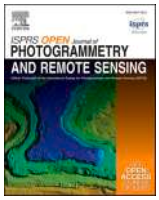
ISPRS Open Journal of Photogrammetry and Remote Sensing 8 (2023) 100031

Contents lists available at ScienceDirect

ISPRS Open Journal of Photogrammetry and Remote Sensing

journal homepage: www.journals.elsevier.com/isprs-open-journal-of-photogrammetry-and-remote-sensing



Towards global scale segmentation with OpenStreetMap and remote sensing

Munazza Usmani [a], Maurizio Napolitano [b], Francesca Bovolo [b,*]

[a] University of Trento & Fondazione Bruno Kessler, Italy
[b] Fondazione Bruno Kessler, Italy

ARTICLE INFO

Keywords:
High resolution
Crowd source
Remote sensing
Deep-learning
Image segmentation

ABSTRACT

Land Use Land Cover (LULC) segmentation is a famous application of remote sensing in an urban environment. Up-to-date and complete data are of major importance in this field. Although with some success, pixel-based segmentation remains challenging because of class variability. Due to the increasing popularity of crowd-sourcing projects, like OpenStreetMap, the need for user-generated content has also increased, providing a new prospect for LULC segmentation. We propose a deep-learning approach to segment objects in high-resolution imagery by using semantic crowdsource information. Due to satellite imagery and crowdsource database complexity, deep learning frameworks perform a significant role. This integration reduces computation and labor costs. Our methods are based on a fully convolutional neural network (CNN) that has been adapted for multi-source data processing. We discuss the use of data augmentation techniques and improvements to the training pipeline. We applied semantic (U-Net) and instance segmentation (Mask R-CNN) methods and, Mask R–CNN showed a significantly higher segmentation accuracy from both qualitative and quantitative viewpoints. The conducted methods reach 91% and 96% overall accuracy in building segmentation and 90% in road segmentation, demonstrating OSM and remote sensing complementarity and potential for city sensing applications.

1. Introduction

Urban land use and land cover (LULC) information is changing rapidly and exploring it is very important for analyses of the environment, land resource management, yield estimation, city planning, change detection analysis, and emergency/disaster situations in urban environments (Huang et al., 2018). Remote sensing data are an effective way to extract the LULC information automatically. After the acquisition, satellite data are processed and geo-referenced for many applications. Satellite images are identified as a valuable source and high-resolution images have received strong attention in recent years for the updated LULC mapping (Zhang et al., 2016). However, the update of LULC maps is a non-trivial task as it requires a large amount of time and cost to acquire and manipulate the very high-resolution (VHR) remote sensing data.

Although advancement in remote sensing, data processing for LULC is complex, and sometimes image processing methods fail in critical applications like disaster management (Geiß et al., 2017). Often methods end up with limited accuracy because LULC-supervised classification methods such as artificial intelligence and advanced strategies being edge-based, shadow-based, and object-based (Lambers and Traviglia, 2016; Ok, 2013) require optimal training datasets. In supervised classification, human users manually draw labels to train the algorithm. To have high-quality training, testing, and validation datasets the preparation requires time and implies costs. Also, for validation of classification maps, reference data are collected by humans using survey techniques including Global Positioning System (GPS) and Geographic Information System (GIS) maps. Unfortunately, training samples with high quality and quantity are seldom available and the development of a high-level automatizing procedure for extracting information from remote sensing imagery remains a major challenge. The absence of both unique joint learning techniques and high-quality training samples were noted as two important obstacles to efficient deep learning for remote sensing at a large scale (Li et al., 2022; Hong et al., 2020). There is still some accessible information such as Google maps and GIS databases, which rely on user-input crowdsourced annotations.

Crowd sourcing is collecting data/information by expert and non-expert volunteers from different geographical locations and freely available with the coverage of the entire world on the internet. With the advancement in Volunteer Geographic Information (VGI) and geospatial

* Corresponding author.
E-mail address: bovolo@fbk.eu (F. Bovolo).

https://doi.org/10.1016/j.ophoto.2023.100031
Received 26 September 2022; Received in revised form 23 January 2023; Accepted 24 January 2023
Available online 16 February 2023

crowd sourcing efforts, the well-known user-generated content is OSM and its first open geodata (Ayala et al., 2021). OSM is a mapping database available to the public as open data, created and managed by volunteers around the world that manually collect data from different sources. In most cases by digitizing the ortho-rectified satellite images, but also by collecting data with GPS, importing data from other sources, or correcting the data with local knowledge. OSM data contains, with semantic information, the outline of the building, the midline of the road, railroads, power transmission channels, etc. Another key feature of OSM is that object-related data is sometimes saved/set in a hierarchical order and can be separated using specific tags or points of interest (POI). We can use this tag to extract or filter out the information related to the application/problem.

The images used in OSM come from various sources for which permission for reuse has been granted. These are not always georeferenced correctly and sometimes not up-to-date; depends from area to area. In most cases, high-resolution satellite images are used. Unlike satellite images which are raster, OSM data is available in vector format with the metadata related to the creation of edit and the name of the author information. Initially, OSM simply focused on mapping ground information, and its members were known as primary consumers of the maps (Mooney and Minghini, 2017). Today, OSM has a broad view in the context of data generation, evaluation, and implementation. It has been getting attention from many private companies and government sectors because of the open data policy. In the meantime, OSM was suffering poor quality control due to huge data, and many areas (small cities) have not yet been mapped/drawn adequately. In addition, we need to take into account the inconsistency and alignment issues between OSM and remote sensing images. So, handling this kind of data is still a challenging task. But crowdsource (OSM) mapping platform could be a good source of data for the LULC classification, especially when data from other sources is not available. We are focusing on the implementation of deep learning algorithms that can learn from noisy open crowdsourced data and also how both datasets are important for creating up-to-date maps automatically. Both sides are crucial and should engage each other to minimize the amount of manual work for learning from a large dataset.

In this work, we will focus on the deep-pixel supervised classification of buildings and roads, as they are paramount features for the impacts of urban growth in any region and are the most important hosting places for human activities.

There are a lot of limitations to extracting these classes. For example, roads are frequently obscured by building shadows and greenery. The colors, widths, structures, textures, and forms of roadways and buildings vary widely between locations. Urban roads, for example, are straighter and broader than rural roads and the same applies to buildings in rural and urban areas. All of these variables make it more difficult to extract features from remote sensing imagery automatically. To overcome the mentioned limitations, our research questions in this study are:

How good are crowd sourcing (OSM) and remote sensing with respect to generating segmentation maps on large scale through existing deep-learning approaches? What are the challenges of pre-processing to improve data quality and the importance of augmentation techniques in computer vision to the proposed task with respect to performance and scalability?

This article outlines the various improvements and outcomes of the proposed approaches. The next Section 2 will focus on the relevant study that evolved to the suggested technique for buildings and roads. Our methodology Section 3 shows two experiments: buildings and roads segmentation work has been explained with a thorough understanding of CNNs models. A brief overview of deep neural networks Mask-RCNN and U-Net is included together with a discussion on how their architectures cope with the problem of image segmentation using the proposed dataset of remote sensing and OSM in Section 3.2. The study area and experimental details with results are described in Sections 4 and 5, respectively. Finally, we conclude the work in Section 6.

