Extracting building outlines based on convolutional neural networks using the property of linear connectivity

Anton Emelyanov[1,2], Vladimir A. Knyaz[1,2], Vladimir V. Kniaz [1,2]

[1] Moscow Institute of Physics and Technology (MIPT), Russia - anton.emelyanov@phystech.edu
[2] State Research Institute of Aviation System (GosNIIAS), 125319 Moscow, Russia - (kniaz.va,kniaz.vv)@mipt.ru

Keywords: Deep learning, Instance segmentation, Boundary regularization, Vectorization, Remote sensing images.

Abstract

For years, researchers have been developing an automated method that can replace humans by drawing the outlines of individual buildings in a vector format, which plays an important role in GIS creation, environmental monitoring, urban planning, population density estimation, and energy supply. There is no doubt that this is an extremely difficult task, not only because of the labor required to develop such a highly intelligent algorithm, but also because of the challenges posed by imperfect imaging conditions, different building structures, and the complexity of the background. One of the current challenges in extracting building outlines is to accurately recreate the polygonal boundaries of the building while extracting vectorized building masks as output for direct use in various applications. This work provides a comprehensive workflow for building extraction and improves the predicted area of buildings through boundary regularization. First, a convolutional neural network is used to train instance segmentation model, then regularization and vectorization processes are performed. The main difference from existing methods is a new regularization method based on the concepts of linear connectivity and convexity of a set of points. This approach can effectively identify and remove points that do not belong to the detected building but were incorrectly segmented by the algorithm. Based on the results of experiments, the algorithm showed a high level of efficiency, comparable to leading methods for extracting building boundaries as PolyWorld.

1. Introduction

For many years, researchers have been developing an automated method that can replace humans for mapping vector format outlines of individual buildings, which play an important role in GIS production, environmental monitoring, urban planning, population density estimation and energy supply. Undoubtedly, this is an extremely difficult task, not only due to the laboriousness of developing such a highly intelligent algorithm, but also due to the challenges associated with imperfect imaging conditions, varied building architecture, and background complexity.

Automatic detection of buildings from aerial photographs has been considered an important means of improving the efficiency of vector map generation for decades (Paparoditis et al., 1998, Persson et al., 2005, Yang et al., 2018). In recent years, with the support of extensive training data and sufficient computing power, deep learning methods such as convolutional neural networks (CNN) (LeCun et al., 1989) and fully convolutional networks (FCN) (Long et al., 2014) significantly improved the accuracy of building detection from remote sensing images (Li et al., 2019, Chen et al., 2020, Šanca et al., 2023). However, automatically generating high-quality vector building maps from aerial photographs is not yet a reality for most communities. This is partly because deep learning-based building detection approaches still face challenges such as low recognition rates of roofs obscured by trees or shadows (Chen et al., 2019) and relatively poor generalization ability for certain geographic regions to others (Maggiori et al., 2017). One of the current challenges in extracting building outlines is to accurately recreate the polygonal boundary of a building while extracting a vectorized building mask as output for direct use in various applications.

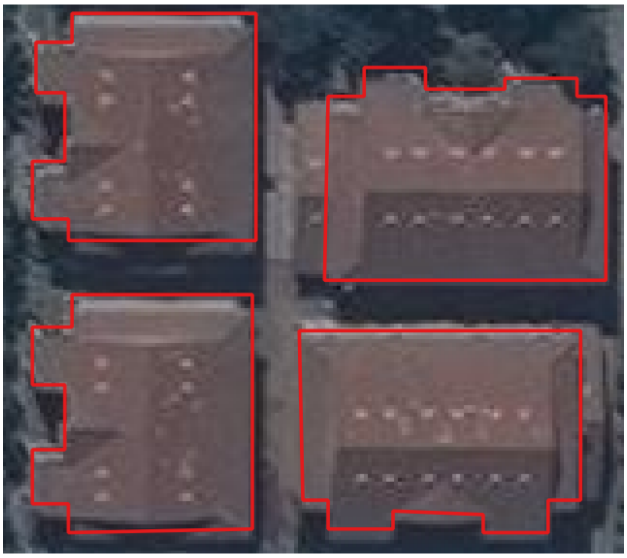

Figure 1. Example of extracting a building boundary.

