END-TO-END SEMANTIC SEGMENTATION AND BOUNDARY REGULARIZATION OF BUILDINGS FROM SATELLITE IMAGERY

Qingyu Li [1,2], *Stefano Zorzi*[3], *Yilei Shi* [4], *Friedrich Fraundorfer*[3,2], *Xiao Xiang Zhu* [1,2]

[1] Data Science in Earth Observation, Technical University of Munich (TUM), Munich, Germany
[2] Remote Sensing Technology Institute (IMF), German Aerospace Center (DLR), Wessling, Germany
[3] Institute of Computer Graphics and Vision, Graz University of Technology (TU Graz), Graz, Austria
[4] Remote Sensing Technology, Technical University of Munich (TUM), Munich, Germany

IGARSS 2021 - 2021 IEEE International Geoscience and Remote Sensing Symposium | 978-1-6654-0369-6/21/$31.00 ©2021 IEEE | DOI: 10.1109/IGARSS47720.2021.9555147

ABSTRACT

Building footprint generation is a vital task of satellite imagery interpretation. However, the segmentation masks of buildings obtained by existing semantic segmentation networks often have blurred boundaries and irregular shapes. In this research, we propose a new boundary regularization network for building footprint generation in satellite images. More specifically, we consider semantic segmentation and boundary regularization in an end-to-end generative adversarial network (GAN). The learned building footprints are regularized by the interplay between the generator and discriminator. By doing so, the straight boundaries and geometric details of the building could be preserved. Experiments are conducted on a collected dataset of Planetscope satellite imagery (spatial resolution: 4.77 m/pixel). Our approach is much superior to the state-of-the-art methods in both quantitative and qualitative results.

Index Terms— semantic segmentation, boundary regularization, building, satellite imagery, generative adversarial network

1. INTRODUCTION

Building footprint generation from satellite imagery is of great interest in the remote sensing community for a wide range of applications, such as real estate management, urban planning, and monitoring. However, different materials and structures of buildings bring challenges to this task. In the early stage, researchers have proposed a variety of methods by exploiting hand-crafted features, but suffer from a problem of poor generalization. Remarkable progress comes from convolutional neural networks (CNNs), which have driven advances in the task of building footprint generation due to their powerful feature extraction capability from raw data. The existing CNNs seem to be able to shown promising results [1] [2] [3] also in large-scale tasks (cf. Fig. 1). However, when we zoom in on some semantic masks of buildings (see results from FC-DenseNet [4] in Fig. 1), it can be clearly seen that building boundaries are blurred, and they often show irregular shapes as there is no constraint on the geometry of the extracted building footprints. Recently, several methodologies have already made an attempt to deal with this problem. Their pipeline consists of two stages, where the building segmentation is implemented as a first step and the building regularization is then followed as a second step. For the regularization stage, they are either based on deep neural networks [5] [6] or utilizing polygon simplification algorithms [7]. In this case, the building regularization results largely rely on the building segmentation results, which can not provide replicable and stable building footprint maps, especially for large-scale applications. Motivated by this observation, in this work, we propose an end-to-end trainable network for automatic building footprint generation, where the building segmentation and regularization are realized simultaneously. More specifically, the proposed method is a Generative Adversarial Network (GAN), which consists of two modules: the generator G and the discriminator D. G learns the regularized segmentation masks which can enhance sharp corners and straight edges. D is utilized to distinguish between generated and ideal footprints.

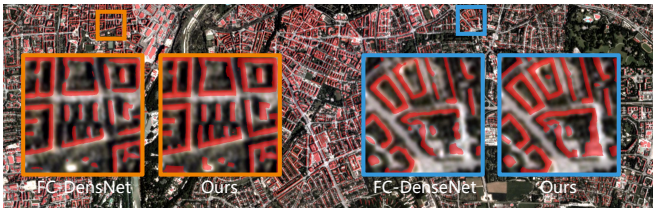

Fig. 1. The building segmentation results obtained from FC-DenseNet [4] in the large scale and two zoomed in areas.

2. METHODOLOGY

2.1. Overview

Fig. 2 provides an overview of the proposed framework, which consists of two modules: the generator G and the

