CONDITIONAL GIS-AWARE NETWORK FOR INDIVIDUAL BUILDING SEGMENTATION IN A VHR SAR IMAGE

Yao Sun[1,2], *Yuansheng Hua*[1,2], *Lichao Mou*[1,2], *Xiao Xiang Zhu*[1,2]

[1]Remote Sensing Technology Institute (IMF), German Aerospace Center (DLR), Germany
[2]Data Science in Earth Observation, Technical University of Munich (TUM), Germany

IGARSS 2021 - 2021 IEEE International Geoscience and Remote Sensing Symposium | 978-1-6654-0369-6/21/$31.00 ©2021 IEEE | DOI: 10.1109/IGARSS47720.2021.9553135

ABSTRACT

In this paper, we propose a network for individual building segmentation from a single VHR SAR image. The proposed network employs building footprints from GIS data in learning multi-level visual features to predict building masks in the SAR image. Experimental results over Berlin show that the proposed network effectively brings improvements with variant backbones. In addition, we propose an approach for generating building labels from an accurate digital elevation model (DEM), which can be used to generate large-scale SAR image datasets.

Index Terms— deep convolutional neural network (CNN), GIS, building segmentation, large-scale, synthetic aperture radar (SAR)

1. INTRODUCTION

Being able to provide data independently of sun illumination and weather conditions, very high resolution (VHR) synthetic aperture radar (SAR) imagery has attracted many researchers in studying buildings in urban regions. However, due to its side-looking imaging geometry and complex backscattering mechanism, the problem is highly challenging. Several works [1, 2] develop algorithms to extract buildings in urban environments, but these methods are difficult to be applied to large-scaled complex scenarios. The TomoSAR technique allows for large-scale extraction and reconstruction [3, 4], however, it requires multiple SAR acquisitions that are often not available. In addition to SAR data, some researchers introduce auxiliary information, such as footprint polygons in GIS data [5–7] for providing exact locations and geometric shapes of buildings in the real world.

In recent years, deep neural networks have been becoming increasingly popular and have shown success in applications of SAR data [8]. One problem of applying deep networks to building analysis tasks using SAR data is the lack of annotations. To address this issue, in [9, 10], building annotations in SAR images are acquired from 3D data, i.e., a TomoSAR point cloud or a DEM, and segmentation networks are trained to extract building areas. However, individual buildings cannot be recognized in both the works due to serious layover effects in urban areas.

In this paper, we introduce GIS building footprints as complementary data and propose a novel conditional GIS-aware network to segment individual buildings in large urban areas. Next, we explain the dataset generation procedure and the proposed network in Section 2. The experimental results are shown and evaluated in Section 3. Section 4 concludes the paper.

2. METHODOLOGY

2.1. Dataset generation

Building annotations and building footprints in SAR images are necessary for training our network. We use an accurate DEM and GIS data in UTM coordinate to generate the dataset.

2.1.1. Data Preparation in the UTM Coordinate System

First, in the UTM coordinate system, we model the scene that is viewed by the radar sensor in two steps: *a. Build a complete point cloud of the scene from the DEM*. The DEM is firstly represented as a nadir-looking point cloud, i.e., each pixel in the DEM is viewed as a 3D point. Then, the vertical data gaps in the point cloud are filled by adding points at these data gaps. *b. Remove sensor-invisible points from the point cloud.* 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 [11] is applied.

Then, for each building, its point clouds in the scene are extracted by selecting points inside its footprint in GIS data.

2.1.2. Dataset Generation in the SAR Image Coordinate System

The point clouds and footprint of each building need to be projected to the SAR image coordinate system. In this work, the coordinate transformation was performed using DLR's Integrated Wide Area Processor.

Then, according to coordinates of building points in the SAR image coordinate system, ground truth masks of build-