This paper proposes the algorithm for automatically extracting building outlines based on instance segmentation, regularization and vectorization. The main difference from existing methods is a new regularization method based on the concepts of linear connectivity and convexity of a set of points. This approach can effectively identify and remove points that do not belong to the detected building but were incorrectly segmented by the algorithm. In summary, the main contributions of this paper are as follows:

- We explore the use of the linear connectivity property to identify “extra” pixels in the instance segmentation process.

- To analyze the results and compare them with existing methods for highlighting building boundaries, one of the most popular datasets for vectorization, CrowdAI (Mohanty et al., 2020), is used.

2. Related work

Currently, the leading building extraction approaches are semantic segmentation and instance segmentation methods.

2.1 Neural network methods

Since building predictions are made at the highest resolution, holes may appear in the large-scale building predictions if the global semantic information is insufficient, while small-scale buildings may be omitted without enough local details. To address these issues, (Wei et al., 2019) introduced a multi-scale aggregation FCN that fuses multi-scale building features to generate final building predictions. The PolygonCNN proposed by (Chen et al., 2020) first performs segmentation to extract initial building outlines. Then, it utilizes a modified PointNet to learn the shape prior and predict polygon vertices to generate precise building vector results by encoding the vertices of building polygons and merging image features extracted from the segmentation step. (Šanca et al., 2023) propose an end-to-end workflow that utilizes binary semantic segmentation, regularization, and vectorization. The novelty of their approach is applying the regularization task on an entirely new building dataset, while adding their own implementation for the vectorization part. The study (Knyaz et al., 2020) proposed masking technique for robust segmentation of the repeated structures in images. Such approach allowed to improve segmentation performance for 11%.

In (Zhao et al., 2018) the authors corrected the segmentation masks produced with Mask R-CNN by first simplifying the detected boundaries using the Douglas-Peucker algorithm and subsequently refining the resulting polygons using a Minimum Descriptor Length method. Aiming at the problem that the quality of detection affects the integrity of the mask, (Zhao et al., 2020) proposed an instance segmentation model for the accuracy of segmentation contours, which used detection and segmentation as a multi-stage process to obtain accurate segmentation edges and improve the geometric regularity of the segmentation results.

This methods create segmentation maps at the pixel level, but the building boundaries produced by the algorithms are usually zigzag and far removed from manual delineation of objects. In addition, the results require extensive post-processing: semantic segmentation cannot distinguish between individual buildings, and the bounding box predicted by the instance segmentation method may contain elements of other buildings, making mask training difficult. However, geographic and cartographic applications typically require precise vector polygons of extracted objects instead of rasterized output. (Zorzi et al., 2022) introduces PolyWorld, a neural network that directly extracts building vertices from an image and connects them correctly to create precise polygons.

A few other studies (Ling et al., 2019, Peng et al., 2020, Liu et al., 2021, Wei et al., 2022) have considered the instance segmentation problem as contour regression, i.e., regressing the vertex coordinates of a contour (in other words, a polygon represented by a series of discrete vertices). The contour-based methods are theoretically advantageous in efficiency since they straightforwardly regress the polygon coordinates, compared to semantic/instance segmentation with a pixelwise operation, and have the potential to get rid of the need for post-processing operations such as raster-to-vector conversion and empirical regularization.

2.2 Datasets

Since the Zeebrugge dataset (Campos-Taberner et al., 2016) was published as part of the 2015 IEEE GRSS Data Fusion Contest, dozens of building detection and segmentation datasets have been released. It is worth noting that the datasets used to evaluate traditional methods are usually small in size, and the training and testing sets are collected from the same local region (or image), resulting in poor generalization ability. In the era of deep learning, more advanced datasets can achieve spatial independence of training and test sets, wider spatial coverage and larger data volume, which corresponds to reality.