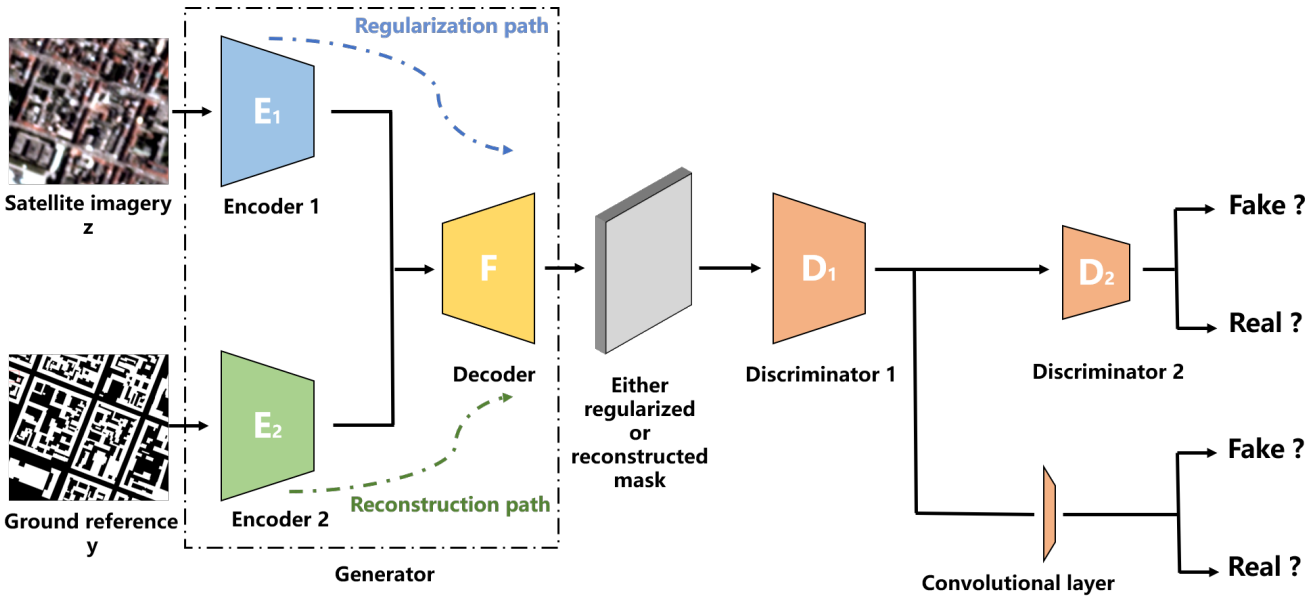

Fig. 2. Overview of the proposed approach.

discriminator D. A residual convolutional encoder-decoder structure [6] is implemented in G. There are two paths in the G: the regularization path and the reconstruction path. The former takes satellite imagery as input and produces the regularized building footprint mask from the encoder E_1 and decoder F. At the same time, the latter encodes and decodes the ideal input mask by the encoder E_2 and decoder F, ensuring to have the same real-valued masks as input to the discriminator. Then two discriminators D_1 and D_2 are introduced in order to distinguish between regularized footprints and ideal ones at different scales, separately, which assure better discrimination. It should be noted that the whole network is trained in an end-to-end fashion.

2.2. Objective Function

The goal of G is to learn a a mapping function from the satellite imagery z to ideal building footprints y, and it is learned with three types of loss functions: adversarial loss L_{GAN}, reconstruction loss L_{REC}, and semantic loss L_{SEG}.

$$L_G = \alpha \cdot L_{GAN} + \beta \cdot L_{REC} + \gamma \cdot L_{SEG} \tag{1}$$

, where we set $\alpha = 0.5$, $\beta = 1$, $\gamma = 10$.

The *adversarial loss* aims to learn the mapping function from z to y, which motivates the E_1 and F to produce footprints similar to the ideal ones that have a constraint for the geometry of buildings. The objective function of this component can be denoted as:

$$L_{GAN} = -E_z[D_1(F(E_1(z)))] - E_z[D_2(D_1(F(E_1(z))))] \tag{2}$$

Noting that D_2 takes the output feature from D_1 as input for further discrimination.

The *reconstruction loss* is introduced to ensure a stable joint training with the common decoder F, and it can be calculated as:

$$L_{REC} = -y \cdot log(F(E_2(y))) - (1-y) \cdot log(1 - F(E_2(y))) \tag{3}$$

Apart from the above two losses, an important loss term *semantic loss* is used to avoid the information loss of the regularization path. The *semantic loss* enables the well-preserved semantic correctness in regularized buildings footprints, and is expressed as:

$$L_{SEG} = -y \cdot log(F(E_1(y))) - (1-y) \cdot log(1 - F(E_1(y))) \tag{4}$$

D tries to discriminate ideal footprints from regularized ones generated by the regularization path. Two discriminators are trained to differentiate between regularized and reconstructed building footprints. The objective function of it is expressed as:

$$\begin{aligned} L_D = &-E_y[D_1(F(E_2(y)))] + E_z[D_1(F(E_1(z)))] \\ &-E_y[D_2(D_1(F(E_2(y))))] + E_z[D_2(D_1(F(E_1(z))))] \end{aligned} \tag{5}$$

Notable that the training of the proposed framework for semantic segmentation and boundary regularization is accomplished in an end-to-end manner, which makes it more efficient and robust.

