Noname manuscript No.
(will be inserted by the editor)

A Novel Adaptive Deep Network for Building Footprint Segmentation

A. Ziaee · R. Dehbozorgi · M. Döller

Received: date / Accepted: date

Abstract Building footprint segmentations for high resolution images are increasingly demanded for many remote sensing applications. By the emerging deep learning approaches, segmentation networks have made significant advances in the semantic segmentation of objects. However, these advances and the increased access to satellite images require the generation of accurate object boundaries in satellite images. In the current paper, we propose a novel network based on Pix2Pix methodology to solve the problem of inaccurate boundaries obtained by converting satellite images into maps using segmentation networks in order to segment building footprints. To define the new network named G2G, our framework includes two generators where the first generator extracts localization features in order to merge them with the boundary features extracting from the second generator to segment all detailed building edges. Moreover, different strategies are implemented to enhance the quality of the proposed networks' results, implying that the proposed network outperforms state-of-the-art networks in segmentation accuracy with a large margin for all evaluation metrics. The implementation is available at https://github.com/A2Amir/A-Novel-Adaptive-Deep-Network-for-Building-Footprint-Segmentation.

Keywords Deep Learning Semantic Segmentation · Building Footprint Segmentation · Conditional Generative Adversarial Networks(CGANs) · Pix2Pix Network.

1 Introduction

Accurate segmentation is an important issue in the field of computer vision. To have an exact segmentation, every pixel of an image should be classified into mul-

Amir Ziaee
The University of Applied Science FH Kufstein E-mail: amir.ziaee@fh-kufstein.ac.at

Raziyeh Dehbozorgi
Iran University of Science and Technology E-mail: r.dehbozorgi2012@gmail.com

Mario Döller
The University of Applied Science FH Kufstein E-mail: Mario.Doeller@fh-kufstein.ac.at

tiple segments. Although, regardless of knowing objects, human visual system can segment all unknown objects of an image such as a satellite image. Computers get into a challenge to conduct such a task. For instance, extracting buildings automatically from an urban image is complex, since roof textures are different in terms of shape and size. Also, the contrast between buildings and its environments is very low [6]. Therefore, efficient extraction of building footprints remains a challenge.

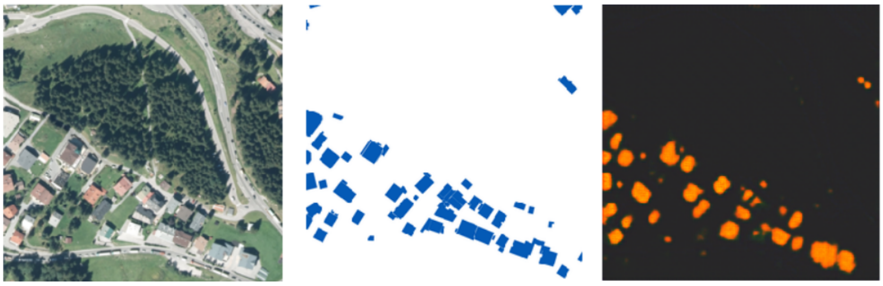
Building detection and footprint extraction with all salient features are of high interest for many applications such as urban planning, building monitoring, sociology, and disaster emergency response. Regarding the variety of the building materials and scales, buildings in aerial/satellite imagery are depicted significantly different. In spite of the abundant research to detect and extract building footprints, a few algorithms have been investigated to deal with the problem of inaccurate edges [3]. Therefore, the quality of results such as precision and resolution of boundaries remain as valid concerns.

Semantic segmentation, representing as one of the granularity levels within the segmentation process, classifies each pixel into a specific class. This level of segmentation is applied in many sorts of analysis such as satellite imagery analysis, recognition of image copies, and human-computer interaction [7, 18, 20]. However, the challenge remains on the correct detection and classification of the individual objects for producing predictions with accurate boundaries [3, 7].

Fully Convolutional Networks (FCNs) was firstly introduced for semantic segmentation [11]. Although, this has been applied to segment satellite images, the limitation of low-resolution predictions leads to further research to obtain better models. Therefore, a number of techniques have been proposed to address this limitation aiming at generating high-resolution and accurate boundaries. In this regard, the first U-Net was built for the segmentation of biomedical images [15], which was extended for many other applications in the field of segmentation, in order to overcome the aforementioned issues [8, 9].

Although, deep learning networks play the most prominent role in the area of semantic segmentation, when analyzing satellite images, they have shown some drawbacks in producing boundaries (Fig. 1). To conquer the problem of inaccurate prediction of boundaries in satellite images, this study proposes a deeper network architecture using two generators based on Pix2Pix network [8], being developed based on the CGAN [12]. Accordingly, current study covers the followings:

- Focusing on Pix2Pix network to perform detailed experiment design for segmentation of buildings. A proof of the concept, implying that using two generators are more powerful than one.
- Representing a new network with a new architecture to improve the accuracy of semantic segmentation networks, which enables us to demonstrate an increased accuracy of approximately 10 percent in the segmentation of building footprints in satellite imagery processing.
- Implementing four-evaluation metrics to prove that our approach significantly exceeds Pix2Pix network and the current state-of-the-art networks in term of accuracy at specified validation without any hyper-parameter optimization.



(a) Satellite Image (b) Ground Truth (c) Prediction

Fig. 1 Prediction with inaccurate boundaries.

Related works

Extracting the momentous features of building footprint are of great importance for creation of an appropriate map with a high local accuracy. Edges are among those features, being rarely investigated due to the various buildings' shapes and scales. Despite the efforts that have been put into developing methodologies based on deep neural networks, to provide an automatic extraction of the building footprints, they are not still producing satisfactory results.

