Frontiers in Scientific Research and Technology

Online ISSN: 2682-297
Print ISSN: 2682-2962
http://fsrt.journals.ekb.eg

FSRT J 4 (2022) 1 - 6 10.21608/fsrt.2022.123397.1058

Building Footprint Extraction from Low-Resolution Satellite Imagery using Instance Segmentation

Ahmed NourEldeen*[1], Mohamed E. Wahed[2], Yasser Fouad[3], Mohamed S. Metwally[1]

[1]Department of Mathematics, Faculty of Science, Suez University, Suez, Egypt

[2]Faculty of Computers and Informatics, Suez Canal University, Ismailia, Egypt

[3]Faculty of Computers and Informatics, Suez University, Suez, Egypt

ARTICLE INFO

Article history:
Received 7 March 2022
Received in revised form 2 April 2022
Accepted 7 April 2022
Available online 14 April 2022

Keywords
Building footprint;
Convolutional neural networks (CNN);
Instance segmentation;
GIS;
Satellite imagery.

ABSTRACT

Extracting building footprint from aerial photos and satellite imagery has played a vital role in change detection, urban development, and detect the Agricultural land encroachments. The deep neural networks have feature extraction capability and provide the methods to detect and extract building footprint from Satellite imagery with high accuracy. Image segmentation, is the process by that we try to segment an image into coherent parts with two type of segmentation. Semantic segmentation is a form of segmentation that attempts to segment an image into meaningful parts or predefined class labels. The pixel-wise classification task can help us determine if a pixel be included in a particular object in a dataset. Instance segmentation is semantic segmentation with the distinction of classifying each instance of an object as itself. The convolutional neural networks (CNN) used in instance and semantic segmentation. Nevertheless, one of the main problems of extracting building footprint is that most approaches use high-resolution imagery in sampling training data and inferencing phases, whereas not free public available or available with high cost. Or use semantic segmentation that not applicable with closely situated or connected buildings. Our proposed approach is extracting building footprint low-resolution satellite imagery using the instance segmentation technique.

1. Introduction

Building footprint extraction from Satellite imagery is used in many geographic information systems (GIS) solutions such as disaster assessment, geospatial analysis, regional planning, population growth estimation, change detection, etc.

Deep Learning models have become the most technique used for computer vision problems in Satellite imagery and the GIS field [1–3]. The deep learning models use a multi-layer neural network architecture to learn the features with different levels of abstraction [4]. The Deep neural networks utilize a complicated combination of linearity and non-linearity operations to form a layer-by-layer connected architecture to encode features from input data to features extraction. The Convolutional Neural Networks (CNN), are commonly used in computer vision tasks and the object detection is one of this tasks.

The R-CNN model was presented in [5]. Which produce region proposals from the image using search algorithm and feeds these regions into the CNN for feature extraction and support vector machine (SVM) to classify bounding boxes [6]. Fast R-CNN [7] use CNN to extract the features and cropping the region proposals with the feature map to generate the region of interest (RoI). Fast R-CNN uses the fully connected layer and softmax for bounding boxes localization and classification [8]. YOLO [9] used other technique that splits the image into grids with a fixed size, and the CNN is applied on each to predict and classify the bounding boxes. SSD [10] additionally uses decreased sizes of convolution layers for pyramid extraction of multiscale features and detect objects of different sizes.

Recent works [11, 12] used CNN to detect main points of objects like points of the corner or points of the center, and then predict the bounding boxes.

In some real world problem such as car auto-drive , building footprint extraction and etc, we need to detect the exact object boundary (masks) so the segmentation technique it will be better than object detection.

Two types of image segmentation are semantic and instance segmentation. Semantic segmentation is to assign class labels to each image pixel used a fully

* Corresponding authors at Suez University

E-mail addresses: ahmednour_cs@yahoo.com (Ahmed NourEldeen)

Convolutional Network (FCN) [13]. Based on FCN architecture, SegNet [14] and U-Net [15] design an encoder and decoder architecture where the encoder is for down-sampling feature and the decoder is for up-sampling of the feature map.