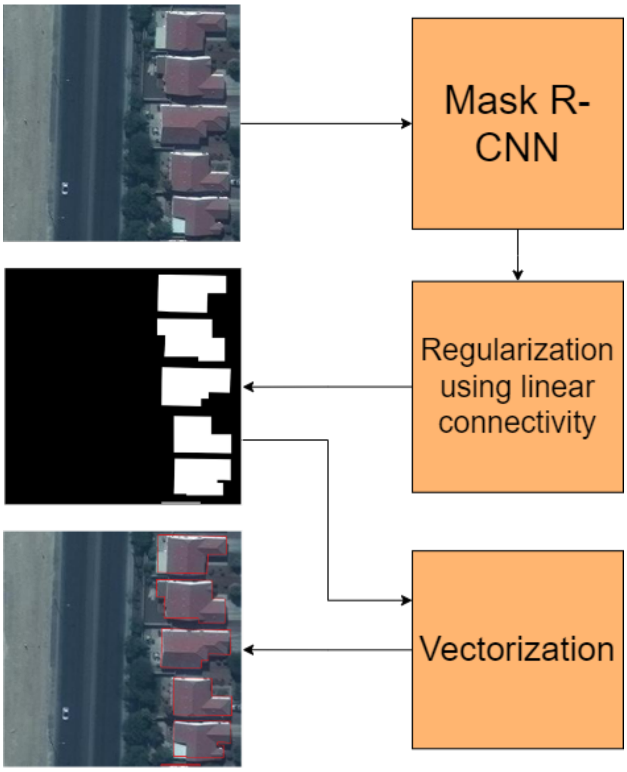

Figure 2. The structure of the proposed algorithm.

Considering the size of buildings ($> 10m^2$), we note some benchmark satellite data/aerial imagery sets (Rottensteiner et al., 2012, Ji et al., 2019, Yang et al., 2022, Tian et al., 2020, Mohanty et al., 2020), most of which have spatial resolution ranging from the centimeter level to 2 m, with the exception of the relatively coarse resolution of SpaceNet 7 (4 m). In addition to the commonly used RGB channels, some datasets also provide additional useful information to further image buildings. In terms of spectral information, the Potsdam and WHU-Satellite datasets have RGB/near-infrared (NIR) bands, and the SpaceNet and SpaceNet 4 datasets consist of eight spectral bands from the WorldView 2/3 sensors. For vertical information, the Potsdam, Vaihingen, Zeebrugge and DFC19-JAX datasets provide airborne LiDAR derived nDSMs, while the SpaceNet 4 dataset consists of 27 unique images for which viewing angles range from -32.5^0 to 54.0^0 (Weir et al., 2019). Several datasets (e.g., DFC19-JAX) have also attempted to improve deep learning networks by combining planar and stereo remote sensing observations. In terms of temporal properties, the WHU Building Change Detection, SECOND, Hi-UCD, and ZKXT_2021 datasets contain multi-temporal remote sensing observations, building contours for each date, and building change records.

3. Method

This work provides a comprehensive workflow for building extraction and improves the predicted area of buildings through boundary regularization. First, a convolutional neural network is used to train instance segmentation model. Second, the property of linear set connectivity is used to organize the predicted contours of buildings and improve their geometry. The final step is the vectorization process, converting the regularized building masks into polygons for use in any application. The scheme of the algorithm is shown in Figure 2.

3.1 Instance segmentation with Mask R-CNN

The initial stage of our methodology involves identifying and delineating the boundaries of buildings depicted in aerial photographs. The neural network Mask R-CNN was used to perform this task.

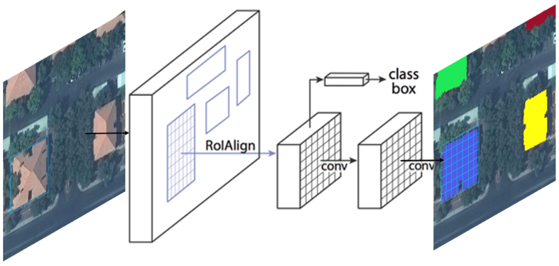

Figure 3. The structure of the Mask R-CNN.

