This manuscript was accepted to be presented at the IEEE International Conference of Pattern Recognition 2020.

Machine-learned Regularization and Polygonization of Building Segmentation Masks

Stefano Zorzi
Institute of Computer Graphics and Vision
Graz University of Technology
stefano.zorzi(at)icg.tugraz.at

Ksenia Bittner
Remote Sensing Technology Institute,
German Aerospace Center (DLR)
ksenia.bittner(at)dlr.de

Friedrich Fraundorfer
Institute of Computer Graphics and Vision
Graz University of Technology
fraundorfer(at)icg.tugraz.at

***Abstract*—We propose a machine learning based approach for automatic regularization and polygonization of building segmentation masks. Taking an image as input, we first predict building segmentation maps exploiting generic *fully convolutional network (FCN)*. A *generative adversarial network (GAN)* is then involved to perform a regularization of building boundaries to make them more realistic, *i.e.*, having more rectilinear outlines which construct right angles if required. This is achieved through the interplay between the discriminator which gives a probability of input image being true and generator that learns from discriminator's response to create more realistic images. Finally, we train the backbone *convolutional neural network (CNN)* which is adapted to predict sparse outcomes corresponding to building corners out of regularized building segmentation results. Experiments on three building segmentation datasets demonstrate that the proposed method is not only capable of obtaining accurate results, but also of producing visually pleasing building outlines parameterized as polygons.**

I. Introduction

The ability to extract vector representations of building polygons from aerial or satellite imagery has become a hot topic in numerous remote sensing applications, such as urban planning and development, city modelling, cartography, *etc*. The interest in and the development of new methodologies was also motivated by the current existence of several public benchmark datasets, like INRIA [1], SpaceNet [2], and CrowdAI [3]. The classical approaches in this research field mostly focused on the assignment of the semantic class to each pixel in the image, obtaining classification masks as output [4–7]. However, for many applications, the more advanced output in form of vector information is under demand. In this work, we aim to provide not only building segmentation results, which outlines follow the realistic building forms, mainly straight lines and right angles, but also to generate a polygonal vector structure for each building instance.

Convolutional neural networks (CNNs) have brought significant contributions to the field of computer vision, establishing themselves as the basis of semantic and instance segmentation. However, while performing the pixel-wise classification with high accuracy, they have problems with delineating the exact and regular building boundaries. To overcome this issue, we apply geometry constraints in the pixel domain using an adversarial loss to regularize the boundaries. Specifically, the generative part of the proposed *generative adversarial network (GAN)*-based architecture takes as input the segmentation results obtained from *residual recursive U-Net (R2U-Net)* or the ideal segments from the dataset's ground truth. By getting the "gradient feedback" from the discriminator which task is to verify if its input comes either from regularized segmentation mask or ideal one, the generator learns to output the improved outline contours of our initial segmentation.

Fig. 1: Building polygon results from our proposed methodology overlaid on top of a sample area from the Inria dataset.

In the literature, several methodologies have already made an attempt to directly predict vertices of object boundaries using CNN paradigm. They are either based on iterative prediction of outline points for one object at a time [8, 9] with possible interaction by users for corrections, or predicting only 4-sided polygons [10]. However, real world buildings are not constrained to a certain amount of corners. Motivated by this ideas, Li *et al.* [11] proposed a *recurrent neural network (RNN)* above the *region proposal network (RPN)* which step by step predicts the possible corners for a single building within every region of interest. In our method, we do not want to be limited to corners prediction for a single building centered inside the input patch. The proposed Mask2Poly network is trained to predict an arbitrary number of corners (depending on structure complexity) for random number of buildings in the image scene from the regularized segmentation results. Some results of polygonal representations after obtaining the

corner predictions from Mask2Poly are shown in Fig. 1.

In Section II, we review state-of-the-art methodologies in the related field. The details of designed architectures and the intuition behind selected objective functions are then presented in Section III. In Section IV, we demonstrate the effectiveness showing qualitative and quantitative results of our approach on three publicly available datasets, *i.e.*, INRIA [1], SpaceNet [2] and CrowdAI [3]. Section V concludes the paper.