2. Related work

The deep learning models for semantic interpretation based on satellite imagery have been researched extensively. Since 2014, many applications including remote sensing image classification, grasp attention towards DL frameworks (Zhu et al., 2017). The fundamental DL network models are recurrent neural networks (RNNs), sparse auto-encoder (AEs), convolutional neural networks (CNNs), deep belief networks (DBNs), and generative adversarial networks (GANs). CNN is the most famous network for image classification and for change detection analysis in remote sensing for being fully convoluted (Ji et al., 2019). Since the early work on information extraction using CNN's (Mnih and Hinton, 2010), several research has looked into deep neural networks for autonomous interpretation of aerial and satellite data. The use of Fully Convolutional Networks (Long et al., 2015), an architecture that was originally designed for semantic segmentation of multimedia displays but has now been successfully applied to remote sensing data at multiple resolutions, has been the topic of recent DL research. On a very high-resolution dataset, this type of deep model produced good results (Li et al., 2017), and it was frequently integrated with multi-scale analysis, a graphical model for post-processing, and boundary or edge detection. The same approaches were successful in extracting building information from aerial scenes at low resolution. Many methods are proposed for image classification or image segmentation possibly for single-class prediction. In (Wang et al., 2020), the efficiency of U-Net architecture for segmentation is described. The network can be trained end-to-end on a small amount of data in a very reasonable time. Accordingly, we used the U-Net approach for handling large areas and buildings with variance aspects. Heterogeneous data fusion was also investigated using end-to-end deep networks in (Hu et al., 2017) for LiDAR and RGB data, and in (Audebert et al., 2017) for ortho-rectified aerial IR-RGB images data, and later utilizing deep features mixed with hand-crafted features for random forest classification (Paisitkriangkrai et al., 2015). High-resolution images have been fused with 3-channel auxiliary input including normalized DSM to get a clearer output in (Gu et al., 2021), with an increase in computational and time cost. With more advancements in computer vision for object recognition, instance segmentation is getting famous for individual instance detection. A series of architectures, such as R–CNN, Fast R–CNN, Faster R–CNN, and Mask R–CNN has been designed for this task (He et al., 2017). Among them, Mask R–CNN is the most recent one. All these networks use Regional Proposal Net (RPN) to detect regions of interest. Other famous architectures for instance segmentation are Single Shot Detector (SSD), You Only Look Once (YOLO), etc., and use one-stage algorithms to predict boundary boxes around objects. Its performance time is fast but without RPN the detection for small targets is not as good as the Mask R–CNN. In addition, sometimes it fails to define boundaries for individual instances (Ruiz-Santaquiteria et al., 2020). The information collected from imagery in computer vision fields has recently been utilized to successfully recover road networks in remote sensing imagery (Cheng et al., 2012). The majority of these approaches are classification-based; they use geometric, photometric, and textural features to extract features from remote sensing data. These approaches based on feature extraction are frequently semi-automatic methods that rely on the manual selection of samples. Developing a large number of annotations is very difficult, and increasing input layers make a network architecture complex and increases computational cost. Since the introduction of OSM and Google Maps in 2004, geographic data has become freely available and widely used in remote sensing applications. They may be utilized as targets for deep learning systems, as demonstrated by Mnih's research work (Mnih, 2013). The data layers can be incorporated into a processing sequence to generate new geospatial data. Despite the fact that the scope and quality of open GIS annotations vary greatly depending on the users' experience and the number of volunteers, this data may contain useful information for particular areas and classes. Active deep learning helps to discover unlabeled objects in

OSM (Chen and Zipf, 2017), while machine learning (ML) technologies (a random forest variant) allow combining remote sensing and volunteered geographic information (VGI) to estimate natural hazard exposure (Geiß et al., 2017) and local climate zones (Danylo et al., 2016). However, very few VGI have been employed as an input (rather than a target) in DL networks. Despite the fact that hand-labeled data is typically correct, the cost of human labeling and the lack of publicly released hand-labeled datasets reduces the availability of remote sensing training and testing sets. While considering remote sensing datasets, crowd-sourced initiatives like OSM could be significant. As a result, datasets that are larger than those that have been hand-labeled can now be created. Although ML and DL approaches are rising, the accuracy of training samples for segmentation in remote sensing is still under analysis and some comprehensive areas have not been researched yet. In this paper, we present the problem of building segmentation with two models: U-Net for semantic segmentation, and Mask R–CNN for instance segmentation.

3. Methodology for buildings and roads segmentation

The proposal includes a straightforward and repeatable process of DL models for deep-pixel-based buildings and road recognition using VHR remote sensing and OSM. Let γ and φ be a remote sensing and crowd source dataset respectively, obtained from different sources and at different times (t_1, t_2). The goal is to develop a robust segmentation framework that could combine γ and φ effectively for object segmentation, class (C_p) and no object class (C_o) in the image, the general workflow is shown in Fig. 1. We demonstrate the capabilities of remote sensing and OSM data fusion, and the capability to generate segmentation maps for areas other than the input training ones. The methodology has two main steps: pre-processing and CNN-based segmentation.

3.1. Pre-processing

Despite the fact that the γ and φ dataset appears to be ready to train DL-based models, a closer look reveals that there is a significant amount of labeling noise owing to the use of open data (φ) and high dimension data (γ). Before producing a suitable training dataset for deep convolutional networks, data pre-processing is a key requirement for accessing the data issues like misalignment and high dimension. The dataset used in this work is generated through the geo-processing tools, like geo-referencing, projection, clipping, translation, and rasterization using the QGIS approach and Osmosis tool; {γ_1, γ_2, γ_3, ….γ_n} and { φ_1, φ_2, φ_3, …..φ_n} are generated by tiling large $(f \times f)$ dimension images in the dataset, where γ_n and φ_n are a total number of tiles of remote sensing and OSM dataset, respectively. The dataset was split into two subsets (training and testing) according to the machine learning guidelines suggested in (Xu and Goodacre, 2018). This makes the quality assessment more reliable than evaluating the dataset as a whole. Here, (γ_i, φ_i) are training inputs for CNN networks and i is the number of patches. The training set was used to train the model parameters, while the test set was used to evaluate the generalization ability of the networks. The OpenStreetMap data dump was downloaded and filtered according to building and road segmentation problems. The planet.osm[1] is called a data dump and includes all semantic information of nodes, ways, and polygons. The whole world or just interesting areas like one country or a small area can be downloaded. We utilize the Osmosis tool to extract the Point Of Interests (POIs) for roads and buildings and store them in a separate file after downloading them for our area of interest. OSM POIs may be connected to many polygons, nodes, or routes. The features that these items represent are described in their tags. The subset of POIs related to buildings and roads has categories as listed in Table 1; motorways, highways, land use, and other categories (more than 10 different categories) are mentioned. About 5537 OSM POIs for buildings and 5937 for road lines for one γ ($(f \times f)$ *image*) were extracted.

The OSM data and the imagery were both projected to the same geographic coordinate system. To keep the images geographically relevant, the imagery was translated to the same coordinate system as the OSM data. For roads, using the same geographic coordinates, the corresponding OSM center line annotations were extracted. Using prior information of the image resolution, the initial road annotations are extrapolated from the center line and apply a buffer to manipulate the pixel values of roads. This is because most road centerlines in VGI data miss accurate width information and hence cannot be easily used to train the models. After rasterization on the same resolution level, all γ and related φ annotation images were clipped using a fixed-size sliding window to create the dataset for training the specified model. Because the proposed CNNs is a segmentation model, it accepts both remote sensing imagery and crowdsourced data as input. We trained two convolutional neural networks for semantic and instance segmentation of (C_p , C_o) at different visual sizes to verify that the proposed dataset has potential to segment buildings accurately, also individually, and used fine-tuned approach for road segmentation.

3.2. CNNs

A state-of-the-art of convolutional neural network (CNN) is presented in (Shrestha and Vanneschi, 2018) for image segmentation and explains how ML and DL models are improving day by day. The CNNs are most widely used in image processing or image segmentation, due to a series of convolutional layers. Pixel-based segmentation through deep learning has been divided into two types known as semantic and instance segmentation. We are considering both types and comparing semantic and instance segmentation methods for our proposed dataset. We utilized two techniques to overcome the problem of building segmentation: U-Net for semantic segmentation and Mask R–CNN for instance segmentation. The U-Net is gradually enhanced with additional characteristics and complexity. Despite the incorporation of more complex characteristics, the performance of some classifiers appears to be limited due to complexity in data handling, boundary separation, and data usage limit. As a result, we look at Mask R–CNN as a second method, to show data compatibility for instance image segmentation. Here, we modified the U-Net and Mask R–CNN architectures, which are considered good approaches for image segmentation, and found these effective and efficient to show the significance of the proposed strategy. Since the Mask R–CNN size for the problem of the building was large and performed well for building segmentation applications, we propose a smaller network, which we call the light-weight classifier. This light-weight classifier is half the size of the original Mask R–CNN architecture. We used fine-tuning approach to train the model with a small dataset that simulates the road environment and provides accurate results for road segmentation without using a huge amount of data and complex architecture. The architecture of the DL models used for training is decided by hyper-parameters being evaluated by examining the training, loss, and accuracy validation. Our approach is an effective method that creates a balance between minimum training time and good accuracy so that the whole process can respond as soon as possible in any emergency response, change detection analysis, or urban planning. The proposed framework is end-to-end and fully automatic.

