ref annotated data set with FCN-8s with CRF-RNN network.

2) Results Analysis: The experimental results have been obtained after applying *staged* training where the results obtained after single-stream and then upgraded to two-stream and three-stream are depicted as FCN-32s (32× upsampled prediction), FCN-16s (16× upsampled prediction), and FCN-8s (8× upsampled prediction), respectively. In each respective stage, the network is learned from end-to-end in a cascaded manner, i.e., all initialization parameters of the previous stage are fed as an input to the subsequent one. Let us denote the automatically generated annotated data set using OSM + TomoSAR point cloud as OSM-Ref and using Optical classification + TomoSAR point cloud as Opt-Ref. Tables II and III depict the results acquired over the whole area of Berlin in different stages of the network architecture. In Table II, for testing/validation, we analyzed the network performance by computing evaluation metrics over (untrained) 5/16 subimage patches annotated using OSM + TomoSAR point cloud (i.e., OSM-Ref). As mentioned earlier, the OSM data set is prone to errors introduced as a consequent of crowdsourcing, and therefore, for a fair evaluation of network architecture, we needed to prepare a more accurate annotated data set (denoted as OSM-GT) for test subimages.

To prepare such a reference annotated data set, we manually inserted missing buildings and removed parts of other structures, e.g., railway tracks misclassified as buildings