3. EXPERIMENT

In this research, the study sites cover five European cities: (1) Munich, Germany; (2) Berlin, Germany; (3) Rome, Italy; (4) Paris, France; (5) Zurich, Switzerland. Planetscope satellite imagery with three bands — Red, Green, Blue (RGB) at 4.77 m spatial resolution is selected as the experimental dataset. In order to fully harness the high-level details of OpenStreetMap and assure the reconstruction path being capable of recovering all those fine details, we firstly resample the satellite imagery to 0.75 m spatial resolution and then rasterize the corresponding building footprints from OpenStreetMap to the same spatial resolution. Afterward, satellite imagery and their corresponding ground reference are cut into patches with a size 256×256. The numbers of training and validation patches covering four European cities: (1) Berlin, Germany; (2) Milan, Italy; (3) Cologne, Germany; (4) Zurich, Switzerland are listed in Table 1. 1600 patches of Planetscope satellite imagery and their corresponding ground reference from the city of Munich, Germany are used as the test set to evaluate the performance of models.

In order to validate the superiority of our methods, we select two methods for comparison: FC-DenseNet [4] and Two stage [6], which are the state-of-the-art semantic segmentation and boundary regularization methods. All methods are implemented within a Pytorch framework on an NVIDIA Quadro P4000 with 8 GB of memory. All models are trained by an optimizer of Adam with a learning rate of 0.0001, and the training batch size of all models is set as 4.

Table 1. The numbers of training and validation patches of the dataset.

City	Training	Validation
Berlin	1606	402
Milan	1596	400
Cologne	1505	377
Zurich	1617	405

4. RESULTS

The performance of all models is evaluated by four metrics. F1-Score and Intersection Over Union (IoU) are selected for mask metrics, while Similarity Index Metric (SIM) and F-Measure are used to measure the quality of building boundaries. Notable that SIM is an evaluation criterion to describe the geometric similarity between the turning functions of the ground reference and estimated polygons [8] [9].

Table 2 and Fig 3 present the quantitative and qualitative results of all methods. Our proposed approach outperforms FC-DenseNet and Two stage in terms of both mask and boundary accuracy. Notable, straight boundaries and sharp corners are preserved in our network, thus, the geometric details of the buildings are well depicted. This indicates that our method is able to increase the quality of buildings boundaries without losing mask accuracy.

Table 2. Accuracy indices of different methods derived from the test set of Munich).

	Mask		Boundary	
Method	F1-Score	IoU	SIM	F-Measure
FC-DenseNet	57.93 %	40.77 %	63.83 %	9.55 %
Two stage	59.90 %	42.75 %	62.88 %	10.77 %
Proposed method	**67.35 %**	**50.78 %**	**68.31 %**	**13.10 %**

5. CONCLUSION

In this paper, we have proposed a novel boundary regularization approach that obtains building footprints based on a GAN model. The regularized segmentation masks are learned by the generator, which can preserve regular shapes of buildings. And the discriminator differentiates the generated footprints from ideal ones. We evaluate our approach on a collected dataset Planetscope satellite imagery and high-resolution ground reference from OpenStreetMap. Experimental results have demonstrated that our method is more competitive when compared with the state-of-the-art semantic segmentation and boundary regularization methods. Notable that the building boundaries generated by our method are fine-grained, and regular shapes of buildings are well preserved. In this regard, further steps such as vectorization that relies much on accurate geometric details will benefit from the proposed approach.

6. ACKNOWLEDGEMENTS

This work is supported by the European Research Council (ERC) under the European Union's Horizon 2020 research and innovation program (grant agreement no.ERC-2016-StG-714087, acronym: So2Sat, www.so2sat.eu), the Helmholtz Association under the framework of the Young Investigators Group "SiPEO" (VH-NG-1018, www.sipeo.bgu.tum.de) and Helmholtz Excellent Professorship "Data Science in Earth Observation - Big Data Fusion for Urban Research".