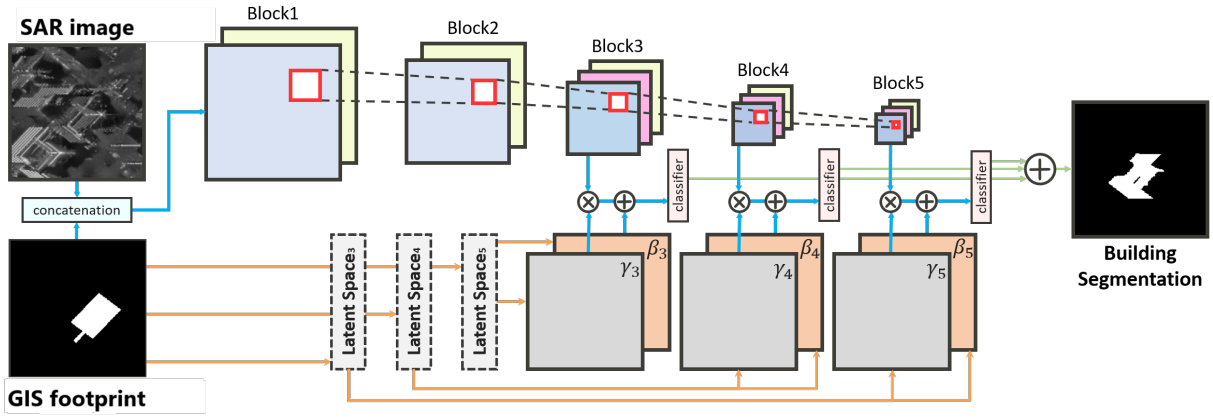

Fig. 1: The architecture of the proposed network.

ings are generated. Building footprint masks in the SAR image coordinate system are also created.

2.2. Conditional GIS-aware network

Our network takes a SAR image and building footprints as inputs and predicts building areas in the SAR image. As building footprints provide geometry and location information, we propose a network module that exploits such information and performs a conditional GIS-aware normalization.

2.2.1. Conditional GIS-aware Normalization Module

To distill the geometry information in GIS data and use it to normalize final predictions, We propose a conditional GIS-aware normalization module (CG module). We use VGG-16 as the backbone of our network to extract features from multiple layers, and the features learned from the last three blocks are fed into the CG module separately (see Fig. 1).

Formally, let $\boldsymbol{m}_{gis}$ be the building footprint mask with a spatial size of $W \times H$, and $\boldsymbol{x}_b$ be feature maps extracted from the b-th convolutional block. W' and H' denotes the width and height of $\boldsymbol{x}_b$, and C' is the number of channels. We consider a naive conditional normalization 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. $\hat{\boldsymbol{x}}_b$ denotes the normalized $\boldsymbol{x}_b$. γ and β are often computed channel-wisely as $\boldsymbol{x}_b$ consists of multiple channels. We denote the c-th channel of $\boldsymbol{x}_b$ with c and rewrite Eq. (1):

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

In this task, to normalize features learned from SAR images and conditioned on GIS data, we reformulate Eq. (2):

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

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 $\boldsymbol{\gamma}_b$ and $\boldsymbol{\beta}_b$ are formatted as matrices with a size of $W' \times H' \times C'$. To implement Eq. (3), we project $\boldsymbol{m}_{gis}$ onto a latent space through 3×3 convolutions and then employ two convolutional layers to learn γ_b and β_b from the encoded $\boldsymbol{m}_{gis}$. then, the element-wise multiplication of $\gamma_b(\boldsymbol{m}_{gis})$ and $\boldsymbol{x}_b$ is performed, and the output is added to $\beta_b(\boldsymbol{m}_{gis})$ pixel by pixel.

2.2.2. Configuration of CG-Net

Fig. 1 illustrates the architecture of the proposed CG-Net. To fully exploit GIS data at multiple scales, three CG modules are appended to the last three convolutional blocks of the backbone. The multi-level feature maps are upsampled to match the spatial resolution of $\boldsymbol{m}_{gis}$ via bilinear interpolation. To reduce the computation overhead of subsequent operations, the number of feature channels is reduced through 1×1 convolutions. Outputs of the CG module are squashed into the number of classes 2 and added via an element-wise addition operation to produce final segmentation results. The proposed CG module is in a plug-and-play fashion and is flexible 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.

3. EXPERIMENTS

3.1. Dataset

With the workflow described in Section 2.1, we generate our dataset using a spotlight TerraSAR-X image over Berlin with the pixel spacing of 0.871 m in the azimuth direction and 0.455 m in the slant range direction, building footprints of

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**

Table 1: Numerical results. The highest values of different metrics are highlighted in **bold**.

the study area[1], and a DEM with a resolution of 7cm/pixel. The dataset contains a 5736×10312 SAR image, footprint masks, and ground truths of individual buildings.

3.2. Training Details

To train the network, the SAR image is cropped into patches of 256×256 pixels with a stride of 150 pixels. There are 30056 buildings in the dataset with three patches each: a SAR image patch, a footprint patch, and a ground truth mask. 19434 are used for training, and the rest for testing. The training and test regions do not overlap. The network is implemented on TensorFlow and trained on one NVIDIA Tesla P100 16GB GPU for 155k iterations. During the training procedure, all weights are updated through back-propagation, and we select Netrov Adam as the optimizer. The loss is defined as binary cross-entropy. The learning rate is initialized as $2e-3$ and reduced by a factor of $\sqrt{10}$ once the loss stops to decrease for two epochs. A small batch size of 5 is used.