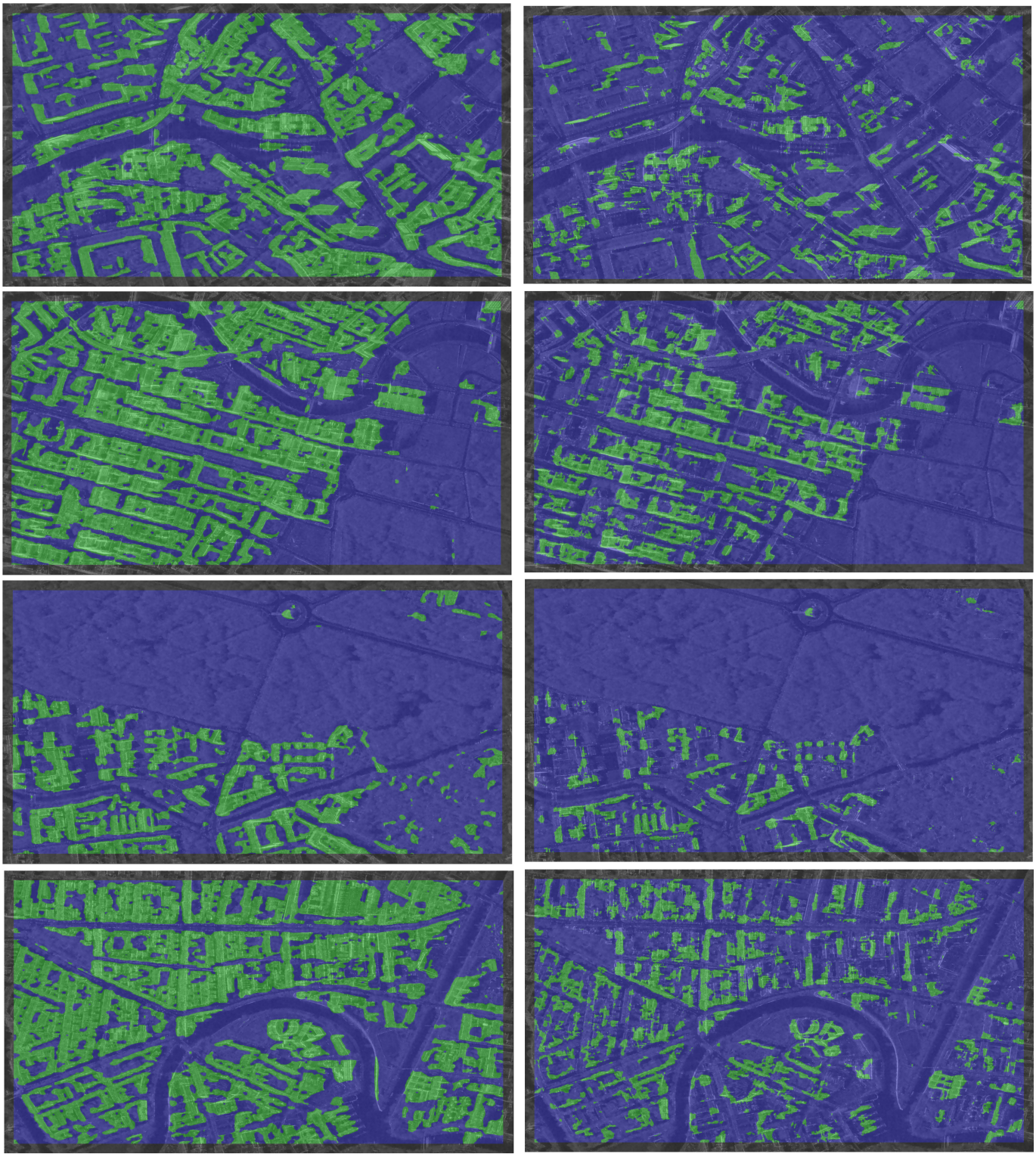


Fig. 18. Close-up views of Figs. 16 (first column) and 17 (second column). The first column depicts the input image with an overlay of the semantic segmentation result. The results have been computed using the OSM-Ref annotated data set with FCN-8s with CRF-RNN network over different test subimage patches. The second column depicts the input image with an overlay of the semantic segmentation results. The results have been computed using the Opt-Ref annotated data set with FCN-8s with CRF-RNN network over different test subimage patches.

(see Figs. 6 and 7). Table III depicts the evaluation results over untrained subimage patches using OSM-GT for testing/validation. For Tables II and III, we see the improvement in network performance in each subsequent stage. In general, the upgraded three-stream FCN-8s with CRF-RNN tends to show a superior performance in distinguishing buildings from nonbuildings. It is important to mention that one may argue here that since the OSM-Ref is used for training,