Some efforts have been made to solve the problem of preserving semantic segmentation boundaries. Bittner et al. [5] introduced an approach to automatically generate a full resolution binary building mask without any assumptions on the scale and shape of the buildings using a Digital Surface Model (DSM) and an FCN architecture. Their solution includes two main steps: i) Training FCN on a large set of patches consisting of normalized DSM (nDSM) as inputs and available ground truth building mask as target outputs. ii) Considering the generated predictions from FCN as unary terms for a fully connected Conditional Random Field (FCRF), which enables them to achieve a final binary building mask. This work was later improved by demonstrating an end-to-end fused-FCN in which the fusion of several networks including three-band (RGB), panchromatic (PAN), and nDSM results in high resolution images [4]. Additionally, an algorithm based on the combination of the robust Classification-convolution Neural Networks (CNN) with an Active Contour Model (ACM) have been introduced to improve the accuracy of current building edge extraction [17]. ACM can be accounted as a useful tool to elevate the accuracy, in case, the footprints of a building are missed in the CNN classification.

To increase the accuracy of semantic segmentation of high-resolution images, even more, Xiaoye Wang et al. [22] also improved Pix2Pix network by adding a controller to have a new Pix2Pix model called ePix2Pix to progress the classification performance and create segmentations that more efficiently match with the ground truths in terms of shape. Another similar work was done by Schuegrafwe and Bittner [16] to develop a deep learning-based algorithm for DSMs and spectral images (PAN and multi-spectral) fusion. They utilized an end-to-end U-Net to combine depth and spectral information within two parallel networks. This led to a combination of the features in the late phase to obtain binary building masks using a

residual block of the neural network.
Zhu et al. [23] developed a new Multi Attending Path Neural Network (MAP-Net) for accurate extraction of building footprints and precise boundaries on multiple levels. They extracted boundary and semantic information through two different networks and merged this information in later stages. Bischke et al. [3] achieved an advancement in this area by proposing an uncertainty weighted and cascaded multi-task loss based on distance transform accompanied, by a deeper network architecture to improve semantic segmentation predictions.
Based on the previous studies, this work design a pre-processing algorithm and a strategy using two different networks and modify Pix2Pix architecture. The new resulting Pix2Pix network termed as G2G network allows us to achieve an improvement of the semantic segmentation in respect to accuracy.

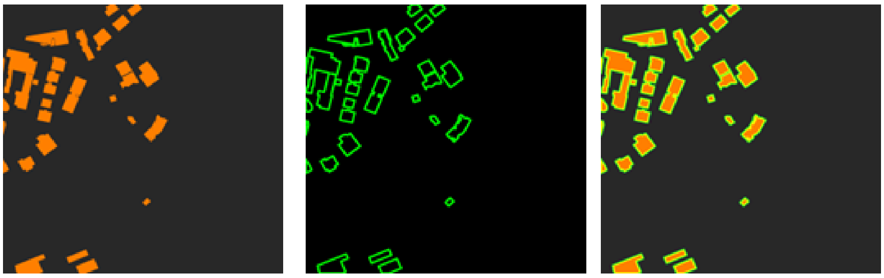
2 Methodology: G2G

The main objective of our approach is to improve the accuracy of segmentation and generate accurate boundaries of building footprints based on deep learning and Pix2Pix networks. The proposed network benefits from the architecture of Pix2Pix network having two generators with almost the same structures, but two different discriminators to improve the accuracy of segmented objects in boundaries. As shown in Fig. 1, majority of the classical algorithms are not able to predict the boundaries, well. Thus, this paper tries to provide a solution by introducing G2G network. To predict the boundaries precisely, building footprint segmentation should consider the following distinct goals;

(i) Localization Information.
(ii) Similarity of the shape and boundaries of a segmented building to ground truth

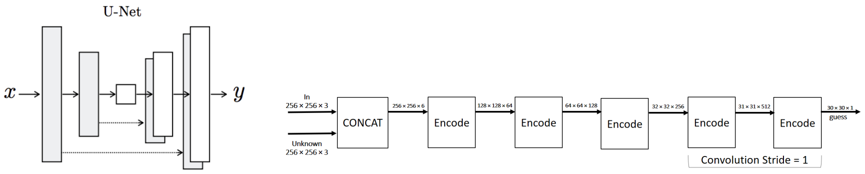
As mentioned before, current networks are able to segment where the buildings are, but they are not able to segment the shapes and edges of segmented buildings. Therefore, boundaries and shapes of the buildings should be taken into consideration, with special focus on how boundaries are drawn. If we assume that the most crucial information of an object are the object boundaries, adding the extracted contours of each object in the ground truth to the corresponding object in the ground truth would certainly improve the shape and boundaries of a segmented building for further research. As seen in Fig. 2, contours were added again to objects of the ground truths after extracting them using the Open CV library.
To go further, a Conditional Generative Adversarial Network (CGAN) called Pix2Pix network was chosen. The network is applicable for image to image translation aims including photos synthesis from label maps, objects reconstruction from edge maps, and images colorization [8, 21].
Pix2Pix network is equipped by two parts (Fig. 3), Generator (G) and Discriminator (D). According to the aim, the generator converts satellite images into building maps, and the discriminator distinguishes real images from fake ones. Although, the Pix2Pix network is generalized for many different tasks without modifying the loss function, there are many issues regarding a successful training of Pix2Pix. Therefore, current researches focus on improving the training of Pix2Pix.