II. Related work

Building segmentation from top view images has been one of the main research topics in remote sensing for decades. Before the deep learning era, the traditional methodologies for building footprint extraction relied on multi-step workflows utilizing detected low-level features to form building hypotheses [12, 13], assumptions that buildings compose of regular rectangular shapes [14, 15] and similarities of spectral reflectance values between building appearances[16, 17]. After the introduction of more powerful hardware, recent approaches began to heavily utilize deep convolutional networks for automatic building delineation providing state-of-the-art results. The task is approached via pixel-wise semantic segmentation applying FCNs on satellite or airborne images using the benefit of their high-resolution spectral information [5, 18]. Some methodologies embedded additional information in forms of heights from *digital surface models (DSMs)* [6, 19] or *OpenStreetMap (OSM)* [20] together with the spectral information to increase the evidence of buildings.

In the last few years, UNet-based architectures became one of the most successful models for segmentation and detection tasks not only in medical images but also in remote sensing. Motivated by recently proposed UNet-based models that achieved state-of-the-art performances in different building extraction challenges [18, 21] , the variant of UNet with residual and recurrent layers [22] is utilized in this work.

Building segmentation regularization has been getting increased attention over the recent years. Because neural networks try to decide for each image pixel whether it belongs to a building or not, they do not consider its geometry. As a result, building segmentation results have very often a blob-like appearance. Therefore, a footprint regularization step is very important to enforce that the resulting outlines not only match the ground truth but also have realistic appearances. Zhao *et al.* [23] proposed to regularize building instances obtained from semantic segmentation networks applying multi-step polygon simplification methods. Marcos *et al.* [24] proposed a more advanced architecture by integrating the classic active contour model of Kass *et al.* [25] into deep CNN to perform a joint end-to-end learning. In the following work, Cheng *et al.* [26] introduced a network based on a polar representation of active contours which prevent self-intersections and enforces outlines to be even closer to the ground truth. Work most related to ours is Zorzi *et al.* [7], which looked at the problem differently. The authors of this paper trained the regularization network in an unsupervised manner using adversarial losses together with Potts [27, 28] and normalized cut [28] regularization losses which embedded additional knowledge about building boundaries from the intensity image to the network. In our work, we extend the algorithm proposed in [7] redefining the training procedure and the architecture of the regularization network to obtain better results both in qualitative and quantitative terms.

Polygon prediction is a difficult but crucial step for multiple disciplines as it provides vector-based data representations. Typically, semantic segmentation results are vectorized employing Douglas-Peucker [29], RANSAC [30] or Hough transform [31] algorithms as a post-processing step. Recent approaches made an attempt to integrate a vectorization procedure into an end-to-end deep learning-based model. The approach of Castrejon *et al.* [8] and the followed work of Acuna *et al.* [9] sequentially produce polygonal vertices around the object boundary based on RNN. Although these methodologies provided impressive results, they are different from our proposed algorithm in terms of the size and amount of polygonized objects (an image crop containing only one object is annotated per procedure). Moreover, a human annotator's interaction is allowed during the prediction of polygonal vertices to correct them if needed. In contrast, we propose a deep learning-based methodology which automatically predicts polygon vertices without any limitation on the amount of objects within an input image.

III. Proposed method

In this paper, we propose a pipeline for building extraction that not only aims to achieve state-of-the-art segmentation accuracy, but also tries to predict visually pleasing building polygons.

The pipeline is composed by three consecutive and independent steps.

As a first step, a FCN is used to detect and segment building footprints given an intensity image. The resulting segmentation can achieve great accuracy in terms of *intersection over union (IoU)*, recall and completeness, but the predicted building boundaries do not have a regular shape since there are no constraints on the building geometry.

In order to produce a more realistic segmentation, we further refine the result through a second CNN trained using a combination of adversarial, reconstruction and regularized losses. As a result, the extracted building footprints have a more regular shape, with sharp corners and straight edges. As we show later in Section IV, this step greatly increases the footprints quality without losing segmentation accuracy.

Finally, we extract a polygon for each building instance detecting the corners from its regularized mask.

In the subsequent sections, we describe in more detail each component of the pipeline.

A. Building detection and segmentation

The first step in the proposed method aims to detect and outline the boundaries of the buildings present in the satellite or aerial image. This task can be solved exploiting one of the many instance or semantic segmentation networks proposed in literature, trained using cross-entropy losses. Since the three