3.3. Evaluation

For evaluation, we calculate the F1 score, the intersection over union (IoU), and overall accuracy (OA):

$$F1 = 2 \cdot \frac{P \cdot R}{P+R}, P = \frac{tp}{tp+fp}, R = \frac{tp}{tp+fn}, \\ IoU = \frac{tp}{tp+fp+fn}, OA = \frac{tp+tn}{tp+tn+fp+fn}. \quad (4)$$

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

Four models are compared in our experiments: FCN, FCN-CG, DeepLabv3, and DeepLabv3-CG, in which 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. Table 1 presents the results of the four models and Fig. 2 shows segmentation results on three examples. As can be seen, with the CG module, the precision improves 1.95% and 3.94% with the backbone, FCN and DeepLabv3, and the IoU increases by 1.50% and 2.16%. The improvements achieved by the CG module in

[1]Downloaded from Berlin 3D-Download Portal: https://www. business-locationcenter.de/downloadportal/

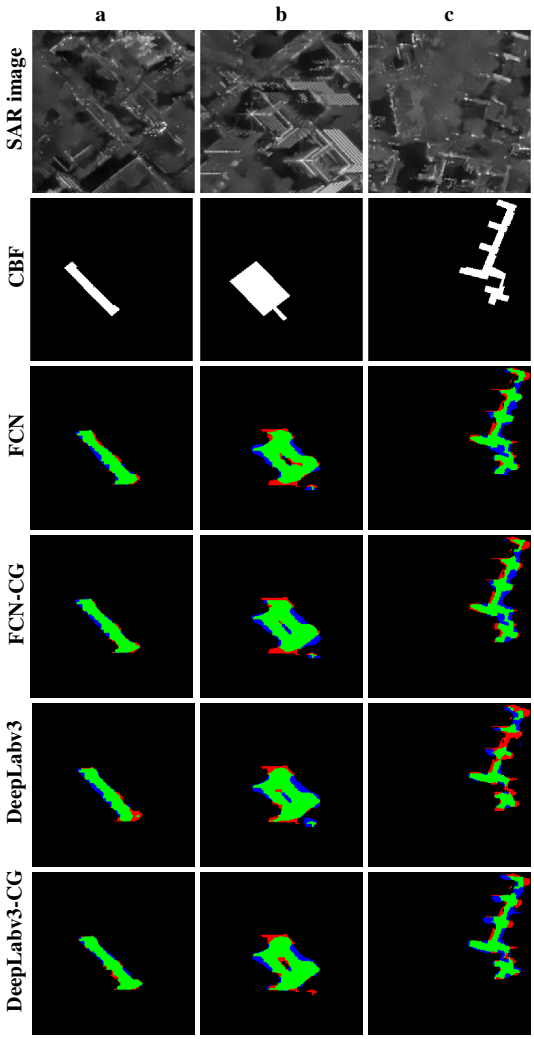

Fig. 2: Examples of segmentation results. Pixel-based true positives, false positives, and false negatives are marked in green, red, and blue, respectively.

FCN-CG and DeepLabv3-CG demonstrate the effectiveness of our proposed network. DeepLabv3-CG achieves the best performance in all four metrics on our dataset. Fig. 3 shows the results obtained with DeepLabv3-CG in the study area.

4. CONCLUSION

This paper proposes a conditional GIS-aware network to segment individual buildings from a VHR SAR image. We also propose an approach to automatically annotate individual building areas in SAR images using an accurate DEM. The proposed methods are validated over the Berlin area using a high-resolution spotlight TerraSAR-X image, and the results demonstrate the effectiveness of the proposed network.