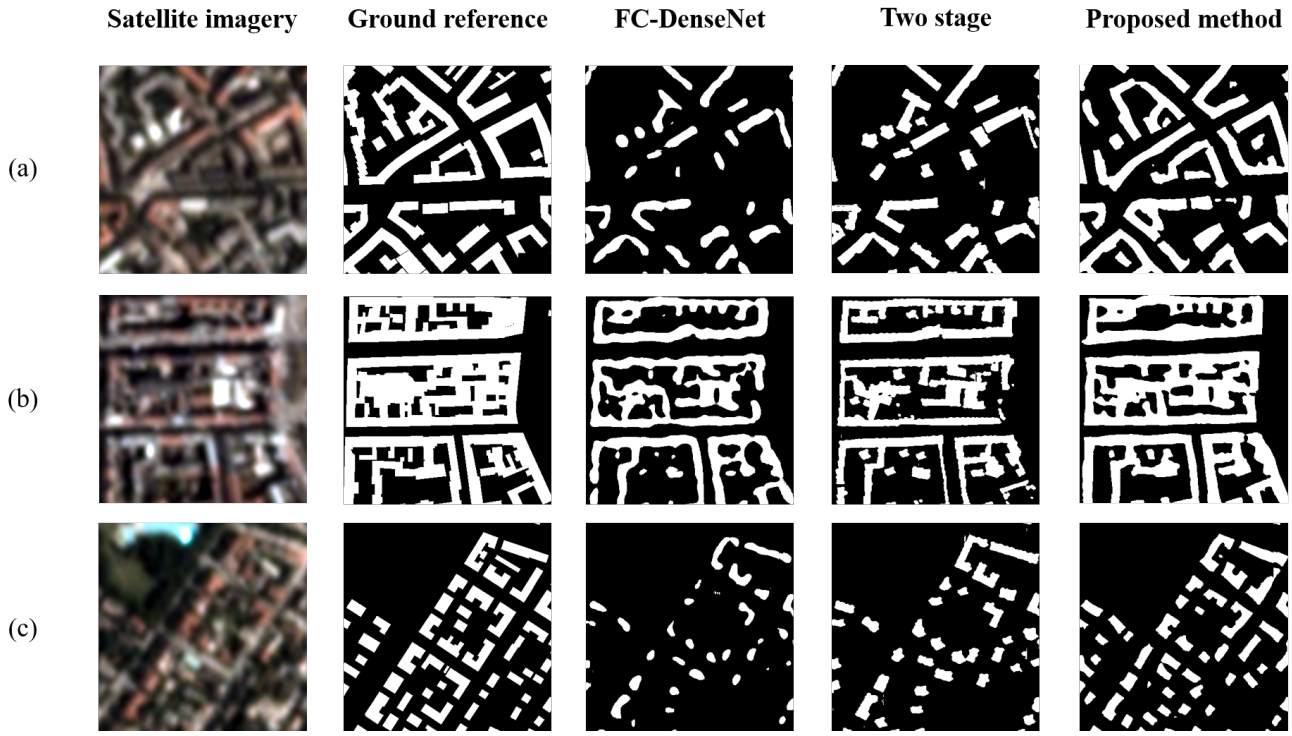

Fig. 3. Satellite imagery, ground reference, and visual results of different methods from example (a), (b), and (c) in test set of Munich.

7. REFERENCES

[1] Qingyu Li, Yilei Shi, Xin Huang, and Xiao Xiang Zhu, "Building footprint generation by integrating convolution neural network with feature pairwise conditional random field (fpcrf)," *IEEE Transactions on Geoscience and Remote Sensing*, 2020.

[2] Yilei Shi, Qingyu Li, and Xiao Xiang Zhu, "Building segmentation through a gated graph convolutional neural network with deep structured feature embedding," *ISPRS Journal of Photogrammetry and Remote Sensing*, vol. 159, pp. 184–197, 2020.

[3] Yilei Shi, Qingyu Li, and Xiao Xiang Zhu, "Building footprint generation using improved generative adversarial networks," *IEEE Geoscience and Remote Sensing Letters*, vol. 16, no. 4, pp. 603–607, 2018.

[4] Simon Jégou, Michal Drozdzal, David Vazquez, Adriana Romero, and Yoshua Bengio, "The one hundred layers tiramisu: Fully convolutional densenets for semantic segmentation," in *Proceedings of the IEEE conference on computer vision and pattern recognition workshops*, 2017, pp. 11–19.

[5] Stefano Zorzi and Friedrich Fraundorfer, "Regularization of building boundaries in satellite images using adversarial and regularized losses," in *IGARSS 2019-2019 IEEE International Geoscience and Remote Sensing Symposium*. IEEE, 2019, pp. 5140–5143.

[6] Stefano Zorzi, Ksenia Bittner, and Friedrich Fraundorfer, "Machine-learned regularization and polygonization of building segmentation masks," *arXiv preprint arXiv:2007.12587*, 2020.

[7] Kang Zhao, Jungwon Kang, Jaewook Jung, and Gunho Sohn, "Building extraction from satellite images using mask r-cnn with building boundary regularization.," in *CVPR Workshops*, 2018, pp. 247–251.

[8] Shenlong Wang, Min Bai, Gellert Mattyus, Hang Chu, Wenjie Luo, Bin Yang, Justin Liang, Joel Cheverie, Sanja Fidler, and Raquel Urtasun, "Torontocity: Seeing the world with a million eyes," *arXiv preprint arXiv:1612.00423*, 2016.

[9] Hongchao Fan, Alexander Zipf, Qing Fu, and Pascal Neis, "Quality assessment for building footprints data on openstreetmap," *International Journal of Geographical Information Science*, vol. 28, no. 4, pp. 700–719, 2014.