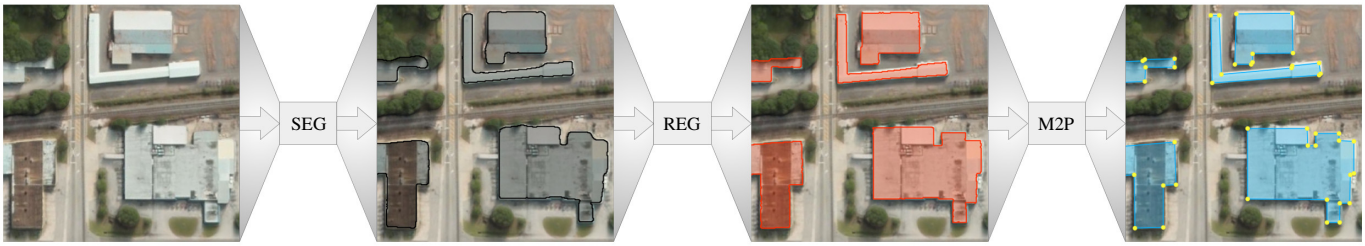

Fig. 2: The schematic overview of the proposed pipeline for automatic extraction of regularized building polygons. Buildings are initially detected and segmented by a *fully convolutional network (FCN)* (result shown in black). A footprint regularization network is then applied to the segmentation mask in the pixel domain (red). Finally, building polygons are extracted from the regularized mask (cyan, vertices highlighted in yellow).

stages of the pipeline are independent from each other, it is possible to choose the instance of semantic segmentation network which is best suited or which performs best on the specific dataset. In this work, we decided to use as segmentation baseline the R2U-Net proposed in [22], a simple but yet precise network which guarantees high building segmentation accuracy.

B. Regularization of the segmentation

The footprints predicted by the segmentation network typically have rounded corners and irregular edges due to the lack of geometric constraints during the prediction. Extracting building polygons from the initial building segmentation is a hard task that could lead to errors in the corners proposal procedure. For this reason, as a second step, we use a CNN for building regularization that aims to produce building footprints with regular and visually pleasing boundaries.

This translation can be successfully achieved training a GAN network composed by two different models. One of these networks is a generator which tries to generate a regularized version of the segmentation mask and the other network is a discriminator that examines generated and ideal footprints and estimates whether they are real or fake. The goal of the generator is to fool the discriminator, and as both networks get better and better at their job over the training, eventually the generator is forced to generate building footprints which become more realistic with each iteration.

The generator aims to learn a mapping function between the domain X, composed by segmented footprints, and the domain Y, made of ideal footprints, given the training samples $\{x_i\}_{i=1}^{N}$ where $x_i \in X$ and $\{y_i\}_{i=1}^{M}$ where $y_i \in Y$. To further improve the results we also exploit the intensity images, $\{z_i\}_{i=1}^{N}$ where $z_i \in Z$, training the model with an additional regularized loss.

The generator performs the regularization $G : \{X, Z\} \rightarrow Y$ exploiting a residual autoencoder structure, as shown in Fig. 3.

The regularized footprint is produced through the path composed by the encoder E_G and the residual decoder F, so the generator G can be seen as their combination $G(x, z) = F(E_G(x, y))$.

The discriminator network D tries to estimate whether the presented images are regularized footprints, generated by G, or ideal ones. The reason behind this path is to derive a reconstructed version of y. However, the adversarial network can easily distinguish two distributions, since the ideal mask is one-hot encoded with zeros and ones and the output of the autoencoder can range between zero and one. Therefore, both reconstructed and regularized image samples are generated using the same network F. Due to the joint training of two autoencoders with the common decoder, the proposed architecture is ensured to be stable and, as a result, escapes the situation where the discriminator wins.

1) Objective Function: Three types of loss functions in the learning procedure are used motivated by the good building footprints produced in [7]: *adversarial loss*, *reconstruction losses* and *regularized loss*.

The *adversarial loss*, introduced in [32], is used to learn the mapping function between the domain X and Y, encouraging the generator G to produce footprints similar to the ideal samples. This component of the objective function acts as a constraint for the geometry boundaries of the buildings and it is expressed as:

$$\mathcal{L}_{GAN}(G, D) = E_{x,z}[\log(1 - D(G(x, z))] \tag{1}$$

The discriminator D is trained to distinguish regularized and reconstructed footprints and its objective function can be expressed as:

$$\begin{aligned} \mathcal{L}_D(G, R, D) = {} & E_y[\log(1 - D(R(y)))] \\ & + E_{x,z}[\log D(G(x, z)] \end{aligned} \tag{2}$$

where the path $R(y) = F(E_R(y))$ encodes and reconstructs the ideal mask and the path $G(x, z) = F(E_G(x, z))$ generates the regularized footprints.