1. *U-Net Architecture*

U-Net has been used in several satellite image segmentation tasks (Chen et al., 2018) and is a popular semantic segmentation model. We modified its layers and added a batch normalization layer for every convolutional layer to extract buildings information. U-Net benefits tremendously due to its deeper architecture, as demonstrated by (Li et al., 2019). The deep layers enable the model to segment data in considerably more detail. The segmentation through U-Net is very

[1] Planet OpenStreetMap.

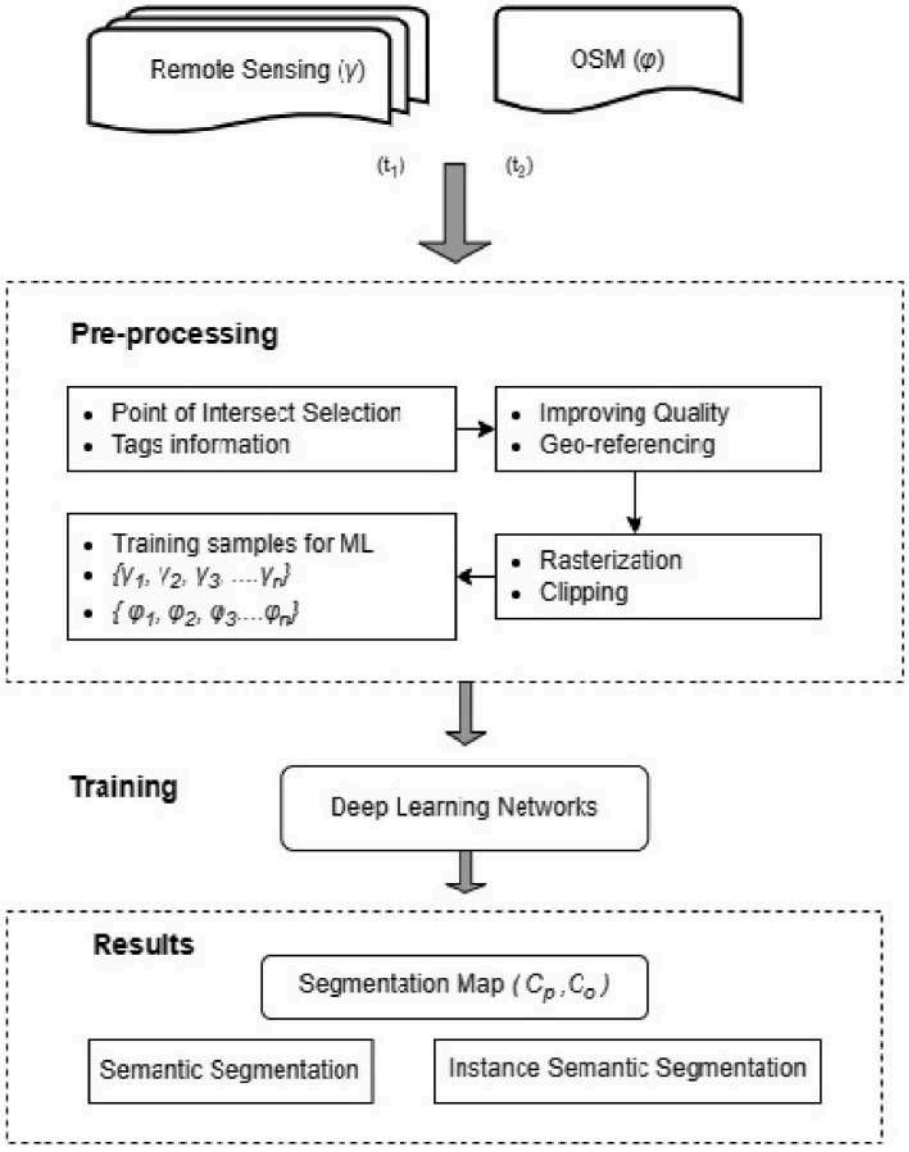


Fig. 1. Proposed workflow to combine RS and OSM information.

Table 1
POI used for this Study.

Key:Value	Quantity
Buildings: *School, Hospitals, Museum, Government, Apartments, Cafe, Company*	5537
Roads: *Highway, Primary Link, Secondary Link, Trunk*	587

accurate, especially at the object borders. A work in (Iglovikov et al., 2017), introduced the deep U-Net architecture, and Fig. 2 depicts the model architecture adopted in it, which takes raw satellite/aerial images and OSM semantic information as input and predicts buildings segmentation maps. The hyper-parameters of the network are described in Table 2. The left half (encoder) of the network is comparable to a CNN that is concerned with constructing a low-dimensional dense set of inputs, while the right half (decoder) up-samples the training feature representations to a shape similar to the input. The skip connections allow data to move from the encoder to the decoder, supporting the defined network in maintaining spatial/feature data. The network is known as a binary classifier because the overall purpose is to detect (C_p , C_o). As a result, we utilize a sigmoid activation function after the last convolution to generate pixel-wise pseudo-probabilities. To match the input and output sizes after down-sampling and up-sampling, we use zero padding and cropping as needed. As mentioned in the pre-processing phase, the training images are subsets of a large dimension high-resolution image. A zero padding would improve the estimations of the building part on the adjacent tile. Under the assumption of binary classification problem (C_p , C_o), binary cross entropy loss function is defined in (Mohanty et al., 2020) as:

$$LOSS = -\varphi_t \log(C_p) - (1-\varphi_t)\log(1-C_p) \quad (1)$$

Where.

φ_t - target class
C_p - predicted class

Every pixel in the image is given a probability of being an object of interest (C_p) by the model. A decision threshold determines the membership to a class. For binary class, probabilities or scores in the range of 0–1 and the threshold is set to 0.5 by default. The distribution of the areas in the dataset was used to determine a reasonable drop off pixel threshold. For example, the strategy can be shown as:

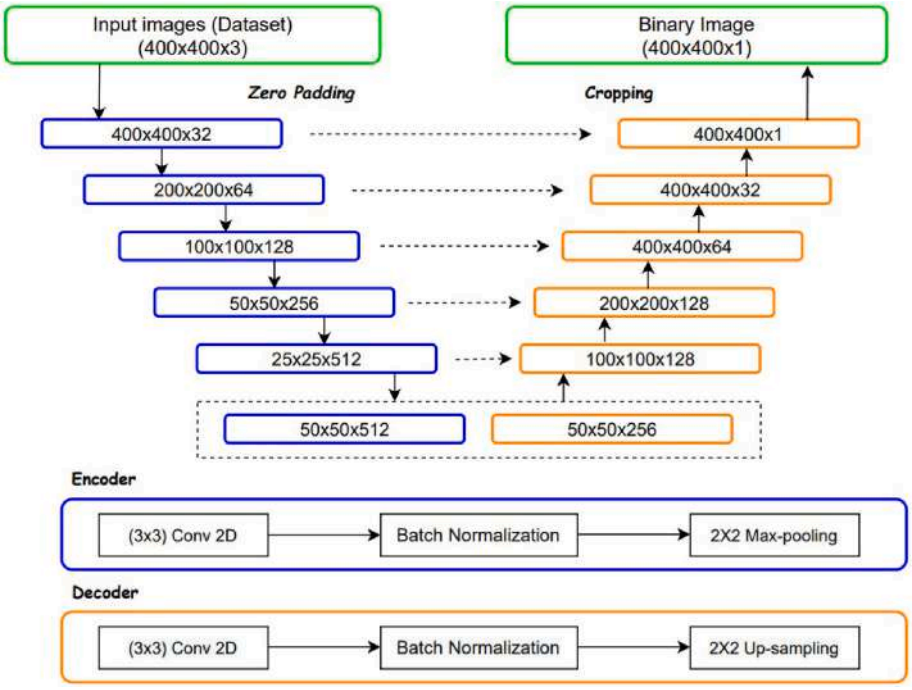


Fig. 2. U-Net based framework: Encoder (in blue) on the left and decoder (in yellow) on the right, dash arrows showing concatenation.

Table 2
U-Net training parameters.