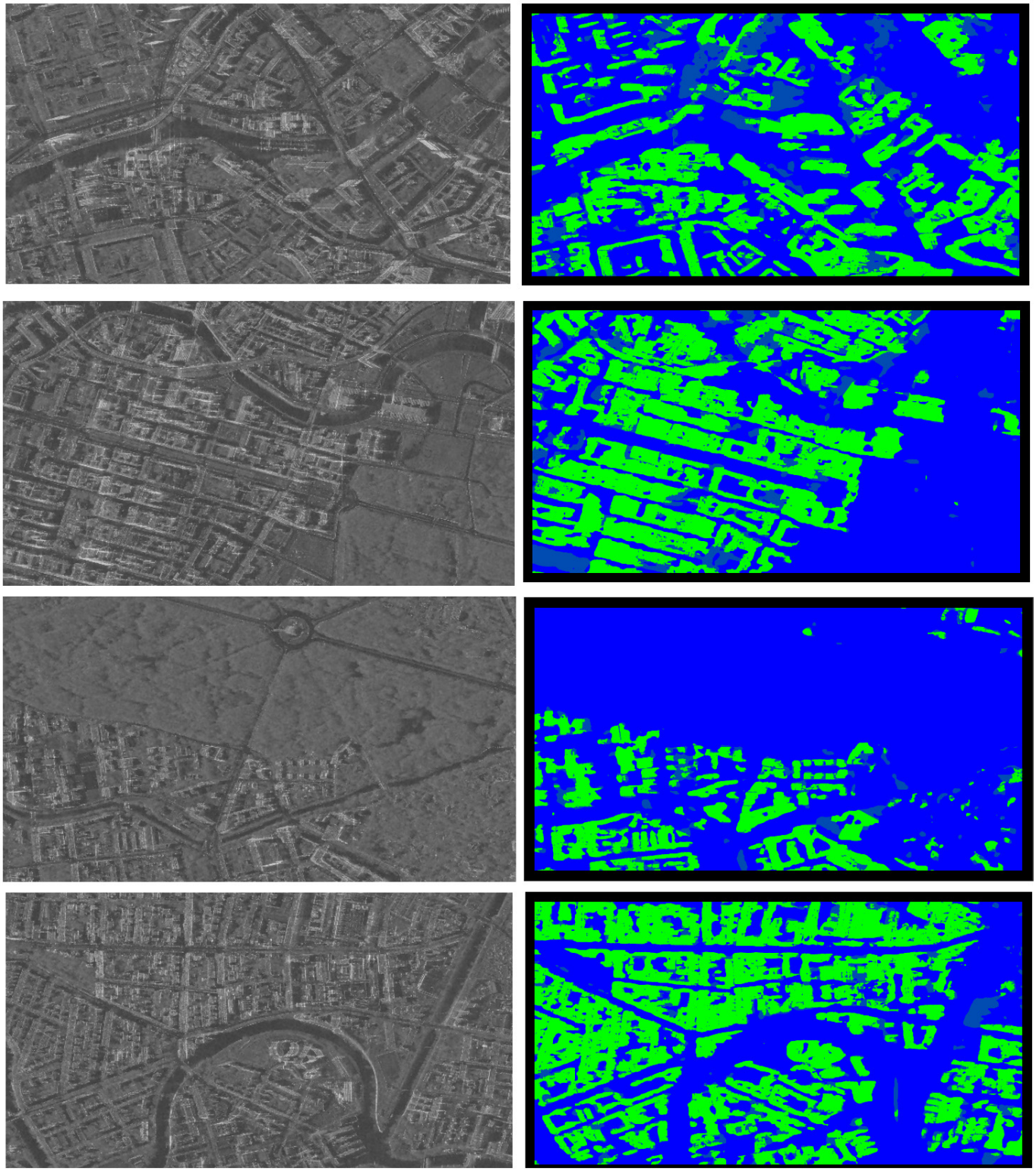


Fig. 19. Results of semantic segmentation computed using the OSM-Ref annotated data set with FCN-8s with CRF-RNN network over different test subimage patches. The first column shows the different SAR test subimage patches, while the second column depicts the difference image of the semantic segmentation result and the manually corrected GT (Training: OSM-Ref; Testing: OSM-GT; Network FCN-8s with CRF-RNN). The light green region in the difference map corresponds to true positives, while the dark green regions are false negatives (negligible here). Light red regions, on the other hand, are the true negatives, while dark red regions correspond to false positives.

the prediction should be more close to the OSM-Ref instead to OSM-GT. The other way around reason is merely due to the fact that the trained CNN architecture correctly recognizes the missing buildings and was able to differentiate the railway tracks from buildings mainly because the training samples contain fewer portions of the railway tracks which were wrongly classified as buildings in the OSM data.