The *reconstruction* term is introduced to force the generator G to produce building footprints having an overall shape and pose similar to the segmentations received as input. The loss is also computed through the reconstruction path R to obtain a reconstructed version of the ideal mask. As reconstruction loss we simply use *binary cross entropy* and two losses can be written as:

$$\begin{aligned} \mathcal{L}_{rec_G}(G) &= -E_{x,z}[x \cdot \log G(x, z)] \\ \mathcal{L}_{rec_R}(R) &= -E_y[y \cdot \log R(y)] \end{aligned} \tag{3}$$

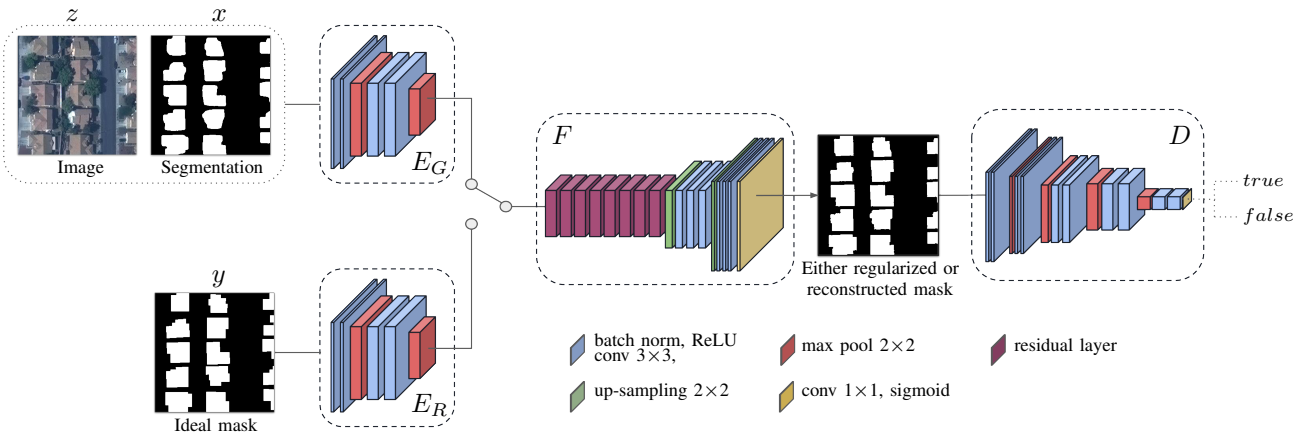

Fig. 3: Workflow of the proposed regularization framework. It is composed of two paths: the generator path ($E_G \rightarrow F$) produces the regularized building footprint mask; the reconstruction path ($E_R \rightarrow F$) encodes and decodes the ideal input mask ensuring to have the same real valued masks as input to the discriminator.

Alongside the adversarial and regularized losses, a soft version of the Potts and Normalized Cut criterions are used to exploit the information of the intensity image to further improve the regularization results. The Potts and the Normalized Cut methods are popular graph clustering algorithms originally proposed for image segmentation. As demonstrated in [7], these terms can be effectively minimized by the generator G. As a result, the final footprints are aligned to the building boundaries observed in the intensity image.

The Potts and the normalized cut losses can be expressed as:

$$\begin{aligned} \mathcal{L}_{Potts}(G) &= E_{x,z} \sum_k S^{k\top} W (1 - S^k) \\ \mathcal{L}_{ncut}(G) &= E_{x,z} \sum_k \frac{S^{k\top} \hat{W} (1 - S^k)}{1^\top \hat{W} S^k} \end{aligned} \tag{4}$$

where $S = G(x, z)$ is the k-way softmax mask generated by the network and S^k describes the vectorization of its k-th channel. W and $\hat{W}$ are matrices of pairwise discontinuity costs and each term describes the weight between two nodes (or pixels) and it is computed using a gaussian kernel over the RGBXY space.

The full objective used to jointly train the generator path G and the reconstruction path R is a linear combination between the adversarial loss, the regularized loss and the reconstruction losses.

$$\begin{aligned} \mathcal{L}(G, R, D) = \alpha \mathcal{L}_{GAN}(G, R, D) \\ + \beta \mathcal{L}_{rec_G}(G) + \gamma \mathcal{L}_{rec_R}(R) \\ + \delta \mathcal{L}_{Potts}(G) + \epsilon \mathcal{L}_{ncut}(G) \end{aligned} \tag{5}$$

It's worth noting that these loss components are obtained by connecting the encoders E_R and E_G to the residual decoder F one at a time. Once the full objective is computed, E_G, E_R and F are updated jointly.