Instance segmentation can produce semantic pixel-wise labels and predict instance-aware labels that distinguish the individual objects in the same class. Two approaches founded for achieving instance segmentation. The first one performs semantic segmentation overall the image then grouping connected pixels to identify individual objects; this presented by DeepMask [16] and SharpMask [17]. The second one was by Mask R-CNN [18]. Mask R-CNN performs instantiation first then segmentation.

2. Related Works

Early works [19] is a building labeling using CNN, it containing only three layers: one convolution layer for feature extraction, the second pooling layer, and the third is a fully connected layer. Compared this method with the traditional algorithms, it shows better than results, but the simple kind of CNN is effected by the hyper parameter. Recent works use complex CNN architecture. [20] designs a multi layer perceptron (MLP) architecture with a skipping layer connection same as the U-Net architecture to aggregate features. The SegNet is used by [21] to train an additional loss for the distance of the building boundary apart from the pixel-wise classification loss. [22] use the U-Net architecture with multiple constraints that restrict the output comparing with the ground-truth images chips. Other works use the data-fusion technique to increase the performance of segmentation and use the semantic segmentation. [23,24] use the U-Net architecture with satellite imagery and GIS maps such as Google Map, OpenStreetMap, and others to utilize vectorized maps.

Generative Adversarial Networks (GAN) are recently applied to semantic segmentation for building footprint extraction problems. [25] designs a matching-GAN architecture, which modifies the basic GAN model to semantic segmentation tasks. [26] uses the U-Net architecture improved by the GAN model to produce more accurate.

For instance segmentation models, the Mask R-CNN architecture is explored in [27] for extracting building footprint and achieves a satisfying performance of instance segmentation.
[28] enhanced the Mask R-CNN architecture by presenting the rotational of predicted bounding boxes to enhance the quality of detected objects.

In instance segmentation, the image labeling requires annotation for each object with its bounding box and pixel-wise segmentation mask for the sample training dataset. Thus, a limitation of the public availability datasets that is suitable for instance segmentation problems. Many publicly labeled datasets exist for other purpose such as LabelMe [29], ImageNet [30], PASCAL [31], Cityscapes [32], Open Images [33], and Creating Common Object in Context (COCO) [34]. The two most popular methods for annotating objects are COCO and Pascal Visual Object Classes (VOC). However, we haven't a large-scale Satellite imagery datasets with suitable annotations for instance segmentation problems.

The semantic segmentation models are more used for building footprint extraction, mainly the U-Net architecture, recognizing small buildings. By contrast, the instance segmentation models are not fully explored and still provide better solutions because the buildings in the largescale images are usually closely situated or connected. So the instance segmentation models can adequately separate them.

It's possible to segment by detecting and segmenting multiple objects within an image. That can perform using a two pipeline stages. The first stage is generating region proposals for objects in the image using RoI (Region of Interest). Second stage predicts the object class within the bounding boxes and the pixel level mask based. This technique use two convolutional layers. The first is any convolutional base network architectures for extracting the features. The second is the final feature map used in classifying the pixels within the segments [18].

This article proposes a new hyper approach for extracting building footprint from low-resolution satellite imagery using instance segmentation. We used instance segmentation Mask RCNN with CNN backbone 'ResNet-152 architecture [35]', CNN 'PointRend architecture[36]' as a segmentation head and finally fine-tuning process to enhance model results.

3. Methodology

3.1 The feature extraction

The Convolution neural networks are widely used in feature extraction process that use the Convolution layer to extract feature from input image and then use the pooling layer to reduce the extracted feature size for faster processing then passing through the fully connected network to classify the features. We used ResNet-152 [35] for feature extraction process.

3.2 Predict bounding-boxes and mask

Mask R-CNN [18] resolve the problem of instance masks predictions by adding a branch for predicting segmentation masks on each Region of Interest (RoI). We used the PointRend architecture [36] to replace Mask R-CNN's default mask and to improve the bounding-boxes and mask prediction.

3.3 Fine-tuning process

The fine-tuning process is an important process for continues enhancing the deep learning model. So we created two sample training datasets for creating and fine-tuning the deep learning model. We applied the hyper approach (will discuss in details in section 3.4) to create a new deep learning model by adopting the Mask R-CNN instance segmentation using ResNet-152 as a backbone architecture for extracting features over the image, the PointRend architecture as a final feature map used in classifying the pixels within the segments to predict bounding-boxes and mask and finally fine tuning process with second dataset to improve the model results.