In a recent effort, where Pix2Pix was used to synthesize an image from semantic



(a) Ground truth (b) Boundaries (c) Combination

Fig. 2 Extracting and re-adding the contours to objects of the ground truths.



(a) The generator (b) The Patchgan discriminator

Fig. 3 Pix2Pix structures of the generator and Patchgan discriminator [15].

labels [21], blurry images were produced. The study claims that the adversarial training of Pix2Pix network might be prone to failure and blurry images for high-resolution image generation task, which was attributed to unstable training. Thus, they attempted to improve Pix2Pix using two generators (Fig. 4) and multi-scale discriminators.

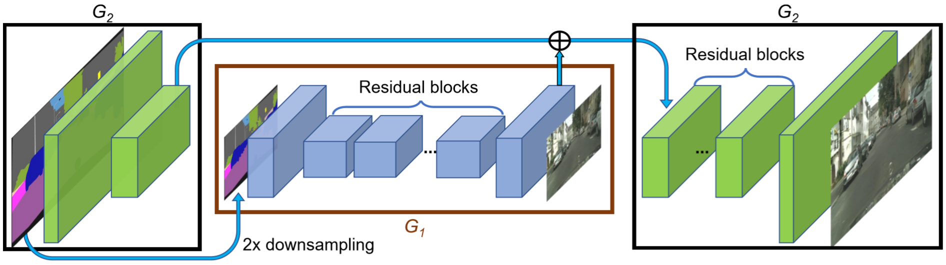


Fig. 4 Architecture of the generators [21].

Another study utilized CGAN in the field of inferring contours from images where normal GAN was used as opposed to the discriminatory structure of Pix2Pix [10]. The study was done based on the idea that although the Pix2Pix discriminator can help other networks to generate satisfactory textures, in this case, it results in many break edges for a single contour of the object.

2.1 Architecture of G2G network

Considering the two major findings [10, 21] explained above, we concluded that the architecture of the G2G network should fulfill the following properties;

1. The structure of the Pix2Pix discriminator should be extended in order to distinguish contours as an indicator to discriminate fake and real images. The similarity of the shapes and edges of a segmented building to the ground truth can be specified with the relevant contours which play remarkably well their role as indicators.

2. Two generators should be used, each having their own specified goal to focus individually on providing accurate localization and boundary features, respectively.

In addition, three types of images (Fig. 5) should be defined to design the new architecture each following their own goals:

Image type 1: Satellite images
Image type 2: Corresponding ground truths of satellite images
Image type 3: Extracted contours overlaid on the corresponding ground truths

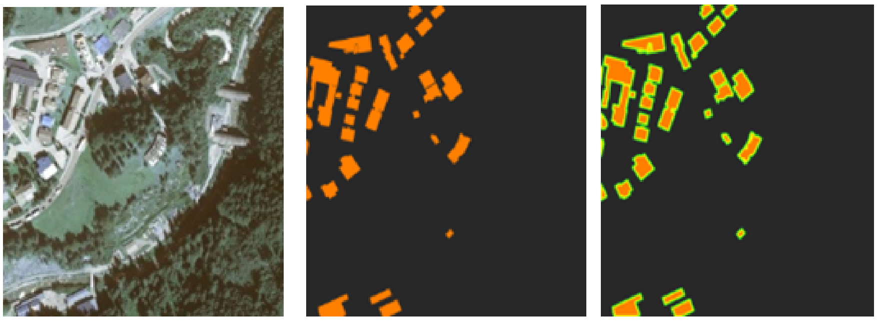

(a) Image type 1 (b) Image type 2 (c) Image type 3

Fig. 5 Three types of images used in G2G network.

Accordingly to the aforementioned properties of G2G network architecture and three available types of images, the architecture of G2G network can be illustrated as follow (Fig 6).

2.1.1 Architecture of the G2G generators

In this study, the global generator was decomposed into two sub-generators: G1, which uses Image 1 as an input and Image 2 as a target, with an objective to detect where the buildings are located, accomplishing the first goal. G2 that utilizes the combination of Image 1 and the first generator's output as an input, and

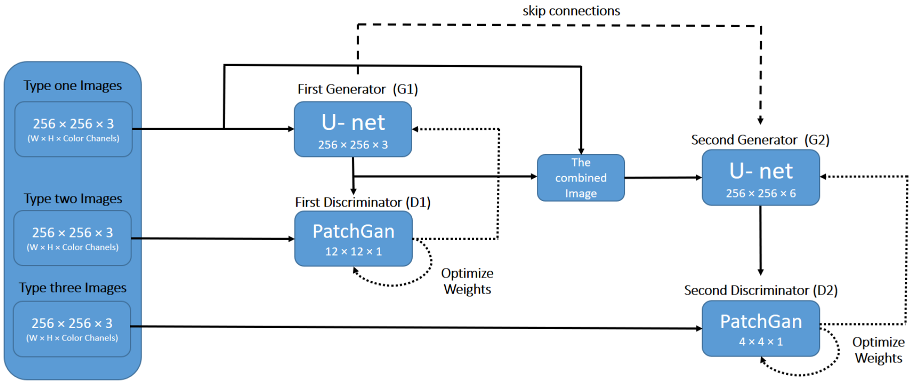


Fig. 6 Architecture of G2G network.

Image 3 as a target. This sub-generator performs as a post-processing algorithm to correct the result of the first generator and at the meantime takes over the edges and contours of the objects, attaining the second goal. An individual discriminator was developed for each generator to focus more efficiently on their main allocated goal. The architecture of the G2G generators is depicted in Fig. 7.