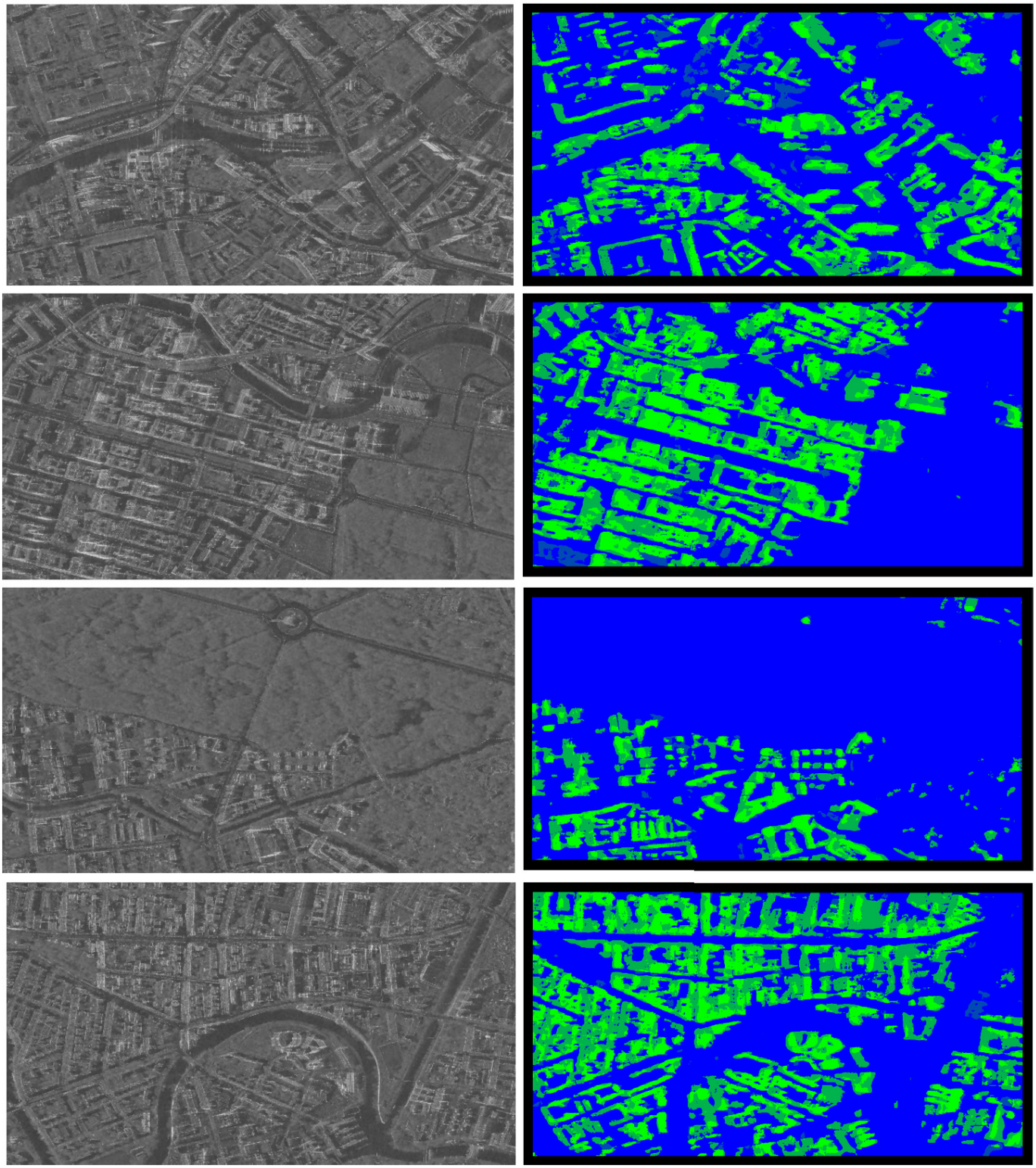


Fig. 20. Results of semantic segmentation computed using the Opt-Ref annotated data set with FCN-8s with CRF-RNN network over different test subimage patches. The first column shows the different SAR test subimage patches, while the second column depicts the difference image of the semantic segmentation result and the manually corrected GT (Training: Opt-Ref; Testing: OSM-GT; Network FCN-8s with CRF-RNN). The light green region in the difference map corresponds to true positives, while the dark green regions are false negatives. Light red, on the other hand, are the true negatives, while dark red regions correspond to false positives.

Similarly, Table IV shows the evaluation results with the FCN-8s with CRF-RNN network architecture with two annotated test images OSM-GT and Opt-Ref where the latter is the result of automatically annotated test subimages generated using Optical classification + TomoSAR point cloud. These quantitative results are obtained using the training sample division, as reported in Section V-D.

TABLE IV

ACCURACY ANALYSIS OF OBTAINED RESULTS USING FCN-8S WITH CRF-RNN NETWORK ARCHITECTURE UTILIZING AUTOMATICALLY GENERATED ANNOTATED DATA USING OPTICAL CLASSIFICATION AND TOMOSAR POINT CLOUD, DENOTED AS OPT-REF, AS TRAINING DATA AND OSM-GT AND OPT-REF AS TESTING/VALIDATION DATA

Training	Testing	PA	MA	MIU	FWIU	FAR	QR
Opt-Ref	OSM-GT	78.35	69.30	56.97	63.74	28.49	64.94
Opt-Ref	Opt-Ref	83.23	79.57	66.45	72.45	20.55	71.57

TABLE V

ACCURACY ANALYSIS OF OBTAINED RESULTS USING FCN-8S WITH CRF-RNN NETWORK ARCHITECTURE UTILIZING OSM-REF AND OPT-REF ANNOTATED MASKS. THE NETWORK PARAMETERS, SUCH AS LEARNING RATE, MOMENTUM, AND WEIGHT DECAY, ARE THE SAME AND PROVIDED IN THE BEGINNING OF SECTION IV. GT CORRESPONDS TO THE MANUALLY PREPARED TESTING GT MASK DEPICTED IN FIG. 12

Training	Testing	PA	MA	MIU	FWIU	FAR	QR
OSM-Ref	GT	82.80	81.25	69.06	70.82	20.77	70.65
Opt-Ref	GT	83	81.58	69.42	71.13	20.45	70.95

Figs. 16 and 17 show the result of FCN-8s with CRF-RNN trained using OSM-Ref and Opt-Ref, respectively, overlaid onto the SAR image of Fig. 6 covering almost the whole region of Berlin. Again, these qualitative results are obtained using the training sample division, as reported in Section V-D. Fig. 18 shows the close-up results over different test subimage patches, while Figs. 19 and 20 show the corresponding difference maps. The light green region in the difference map corresponds to common regions, i.e., true positives, while the dark green regions are buildings that have not been detected by the network, i.e., false negatives. light blue regions, on the other hand, are the true negatives, while dark blue regions correspond to wrongly classified buildings, i.e., false positives. With the OSM-Ref trained network, we hardly see any dark green regions implicitly implying a high degree of completeness (see Fig. 19). In contrast, for Opt-Ref trained network, we have a fair amount of dark green regions depicting miss detections (see Fig. 20). The main reason for this is that the network has been trained with a less number of training samples in case of Opt-Ref compared with OSM-Ref (see Section V-D).

Nevertheless, for comparison and to provide accurate and fair accuracy analysis, we also trained the network separately using both OSM-Ref and Opt-Ref annotated SAR building masks in a controlled manner. We carefully designed the experiment by using the same (geographic) region, network parameters, and the size of the training patches. For evaluation, we used the manually prepared GT testing mask GT depicted in Fig. 12. Table V shows the evaluation results obtained by training the FCN-8s with CRF-RNN network architecture with both OSM-Ref and Opt-Ref annotated masks and tested using GT. The quantitative analysis of this experiment demonstrates that the network trained using both the annotation masks reveals a similar performance. However, since there are completeness issues with OSM data, the Opt-Ref annotation is slightly better. In spite of this, in comparison with Opt-Ref, the generation of OSM-Ref annotation mask is much easier to obtain and has the potential to produce the large-scale SAR annotation masks. The accuracy of such kind of masks can, however, be improved by replacing the OSM data with more accurate cadastral data (2-D footprints), if available from other sources, e.g., city administration and so on.