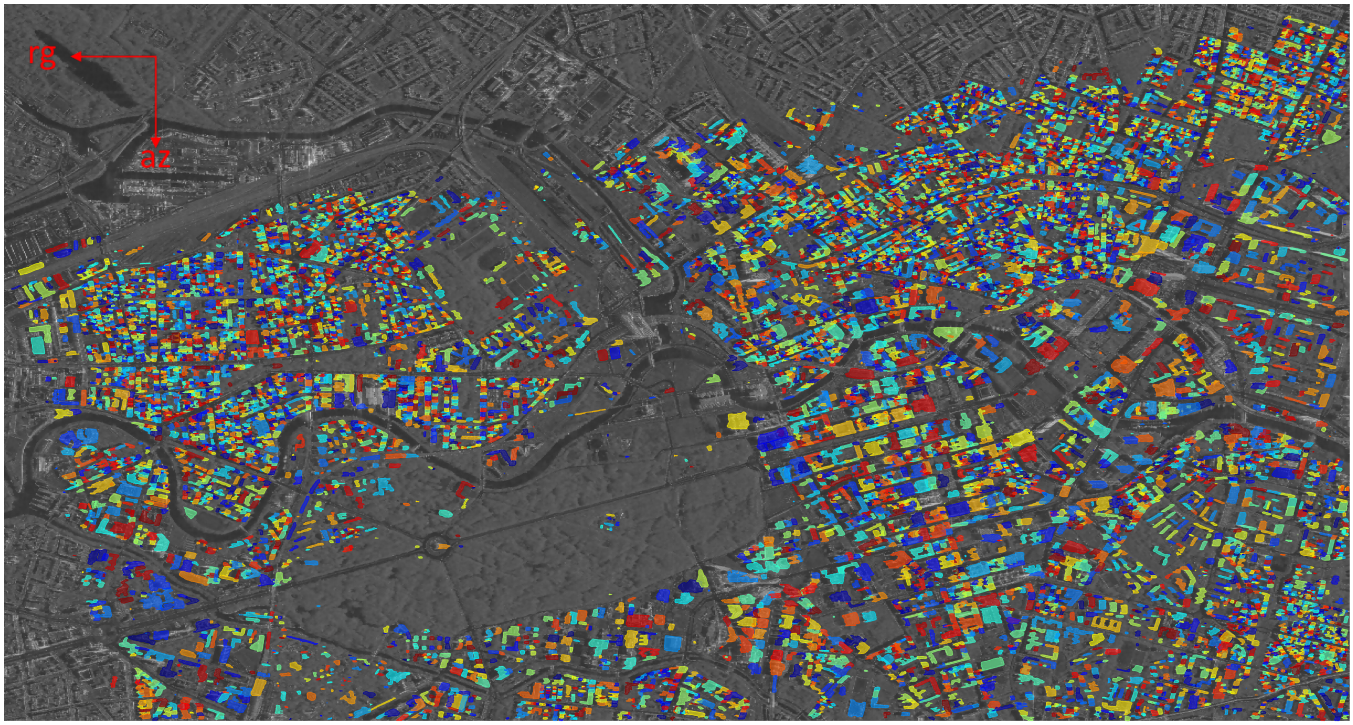

Fig. 3: Results in the study area obtained by DeepLabv3-CG. The segments are plotted in different colors translucently for visualizing overlapping areas between buildings. rg and az denote the range and azimuth direction, respectively.

5. REFERENCES

[1] R. Guida, A. Iodice, and D. Riccio, "Height retrieval of isolated buildings from single high-resolution SAR images," *IEEE Transactions on Geoscience and Remote Sensing*, vol. 48, no. 7, pp. 2967–2979, 2010.

[2] 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 Transactions on Geoscience and Remote Sensing*, vol. 48, no. 3, pp. 1487–1504, 2010.

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

[4] Y. Sun, "3D building reconstruction from spaceborne TomoSAR point cloud," Master's thesis, Technical University of Munich, Munich, Germany, 2016.

[5] L. Wen and F. Yamazaki, "Building height detection from high-resolution TerraSAR-X imagery and GIS data," in *Joint Urban Remote Sensing Event (JURSE)*, 2013.

[6] Y. Sun, M. Shahzad, and X. X. Zhu, "Building height estimation in single SAR image using OSM building footprints," in *Joint Urban Remote Sensing Event (JURSE)*, 2017.

[7] Y. Sun, S. Montazeri, Y. Wang, and X. X. Zhu, "Automatic registration of a single sar image and gis building footprints in a large-scale urban area," *ISPRS Journal of Photogrammetry and Remote Sensing*, vol. 170, pp. 1–14, 2020.

[8] X. X. Zhu, S. Montazeri, M. Ali, Y. Hua, Y. Wang, L. Mou, Y. Shi, F. Xu, and R. Bamler, "Deep learning meets SAR," *IEEE Geoscience and Remote Sensing Magazine*, vol. pp, no. pp, pp. 1–26, 2021.

[9] M. Shahzad, M. Maurer, F. Fraundorfer, Y. Wang, and X. X. Zhu, "Buildings detection in VHR SAR images using fully convolution neural networks," *IEEE Transactions on Geoscience and Remote Sensing*, vol. 57, no. 2, pp. 1100–1116, 2019.

[10] 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 *Joint Urban Remote Sensing Event (JURSE)*, 2019.

[11] S. Katz, A. Tal, and R. Basri, "Direct visibility of point sets," *ACM Transactions on Graphics*, vol. 26, no. 3, p. 24, 2007.