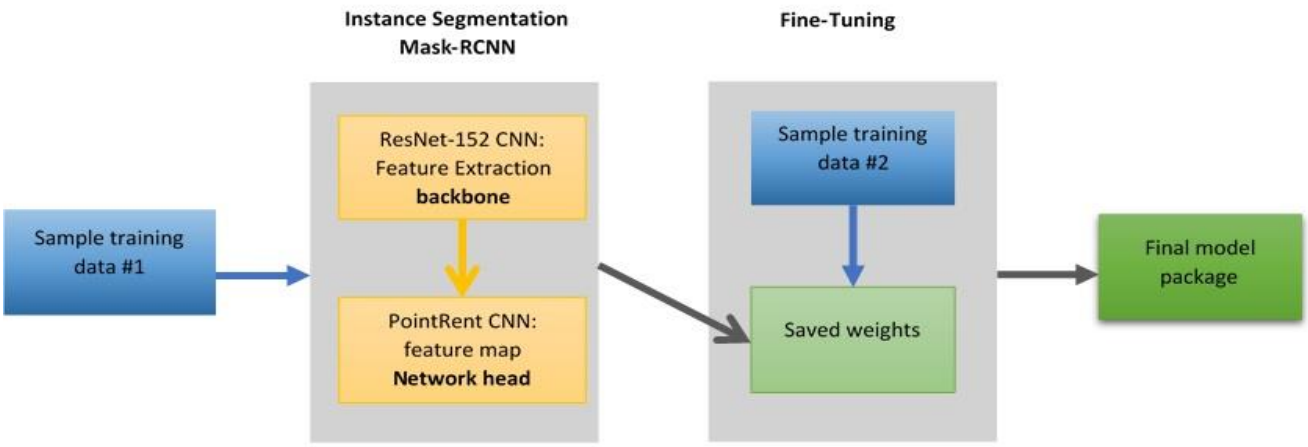

Fig. 1: Proposed model

3.4. Proposed Approach

Figure 1 depicts the proposed model's architecture. The paper methodology is based on four primary phases: phase1 (Input Dataset) and phase2 (feature extraction). Then, in phase3 (feature map technique), phase4 (fine-tuning technique and performance evaluation metrics).

Phase1: Images dataset (First Sample training data)

The used sample training tiles were 400 x 400 pixel image chips. A total of 242 image chips were generated with 3450 building footprint features. We split the dataset to 80 percent for training and 20 percent for validation, and for testing, we used different satellite imagery areas to ensure our model outcomes.

Phase2: Feature extraction

Feature extraction is the core phase in data classification. The ResNet-152 is will be for extracting the features; it is more accurate than other ResNet architectures[34].

Phase 3: Feature map technique,

A feature map is a network head for bounding-box and masks prediction that applied to each RoI. We use PointRent as a feature map architecture with a high outcome with Mask R-CNN.

Phase 4: Fine-tuning technique and performance evaluation metrics,

In this phase, we used saved model weights from training the model during all previous phases and fine-tuning (re-train) the model using a new sample data (Second Sample Dataset) that contains 21 image chips 400 x400 pixels were generated with 199 building footprint features. We split the dataset to 80 percent for training and 20 percent for validation.

For evaluation our model we used standard COCO metric is Average Precision (AP) [18,34,35] and F1 Score. AP is the most commonly used matrix in instance segmentation architectures, that included in the original Mask RCNN research [18].

The average precision value (AP) was calculated using the mean value from 10 IoU thresholds, starting from 0.5 up-to 0.95 with 0.05 steps size. The better model if AP closer to 1. IoU is intersection pixels between the target and prediction masks divided by the entire amount of existing pixels across both masks figure2.