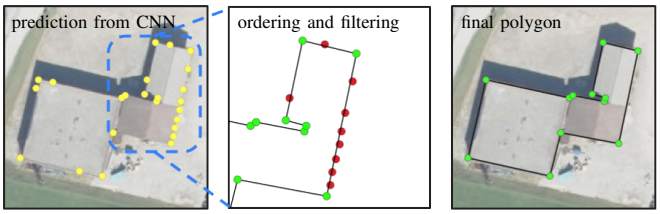

Fig. 4: Polygon extraction steps: given the regularized building footprint, a CNN model detects all the building corners candidates (yellow vertices). The vertices are then sorted to produce a valid set of polygon coordinates. Points which lie too close to a building edge are filtered (in red). The final set of coordinates which describes the polygon is highlighted in green.

C. Polygon extraction

Once the building footprints have been regularized, we extract a polygon for each building instance.

This task is accomplished using a simple CNN for corner detection. The model receives the regularized mask as input and produces a corner proposal probability map. Pixels with a value higher than a certain threshold in the probability map can be considered valid corners for the building polygon.

During inference each regularized footprint is evaluated by the corner detection network independently. The detected points are then ordered clockwise moving along the perimeter of the regularized footprint in order to produce a valid set of coordinates for the polygon. As a final step, we filter redundant points that lie close to an edge as shown in Fig. 4.

IV. Experiments

A. Experimental setup

1) Dataset: The proposed pipeline has been evaluated on several aerial and satellite building segmentation datasets: INRIA [1], CrowdAI [3], and SpaceNet [2].

The INRIA dataset is an aerial dataset which covers a wide range of urban settlement appearances from different geographic locations. The particularity of this dataset is that the cities included in the test set are different from those of the training set, and it is composed of 180 training and 180 testing 5000×5000 orthorectified images with a resolution of 30 cm. The CrowdAI dataset consists of 280,000 satellite images for training and 60,000 images for testing with an image resolution of 300×300 pixels. During the test set inference over 500,000 building instances are extracted and regularized. The SpaceNet dataset is composed of 30-50 cm pan-sharpened RGB satellite images from two cities in Florida: Jacksonville and Tampa. The dataset is split into 62 images for the test set and 174 images for the training set. The provided images have 2048×2048 pixels size.

All these datasets have a wide variety of buildings with different sizes, shapes and complexities that make the extraction of regularized polygons challenging.

2) Network Architecture: The **regularization network** has a residual autoencorder structure as shown in Fig. 3. The encoders E_G and E_R are a sequence of 3×3 convolutional layers followed by batch normalization [33] and 2×2 max-pooling layers. After every down-sampling operation the number of convolutional filters is doubled, while the tensor size is halved. The decoder F is composed by a chain of 8 residual layers [34] followed by 3×3 convolutions, batch normalization layers and 2×2 up-sampling operations. Compared to the architecture proposed in [7], our encoders only have two pooling layers in order to keep trace of fine details of the input mask. As shown in Section IV, this choice allows the decoder F to reconstruct with more accuracy the buildings received as input and at the same time it can regularize them effectively, regardless their shape and complexity. The discriminator D shares the same layer combination of the encoders E_G and E_R but it has a deeper architecture, with 4 max-pooling operations in total.

For the **corner detection network** we just simply exploit the architectural model of the network G used for the building regularization but using only 4 residual layers.

3) Training Details: Unlike the training approach proposed in [7] where building instances are scaled and forward-propagated through the regularization network one by one, we train our GAN using 256×256 patches cropped from the dataset samples. This helps to learn a generator and discriminator aware of the shape differences between small, medium and big buildings. As ideal masks we exploit the accurate and good looking building footprints present in the ground truth of the chosen datasets. The model is trained with batch size of 4 for 140,000 iterations. We set $\alpha = 3$, $\beta = 1$, $\gamma = 3$ in Eq. (5). ϵ and δ are kept to 0 for the first 40,000 batches, then they are linearly increased to 1 and 175, respectively, in the following 40,000 batches to keep the learning more stable. The weight matrix W and $\hat{W}$ for *Potts loss* and *normalized cut loss* in the Eq. (4) are computed using the same expression and hyper-parameters described in [7].

Since the datasets we use for evaluation provide the ground truth already rasterized, the CNN used to detect building corners is trained using the building polygons available in OpenStreetMap for the cities of Chicago and Jacksonville.

Fig. 5: On the left side: satellite image with occluded constructions. On the right side: result of the regularization network. Extracted footprints with wrong pose are highlighted in red.

For the initial building segmentation we used R2U-Net trained with 448×448 patches randomly cropped from the SpaceNet and INRIA image samples. In CrowdAI we directly train the model using the 300×300 images provided in the dataset. Also, we provide some results using Mask R-CNN [35] as baseline using the pre-trained weights available in [3].