Hyper Parameters	Value
Iteration	150
LEARNING RATE	0.0001
Batch Size	4
Non-Linear Activation	ReLU/Sigmoid
Optimizer	ADAM

Prediction $>= 0.5 = C_p$

Prediction $< 0.5 = C_o$

In semantic segmentation, true positive (TP) are pixels that are correctly predicted as a class (C_p), false positive (FP) are misclassified, and false negative (FN) are mixed with background pixels, like C_p wrongly predicted as C_o.

2. *Mask R–CNN Architecture*

As a basic network, for buildings and roads, we adopted the Mask R–CNN (Ren et al., 2015) because of its simplicity in the network structure and hyper-parameter tuning. Mask R–CNN is a framework that can identify objects in an image quickly and accurately, the framework shown in Fig. 3 described by (Zhao et al., 2018). It can also generate high-quality segmentation masks for each instance. The Resnet-101 was chosen as the backbone network in the training of Mask R–CNN, which is a network-based solution for tackling the degradation problem. During training, the Mask R–CNN features as a Feature Pyramid Network (FPN), which is used to improve the detection of multi-scale objects. It works seamlessly with the head network. To minimize the complexity of building detection and segmentation, we trained Mask R–CNN in three steps and modify the model from the multi-class object to single class (building) detection. For roads, we reduce the network size and fine-tune it.

Data augmentation techniques are used to improve the quality and quantity of training data for a network. It helps in generating robust and sufficient training data for a given problem and to improve the robustness of the network for multi-scale building detection. We adopted various transformations such as horizontal or vertical flipping, blurring, and noise reduction during training. In the final implementation, When the Intersection Over Union (IOU) is less than 0.3, negative anchors are assigned, and positive anchors are assigned (IOU) is greater than or equal to 0.7, IOU was calculated as 2:

$$IoU = \frac{TP}{TP + FP + FN} \tag{2}$$

In instance segmentation, TP are objects/instances that are correctly predicted instead of pixels, FP are misclassified instances, and FN are mixed with background instances like C_p wrongly predicted as C_o.

We employ a mix of binary cross entropy and a soft Jaccard loss (Iglovikov et al., 2017) as a loss function. The mechanism proposed by (Iglovikov et al., 2018) to generalize discrete Jaccard index into a differentiable version. As a result, the network may directly optimize the loss throughout the training phase.

$$J(A, B) = \frac{|A \cap B|}{|A| + |B| - |A \cup B|} \tag{3}$$

$$J = \frac{1}{n} \sum_{c=1}^{2} w_c \sum_{i=1}^{n} \left(\frac{\varphi_i^c \widehat{C}_i^c}{C_i^c + \widehat{C}_i^c - \varphi_i^c \widehat{C}_i^c} \right) \tag{4}$$

Where, φ_i^c binary value (label), $\widehat{C_i^c}$ the corresponding predicted probability for the pixel i of the class c.

During the training phase, a loss is calculated and defined as classification loss, bounding box regression loss, and mask loss as (He et al., 2017):

$$L = L_{cls} + L_{box} + L_{mask} \tag{5}$$

The classification loss is considered the same as bounding-box loss. The model tries to learn a mask for each class using a sigmoid function and the average binary cross entropy loss is known as mask loss. We applied a per-pixel sigmoid activation function and measured mask loss as the average binary cross-entropy loss. Regarding implementation, both networks (U-Net & Mask R–CNN) were trained with different

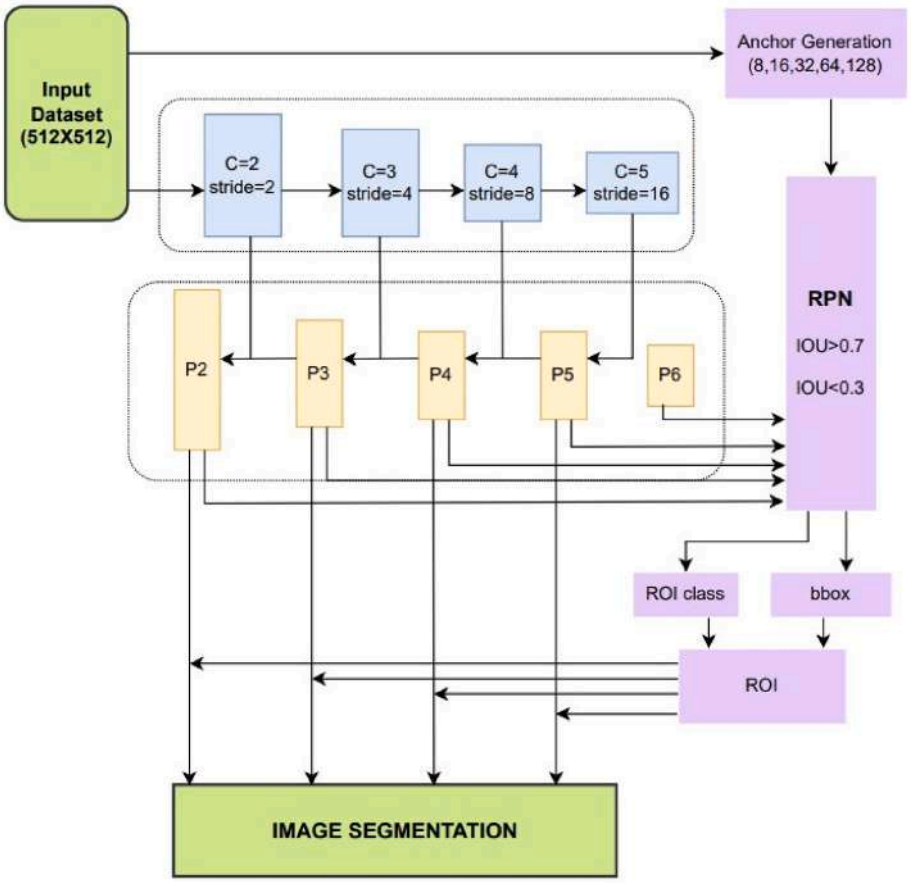


Fig. 3. Mask R–CNN stages: RPN (selects the regions of interest), ROI (aligns the feature maps).

parameters and hyper-parameters, and the details are shown in Tables 2 and 3, respectively. The purpose is to create a thorough semantic segmentation of a satellite image.

4. Study area and data set

To verify the proposed approach of fusion of remote sensing and crowd source dataset for building detection, two deep learning models, for semantic and for instance segmentation, are utilized. In the semantic segmentation model, the dataset originated from two sources: the SpaceNet (Van Etten et al., 2018), and the Trento airborne high-resolution imagery. The SpaceNet dataset is provided in the DeepGlobe challenge and contains satellite images of different cities having urban and suburban regions with 30 cm resolution. On the other side, Trento imagery was collected through an airborne campaign with a high spatial resolution (100 cm) and covers about 50 km^2. The high-dimension images are processed using a sliding window of 400x400. In total, about 6000 tiles were used in the first dataset, which was divided randomly into 80% and 20% for training and validation datasets for the model.

In the instance segmentation model, we used the airborne imagery of Trento, Italy with a spatial resolution of 100 cm. The 512x512 sliding window size has been used to process the high dimension images, which were divided into 70% and 30% for training and validation of the model. The study area chosen for this is shown in Fig. 4.

In both datasets, the remote sensing imagery was used with three spectral bands (RGB) information. The idea is that after we have shown that a method works effectively on RGB imagery, we can work on expanding it to multi-channel remote sensing data. The idea behind using different augmentation techniques is to increase the model's robustness and make it suitable for building extraction on different and large-scale areas. As a result, we will be able to test the suggested technique on geographically dispersed complicated urban scenes with a large variety.

The corresponding semantic information was downloaded from OSM (for both datasets) which is publicly released. Portals like bbbike[2] and Geofabrik[3] integrate OSM data in order to provide free geodata to the community. For roads, we used a small basic dataset, only Trento imagery, which was collected through an airborne campaign with a high spatial resolution (100 cm) and covers about 20 km^2. The high-dimension images are processed using a sliding window of 512x512. After augmentation, about 3000 tiles were processed and we divided 80% into the training set and 20% into the validation set. Study area details are listed in Table 4, showing the total number of OSM attributes (polygons and ways) and the corresponding size in pixels for each

Table 3
Mask R–CNN training parameters.