The IoU calculated in other way based on false negative (FN), true negative (TN), false positive (FP), and true positive (TP) as follows:

$$IoU = \frac{TP}{TP + TF + TN}$$

The Intersection over Union (IoU) bounding boxes Figure 2:

Fig .2 Intersection over Union bounding boxes

$$IoU(A,B) = \frac{|A \cap B|}{|A \cup B|}$$

Where A: actual bounding box, green box

B: predicted bounding box, red box

The other matrices used to evaluate the model are recall, precision, F1-score as follows:

$$\text{recall} = \frac{TP}{TP + FN}$$

$$\text{precision} = \frac{TP}{TP + FP}$$

$$\text{F1} - \text{score} = \frac{2 * \text{precision} * \text{recall}}{\text{precision} + \text{recall}}$$

4. Results and Discussion

4.1 Datasets

Since there are no publicly available datasets for the building footprint problem [36] [37], we consider manually annotating of the image tiles to generate a building footprint class. We use a 30m pixel resolution satellite imagery to generate our custom first dataset 400 x 400 image chips Figure 3 (a) and (b). First Dataset contains 242 image chips were generated with 3450 building footprint features. For second dataset we used a 50cm pixel resolution aerial photo to generate 400 x 400 image chips Figure 3 (c) and (d). Second dataset contains 21 image chips were generated with 199 building footprint features.

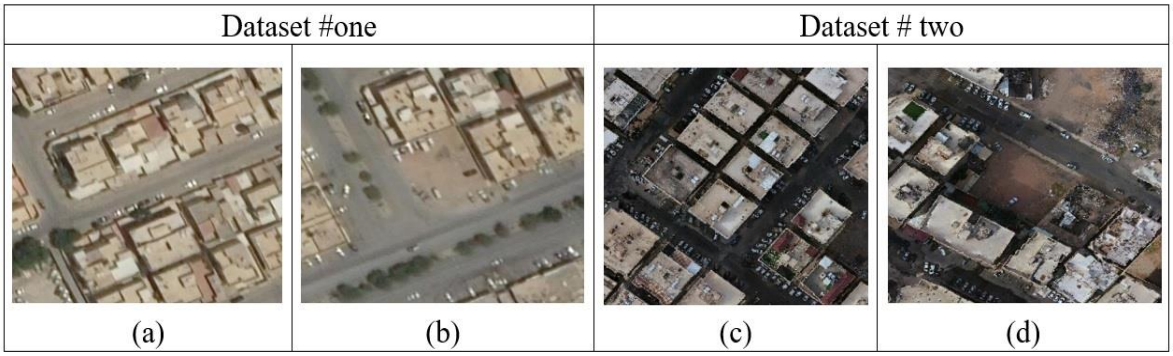

Fig. 3 Training datasets (a) and (b) sample images from first dataset, (c) and (d) sample images from second dataset.

4.2 Training

The experiments machine was an Intel Core i7, 32 GB RAM CPU machine with an onboard NVIDIA GeForce GTX1070 GPU 8 GB memory.

The proposed hyper approach was trained for 50 epochs with batch sizes 4. The cross-entropy loss function and Stochastic Gradient Decent (SGD) were applied for optimization with a momentum of 0.9.

4.3 Evaluation

Our model was tested and on the different geographical areas using low-resolution satellite imagery, and the model outcome was high accuracy for detecting building footprint Figure 4.

We evaluate the proposed model using the primary challenge metric Average Precision value , F1 Score and time for training phase Table 1.