During the training of all the networks, we applied standard data augmentation to the images (random rotations and flipping) and we trained all the pipeline models using Adam [36] optimizer with learning rate set to 0.0001.

B. Results

In the **INRIA** and the **SpaceNet** datasets we compare against the baseline method and Zorzi *et al.* [7]. The baseline exploits R2U-Net as backbone to perform the initial building segmentation. The results are then processed by the regularization method described in [7] and by our building extraction method to produce the final footprints. The final scores, based on IoU and accuracy, are shown in Table I and Table II. Our building refinement can achieve quantitative results comparable or, in some test areas, even higher than the pure baseline. Our approach, in fact, gets the higher IoU values in the test areas of Bellingham, Bloomington and Tyrol from the INRIA dataset and demonstrates to achieve accuracies very close to the pure baseline solution in the SpaceNet dataset. This is a sign that the pipeline, made by multiple modules connected in cascade, does not lead to a significant drop in performance. It is worth noting that the method Zorzi *et al.* [7] has a significant IoU drop in these two datasets. This is caused by the network architecture which is not capable to generalize well for big and complex buildings as shown in the results in Fig. 6.

In **CrowdAI** we test both R2U-Net and Mask R-CNN as baseline networks for the initial segmentation. Again, the proposed regularization can achieve results close to the pure segmentation network. The IoU and accuracy scores achieved by Zorzi *et al.* [7] are explainable considering that the CrowdAI dataset is mainly composed of midsize and small size constructions, with a low number of corners.

	INRIA											
	Bellingham		Bloomington		Innsbruck		San Francisco		Tyrol		Overall	
	IoU	Acc	IoU	Acc	IoU	Acc	IoU	Acc	IoU	Acc	IoU	Acc
R2UNet	70.30	**97.04**	72.94	**97.40**	**73.48**	**96.85**	**76.29**	**91.85**	75.92	**97.84**	**74.57**	**96.20**
Zorzi *et al.* [7]	63.90	96.37	63.65	96.51	60.20	95.23	55.97	84.60	65.56	96.88	59.81	93.92
Ours	**70.36**	96.99	**73.01**	97.36	73.34	96.77	75.88	91.55	**76.15**	**97.84**	74.40	96.10

TABLE I: Quantitative evaluation of building extraction and regularization results on the INRIA dataset. Scores are obtained by submissions of the predictions to https://project.inria.fr/aerialimagelabeling/.

	SpaceNet											
	Jacksonville				Tampa				Overall			
	IoU		Acc		IoU		Acc		IoU		Acc	
	μ	σ	μ	σ	μ	σ	μ	σ	μ	σ	μ	σ
R2UNet	**72.85**	7.077	**96.54**	1.105	**70.74**	6.056	**94.90**	1.219	**71.80**	6.670	**95.75**	1.406
Zorzi *et al.* [7]	59.17	5.348	94.73	1.693	57.99	6.892	92.58	2.317	58.58	6.197	93.65	2.296
Ours	70.90	7.551	96.29	1.169	69.04	6.587	**94.90**	1.286	69.97	7.146	95.50	1.463

TABLE II: Quantitative evaluation of building extraction and regularization results on the SpaceNet dataset

Dataset		CrowdAI			
Method		IoU		Acc	
Baseline	**Regularization**	μ	σ	μ	σ
R2U-Net	-	**80.44**	16.10	**95.86**	5.20
R2U-Net	Zorzi *et al.* [7]	76.95	15.34	94.75	5.473
R2U-Net	Ours	79.87	15.93	95.57	5.281
Mask R-CNN	-	73.22	17.84	**94.38**	4.778
Mask R-CNN	Zorzi *et al.* [7]	71.72	17.32	93.88	4.822
Mask R-CNN	Ours	**73.57**	17.65	94.34	4.749

TABLE III: Quantitative evaluation of building extraction and regularization results on the CrowdAI dataset

1) Qualitative results: We visualize some building footprints generated with different approaches in Fig. 6. Building footprints extracted with [7] are accurate and visually pleasing if the building has a low number of vertices. Vice versa, if the construction is complex, the network fails on producing a decent building boundary.

The algorithm proposed in this paper overcomes this problem producing accurate and realistic footprints regardless of the building size and complexity. It is worth noting that our polygon extraction algorithm can also deal with inner courtyards creating a polygon for each building perimeter, as shown in the second row of Fig. 6.