Hyper Parameters	Value
BACKBONE	Resnet101
Iteration	200
LEARNING RATE	0.001
RPN NMS THRESHOLD	0.7
WEIGHT DECAY	0.0001
DETECTION-MAX INSTANCES	100
RPN ANCHOR RATIOS	[0.5, 1, 2]
RPN ANCHOR SCALES	(8, 16, 32, 64, 128)
Optimizer	Jaccard

[2] BBBike extracts OpenStreetMap.
[3] GEOFABRIK.

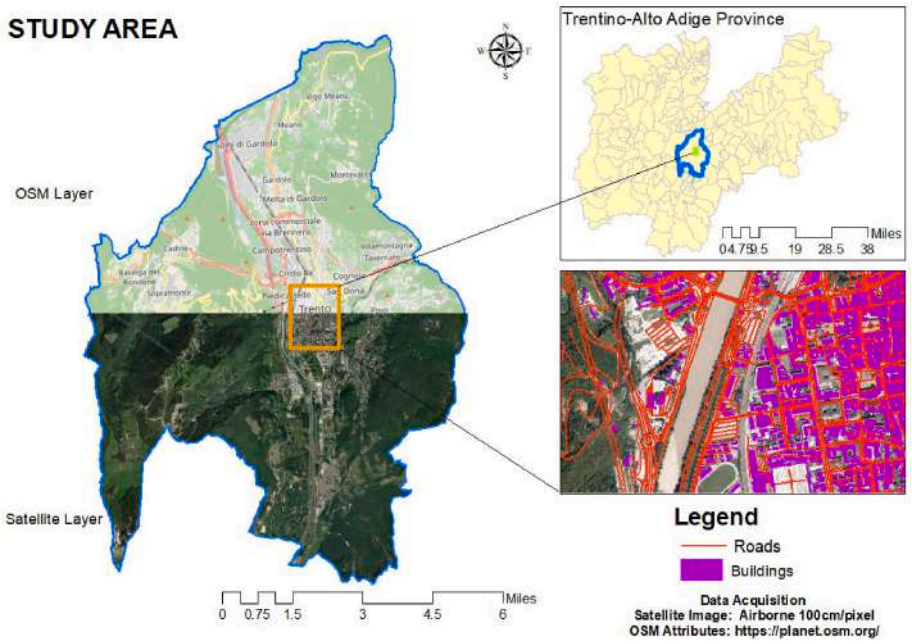


Fig. 4. Geographical location of Trento: Province boundary (left), the airborne image of the area, and OSM annotations (bottom right).

Table 4
Study areas details.

	Type						
Dataset	Area(km^2)	Resolution(cm)	Dimension	Tiles	Buildings	Roads	Set
Space Net	195	50	406x403	1100	6800	–	Train/Test
Trento	50	100	3490x3070	4400	21,320	5937	Train/Test
Beirut & GoogleMaps	–	–	–	–	–	–	Test

dataset. Our proposed segmentation models are evaluated on the same validation data of Trento. To evaluate the proposed approaches, 100 patches of Trento were manually annotated as reference samples due to ground truth data unavailability.

5. Experimental setup and results

For buildings, the number of filters in the convolutional layers for the CNN model (U-Net) that recognizes buildings with predetermined forms were 16, 32, 64, 128, 256, and 512 with a kernel size of 3x3. After the first two convolutional layers, we used max-pooling (zero padding), ReLU as an activation function, and batch normalization. As it is a binary class segmentation, we utilize a sigmoid function in the last output layer. We train the first model (U-Net) for 150 iterations with Adam optimizer using a learning rate of 0.0001 and the second model (Mask R–CNN) for 200 iterations at 0.0001 learning rate with the backbone of RESNET 101 and Jaccard optimizer. The anchor scales of 8, 16, 32, 64, and 128 were selected for the segmentation of buildings in the Mask R–CNN network. Early stopping is utilized to reduce overfitting and to stop training if validation loss decreases in five consecutive epochs. We implemented the proposed method using the Tensor-flow and Keras frameworks. By considering the GPU memory and cost issues, we tested our code on a Google Colab machine equipped with a low-end GPU: 1xTesla K80, having 2496 CUDA cores, 12 GB GDDR5 VRAM.

For roads, we present a lightweight U-Net classifier and investigate the impact of fine-tuning approach with limited OSM data and time for semantic segmentation. The classifier was implemented utilizing Google open-source TensorFlow and python framework. All other tuning hyper-parameters are similar to the main U-Net network. The code was run on GPU in a Google Colab environment for 50 epochs. All of the pre-processed data were fed into the model as an input, and the trained model produced a two-category classification map as output: (C_p , C_o). In the last phase, the parameters are normalized, and the segmentation is done at the pixel level, some samples of results are shown in Fig. 7.

For the performance evaluation of classifiers, we employed some key assessment parameters to calculate the effectiveness of suggested networks in binary segmentation: Recall (υ), Precision (ρ), F1 Score, Overall Accuracy (OA), and Mean Average Precision (mAP) (Ghasemkhani et al., 2020), shown in Table 5. These parameters are calculated for both instance and semantic segmentation on the same validation dataset.

A patch (γ_i, φ_i) with *(fxf)* was fed into the network and that was classified into (C_p, C_o). In addition, we used to make predictions on different areas other than training one (described in the next section) and results show the robustness of the proposed approach and the potential of DL models for the fusion of remote sensing and crowdsource dataset as input layers. The evaluation has been done using the formulas described in equations (6)–(9):

$$\rho = \frac{TP}{TP + FP} \tag{6}$$

$$\upsilon = \frac{TP}{TP + FN} \tag{7}$$

$$OA = \frac{TP + TN}{TP + FP + FN + TN} \tag{8}$$

$$F = 2 \times \frac{\rho \times \upsilon}{\rho + \upsilon} \tag{9}$$

Precision is a positive predicted value and is measured by a fraction of relevant instances to the retrieved instances. While the recall is

Table 5
Evaluation matrix for semantic and instance segmentation on validation dataset.

Segmentation	Method	Class	Training Dataset	Precision	Recall	F1 Score	OA
Semantic	K-means	Buildings	Trento	0.67	0.61	0.64	0.69
	Object-based	Buildings/Roads	Trento	0.73	0.60	0.65	0.82
	Deeplabv3+	Buildings	SpaceNet	0.81	0.90	0.85	0.93
	U-Net (Proposed)	Buildings	Trento & SpaceNet	0.74	0.83	0.82	**0.91**
	U-Net (Proposed)	Roads	Trento	0.66	0.61	0.63	**0.89**
Instance	Mask R–CNN–ECA	Buildings	Trento	0.77	0.62	0.69	0.84
	Mask R–CNN (Proposed)	Buildings	Trento	0.96	0.96	0.96	**0.96**

calculated by all relevant instances and is known as sensitivity.

To answer the research problem mentioned in section 1, we trained the neural networks proposed in section 3.2 on the dataset provided in Table 4, and consider three scenarios to evaluate the performance of baseline models. Below, there are the three setups, against which the results were compared:

Set-up I: The building case corresponds to the semantic and instance segmentation of an area similar to the training one but not included in the training phase. The results are produced using U-Net and Mask R–CNN proposed models and shown in Figs. 5 and 6. Building footprints with sharp boundaries and uniform geometries are delineated precisely using the proposed strategy. Our fusion-based strategy surpasses the other models like (Zhao et al., 2018), (Li et al., 2020); where they are using the traditional classification approach (hand-labeled data) and multi-source GIS data (Google Maps) for buildings detection. Our approach is achieving a significant recall score (91%, 96% for U-Net and Mask R–CNN, respectively) as well as an overall rise in the F1 measure, as seen in Table 5. Both models handle the specific size issues of small and large buildings. The results indicate that many of the accurate predictions are shared by both models. Despite buildings in these areas may be easier to detect than other areas or scenes, our network has a tendency to correctly map/segment buildings in different areas other than the training one, as expected. In addition, a small case study has been generated on a resolution different from the training remote sensing data one (i.e., 100 cm) for the evaluation of the approach in different conditions (details in set-up II).

The road results show better performance for the smaller network even with fine-tune approach. It achieves close performance (89% overall accuracy) compared with a fully supervised model by only using small OSM center-line data as input. This embraces the proposed approach and data-handling strategy during the training procedure. Moreover, the proposed method could be used with any other attribute data.