3) Analysis of the Network: Fig. 15 shows the learning curves across different stages of the network using the ReLU activation function. As shown in Fig. 15, we can make use of a fairly high learning rate to train the staged network for detecting building regions without the risk of divergence. The jumps between different stages of the network architecture originate by the fact of having a new part at the end of the network that is just initialized but not trained at all (due to staged training).

4) Hardware and Processing Time: All the experiments have been conducted on a GPU equipped personal computer with the following details: Intel Core i7 @ 3.7-GHz and 32-GB RAM. For one test image of the dimension $20\,626 \times 11\,472$ covering almost the whole area of Berlin, it took in average around 259 s. Such an evident fast speed of deep learning architectures is very important in practical scenarios. In addition, the downside of deep learning architectures (i.e., long training times) is becoming increasingly ignorable with rapid development in the hardware technology particularly in GPUs.

VI. DISCUSSION

The experiments presented in this paper show a variety of things.

1) It demonstrated that it is possible to automatically generate reference data sets with the potential to be produced globally opening new perspectives of producing benchmark SAR reference data sets. Another method of choice to generate such a reference data set may be obtained by exploiting simulation-based methods as proposed, e.g., in [26] and [27]. However, such methods have their own limitations in a sense, and they typically require the accurate models (3-D building models and/or accurate digital surface models) to precisely generate such GT data which, in most cases, is not available.
2) Deep learning architectures are greedy in terms of training data limiting their potential application. However, with the possibility of producing large-scale annotated data sets, the application of different deep learning network architectures is possible for the classification of built-up areas in SAR images.
3) In the case of OSM data, although the completeness and correctness of the OSM data are fairly good but have not yet reached to a level where it covers the whole globe. Nevertheless, at least in the developed countries, such data have the potential to be used either directly as the reference/GT data set or to generate training data (i.e., the labeled buildings masks as demonstrated in our case) where it is difficult to obtain such information with other interactive/expert methods.
4) It is also worth to mention that both the automatic annotation results are produced using TomoSAR

point clouds. Due to complex multiple scattering and different microwave scattering properties of the objects in the scene which possess different geometrical and material features, TomoSAR point clouds exhibit some special characteristics, such as low positioning accuracy, a high number of outliers, gaps in the data, and rich façade information due to the side-looking geometry. These properties make classification of TomoSAR point clouds a challenging task. With the aid of additional auxiliary information, the problem is rectified.

5) Hypothetically, the automatic annotation results could be improved with higher density of TomoSAR points because when projected to SAR coordinates (azimuth and range), a denser map of the reference would be generated. TomoSAR point density is, however, dependent on several factors, e.g., the geometrical properties of objects appearing in the scene, the number of SAR images used for tomographic reconstruction, and so on. In the current scenario, the effect of low point density is reduced by densifying the resulting building mask using the mathematical image dilation operation.
6) Last, the capability to produce automatic large area annotations together with their exploitation to detect buildings in SAR imagery may benefit the field of SAR-based (e.g., D-InSAR or TomoSAR) risk management against potential threats (including subsidence, landslides, and so on) by performing building damage/vulnerability analysis, e.g., as depicted in [53] and [54].

VII. Conclusion

In this paper, we have presented a deep learning-based network architecture that is able to classify buildings from nonbuildings in SAR images. Two automated annotation methods able to generate reference building masks for training and testing the classifier have been presented. The methods of automated annotation are generic and have the potential toward generation of large-scale SAR reference data sets. The annotated building masks have been utilized to construct and train the deep fully CNNs with an additional CRF represented as an RNN to detect building regions in a single (nonlocally filtered) SAR image with an MA of around 93.84%. The presented results are expected to further stimulate the research interest in exploiting SAR imagery using deep learning network architectures.

The results of this paper are promising, but still there are things that could be addressed in the future. For instance, the heights of individual buildings could be retrieved/estimated by identifying layover regions in the obtained CNN-based detection results. One application of such estimation is in reducing the number of images required for accurate tomographic reconstruction as demonstrated in [4]. In addition, in this paper, we aimed at detecting buildings for which we utilized/generated OSM-based annotated building masks for training and testing/validation. In the future, such annotated masks could also be produced using other objects in the OSM data set, e.g., roads, coastlines, and so on.

Acknowledgment

The authors would like to thank G. Baier from the German Aerospace Center, Germany, to perform nonlocal filtering of the synthetic aperture radar image of Berlin used in this paper. They would also like to thank the Gauss Centre for Supercomputing e.V for providing computing time at the GCS Supercomputer SuperMUC, Leibniz Supercomputing Centre (Project ID: pr53ya) [25], [28].

References

[1] J. D. Wegner, R. Hänsch, A. Thiele, and U. Soergel, "Building detection from one orthophoto and high-resolution InSAR data using conditional random fields," *IEEE J. Sel. Topics Appl. Earth Observ. Remote Sens.*, vol. 4, no. 1, pp. 83–91, Mar. 2011.