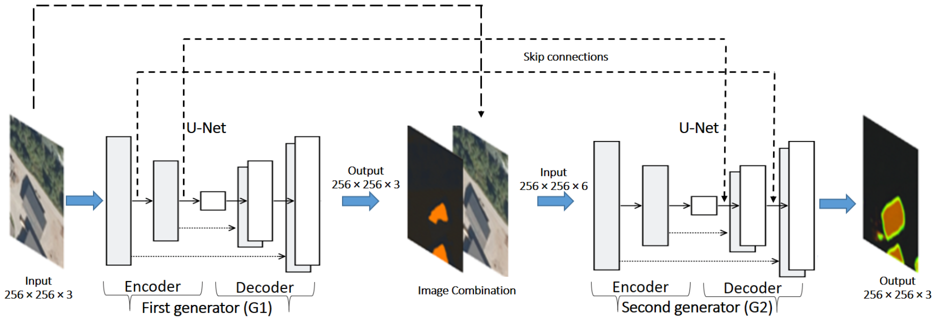


Fig. 7 Architecture of the G2G generators.

The generators were built based on the architecture of the U-Net structure, as one of the most common and effective structures in the field of segmentation. It consists of two components: one encoder and one decoder with skip connections. In contrast to the first generator of G2G network, which has the input size of $256 \times 256 \times 3$ (height, width, channels), the second generator (Fig. 8) has the input size of $256 \times 256 \times 6$, representing a combination of the image generated by the first generator and the corresponding satellite image. The first generator was connected to the second generator by adding the skip connections from the initial layers of the first generator to the end layers of the second generator.

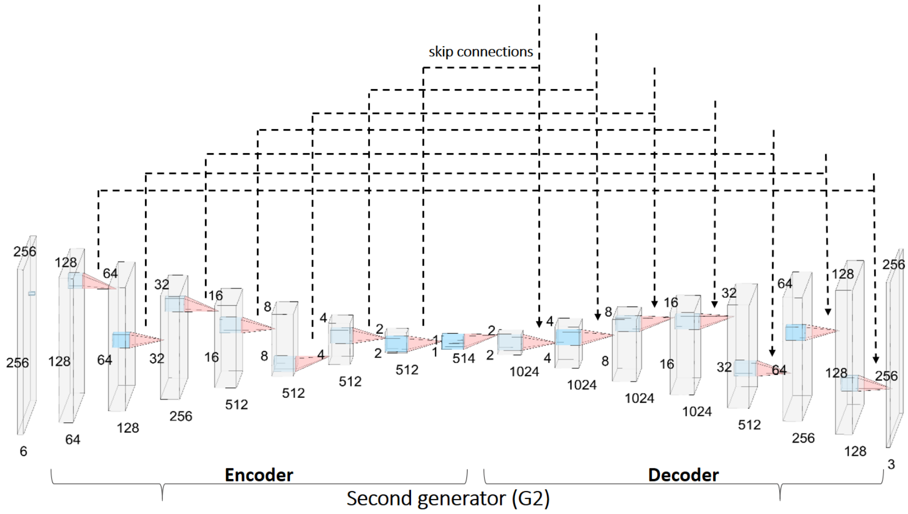
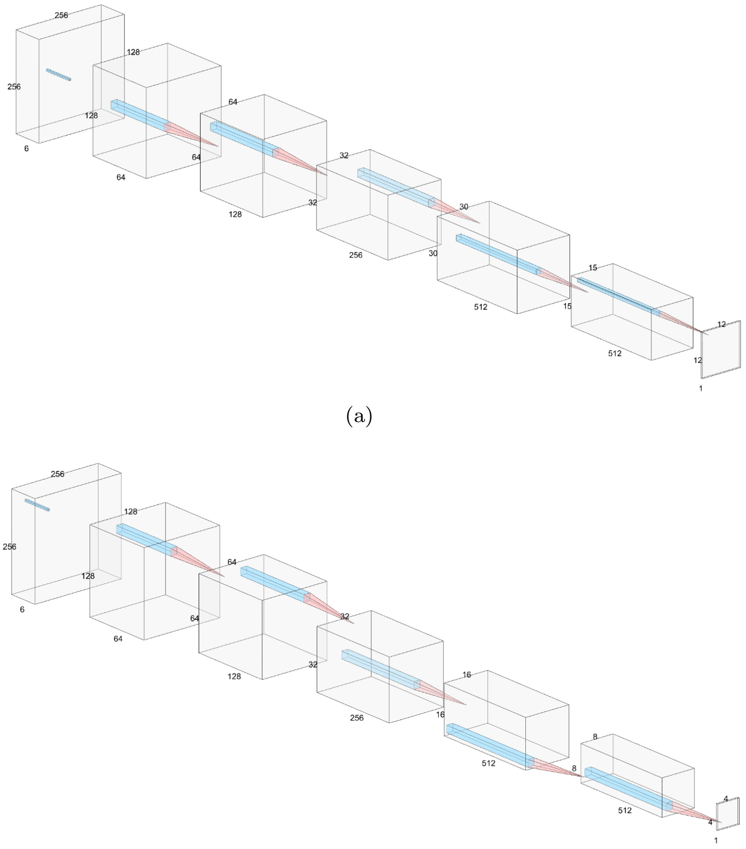


Fig. 8 Structure of the second generator.

2.1.2 Architecture of the G2G discriminators

Two different discriminators with two distinct network structures were designed to operate on identical image scales which are denoted as D1 and D2 for the first and second discriminators, respectively. Due to using Zero Padding layer in the structure of the original Pix2Pix discriminator, a lot of information is lost and the structure of the two proposed discriminators is built without any Zero Padding layer Fig. 9(a). The first layer of D1 is a concatenated layer of inputs followed by the three convolutional layers consisting of convolutional layer, batch normalization, and leaky rectified linear unit (leaky ReLU) as an activation function which has a small slope for negative values, instead of zero values.