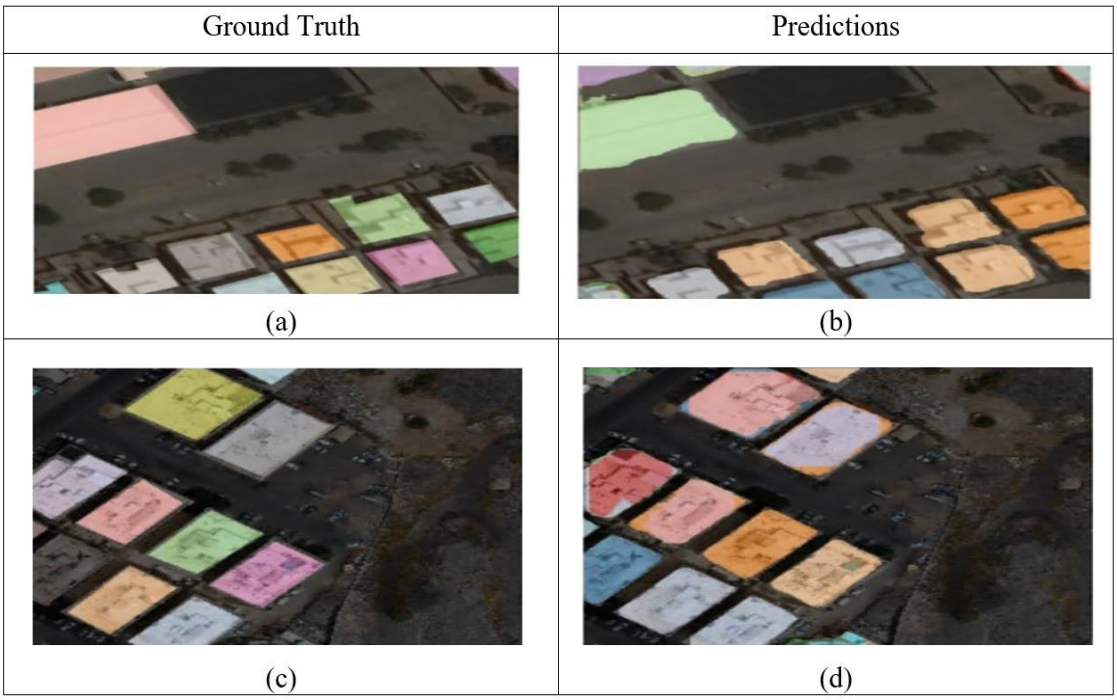

Fig. 4 The deep learning model comparison outcome (a) and (c) are the ground truth images, (b) and (d) are the model prediction images.

Table 1: The model details (number of epochs, AP score , F1 score and training time)

Model	First Dataset			Second Dataset		
Number of epochs	**50 epochs**			**50 epochs**		
	Average Precision (AP%)	F1 Score (%)	Training Time	Average Precision (AP %)	F1 Score (%)	Training Time
Ours	77.567	83.820	104.5 Mins	93.074	91.666	7.5 Mins

Table 2: Comparative AP score between Mask R-CNN with different backbones CNNs

Backbone	Average Precision (AP%)
ResNet50-FPN	70.567
ResNet50-DC5	65.28
ResNet50-C4	67.835
ResNet101-FPN	75.213
ResNet101-DC5	74.408
ResNet101-C4	74.776

Table 3: Comparative AP score between our approach and the most recent approach

Model	Backbone	Average Precision (AP%)
H. Su, S. Wei, etc. [37]	ResNet101	64.801
Carvalho, O.L.F.d. [38]	ResNeXt101	74.776
Ours (dataset # 1)	ResNet-152	77.567
Ours (dataset #2 fine tuning)	ResNet-152	93.074

Table 2 shows the AP scores of building footprint extraction using Mask R-CNN with different backbones CNNs.

4.4 Comparison with Similar works

Table 3 shows the AP score comparison between our model and the most recent approach, our model outperformed the most recent approaches. With a combination Mask RCNN, ResNet-152, PointRent and fine tuning technique, it achieved an AP of 93.074 percent.

5. Conclusion

We proposed a hyper approach for extracting building footprint method. It is a combining an instance segmentation method using Mask R-CNN with ResNet-152 as a backbone architecture used for extracting features over the image and the PointRend architecture as a final feature map is used in classifying the pixels within the segments. The experimental results presented have verified that our proposed model with relatively low training time has achieved an appreciable Average Precision (AP 93.074%), which will have a great potential in remote sensing building footprint extraction.

References

[1] Zhu, X.X.; Tuia, D.; Mou, L.; Xia, G.; Zhang, L.; Xu, F.; Fraundorfer, F. 2017, Deep Learning in Remote Sensing: A Comprehensive Review and List of Resources. IEEE Geosci. Remote Sens. Mag

[2] Liu, L.; Ouyang, W.; Wang, X.; Fieguth, P.; Chen, J.; Liu, X.; Pietikäinen, M. 2020, Deep Learning for Generic Object Detection: A Survey. Int. J. Comput. Vis., 128, 261–318.

[3] Zhang, L.; Zhang, L.; Du, B. 2016, Deep learning for remote sensing data: A technical tutorial on the state of the art. IEEE Geosci. Remote Sens. Mag. 4, 22–40.

[4] Y. LeCun, Y. Bengio, and G. Hinton. 2015, Deep learning, Nature, vol. 521, pp. 436– 444

[5] R. Girshick, J. Donahue, T. Darrell, and J. Malik. Rich. 2014, feature hierarchies for accurate object detection and semantic segmentation, in The IEEE Conference on Computer Vision and Pattern Recognition (CVPR).