[2] 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.

[3] J. Tao, G. Palubinskas, and P. Reinartz, "Automatic interpretation of high resolution SAR images: First results of SAR image simulation for single buildings," in *Proc. ISPRS Hannover Workshop*, 2011, pp. 313–317.

[4] X. X. Zhu, N. Ge, and M. Shahzad, "Joint sparsity in SAR tomography for urban mapping," *IEEE J. Sel. Topics Signal Process.*, vol. 9, no. 8, pp. 1498–1509, Dec. 2015.

[5] 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.

[6] 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.

[7] 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.

[8] 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, p. 56, Dec. 2013.

[9] Y. Cao, C. Su, and G. Yang, "Detecting the number of buildings in a single high-resolution SAR image," *Eur. J. Remote Sens.*, vol. 47, no. 1, pp. 513–535, 2014.

[10] 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.

[11] 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.

[12] L. Deng and C. Wang, "Improved building extraction with integrated decomposition of time-frequency and entropy-alpha using polarimetric SAR data," *IEEE J. Sel. Topics Appl. Earth Observ. Remote Sens.*, vol. 7, no. 10, pp. 4058–4068, Oct. 2014.

[13] Q. Zhao and J. C. Principe, "Support vector machines for SAR automatic target recognition," *IEEE Trans. Aerosp. Electron. Syst.*, vol. 37, no. 2, pp. 643–654, Apr. 2001.

[14] Y. Sun, Z. Liu, S. Todorovic, and J. Li, "Adaptive boosting for SAR automatic target recognition," *IEEE Trans. Aerosp. Electron. Syst.*, vol. 43, no. 1, pp. 112–125, Jan. 2007.

[15] M. Li, Y. Wu, and Q. Zhang, "SAR image segmentation based on mixture context and wavelet hidden-class-label Markov random field," *Comput. Math. Appl.*, vol. 57, no. 6, pp. 961–969, 2009.

[16] 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.

[17] L. Mou and X. X. Zhu, "Vehicle instance segmentation from aerial image and video using a multitask learning residual fully convolutional network," *IEEE Trans. Geosci. Remote Sens.*, pp. 1–13, 2018.

[18] L. H. Hughes, M. Schmitt, L. Mou, Y. Wang, and X. X. Zhu, "Identifying corresponding patches in SAR and optical images with a pseudo-siamese CNN," *IEEE Geosci. Remote Sens. Lett.*, vol. 15, no. 5, pp. 784–788, May 2018.

[19] 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.

[20] A. Profeta, A. Rodriguez, and H. S. Clouse, "Convolutional neural networks for synthetic aperture radar classification," in *Proc. SPIE, Algorithms Synth. Aperture Radar Imag. XXIII*, vol. 9843, p. 98430M, May 2016.

[21] J. Li, R. Zhang, and Y. Li, "Multiscale convolutional neural network for the detection of built-up areas in high-resolution SAR images," in *Proc. IEEE Int. Geosci. Remote Sens. Symp. (IGARSS)*, Jul. 2016, pp. 910–913.

[22] D. Malmgren-Hansen and M. Nobel-Jørgensen, "Convolutional neural networks for SAR image segmentation," in *Proc. IEEE Int. Symp. Signal Process. Inf. Technol. (ISSPIT)*, Dec. 2015, pp. 231–236.

[23] 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.

[24] J. Zhao, W. Guo, S. Cui, Z. Zhang, and W. Yu, "Convolutional neural network for SAR image classification at patch level," in *Proc. IEEE Int. Geosci. Remote Sens. Symp. (IGARSS)*, Jul. 2016, pp. 945–948.

[25] Z. Xu, R. Wang, H. Zhang, N. Li, and L. Zhang, "Building extraction from high-resolution SAR imagery based on deep neural networks," *Remote Sens. Lett.*, vol. 8, no. 9, pp. 888–896, 2017.

[26] S. Auer, S. Hinz, and R. Bamler, "Ray-tracing simulation techniques for understanding high-resolution SAR images," *IEEE Trans. Geosci. Remote Sens.*, vol. 48, no. 3, pp. 1445–1456, Mar. 2010.

[27] J. Tao, S. Auer, G. Palubinskas, P. Reinartz, and R. Bamler, "Automatic SAR simulation technique for object identification in complex urban scenarios," *IEEE J. Sel. Topics Appl. Earth Observ. Remote Sens.*, vol. 7, no. 3, pp. 994–1003, Mar. 2014.

[28] 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.

[29] M. Shahzad, M. Maurer, F. Fraundorfer, Y. Wang, and X. Zhu, "Extraction of buildings in vhr SAR images using fully convolution neural networks," in *Proc. IEEE Int. Geosci. Remote Sens. Symp. (IGARSS)*, 2018.

[30] 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.