The output of the third convolutional layer passes through the fourth layer, which is a series of convolutional and batch normalized layers without leaking ReLU. Then, the output of this layer was fed to the next one, including max pooling and convolutional layers. The output size of the final layer is 12 × 12 × 1, which corresponds to the 28 × 28 patch of the input image, in contrast to the original Pix2Pix discriminator where each pixel of the output discriminator (30 × 30) corresponds to the 70 × 70 patch of the input image. As shown in Fig. 9(b), in the second discriminator structure, the first six layers have the same structure as the first four layers of D1 but the output of the fifth layer is reduced by half and followed by that, it is fed to the sixth layer. To decrease variance and complexity of computation, a max pooling layer was added at the end of the D2 structure in G2G network, to extract low-level and important features including edges from the neighborhood.



(a)

(b)

Fig. 9 Structures of the two discriminators, (a) Structure of the first discriminator (D1), (b) Structure of the second discriminator (D2).

The output size of the final layer is $4 \times 4 \times 1$, which corresponds to the 9×9 patch of the input image. Unlike the first discriminator, where each pixel of the output (12×12 image) is relevant to the 28×28 patch of the input image, in the second discriminator, each pixel of the output corresponds to the 9×9 patch of the input image resulting in more accurate object segmentation.

2.2 Training phase of G2G network

To adjust the weights of G2G, two steps are performed (Fig. 10). In the first step, the first discriminator takes the input (Image 1)/ target (Image 2) and then input (Image 1)/ output (the generated output of the first generator) pairs and makes its estimate on how realistic they look. Next, based on the differences, the weights of the first discriminator will be adjusted. At the end, the combination of the first generator's output with the corresponding satellite image (Image 1) passes through the second generator as input. Thereafter, the output of the second generator will be calculated and inserted into the second discriminator, which compares the input (the combined image)/ target (Image 3) then, the input (the combined image)/output (the generated output of the second generator) pairs and makes its estimate on how realistic they look in order to adjust the weights of the second discriminator.

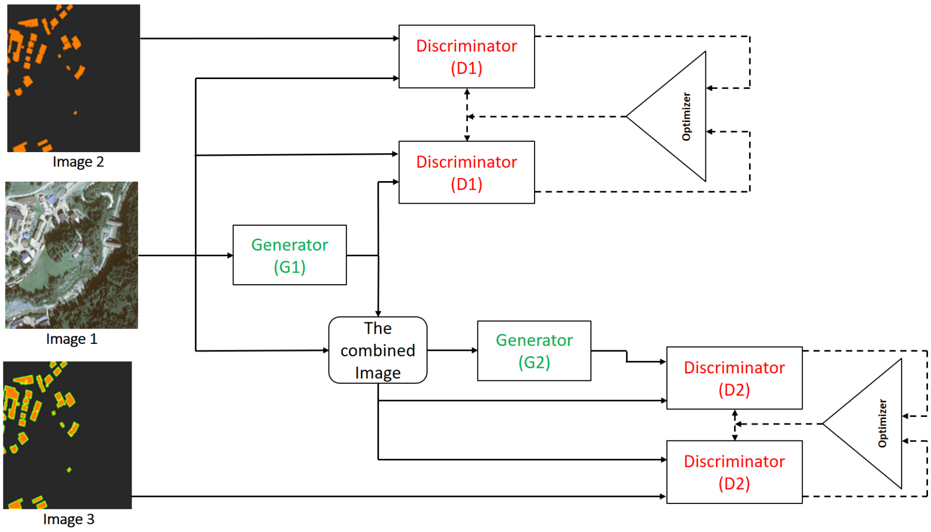


Fig. 10 Flowchart of adjusting the weights of G2G.

In the second step, the weights of the generators were adjusted according to the outputs of the discriminator and differences between the objectives and the outputs. In this study, the same objective function of Pix2Pix network (1), was employed because of its efficiency for segmentation tasks;

$$G^* = \arg\min_G \min_D \Big(\mathcal{L}_{cGAN}(G, D) + \lambda \mathcal{L}_{L1}(G)\Big),$$

where the loss function $L1$ called Manhattan distance [2] measures the standard distance between the generated output and the target.

2.3 Training strategy

Following the definition of the training phase for training G2G network, a strategy was developed to train the network, which includs the following two steps:

1. Simultaneous training of the generators and discriminators with a nearly high learning rate of $10e-3$ recommended in previous studies on Pix2Pix network [8] and 200 epochs.

2. Training the second generator and discriminator with a nearly low learning rate of $10e-6$ and 200 epochs.

3 Experimental results

This section is devoted to discuss the dataset used in the training phase of G2G network and the evaluation criteria which will be selected to score our approach and results.

3.1 Dataset

The Orthofoto Tirol dataset [1] was used to investigate the performance of the present approach. The dataset consists of two categories; satellite images and ground truths. Main characteristics in terms of Number, size, and type of the images are as presented in Table 1.

Table 1 Characteristics of the Orthofoto Tirol dataset.