Mask R-CNN (He et al., 2018) extends Faster R-CNN (Ren et al., 2016) by adding a branch for predicting segmentation masks on each Region of Interest (RoI), in parallel with the existing branch for classification and bounding box regression. The mask branch is a small fully convolutional network (FCN) applied to each RoI, predicting a segmentation mask in a pixel-to-pixel manner. Mask R-CNN is simple to implement and train due to the Faster R-CNN framework, which facilitates a wide range of flexible architecture designs. Additionally, the mask branch only adds a small computational overhead, enabling a fast system and rapid experimentation.

The Adam optimizer with Binary Cross Entropy Loss with logits was used during training to measure the difference between the predicted result and the ground truth. The loss function is defined as:

$$L = -\frac{1}{N}\sum_{i=1}^{N}[x_i log(\sigma(y_i)) + (1 - x_i) log(1 - \sigma(y_i))] \quad (1)$$

where N is the batch size, x_i is the ground truth image for sample i, y_i is the logit output of the model for sample i and σ is the Sigmoid function. A Sigmoid function is any mathematical function whose graph has a characteristic S-shaped curve (sigmoid curve). For the sigmoid function we use logistic function, which is defined as:

$$\sigma(x) = \frac{1}{1 + e^{-x}} \quad (2)$$

3.2 Applying regularization on predictions

Once the predictions are generated using the trained model, a post-processing step applies regularization to further improve the geometry and accuracy of the predicted building masks. Since pixel-based classification results in rounded corners and closed-edge predictions, regularization is an important step to further improve predictions. Also, after the segmentation process, the predicted bounding box may contain additional instances, which makes it difficult to train the mask head of the network.

Considering that the identification of building boundaries is carried out on remote sensing images, we will assume that the buildings in the images do not intersect or overlap each other. In this case, each building is a closed limited set of pixels. Thus, using the property of linear connectivity, unnecessary points that do not belong to the main object are removed from the bounding rectangles.

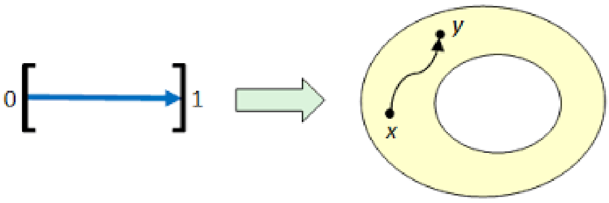

Figure 4. Schematic definition of path-connected set.

In topology and related branches of mathematics, a connected space is a topological space that cannot be represented as the union of two or more disjoint non-empty open subsets. A path-connected space is a stronger notion of connectedness, requiring the structure of a path. A path from a point x to a point y in a topological space X is a continuous function f from the unit interval [0,1] to X with $f(0) = x$ and $f(1) = y$. A path-component of X is an equivalence class of X under the equivalence relation which makes x equivalent to y if there is a path from x to y. The space X is said to be path-connected if there is exactly one path-component. For non-empty spaces, this is equivalent to the statement that there is a path joining any two points in X. The definition of a path-connected set is similar to the definition for a space.

Thus, the developed algorithm has the following structure:

1. Detection of objects of the ”building” class in the image
2. Segmentation of objects of the ”building” class in each identified bounding box
3. Find and remove points that do not belong to the main building in the bounding box, but are segmented as ”building”
4. Vectorizing the resulting images

It can be presented as Algorithm 1.