[6] C. Cortes and V. Vapnik. 1995, Support-vector networks, in Machine Learning.

[7] R. Girshick. 2015, Fast R-CNN. in The IEEE International Conference on Computer Vision (ICCV).

[8] J. S. Bridle. 1990, Training stochastic model recognition algorithms as networks can lead to maximum mutual information

estimation of parameters. Advances in Neural Information Processing Systems 2.

[9] J. Redmon, S. Divvala, R. Girshick, and A. Farhadi. 2016, You only look once: Unified, real-time object detection. in The IEEE Conference on Computer Vision and Pattern Recognition (CVPR).

[10] W. Liu, D. Anguelov, D. Erhan, C. Szegedy, S. Reed, C.-Y. Fu, and A. C. Berg. 2016, Ssd: Single shot multibox detector. in The European Conference on Computer Vision (ECCV).

[11] H. Law and J. Deng. 2018, Cornernet: Detecting objects as paired keypoints. in The European Conference on Computer Vision (ECCV).

[12] X. Zhou, J. Zhuo, and P. Krahenbuhl. 2019, Bottom-up object detection by grouping extreme and center points. in The IEEE Conference on Computer Vision and Pattern Recognition (CVPR),.

[13] E. Shelhamer, J. Long, and T. Darrell. 2017, Fully convolutional networks for semantic segmentation. IEEE Transactions on Pattern Analysis and Machine Intelligence, vol. 39, no. 4, pp. 640–651.

[14] V. Badrinarayanan, A. Kendall, and R. Cipolla. Segnet, 2017. A deep convolutional encoder-decoder architecture for image segmentation. IEEE Transactions on Pattern Analysis and Machine Intelligence,.

[15] O. Ronneberger, P.Fischer, and T. Brox. U-net, 2015. Convolutional networks for biomedical image segmentation. in Medical Image Computing and ComputerAssisted Intervention (MICCAI), ser. LNCS, vol. 9351,.

[16] P. O. Pinheiro, R. Collobert, and P. Dollar, , 2015. Learning to segment object candidates. in Advances in Neural Information Processing Systems.

[17] P. O. Pinheiro, T.-Y. Lin, R. Collobert, and P. Dollar,2016. Learning to refine object segments. in The European Conference on Computer Vision (ECCV),.

[18] K. He, G. Gkioxari, P. Dollar, and R. B. Girshick, 2017. Mask r-cnn. in The IEEE International Conference on Computer Vision (ICCV),.

[19] V. Mnih, 2013. Machine learning for aerial image labeling. Ph.D. dissertation, University of Toronto.

[20] E. Maggiori, Y. Tarabalka, G. Charpiat, and P. Alliez, 2017. High-resolution aerial image labeling with convolutional neural networks. IEEE Transactions on Geoscience and Remote Sensing, vol. 55, no. 12, pp. 7092–7103.

[21] B. Bischke, P. Helber, J. Folz, D. Borth, and A. Dengel. Multi-task learning for segmentation of building footprints with deep neural networks. ArXiv, vol. abs/1709.05932, 2017.

[22] G. Wu, X. Shao, Z. Guo, Q. Chen, W. Yuan, X. Shi, Y. Xu, and R. Shibasaki. Automatic building segmentation of aerial imagery using multi-constraint fully convolutional networks. Remote Sensing, vol. 10, no. 3, 2018.

[23] N. Audebert, B. Le Saux, and S. Lefevre. Joint learning from earth observation and openstreet map data to get faster better semantic maps. in The IEEE Conference on Computer Vision and Pattern Recognition Workshops (CVPRW), 2017.

[24] W. Li, C. He, J. Fang, J. Zheng, H. Fu, and L. Yu. Semantic segmentation based building footprint extraction using very high-resolution satellite images and multi-source gis data. Remote Sensing, vol. 11, no. 4, 2019.

[25] G. Mattyus and R. Urtasun. Matching adversarial networks. in The IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2018.