	Name of the category	Number of images	Size of images	Type of images
1	Satellite Images	21,076	4053 × 4053	PNG
2	Ground truths	21,076	4053 × 4053	PNG

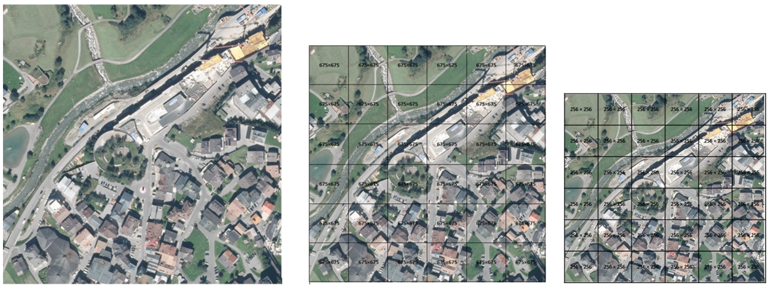
Since segmentation models are using every single pixel of images with specific input sizes, all pixels of an image are of equal importance. In case that the number of the pixels have to be reduced due to the limited size of a model input, the image will be much smaller and this results in loss of pixels. Therefore, it is important to retain more information of images by resizing them without losing the pixels. Images extracted from the Orthofoto Tirol dataset needed to be cut into 256×256, since the G2G is built based on Pix2Pix network, which only takes 256×256 images. In the first attempt, the size of the satellite and corresponding ground truth images was scaled down all at once to 256×256, which led to loss of 16.361.273 pixels. Since each pixel matters in precise detection of the edges of buildings, to avoid loosing of the main pixels, a strategy with the following steps was planned:

1. Images with high building density were selected from the entire dataset.

2. The size of each selected image (4053 × 4053 pixels) was changed to 4050 × 4050 pixels, Fig. 11(a).

3. All images from the previous step were cropped to the size of 675 × 675 pixels, Fig. 11(b).

4. All cropped images were scaled down to 256 × 256 pixels, hence not a big deal of information was lost. Number of 390.089 pixels were lost, which is far fewer than 16.361.273 pixels, Fig. 11(c).



(a) Step 2 (b) Step 3 (c) Step 4

Fig. 11 Three last steps of the planned strategy for creating training, validation and test datasets.

This strategy was performed to create datasets for training, validation, and testing purposes, where the testing dataset is used to assess the accuracy of the model and the validation dataset is needed to optimize the model during the training period, in order to minimize overfitting and fine-tuning of the hyper-parameters of the model. Properties of each created dataset are indicated in Table 2.

Table 2 Characteristics of the training, validation and test datasets derived from the Orthofoto Tirol dataset.

Dataset	Selected images	Cropped images for each image	Entire images
Training dataset	191	36	6876
Validation dataset	21	36	756
Test dataset	22	36	792

3.2 Evaluation criteria

Evaluation of the quality of a segmentation model is essential specifically for image processing in security cases such as autonomous vehicles. Although, there are many evaluation criteria developed for evaluating of segmentation models, there is no conclusive effective technique for selecting the best criteria [13, 19]. Dice similarity coefficient is one of the common criteria which determines the similarity between the results of segmentation and its corresponding ground truths in a slightly different way [24]. Considering the ground truth and prediction boxes (Fig. 12), the overlapping area between the two mentioned boxes showed in blue rectangle display where the pixels of the ground truth and prediction match, and it is called true positive (TP). In addition, the red region includes the pixels called false positives (FP).

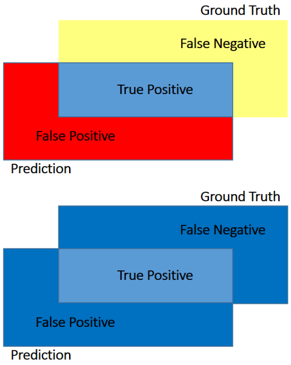


Fig. 12 An Intersection over Union example for the Dice similarity coefficient.

Yellow region in Fig. 12, possesses the pixels, which are missed from segmentation and called false negatives (FN). When the overlapping area completely covers the union area, the segmentation is flawless, the value of the Intersection over Union (IoU) criteria is 1, and the values of FP, TP, and FN are zero. Therefore, the IoU criteria formula was rewritten as follow [14]:

$$IoU = \frac{TP}{(TP + FP + FN)} \tag{1}$$

Accordingly, Dice similarity coefficient was calculated as follow [14]:

$$Dice = \frac{2 \times TP}{(TP + FP) + (TP + FN)} \tag{2}$$

In what follows, we receive hand from the following four evaluation metrics inspired by the outstanding work [11] to evaluate the efficiency of the G2G segmentation network. The metrics are extracted based on pixel accuracy and IoU metric, assuming:

n_{cl}: the number of classes,
t_i: the total number of pixels in class i,
n_{ij}: the number of pixels of class we predicted to belong to class j,
n_{ii}: the number of correctly classified pixels (true positives),
n_{ij}: the number of pixels wrongly classified (false positives),
n_{ji}: the number of pixels wrongly not classified (false negatives).

I) Pixel Accuracy (PA): The overall accuracy that is calculated as follow [11]:

$$PA = \frac{\sum_i n_{ii}}{\sum_i t_i}. \tag{3}$$

II) Mean Accuracy (MA): The average accuracy among all the classes related to ground truths. This metric is calculated as follow[11]:

$$MA = \frac{1}{n_{cl}} \sum_i \frac{n_{ii}}{t_i}. \tag{4}$$

III) Mean Intersection over Union (MIoU): Commonly used for semantic segmentation performance evaluation which is calculated based on Dice similarity as follow [11]:

$$MIoU = \frac{1}{n_{cl}} \sum_i \frac{n_{ii}}{t_i + \sum_j n_{ji} - n_{ii}}. \tag{5}$$

IV) Frequency weighted intersection over union (FWIoU): The metric considers the number of data points in each class [11] which is calculated as below:

$$FWIoU = \left(\sum_k t_k\right)^{-1} \sum_i \frac{t_i n_{ii}}{t_i + \sum_j n_{ji} - n_{ii}}. \tag{6}$$

Unlike the MIoU and FWIoU, the other two aforementioned metrics are not susceptible to unbalanced datasets [13, 19].