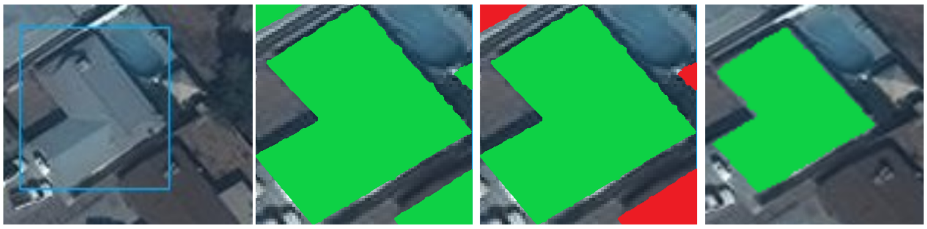

Figure 5. Using the linear connectivity property to remove unnecessary points from images. In the first image, an object of the “building” class is detected. On the second, objects of the “building” class are segmented in the bounding box. On the third, pixels that do not belong to the main object, but are erroneously segmented, are highlighted (in red) and deleted. The last image contains the final result of segmentation and regularization.

```
Algorithm 1: Regularization using properties of a path-connected set of points
Input:
 An image with set of points belonging to the class
 ”building” $\mathbf{B} = \{x_i\}$,
Output:
 An image with set of points belonging to the class
 ”building” $\mathbf{B} = \{x_i'\}$,

Search and removal unnecessary points from the set of points B ;

Procedure Search($\boldsymbol{B}$):
    for each point $x_i$ of class ”building” B do
        Construct a straight line $L_i$ passing through points
          $x_i$ and $x_0$, where $x_0$ is the central point of the
          bounding box
        if $\exists x_{extra} \notin \boldsymbol{B}_0$, but $x_{extra} \in L_i$ then
            draw a straight line $L_{i_1}$ to the last one,
              belonging to the class “building” $x_1$
            take points $x_{1_n}$ from the unit neighborhood $x_1$,
              belonging to the class “building” ;
            draw straight lines to from points $x_{1_n}$ to $x_i$;
            repeat until polygonal chain PC appears
              connecting $x_0$ and $x_i$;
            if $\exists PC$ then
                skip;
            else
                delete $x_i$ from B
        else
            skip;
    return B;
```

4. Results

4.1 Evaluation metrics

Similarly to (Zorzi et al., 2022) we use the following evaluation metrics.

Intersection-over-Union (IoU) or the Jaccard index, is the ratio of the intersection area of the predicted and ground truth mask to their union:

$$IoU = \frac{Intersection}{Union} = \frac{TP}{TP + FP + FN} \quad (3)$$

Also precision and recall were calculated to determine average precision (AP) and average recall (AR) values:

$$Precision = \frac{TP}{TP + FP} \quad (4)$$

$$Recall = \frac{TP}{TP + FN} \quad (5)$$

where TP , FP , FN are the true positive, false positive and false negative of the building class.

4.2 Experiment

The developed algorithm was trained on the open CrowdAI Mapping Challenge database (Mohanty et al., 2020), which is composed of over 280k satellite images for training and 60k images for testing. The training images were divided into two parts: 80% of the images were used to train the algorithm, 20% for validation. The training was performed locally with CUDA 11.7 on an NVIDIA GeForce RTX 3070 graphics card with 8 GB of memory.

Figure 6. Example images from the CrowdAI Mapping Challenge dataset.

The meaning of the calculated metrics is given in Table 1. To understand the level of efficiency of the algorithm, the table also includes the results of the leading methods on similar data.

Method	AP	AR	IoU
Mask R-CNN	41.9	47.6	-
PolyMapper	55.7	62.1	-
PolyWorld	63.3	75.4	91.3
LC(our method)	65.2	74.9	91.4

Table 1. Results on the CrowdAI test dataset for all the building extraction and polygonization experiments.