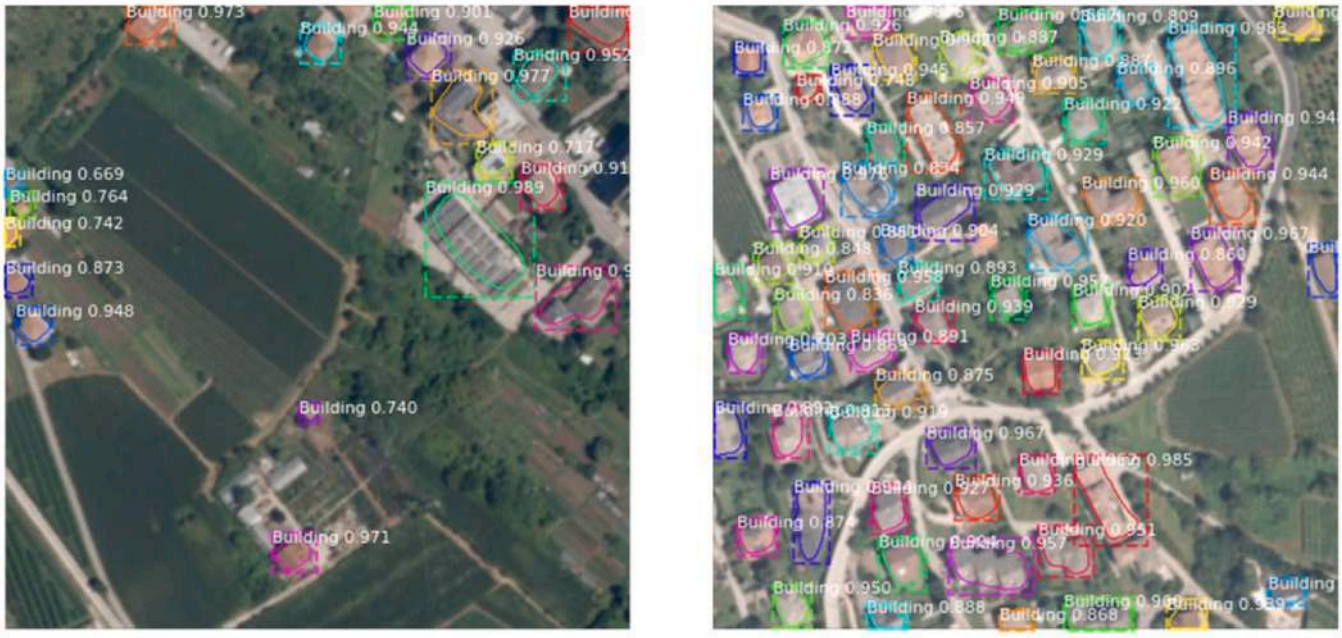
Set-up II: Following the set-up I query, this part shows the instance semantic segmentation of buildings for two different areas which are spatially disconnected and differ from the training area. The first area is downloaded randomly from a publicly available dataset and used as an input to the Mask R–CNN model to produce building instance segmentation (Fig. 8). The second area is from Beirut, Lebanon with a spatial resolution of 30 cm. Results are shown in Fig. 9. This area was downloaded using the Base Map Server (BPS), an online map server in ArcGIS. Both areas have buildings with different properties in terms of texture, shape, size, and spatial resolution than the training one. The idea behind testing these areas is to check the generalization capability of the proposed method. The baseline model used to test the approach is Mask-RCNN due to advancement in the structure. Based on the visual comparison, the approach can accurately extract the majority of the structures, even though in different scenes and resolutions.

Set-up III: In this set-up, we compared the proposed approaches to state-of-the-art (SOA) methods from basic supervised and unsupervised approaches to advance CNN-based methods. For the semantic segmentation model, we chose the K-means algorithm which is known as a famous unsupervised classification method (Wu, 2017) and supervised object-based classification (Blaschke, 2010) (here training samples were manually drawn). We also choose a widely used CNN model for image classification, deeplabv3+ (Liu et al., 2021). For the instance segmentation model, Mask R–CNN with Efficient Channel Attention (ECA) (Wang et al., 2020) is compared with simple Mask R–CNN. Table 5 summarizes the quantitative performance of the proposed approaches and SOA methods. The OA for the K-means method on building class reached 69% due to considerable misclassification. The K-means algorithm is strongly affected by the number of clusters that leads to under-segmentation or over-segmentation if inaccurate. Object-based classification reaches an OA of 82%, but it is time-consuming and hectic to draw/collect the samples to train a classifier with good accuracy. Moreover, sample collection is to be repeated for any new area to be classified.

If we compare the instance segmentation architecture based on Mask R–CNN with SOA approaches it achieves the highest OA (i.e., 96%). Among the semantic segmentation SOA methods, Deeplabv3+, recently

Fig. 5. Buildings semantic segmentation with U-Net method on Trento area (satellite image, OSM and semantic segmentation).

(a)



(b)

Fig. 6. Buildings instance segmentation with Mask R–CNN method on Trento area: ((a) RS and OSM (b) instance segmentation).

introduced as a good segmentation model, achieved an OA of 93%, only. The proposed U-Net architecture is just shortly behind the Deeplabv3+ with a lower OA and F1-score of about 2% and 3%, respectively. However, this is achieved with a much light-weight architecture. For Mask R–CNN with ECA, we adopted the same hyper-parameters setting as in the proposed Mask R–CNN–based architecture, and it reaches 84% accuracy staying 12% behind Mask R–CNN. By adding a channel attention mechanism, the Mask R–CNN with ECA increases the computational and memory requirements and makes convergence slower. Network performance is correlated to the network architecture in terms of layers depth, parameters, batch size, and thus with the computational and time costs. So accordingly, ECA-Mask R–CNN took a longer time (about 44 h) to train than other CNN-based methods with lower performance. Deeplabv3+ and the proposed U-net architecture took about 8–10 h for training and inference, with similar performance. Whereas the Mask R–CNN–based architecture took slightly longer than the latter two

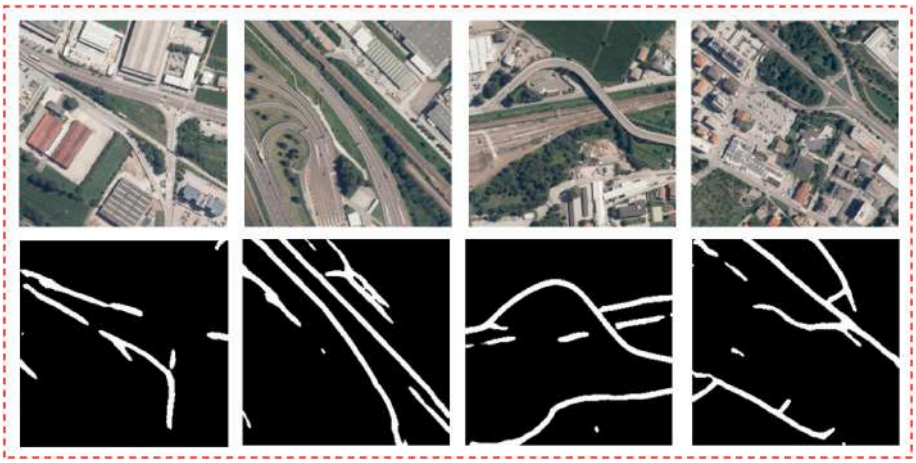


Fig. 7. Roads segmentation on Trento area: Satellite input and roads semantic segmentation.

