Pixels relationship analysis for extracting building footprints

Anton Emelyanov[1,2], Vladimir Knyaz[1,2], Vladimir Kniaz[1,2], Dana Artist[3]

[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
[3] Moscow State University of Geodesy and Cartography (MIIGAiK), Moscow, Russia - dana_artist@hotmail.com

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

Abstract

Currently, remote sensing research is focused on developing an automated algorithm that can compete with empirical methods for mapping the contours of individual buildings. Despite facing numerous challenges related to suboptimal imaging conditions, diverse building architecture, and complex backgrounds, creating such an algorithm is essential for monitoring urban and natural areas, generating 3D city models, managing disasters, and estimating population density. Obtaining the polygonal boundary of a building and extracting a vectorized building mask as output for direct use is one of the current challenges in drawing building outlines. This work provides a comprehensive workflow for building extraction and improves their predicted area by regularization the boundaries of buildings. First, a convolutional neural network is used to train the binary semantic segmentation model and then regularization and vectorization processes are performed. The main difference from existing methods is a new regularization method based on compiling a neighborhood matrix for each point belonging to the “building” class. According to the experimental results, the algorithm shows high efficiency: IOU (intersection over union) = 91.2%, AP (average precision) = 64.1% and AR (average recall) = 75.1%, comparable to leading building boundary extraction methods such as PolyWorld.

1. Introduction

Leading contemporary research in remote sensing is aimed at developing an automated algorithm that can effectively compete with empirical methods for mapping the contours of individual buildings. Despite a number of challenges associated with imperfect imaging conditions, varied building architecture, and background complexity, the development of such an algorithm is critical for monitoring urban and natural areas, 3D city modeling, disaster management, and population density estimation.

Aerial photography has long been considered essential for detecting buildings and improving the efficiency of vector map generation for many years (Paparoditis et al., 1998, Persson et al., 2005, Yang et al., 2018). Recent advancements in remote sensing technology have greatly increased the precision of building detection from images (Li et al., 2019, Chen et al., 2020, Šanca et al., 2023). This improvement can be attributed to the implementation of deep learning techniques such as convolutional neural networks (CNN) (LeCun et al., 1989) and fully convolutional networks (FCN) (Long et al., 2014), supported by extensive training data and computational capabilities. Still, creating accurate vector maps of buildings from aerial images using automation is not possible in most urban areas. One reason for this challenge is the limitations of deep learning-based building detection techniques. These methods struggle with identifying roofs covered by trees or shadows (Chen et al., 2019) and may have difficulty adapting to different geographic regions (Maggiori et al., 2017). There is an additional concern that has not been given sufficient focus, which is that small missed or false detections may still occur along the boundaries of accurately detected buildings. This can result in inaccurate and irregular vector shapes when polygon simplification techniques are applied. A key challenge in extracting building outlines is accurately recreating the polygonal boundary of a building in order to produce a vectorized building mask for use in different applications.

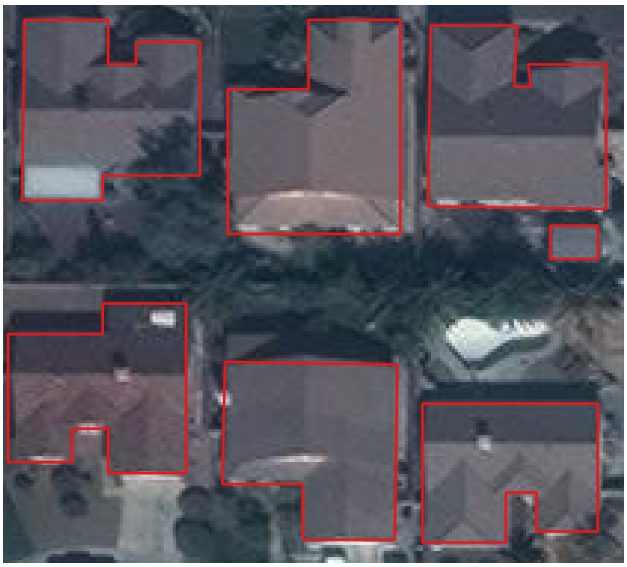

Figure 1. Example of extracting a building boundary.