Despite the good results obtained in most of the circumstances, the proposed method is still not capable to extract sufficient context information to perform a correct regularization in the presence of occlusions. In Fig. 5 is shown a residential area evaluated by Mask2Poly. The presence of the road in front of the constructions arranged in a line would suggest that the occluded buildings are also facing the street, in opposition with the extracted footprints. Embedding a constraint about the disposition and the orientation of all the constructions in the scene would help the regularization network producing a coherent cartographic map of the the buildings from satellite or aerial images.

V. Conclusion

In this paper, we presented an approach for building segmentation and regularized polygon extraction, composed of three different and independent neural network modules.

The combination of the adversarial and the regularized losses results in a effective geometry constrain for the constructions, and encourages our predicted footprints to match building boundaries. Furthermore, the regularization allows us to extract precise building polygons using a simple but effective *fully convolutional network (FCN)* for corners detection.

The proposed method has proved to be capable not only of achieving equivalent or even higher results in terms of IoU and accuracy compared to state-of-the-art segmentation networks, but also of generating realistic and visually pleasing construction outlines that can be used in many cartographic and engineering applications.

References

[1] E. Maggiori, Y. Tarabalka, G. Charpiat, and P. Alliez, "Can semantic labeling methods generalize to any city? the inria aerial image labeling benchmark," in *IEEE International Geoscience and Remote Sensing Symposium (IGARSS)*, IEEE, 2017.

[2] H. Goldberg, M. Brown, and S. Wang, "A benchmark for building footprint classification using orthorectified rgb imagery and digital surface models from commercial satellites," in *Proceedings of IEEE Applied Imagery Pattern Recognition Workshop 2017*, 2017.

[3] S. P. Mohanty, *Crowdai mapping challenge 2018 dataset*, https://www.crowdai.org/challenges/mapping-challenge, 2019 (accessed November 10, 2019).

(a) Segmentation (b) Zorzi *et al.* [7] (c) Proposed regularization (d) Extracted polygons

Fig. 6: Buildings extraction results overlaid on top of a sample areas from Inria, CrowdAI and SpaceNet datasets.

[4] P. Tokarczyk, J. D. Wegner, S. Walk, and K. Schindler, "Beyond hand-crafted features in remote sensing," *ISPRS Annals of Photogrammetry, Remote Sensing and Spatial Information Sciences*, vol. 1, no. 1, pp. 35–40, 2013.

[5] J. Yuan, "Automatic building extraction in aerial scenes using convolutional networks," *ArXiv preprint arXiv:1602.06564*, 2016.

[6] K. Bittner, F. Adam, S. Cui, M. Körner, and P. Reinartz, "Building footprint extraction from vhr remote sensing images combined with normalized dsms using fused fully convolutional networks," *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing*, vol. 11, no. 8, pp. 2615–2629, 2018.

[7] S. Zorzi and F. Fraundorfer, "Regularization of building boundaries in satellite images using adversarial and regularized losses," *IEEE International Geoscience and Remote Sensing Symposium (IGARSS)*, 2019.

[8] L. Castrejon, K. Kundu, R. Urtasun, and S. Fidler, "Annotating object instances with a polygon-rnn," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2017, pp. 5230–5238.

[9] D. Acuna, H. Ling, A. Kar, and S. Fidler, "Efficient interactive annotation of segmentation datasets with polygon-rnn++," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2018, pp. 859–868.

[10] N. Girard and Y. Tarabalka, "End-to-end learning of polygons for remote sensing image classification," in *IGARSS 2018-2018 IEEE International Geoscience and Remote Sensing Symposium*, IEEE, 2018, pp. 2083–2086.

[11] Z. Li, J. D. Wegner, and A. Lucchi, "Topological map extraction from overhead images," in *Proceedings of the IEEE International Conference on Computer Vision*, 2019, pp. 1715–1724.

[12] A. Huertas and R. Nevatia, "Detecting buildings in aerial images," *Computer Vision, Graphics, and Image Processing*, vol. 41, no. 2, pp. 131–152, 1988.

[13] R. Guercke and M. Sester, "Building footprint simplification based on hough transform and least squares adjustment," in *Proceedings of the 14th Workshop of the ICA Commission on Generalisation and Multiple Representation, Paris, France*, vol. 30, 2011.

[14] Z. Kim and R. Nevatia, "Uncertain reasoning and learning for feature grouping," *Computer Vision and Image Understanding*, vol. 76, no. 3, pp. 278–288, 1999.