5. Conclusion

The main goal of our study was to develop an end-to-end workflow for extracting building outlines using instance segmentation, linear connectivity-based regularization, and vectorization. We concluded that regularization using the linear connectivity property improves segmentation accuracy by an average of 23.3 in AP(average precision) and 27.3 in AR(average

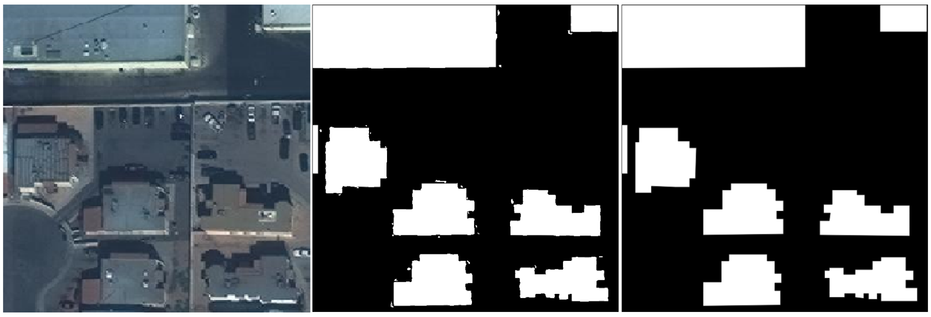

Figure 7. Example of the resulting segmented images before and after the regularization process.

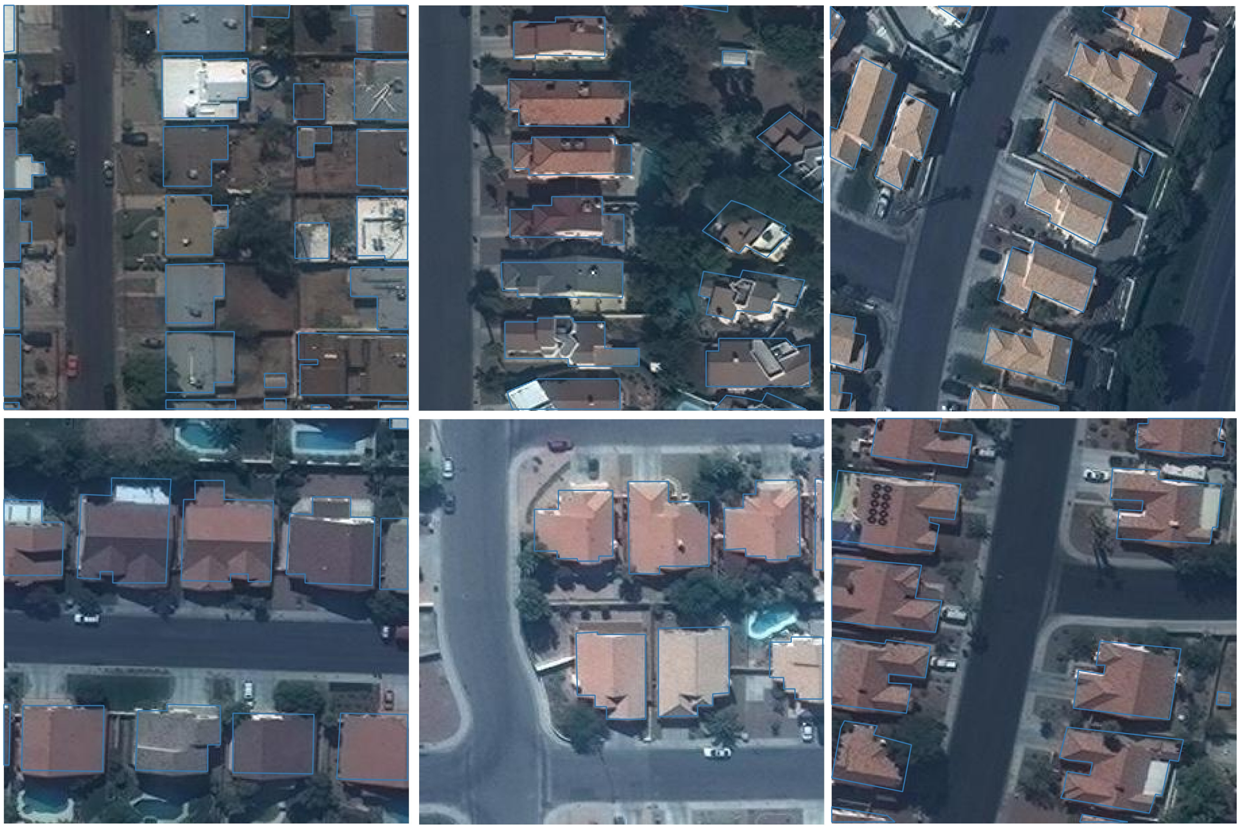

Figure 8. Experiment results.

recall). Regularization not only improves predictions, but also improves the geometric shape of building outlines. Based on the results of experiments, the algorithm showed a high level of efficiency, comparable to leading methods for extracting building boundaries as PolyWorld (Zorzi et al., 2022).