3.3 Evaluation results

There are several strategies to evaluate the outputs of G2G. However, current study developed two different strategies to quantify the quality of the G2G's results. In the first strategy, all four evaluation metrics were used which produced more accurate results in the field of segmentation. Regarding this strategy, ground truths and images predicted by the G2G were fully compared using the four evaluation metrics.

Finally, to provide a global mean value of all test images, values of each metric were averaged. The second evaluation strategy was training of original Pix2Pix network with its corresponding recommended hyper-parameters which were used to train the G2G. The Pix2Pix network was evaluated on the same test dataset using the four evaluation metrics in order to obtain a global mean value of all the test images. Following determining of the global mean values of the G2G and

Pix2Pix network, they were compared to see which network outperforms the other. Results driven from the two strategies (Table 3) demonstrate that the G2G performed more efficient than Pix2Pix in regards to representing a large margin for all the evaluation metrics. Furthermore, comparison between the mean IoU of G2G and the recent multi-task learning study [3] (Table 3) verifies the superiority of the present network.

Table 3 Comparison between G2G, Pix2Pix and multi-task learning [3] networks in terms of the percentage values of the four evaluation metrics.

Metrics	G2G	Pix2Pix	Multi-task [3]
Mean pixel accuracy	0.96	0.89	0.95
Mean accuracy	0.89	0.74	–
Mean IoU	0.83	0.65	0.70
Mean frequency weighted IU	0.90	0.82	–

In line with the applied strategies, G2G network was tested on validation dataset and the second generator of G2G network acted as a post-processing algorithm (Fig. 13) trying to improve the output of the first generator and making predictions that looked like ground truth from the perspective of boundaries and edges.

4 Conclusion

Current study developed a new network called G2G, based on Pix2Pix network, to solve the problem of imprecise boundaries of the building footprint segmentation, which was mainly conducted through conversion of high-resolution satellite images into maps of the buildings. This novel network is based on employing two generators with certain individual purposes to improve the accuracy of building footprint segmentation. The network preserves the local and boundary information which gives it the opportunity to achieve a high accuracy in the segmentation analysis. Different criteria verify the superiority of G2G network against other pre-existing networks.

References

1. Orthofoto Tirol dataset. https://www.data.gv.at/katalog/dataset/35691b6c-9ed7-4517b4b3-688b0569729a
2. Bhatia, P.: Data mining and data warehousing: principles and practical techniques. Cambridge University Press (2019)
3. Bischke, B., Helber, P., Folz, J., Borth, D., Dengel, A.: Multi-task learning for segmentation of building footprints with deep neural networks. In: 2019 IEEE International Conference on Image Processing (ICIP), pp. 1480–1484. IEEE (2019)