[26] X. Pan, F. Yang, L. Gao, Z. Chen, B. Zhang, H. Fan, and J. Ren. Building extraction from high-resolution aerial imagery using a generative adversarial network with spatial and channel attention mechanisms. Remote Sensing, vol. 11, no. 8, 2019.

[27] K. Zhao, J. Kang, J. Jung, and G. Sohn. Building extraction from satellite images using mask r-cnn with building boundary regularization. in The IEEE Conference on Computer Vision and Pattern Recognition Workshops (CVPRW), 2018.

[28] Q. Wen, K. Jiang, W. Wang, Q. Liu, Q. Guo, L. Li, and P. Wang. Automatic building extraction from google earth images under complex backgrounds based on deep instance segmentation network. Sensors, vol. 19, no. 2, 2019.

[29] Russell, B.C.; Torralba, A.; Murphy, K.P.; Freeman, W.T. LabelMe: A Database and Web-Based Tool for Image Annotation. Int. J. Comput. Vis. 2008, 77, 157–173.

[30] Deng, J.; Dong, W.; Socher, R.; Li, L.-J.; Kai, L.; Fei-Fei, L. ImageNet: A large-scale hierarchical image database. In Proceedings of the 2009 IEEE Conference on Computer Vision and Pattern Recognition, Miami, FL, USA, 22–24, 2009; pp. 248–255.

[31] Everingham, M.; Eslami, S.M.A.; Van Gool, L.; Williams, C.K.I.; Winn, J.; Zisserman, A. The Pascal Visual Object Classes Challenge: A Retrospective. Int. J. Comput. Vis. 2015, 111, 98–136.

[32] Cordts, M.; Omran, M.; Ramos, S.; Rehfeld, T.; Enzweiler, M.; Benenson, R.; Franke, U.; Roth, S.; Schiele, B. The Cityscapes Dataset for Semantic Urban Scene Understanding. IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Las Vegas, NV, USA,2016; Volume 29, pp. 3213–3223.

[33] Kuznetsova, A.; Rom, H.; Alldrin, N.; Uijlings, J.; Krasin, I.; Pont-Tuset, J.; Kamali, S.; Popov, S.; Malloci, M.; Kolesnikov, A.; et al. The Open Images Dataset V4. Int. J. Comput. Vis. 2020, 128, 1956–1981.

[34] Lin, T.-Y.; Maire, M.; Belongie, S.; Hays, J.; Perona, P.; Ramanan, D.; Dollár, P.; Zitnick, C.L. Microsoft COCO: Common Objects in Context. In Proceedings of the 13th European Conference on Computer Vision (ECCV 2014), Zurich, Switzerland, 6–12 2014; Volume 8693, pp. 740–755, ISBN 978-3-319-10601-4.

[35] Kaiming He, Xiangyu Zhang, Shaoqing Ren, Jian Sun. Deep Residual Learning for Image Recognition. arXiv:1512.03385v1 [cs.CV] 10 2015.

[36] Alexander Kirillov, Yuxin Wu, Kaiming He, Ross Girshick. PointRend: Image Segmentation as Rendering. arXiv:1912.08193v2 [cs.CV] 2020.

[37]H. Su, S. Wei, M. Yan, C. Wang, J. Shi and X. Zhang, "Object Detection and Instance Segmentation in Remote Sensing Imagery Based on Precise Mask R-CNN," IGARSS 2019 - 2019 IEEE International Geoscience and Remote Sensing Symposium, 2019, pp. 1454-1457,

[38] Carvalho, O.L.F.d.; de Carvalho Júnior, O.A.; Albuquerque, A.O.d.; Bem, P.P.d.; Silva, C.R.; Ferreira, P.H.G.; Moura, R.d.S.d.; Gomes, R.A.T.; Guimarães, R.F.; Borges, D.L. Instance Segmentation for Large, Multi-Channel Remote Sensing Imagery Using Mask-RCNN and a Mosaicking Approach. Remote Sens. 2021, 13, 39. https://dx.doi.org/10.3390/rs13010039.

[39] I. M. Bello, K. Zhang, J. Wang and M. A. Aslam. A Multiscale Segmentation Framework for Uncompleted Building Footprint Extraction from Remote Sensing Imagery. IEEE Asia-Pacific Conference on Geoscience, Electronics and Remote Sensing Technology (AGERS), 2021, pp. 119-124.