6. Acknowledgement

The research was carried out at the expense of a grant from the Russian Science Foundation No. 24-21-00269, https://rscf.ru/project/24-21-00269/

References

Campos-Taberner, M., Romero-Soriano, A., Gatta, C., Camps-Valls, G., Lagrange, A., Le Saux, B., Beaupère, A., Boulch, A., Chan-Hon-Tong, A., Herbin, S., Randrianarivo, H., Ferecatu, M., Shimoni, M., Moser, G., Tuia, D., 2016. Processing of Extremely High-Resolution LiDAR and RGB Data: Outcome of the 2015 IEEE GRSS Data Fusion Contest–Part A: 2-D Contest. *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing*, 9(12), 5547-5559.

Chen, Q., Wang, L., Waslander, S. L., Liu, X., 2020. An end-to-end shape modeling framework for vectorized building outline generation from aerial images. *ISPRS Journal of Photogrammetry and Remote Sensing*, 170, 114-126. https://www.sciencedirect.com/science/article/pii/S092427162030280X.

Chen, Q., Wang, L., Wu, Y., Wu, G., Guo, Z., Waslander, S. L., 2019. TEMPORARY REMOVAL: Aerial imagery for roof segmentation: A large-scale dataset towards automatic mapping of buildings. *ISPRS Journal of Photogrammetry and Remote Sensing*, 147, 42-55. https://www.sciencedirect.com/science/article/pii/S0924271618303083.

He, K., Gkioxari, G., Dollár, P., Girshick, R., 2018. Mask r-cnn.

Ji, S., Wei, S., Lu, M., 2019. Fully Convolutional Networks for Multisource Building Extraction From an Open Aerial and Satellite Imagery Data Set. *IEEE Transactions on Geoscience and Remote Sensing*, 57(1), 574-586.

Knyaz, V. A., Kniaz, V. V., Remondino, F., Zheltov, S. Y., Gruen, A., 2020. 3D Reconstruction of a Complex Grid Structure Combining UAS Images and Deep Learning. *Remote Sensing*, 12(19). https://www.mdpi.com/2072-4292/12/19/3128.

LeCun, Y., Boser, B., Denker, J. S., Henderson, D., Howard, R. E., Hubbard, W., Jackel, L. D., 1989. Backpropagation Applied to Handwritten Zip Code Recognition. *Neural Computation*, 1(4), 541-551. https://doi.org/10.1162/neco.1989.1.4.541.

Li, Z., Wegner, J. D., Lucchi, A., 2019. Topological map extraction from overhead images. *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*.

Ling, H., Gao, J., Kar, A., Chen, W., Fidler, S., 2019. Fast interactive object annotation with curve-gcn. *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*.

Liu, Z., Liew, J. H., Chen, X., Feng, J., 2021. Dance: A deep attentive contour model for efficient instance segmentation. *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV)*, 345–354.

Long, J., Shelhamer, E., Darrell, T., 2014. Fully Convolutional Networks for Semantic Segmentation. *CoRR*, abs/1411.4038. http://arxiv.org/abs/1411.4038.

Maggiori, E., Tarabalka, Y., Charpiat, G., Alliez, P., 2017. Can semantic labeling methods generalize to any city? the inria aerial image labeling benchmark. *2017 IEEE International Geoscience and Remote Sensing Symposium (IGARSS)*, 3226–3229.

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. *Frontiers in Artificial Intelligence*, 3.