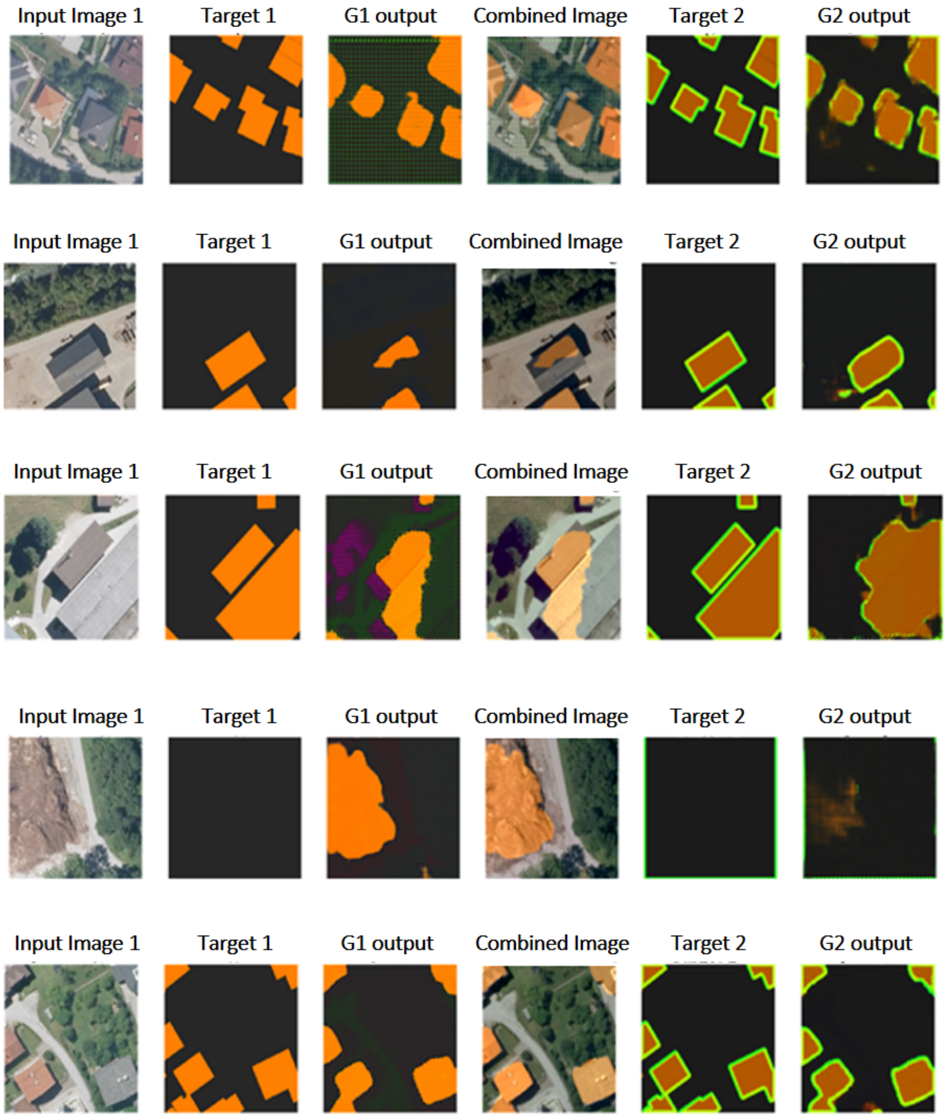


Fig. 13 Results of the G2G on the validation dataset.

4. Bittner, K., Adam, F., Cui, S., Körner, M., Reinartz, P.: 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 **11**(8), 2615–2629 (2018)
5. Bittner, K., Cui, S., Reinartz, P.: Building extraction from remote sensing data using fully convolutional networks. International Archives of the Photogrammetry, Remote Sensing and Spatial Information Sciences-ISPRS Archives **42**(W1), 481–486 (2017)

6. Gavankar, N.L., Ghosh, S.K.: Object based building footprint detection from high resolution multispectral satellite image using k-means clustering algorithm and shape parameters. Geocarto International **34**(6), 626–643 (2019)
7. Guo, Y., Liu, Y., Georgiou, T., Lew, M.S.: A review of semantic segmentation using deep neural networks. International journal of multimedia information retrieval **7**(2), 87–93 (2018)
8. Isola, P., Zhu, J.Y., Zhou, T., Efros, A.A.: Image-to-image translation with conditional adversarial networks. In: Proceedings of the IEEE conference on computer vision and pattern recognition, pp. 1125–1134 (2017)
9. Kaczmarek, K.: Mapping challenge winning solution. https://towardsdatascience.com/mapping-challenge-winning-solution-1aa1a13161b3 (2018)
10. Li, M., Lin, Z., Mech, R., Yumer, E., Ramanan, D.: Photo-sketching: Inferring contour drawings from images. In: 2019 IEEE Winter Conference on Applications of Computer Vision (WACV), pp. 1403–1412. IEEE (2019)
11. Long, J., Shelhamer, E., Darrell, T.: Fully convolutional networks for semantic segmentation. In: Proceedings of the IEEE conference on computer vision and pattern recognition, pp. 3431–3440 (2015)
12. Mirza, M., Osindero, S.: Conditional generative adversarial nets. arXiv preprint arXiv:1411.1784 (2014)
13. Möller, M., Lymburner, L., Volk, M.: The comparison index: A tool for assessing the accuracy of image segmentation. International Journal of Applied Earth Observation and Geoinformation **9**(3), 311–321 (2007)
14. Parsad., N.M.: Deep learning in medical imaging V. https://medium.com/datadriveninvestor/deep-learning-in-medical-imaging-3c1008431aaf (2018)
15. Ronneberger, O., Fischer, P., Brox, T.: U-net: Convolutional networks for biomedical image segmentation. In: International Conference on Medical image computing and computer-assisted intervention, pp. 234–241. Springer (2015)
16. Schuegraf, P., Bittner, K.: Automatic building footprint extraction from multi-resolution remote sensing images using a hybrid fcn. ISPRS International Journal of Geo-Information **8**(4), 191 (2019)
17. Sun, Y., Zhang, X., Zhao, X., Xin, Q.: Extracting building boundaries from high resolution optical images and lidar data by integrating the convolutional neural network and the active contour model. Remote Sensing **10**(9), 1459 (2018)
18. TAbhishek, Jindal, N.: Copy move and splicing forgery detection using deep convolution neural network, and semantic segmentation. Multimedia Tools and Applications
19. Taha, A.A., Hanbury, A., del Toro, O.A.J.: A formal method for selecting evaluation metrics for image segmentation. In: 2014 IEEE international conference on image processing (ICIP), pp. 932–936. IEEE (2014)
20. Vo, D.M., Lee, S.W.: Semantic image segmentation using fully convolutional neural networks with multi-scale images and multi-scale dilated convolutions. Multimedia Tools and Applications **77**(14), 18689–18707 (2018)
21. Wang, T.C., Liu, M.Y., Zhu, J.Y., Tao, A., Kautz, J., Catanzaro, B.: High-resolution image synthesis and semantic manipulation with conditional gans. In: Proceedings of the IEEE conference on computer vision and pattern recog-

nition, pp. 8798–8807 (2018)

22. Wang, X., Yan, H., Huo, C., Yu, J., Pant, C.: Enhancing pix2pix for remote sensing image classification. In: 2018 24th International Conference on Pattern Recognition (ICPR), pp. 2332–2336. IEEE (2018)
23. Zhu, Q., Liao, C., Hu, H., Mei, X., Li, H.: Map-net: Multi attending path neural network for building footprint extraction from remote sensed imagery. arXiv preprint arXiv:1910.12060 (2019)
24. Zou, K.H., Warfield, S.K., Bharatha, A., Tempany, C.M., Kaus, M.R., Haker, S.J., Wells III, W.M., Jolesz, F.A., Kikinis, R.: Statistical validation of image segmentation quality based on a spatial overlap index1: scientific reports. Academic radiology **11**(2), 178–189 (2004)