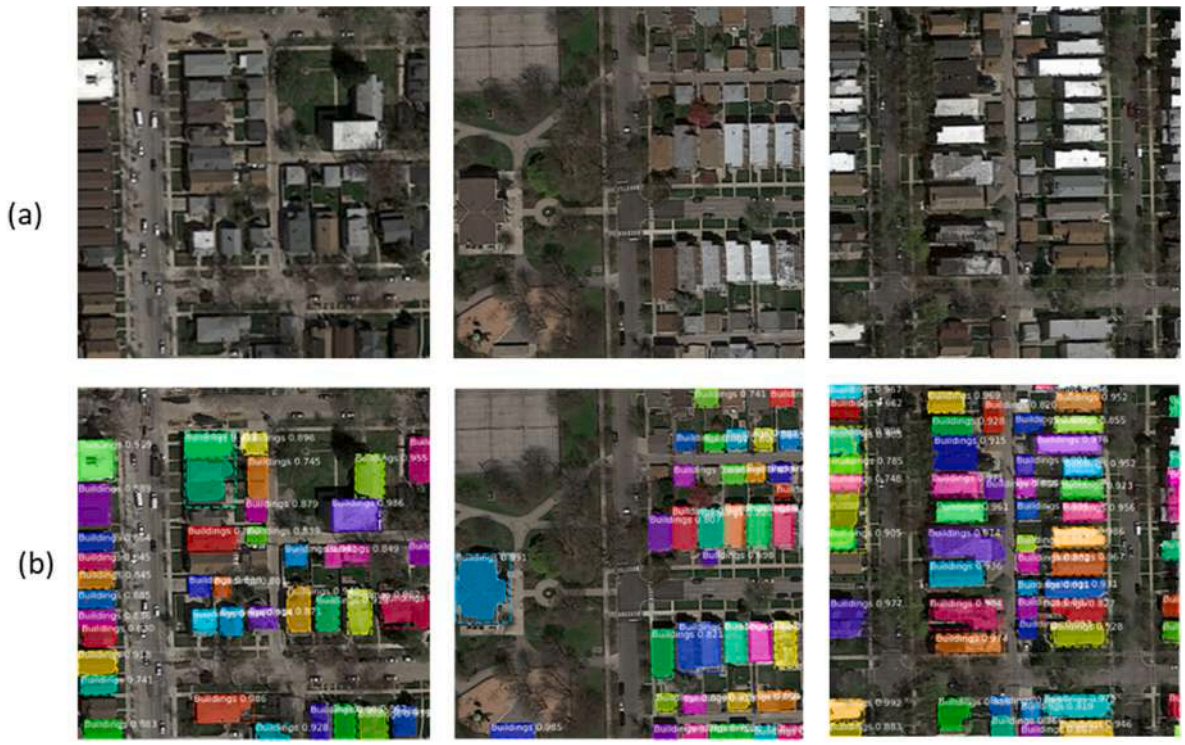


Fig. 8. Test Area-I (GoogleMaps): (a)satellite input (b) buildings segmentation.

(about 16–18 h) which is a reasonable trade-off with respect to the accuracy achieved on the large-scale areas. Further, this is obtained with a low number of parameters.

For qualitative analysis, Fig. 10 shows the segmentation results of the CNN-based methods for three samples of the SpaceNet dataset. The first two columns show the original images and corresponding ground truth masks while the next columns show the segmentation output produced by each network. In the building segmentation setup, the Deeplabv3+ often failed to differentiate between roads and buildings. The skip connections in U-Net improved effectiveness in exploiting the high resolution and geometrical details of the image in the expansion step and achieved better qualitative results as compared to Deeplabv3+. We observe that the Deeplabv3+ has a poor representation of boundaries and class transitions in the prediction. The effect is clearly visible in Fig. 10, line (c). ECA Mask R–CNN approach shows good performance in instance segmentation but misses many instances. The proposed Mask R–CNN–base architecture shows instance segmentation with a maximum number of detected buildings in comparison to all methods and accurate boundary delineation in terms of geometrical details.

6. Conclusion

In this paper, we presented the approaches for building deep-pixel segmentation in VHR images. We considered two types of segmentation: (i) U-Net (semantic segmentation), and (ii) Mask R–CNN (instance segmentation) to handle the problem of binary class segmentation. The method integrates two heterogeneous datasets (remote sensing and crowd source) as target layers for LULC deep-pixel segmentation using DL approaches. The approach accurately segments different kinds of buildings, e.g., the residential and business ones. The results have been compared with some SOA methods (both semantic and instance) showing that the proposed approaches reach better qualitative and

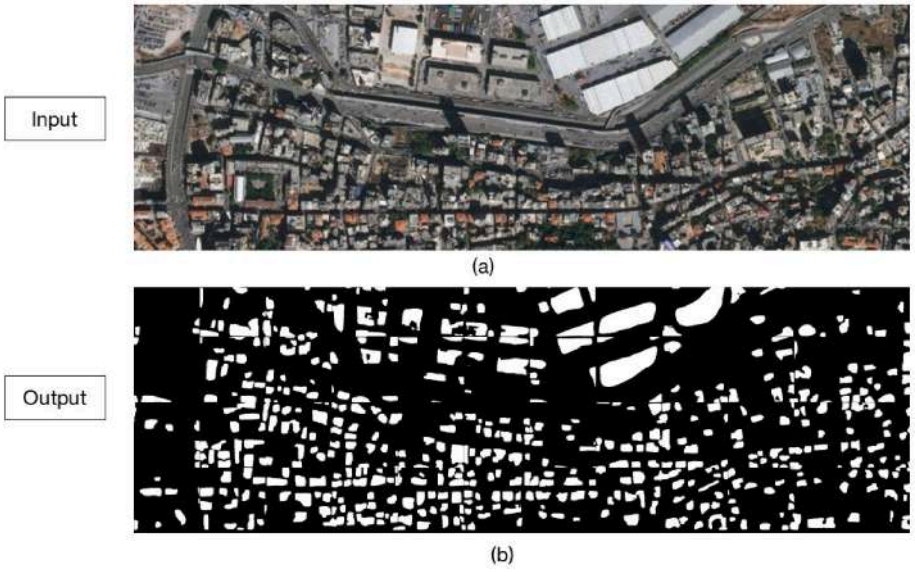


Fig. 9. Test Area-II (Beirut): (a) satellite Input (b) buildings segmentation.

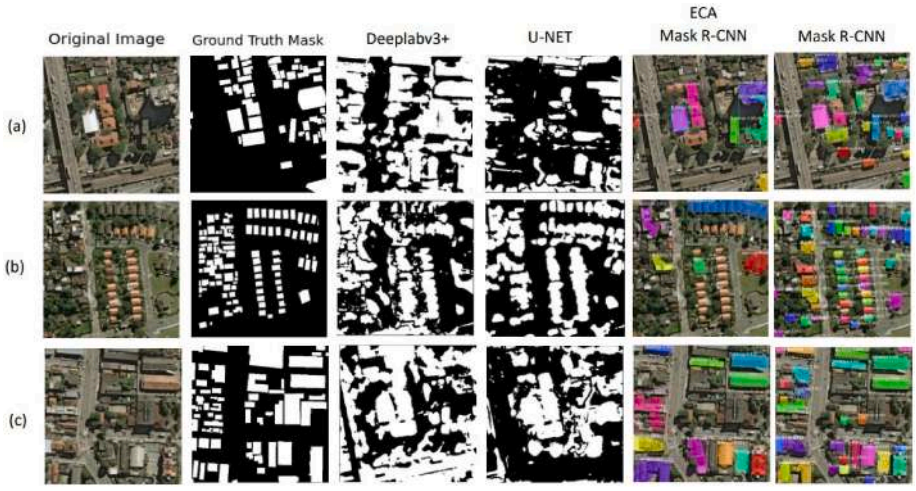


Fig. 10. Comparison of CNN methods on SPACENET dataset.

quantitative results. We have shown that CNN can be trained to attain SOA segmentation performance with noisy OSM and minimal human supervision. The best-proposed model, Mask R–CNN with OA 96% delivers an excellent performance in two aspects: (i) doing multi-source dataset fusion, (ii) generalizing the model in different terrains and environments, we tested our approach for four different datasets (Beirut, GoogleMaps, Trento, SpaceNet) and the results illustrate good performance. This is dealing with noise in the input dataset and shows a fusion approach for binary class instance segmentation. In addition, both models achieve more accuracy (quality and quantity) as compared to SOA methods. So, the proposed approaches can handle segmentation problems on a large scale. The module can refine the features and provide more areas of coverage, with better accuracy. The approach has been tested on road segmentation as well, with small data and minimum computing cost, to show the performance of the proposed method in different classes other than building and the results are promising as well. In future work, we plan to use the method to analyze critical regions with limited building and road OSM data or incomplete crowdsourcing data. Further, we plan to extend the method to other target classes else than buildings and roads.

Disclosure statement

The authors declare no conflict of interest.

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

References

Audebert, N., Le Saux, B., Lefèvre, S., 2017. 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, pp. 67–75.

Ayala, C., Sesma, R., Aranda, C., Galar, M., 2021. A deep learning approach to an enhanced building footprint and road detection in high-resolution satellite imagery. Rem. Sens. 13 (16), 3135.