Paparoditis, N., Cord, M., Jordan, M., Cocquerez, J.-P., 1998. Building Detection and Reconstruction from Mid- and High-Resolution Aerial Imagery. *Computer Vision and Image Understanding*, 72(2), 122-142. https://www.sciencedirect.com/science/article/pii/S1077314298907226.

Peng, S., Jiang, W., Pi, H., Li, X., Bao, H., Zhou, X., 2020. Deep snake for real-time instance segmentation. *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*.

Persson, M., Sandvall, M., Duckett, T., 2005. Automatic building detection from aerial images for mobile robot mapping. *2005 International Symposium on Computational Intelligence in Robotics and Automation*, 273–278.

Ren, S., He, K., Girshick, R., Sun, J., 2016. Faster r-cnn: Towards real-time object detection with region proposal networks.

Rottensteiner, F., Sohn, G., Jung, J., Gerke, M., Baillard, C., Benitez, S., Breitkopf, U., 2012. THE ISPRS BENCHMARK ON URBAN OBJECT CLASSIFICATION AND 3D BUILDING RECONSTRUCTION. *ISPRS Annals of the Photogrammetry, Remote Sensing and Spatial Information Sciences*, I-3, 293–298. https://isprs-annals.copernicus.org/articles/I-3/293/2012/.

Tian, S., Ma, A., Zheng, Z., Zhong, Y., 2020. Hi-ucd: A large-scale dataset for urban semantic change detection in remote sensing imagery.

Šanca, S., Jyhne, S., Gazzea, M., Arghandeh, R., 2023. AN END-TO-END DEEP LEARNING WORKFLOW FOR BUILDING SEGMENTATION, BOUNDARY REGULARIZATION AND VECTORIZATION OF BUILDING FOOTPRINTS. *The International Archives of the Photogrammetry, Remote Sensing and Spatial Information Sciences*, XLVIII-4/W7-2023, 169–175. https://isprs-archives.copernicus.org/articles/XLVIII-4-W7-2023/169/2023/.

Wei, S., Ji, S., Lu, M., 2019. Toward automatic building footprint delineation from aerial images using CNN and regularization. *IEEE Transactions on Geoscience and Remote Sensing*, 58(3), 2178–2189.

Wei, S., Zhang, T., Ji, S., 2022. A Concentric Loop Convolutional Neural Network for Manual Delineation-Level Building Boundary Segmentation From Remote-Sensing Images. *IEEE Transactions on Geoscience and Remote Sensing*, 60, 1-11.

Weir, N., Lindenbaum, D., Bastidas, A., Etten, A. V., McPherson, S., Shermeyer, J., Kumar, V., Tang, H., 2019. Spacenet mvoi: A multi-view overhead imagery dataset. *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*.

Yang, H. L., Yuan, J., Lunga, D., Laverdiere, M., Rose, A., Bhaduri, B., 2018. Building Extraction at Scale Using Convolutional Neural Network: Mapping of the United States. *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing*, 11(8), 2600-2614.

Yang, K., Xia, G.-S., Liu, Z., Du, B., Yang, W., Pelillo, M., Zhang, L., 2022. Asymmetric Siamese Networks for Semantic Change Detection in Aerial Images. *IEEE Transactions on Geoscience and Remote Sensing*, 60, 1-18.

Zhao, K., Kang, J., Jung, J., Sohn, G., 2018. Building extraction from satellite images using mask r-cnn with building boundary regularization. *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR) Workshops*.

Zhao, W., Persello, C., Stein, A., 2020. Building instance segmentation and boundary regularization from high-resolution remote sensing images. *IGARSS 2020 - 2020 IEEE International Geoscience and Remote Sensing Symposium*, 3916–3919.

Zorzi, S., Bazrafkan, S., Habenschuss, S., Fraundorfer, F., 2022. Polyworld: Polygonal building extraction with graph neural networks in satellite images. *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 1848–1857.

This contribution has been peer-reviewed.
https://doi.org/10.5194/isprs-archives-XLVIII-1-2024-147-2024 | © Author(s) 2024. CC BY 4.0 License.