[31] G. Fornaro, F. Lombardini, and F. Serafino, "Three-dimensional multipass SAR focusing: Experiments with long-term spaceborne data," *IEEE Trans. Geosci. Remote Sens.*, vol. 43, no. 4, pp. 702–714, Apr. 2005.

[32] S. Zheng *et al.*, "Conditional random fields as recurrent neural networks," in *Proc. IEEE Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2015, pp. 1529–1537. [Online]. Available: https://arxiv.org/abs/1502.03240

[33] G. Fornaro, F. Serafino, and F. Soldovieri, "Three-dimensional focusing with multipass SAR data," *IEEE Trans. Geosci. Remote Sens.*, vol. 41, no. 3, pp. 507–517, Mar. 2003.

[34] X. Zhu, *Very High Resolution Tomographic SAR Inversion for Urban Infrastructure Monitoring: A Sparse and Nonlinear Tour* (Deutsche Geodätische Kommission bei der Bayerischen Akademie der Wissenschaften: Dissertationen: Reihe C), vol. 666. Berlin, Germany: Verlag der Bayerischen Akademie der Wissenschaften, 2011, p. 160.

[35] X. X. Zhu, Y. Wang, S. Gernhardt, and R. Bamler, "Tomo-GENESIS: DLR's tomographic SAR processing system," in *Proc. Joint Urban Remote Sens. Event (JURSE)*, Apr. 2013, pp. 159–162.

[36] Y. Wang, X. X. Zhu, B. Zeisl, and M. Pollefeys, "Fusing meter-resolution 4-D InSAR point clouds and optical images for semantic urban infrastructure monitoring," *IEEE Trans. Geosci. Remote Sens.*, vol. 55, no. 1, pp. 14–26, Jan. 2017.

[37] 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.

[38] M. Haklay, "How good is volunteered geographical information? A comparative study of OpenStreetMap and ordnance survey datasets," *Environ. Planning B, Planning Design*, vol. 37, no. 4, pp. 682–703, Aug. 2010.

[39] M. Shimrat, "Algorithm 112: Position of point relative to polygon," *Commun. ACM*, vol. 5, no. 8, p. 434, Aug. 1962, doi: 10.1145/368637.368653.

[40] F. P. Preparata and M. I. Shamos, *Computational Geometry: An Introduction*. New York, NY, USA: Springer-Verlag, 1985.

[41] G. Csurka, C. R. Dance, L. Fan, J. Willamowski, and C. Bray, "Visual categorization with bags of keypoints," in *Proc. Workshop Stat. Learn. Comput. Vis. (ECCV)*, Prague, Czech Republic, vol. 1, 2004, pp. 1–16.

[42] C. Cortes and V. Vapnik, "Support-vector networks," *Mach. Learn.*, vol. 20, no. 3, pp. 273–297, 1995.

[43] A. Vedaldi and B. Fulkerson, "Vlfeat: An open and portable library of computer vision algorithms," in *Proc. 18th ACM Int. Conf. Multimedia*, New York, NY, USA, 2010, pp. 1469–1472, doi: 10.1145/1873951.1874249.

[44] *Generate 2D and 3D Information, Purely From Images With Pix4d*. Accessed: Jan. 17, 2018. [Online]. Available: https://pix4d.com/

[45] I. Sobel and G. Feldman, "A 3×3 isotropic gradient operator for image processing, presented at a talk at the stanford artificial project," in *Pattern Classification and Scene Analysis*, R. Duda and P. Hart, Eds. Hoboken, NJ, USA: Wiley, 1968, pp. 271–272.

[46] J. Shotton, J. Winn, C. Rother, and A. Criminisi, "TextonBoost for image understanding: Multi-class object recognition and segmentation by jointly modeling texture, layout, and context," *Int. J. Comput. Vis.*, vol. 81, no. 1, pp. 2–23, Dec. 2007.

[47] J. Shotton, M. Johnson, and R. Cipolla, "Semantic texton forests for image categorization and segmentation," in *Proc. CVPR*, Jun. 2008, pp. 1–8.

[48] S. Schulter, P. Wohlhart, C. Leistner, A. Saffari, P. M. Roth, and H. Bischof, "Alternating decision forests," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2013, pp. 508–515.

[49] J. Long, E. Shelhamer, and T. Darrell, "Fully convolutional networks for semantic segmentation," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2015, pp. 3431–3440.

[50] E. Shelhamer, J. Long, and T. Darrell, "Fully convolutional networks for semantic segmentation," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 39, no. 4, pp. 640–651, Apr. 2017.

[51] P. Krähenbühl and V. Koltun, "Efficient inference in fully connected CRFs with Gaussian edge potentials," in *Proc. Adv. Neural Inf. Process. Syst.*, 2011, pp. 109–117.

[52] 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. 11th Eur. Conf. Synth. Aperture Radar (EUSAR)*, Jun. 2016, pp. 1–4.