Blaschke, T., 2010. Object based image analysis for remote sensing. ISPRS J. Photogrammetry Remote Sens. 65 (1), 2–16.

Chen, J., Zipf, A., 2017. Deepvgi: deep learning with volunteered geographic information. In: Proceedings of the 26th International Conference on World Wide Web Companion, pp. 771–772.

Chen, L.-C., Zhu, Y., Papandreou, G., Schroff, F., Adam, H., 2018. Encoder-decoder with atrous separable convolution for semantic image segmentation. In: Proceedings of the European Conference on Computer Vision. ECCV, pp. 801–818.
Cheng, J., Ding, W., Ku, X., Sun, J., 2012. Road extraction from high-resolution sar images via automatic local detecting and human-guided global tracking. Int. J. Antenn. Propag. 2012.
Danylo, O., See, L., Bechtel, B., Schepaschenko, D., Fritz, S., 2016. Contributing to wudapt: a local climate zone classification of two cities in Ukraine. IEEE J. Sel. Top. Appl. Earth Obs. Rem. Sens. 9 (5), 1841–1853.
Geiß, C., Schauß, A., Riedlinger, T., Dech, S., Zelaya, C., Guzmán, N., Hube, M.A., Arsanjani, J.J., Taubenböck, H., 2017. Joint use of remote sensing data and volunteered geographic information for exposure estimation: evidence from valparaíso, Chile. Nat. Hazards 86, 81–105.
Ghasemkhani, N., Vayghan, S.S., Abdollahi, A., Pradhan, B., Alamri, A., 2020. Urban development modeling using integrated fuzzy systems, ordered weighted averaging (owa), and geospatial techniques. Sustainability 12 (3), 809.
Gu, Y., Hao, J., Chen, B., Deng, H., 2021. Top-down pyramid fusion network for high-resolution remote sensing semantic segmentation. Rem. Sens. 13 (20), 4159.
He, K., Gkioxari, G., Dollár, P., Girshick, R., 2017. Mask r-cnn. In: Proceedings of the IEEE International Conference on Computer Vision, pp. 2961–2969.
Hong, D., Gao, L., Yokoya, N., Yao, J., Chanussot, J., Du, Q., Zhang, B., 2020. More diverse means better: multimodal deep learning meets remote-sensing imagery classification. IEEE Trans. Geosci. Rem. Sens. 59 (5), 4340–4354.
Hu, J., Mou, L., Schmitt, A., Zhu, X.X., 2017. Fusionet: a two-stream convolutional neural network for urban scene classification using polsar and hyperspectral data. In: 2017 Joint Urban Remote Sensing Event (JURSE). IEEE, pp. 1–4.
Huang, B., Zhao, B., Song, Y., 2018. Urban land-use mapping using a deep convolutional neural network with high spatial resolution multispectral remote sensing imagery. Remote Sens. Environ. 214, 73–86.
Iglovikov, V., Mushinskiy, S., Osin, V., 2017. Satellite Imagery Feature Detection Using Deep Convolutional Neural Network: A Kaggle Competition. *arXiv preprint arXiv: 1706.06169*.
Iglovikov, V., Seferbekov, S., Buslaev, A., Shvets, A., 2018. Ternausnetv2: fully convolutional network for instance segmentation. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops, pp. 233–237.
Ji, S., Shen, Y., Lu, M., Zhang, Y., 2019. Building instance change detection from large-scale aerial images using convolutional neural networks and simulated samples. Rem. Sens. 11 (11), 1343.
Lambers, K., Traviglia, A., 2016. Automated detection in remote sensing archaeology: a reading list. In: AARGnews-The newsletter of the Aerial Archaeology Research Group, 53, pp. 25–29.
Li, H., Zech, J., Hong, D., Ghamisi, P., Schultz, M., Zipf, A., 2022. Leveraging openstreetmap and multimodal remote sensing data with joint deep learning for wastewater treatment plants detection. Int. J. Appl. Earth Obs. Geoinf. 110, 102804.
Li, Q., Shi, Y., Huang, X., Zhu, X.X., 2020. Building footprint generation by integrating convolution neural network with feature pairwise conditional random field (fpcrf). IEEE Trans. Geosci. Rem. Sens. 58 (11), 7502–7519.
Li, W., He, C., Fang, J., Zheng, J., Fu, H., Yu, L., 2019. Semantic segmentation-based building footprint extraction using very high-resolution satellite images and multi-source gis data. Rem. Sens. 11 (4), 403.
Li, Y., Qi, H., Dai, J., Ji, X., Wei, Y., 2017. Fully convolutional instance-aware semantic segmentation. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 2359–2367.
Liu, M., Fu, B., Xie, S., He, H., Lan, F., Li, Y., Lou, P., Fan, D., 2021. Comparison of multi-source satellite images for classifying marsh vegetation using deeplabv3 plus deep learning algorithm. Ecol. Indicat. 125, 107562.
Long, J., Shelhamer, E., Darrell, T., 2015. Fully convolutional networks for semantic segmentation. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 3431–3440.
Mnih, V., 2013. Machine Learning for Aerial Image Labeling. University of Toronto, Canada.
Mnih, V., Hinton, G.E., 2010. Learning to detect roads in high-resolution aerial images. In: Computer Vision–ECCV 2010: 11th European Conference on Computer Vision, Heraklion, Crete, Greece, September 5-11, 2010, Proceedings, Part VI, vol. 11. Springer, pp. 210–223.
Mohanty, S.P., Czakon, J., Kaczmarek, K.A., Pyskir, A., Tarasiewicz, P., Kunwar, S., Rohrbach, J., Luo, D., Prasad, M., Fleer, S., et al., 2020. Deep learning for understanding satellite imagery: an experimental survey. Front Artificial Intelligence. 3, 534696.
Mooney, P., Minghini, M., et al., 2017. A Review of Openstreetmap Data. *Mapping and the Citizen Sensor*, pp. 37–59.
Ok, A.O., 2013. Automated detection of buildings from single vhr multispectral images using shadow information and graph cuts. ISPRS J. Photogrammetry Remote Sens. 86, 21–40.
Paisitkriangkrai, S., Sherrah, J., Janney, P., Hengel, V.-D., et al., 2015. Effective semantic pixel labelling with convolutional networks and conditional random fields. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops, pp. 36–43.
Ren, S., He, K., Girshick, R., Sun, J., 2015. Faster r-cnn: towards real-time object detection with region proposal networks. Adv. Neural Inf. Process. Syst. 28.
Ruiz-Santaquiteria, J., Bueno, G., Deniz, O., Vallez, N., Cristobal, G., 2020. Semantic versus instance segmentation in microscopic algae detection. Eng. Appl. Artif. Intell. 87, 103271.
Shrestha, S., Vanneschi, L., 2018. Improved fully convolutional network with conditional random fields for building extraction. Rem. Sens. 10 (7), 1135.
Van Etten, A., Lindenbaum, D., Bacastow, T.M., 2018. Spacenet: A Remote Sensing Dataset and Challenge Series. *arXiv preprint arXiv:1807.01232*.
Wang, Q., Wu, B., Zhu, P., Li, P., Zuo, W., Hu, Q., 2020. Eca-net: efficient channel attention for deep convolutional neural networks. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 11534–11542.
Wu, Q., 2017. Gis and Remote Sensing Applications in Wetland Mapping and Monitoring.
Xu, Y., Goodacre, R., 2018. On splitting training and validation set: a comparative study of cross-validation, bootstrap and systematic sampling for estimating the generalization performance of supervised learning. J. Anal. Test. 2 (3), 249–262.
Zhang, P., Gong, M., Su, L., Liu, J., Li, Z., 2016. Change detection based on deep feature representation and mapping transformation for multi-spatial-resolution remote sensing images. ISPRS J. Photogrammetry Remote Sens. 116, 24–41.
Zhao, K., Kang, J., Jung, J., Sohn, G., 2018. Building extraction from satellite images using mask r-cnn with building boundary regularization. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops, pp. 247–251.
Zhu, X.X., Tuia, D., Mou, L., Xia, G.-S., Zhang, L., Xu, F., Fraundorfer, F., 2017. Deep learning in remote sensing: a comprehensive review and list of resources. IEEE Geosci. Remote Sense Magazine. 5 (4), 8–36.