[15] M. Brédif, O. Tournaire, B. Vallet, and N. Champion, "Extracting polygonal building footprints from digital surface models: A fully-automatic global optimization framework," *ISPRS journal of photogrammetry and remote sensing*, vol. 77, pp. 57–65, 2013.

[16] X. Huang and L. Zhang, "Morphological building/shadow index for building extraction from high-resolution imagery over urban areas," *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing*, vol. 5, no. 1, pp. 161–172, 2012.

[17] H. Baluyan, B. Joshi, A. Al Hinai, and W. L. Woon, "Novel approach for rooftop detection using support vector machine," *ISRN Machine Vision*, vol. 2013, 2013.

[18] R. Hamaguchi and S. Hikosaka, "Building detection from satellite imagery using ensemble of size-specific detectors," in *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*, IEEE, 2018, pp. 223–2234.

[19] A. Lagrange, B. Le Saux, A. Beaupere, A. Boulch, A. Chan-Hon-Tong, S. Herbin, H. Randrianarivo, and M. Ferecatu, "Benchmarking classification of earth-observation data: From learning explicit features to convolutional networks," in *2015 IEEE International Geoscience and Remote Sensing Symposium (IGARSS)*, IEEE, 2015, pp. 4173–4176.

[20] N. Audebert, B. Le Saux, and S. Lefèvre, "Joint learning from earth observation and openstreetmap data to get faster better semantic maps," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops*, 2017, pp. 67–75.

[21] V. Iglovikov, S. S. Seferbekov, A. Buslaev, and A. Shvets, "Ternausnetv2: Fully convolutional network for instance segmentation.," in *CVPR Workshops*, 2018, pp. 233–237.

[22] M. Z. Alom, M. Hasan, C. Yakopcic, T. M. Taha, and V. K. Asari, "Recurrent residual convolutional neural network based on u-net (r2u-net) for medical image segmentation," *ArXiv preprint arXiv:1802.06955*, 2018.

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

[24] D. Marcos, D. Tuia, B. Kellenberger, L. Zhang, M. Bai, R. Liao, and R. Urtasun, "Learning deep structured active contours end-to-end," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2018, pp. 8877–8885.

[25] M. Kass, A. Witkin, and D. Terzopoulos, "Snakes: Active contour models," *International journal of computer vision*, vol. 1, no. 4, pp. 321–331, 1988.

[26] D. Cheng, R. Liao, S. Fidler, and R. Urtasun, "Darnet: Deep active ray network for building segmentation," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2019, pp. 7431–7439.

[27] M. Tang, F. Perazzi, A. Djelouah, I. Ben Ayed, C. Schroers, and Y. Boykov, "On regularized losses for weakly-supervised cnn segmentation," in *Proceedings of the European Conference on Computer Vision (ECCV)*, 2018, pp. 507–522.

[28] M. Tang, A. Djelouah, F. Perazzi, Y. Boykov, and C. Schroers, "Normalized cut loss for weakly-supervised cnn segmentation," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2018, pp. 1818–1827.

[29] D. H. Douglas and T. K. Peucker, "Algorithms for the reduction of the number of points required to represent a digitized line or its caricature," *Cartographica: The international journal for geographic information and geovisualization*, vol. 10, no. 2, pp. 112–122, 1973.

[30] M. A. Fischler and R. C. Bolles, "Random sample consensus: A paradigm for model fitting with applications to image analysis and automated cartography," *Communications of the ACM*, vol. 24, no. 6, pp. 381–395, 1981.

[31] R. O. Duda and P. E. Hart, "Use of the hough transformation to detect lines and curves in pictures," Sri International Menlo Park Ca Artificial Intelligence Center, Tech. Rep., 1971.

[32] I. Goodfellow, J. Pouget-Abadie, M. Mirza, B. Xu, D. Warde-Farley, S. Ozair, A. Courville, and Y. Bengio, "Generative adversarial nets," in *Advances in neural information processing systems*, 2014, pp. 2672–2680.

[33] S. Ioffe and C. Szegedy, "Batch normalization: Accelerating deep network training by reducing internal covariate shift," *ArXiv preprint arXiv:1502.03167*, 2015.

[34] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2016, pp. 770–778.

[35] K. He, G. Gkioxari, P. Dollár, and R. Girshick, "Mask r-cnn," in *Proceedings of the IEEE international conference on computer vision*, 2017, pp. 2961–2969.

[36] D. P. Kingma and J. Ba, "Adam: A method for stochastic optimization," *ArXiv preprint arXiv:1412.6980*, 2014.