[53] L. Cascini *et al.*, "Detection and monitoring of facilities exposed to subsidence phenomena via past and current generation SAR sensors," *J. Geophys. Eng.*, vol. 10, no. 6, p. 064001, 2013.

[54] D. Peduto, S. Ferlisi, G. Nicodemo, D. Reale, G. Pisciotta, and G. Gullà, "Empirical fragility and vulnerability curves for buildings exposed to slow-moving landslides at medium and large scales," *Landslides*, vol. 14, no. 6, pp. 1993–2007, Dec. 2017.

Muhammad Shahzad (S'12–M'16) received the B.E. degree in electrical engineering from the National University of Sciences and Technology (NUST), Islamabad, Pakistan, the M.Sc. degree in autonomous systems (robotics) from the Bonn Rhein Sieg University of Applied Sciences, Sankt Augustin, Germany, and the Ph.D. degree in radar remote sensing and image anaylsis from the Department of Signal Processing in Earth Observation, Technische Universität München, Munich, Germany, in 2004, 2011, and 2016, respectively. His Ph.D. dissertation was on automatic 3-D reconstruction of objects from point clouds retrieved from spaceborne synthetic aperture radar image stacks. He has attended twice two weeks professional thermography training course at the Infrared Training Center, North Billerica, MA, USA, in 2005 and 2007.

He was a Guest Scientist with the Institute for Computer Graphics and Vision, Technical University of Graz, Graz, Austria, from 2015 to 2016. Since 2016, he has been an Assistant Professor with the School of Electrical Engineering and Computer Science, National University of Sciences and Technology. His research interests include processing both unstructured/structured 3-D point clouds, optical RGBD data, and very high-resolution radar images.

Michael Maurer received the B.S. and M.S. degrees (Hons.) in telematics (computer science and electronics) from the Graz University of Technology, Graz, Austria, in 2008 and 2010, respectively, with a focus on computer vision and mobile robots, where he is currently pursuing the Ph.D. degree with the Institute of Computer Graphics and Vision.

Next to his Ph.D. studies, he is managing and leading industrial research projects and is the representative of the institute at the Disaster Competence Network Austria. He has authored (co-authored) about 20 scientific publications, including journal papers and peer-reviewed conference papers. His research interests include 3-D semantics, deep learning, 3-D reconstruction, visual navigation, image-based localization and mapping, and image acquisition systems and camera drones.

Mr. Maurer is an active member of the Computer Vision Community which led to being an Invited Speaker at the 169. DVW-Seminar, Germany, in 2018. He was nominated for the Best Video Award at the International Conference on Robotics and Automation 2012 and a finalist at the DJI Challenge (Search and Rescue using a Camera Drone) in 2016.

Friedrich Fraundorfer received the Ph.D. degree in computer science from the Institute of Computer Graphics and Vision, Graz University of Technology (TU Graz), Graz, Austria, in 2006.

He is currently an Assistant Professor with TU Graz. His research interests include 3-D computer vision, robot vision, multiview geometry, visual-inertial fusion, microaerial vehicle, autonomous systems, and aerial imaging.

Yuanyuan Wang (S'10–M'14) received the B.Eng. degree (Hons.) in electrical engineering from The Hong Kong Polytechnic University, Hong Kong, in 2008, and the M.Sc. and Dr. Ing. degrees from the Technical University of Munich (TUM), Munich, Germany, in 2010 and 2015, respectively.

In 2014, he was a Guest Scientist with the Institute of Visual Computing, ETH Zurich, Zürich, Switzerland. He is currently with the Signal Processing in Earth Observation, TUM. He is devoting himself in coordinating and developing key algorithms in the European project So2Sat: Big Data for 4D Global Urban Mapping—1016 Bytes from Social Media to Earth Observation Satellites. His research interests include optimal and robust parameters estimation in multibaseline interferometric synthetic aperture radar (InSAR) techniques, multisensor fusion algorithms of InSAR and optical data, nonlinear optimization for complex numbers, and the applications of these techniques in urban and volcanic areas.

Dr. Wang serves the community as a reviewer for several remote sensing journals and a Reviewer for the European Research Council and the French National Research Agency. He is one of the best reviewers of the IEEE TRANSACTIONS ON GEOSCIENCE AND REMOTE SENSING in 2016.

Xiao Xiang Zhu (S'10–M'12–SM'14) 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 was a Guest Scientist or a Visiting Professor with the Institute for Electromagnetic Sensing of Environment, Italian National Research Council, 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. She is currently the Professor for Signal Processing in Earth Observation, TUM, and the Head of the Department of EO Data Science, Earth Observation Center of the German Aerospace Center (DLR), and the Helmholtz Young Investigator Group SiPEO, DLR, and TUM, Germany. Her research interests include 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.