This paper presents an algorithm that automatically extracts building outlines through a combination of binary segmentation, regularization, and vectorization techniques. A new regularization method has been introduced, which involves creating a neighborhood matrix for each point classified as belonging to the ”building” category. This sets it apart from other existing methods. This matrix assists in efficiently identifying and eliminating incorrectly segmented points, while also emphasizing important points like vertices and building boundaries. In summary, the main contributions of this paper are as follows:

- We investigate how the neighborhood matrix can help pinpoint additional pixels during binary segmentation.

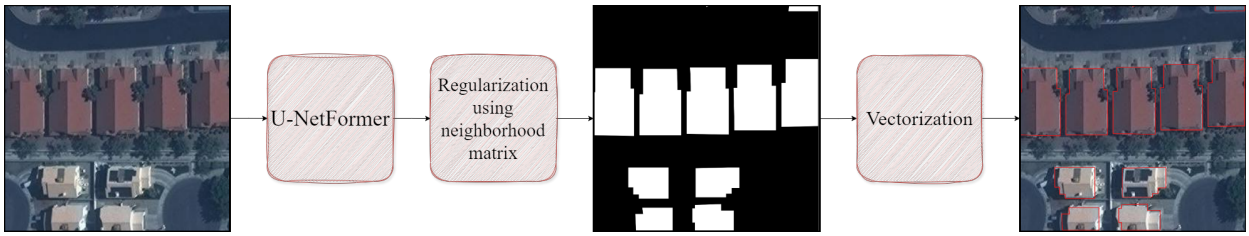

Figure 2. The proposed algorithm’s structure.

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

2. Related work

Semantic segmentation (Wei et al., 2019, Chen et al., 2020, Šanca et al., 2023) and instance segmentation (Zhao et al., 2020, Emelyanov et al., 2024) methods are currently the two most prominent building extraction approaches. Because predictions for building locations are done at such a detailed level, gaps may occur in the overall predictions if there is not enough global context provided. Additionally, smaller buildings might not be included if there are not sufficient local details. In order to tackle these challenges, (Wei et al., 2019) proposed a multi-scale aggregation FCN that combines different scales of building features to produce accurate building predictions. (Šanca et al., 2023) offers a comprehensive workflow that incorporates binary semantic segmentation, regularization, and vectorization techniques. Their innovation lies in implementing the regularization task on a fresh building dataset, along with introducing their approach to vectorization. (Knyaz et al., 2020) suggested a masking technique to effectively segment repeated structures in images. Using this approach resulted in a 11% improvement in segmentation performance.

(Zhao et al., 2020) suggested an instance segmentation model that addressed the issue of how detection quality impacts mask integrity. This model utilized a multi-stage process involving both detection and segmentation to enhance the accuracy of segmentation edges and enhance the geometric consistency of the results. (Emelyanov et al., 2024) proposes the algorithm for automatically extracting building outlines based on instance segmentation, regularization and vectorization. The main difference from existing methods is a regularization method based on the concepts of linear connectivity and convexity of a set of points.

Several other studies (Huang et al., 2021, Liu et al., 2021) have examined the challenge of instance segmentation by approaching it as contour regression. This involves predicting the vertex coordinates of a contour, or in simpler terms, determining the position of the corners of a polygon. Active contour models are used in traditional methods (Kass et al., 1988, Chan and Vese, 2001) to extract the contours of objects by optimizing an empirically developed energy function. The amalgamation of an active contour model and a CNN in recent methods (Marcos et al., 2018, Hatamizadeh et al., 2019) has enhanced the model’s robustness. Researchers have recently started using a unified deep neural network framework for contour extraction in their studies. A few researchers (Liu et al., 2022, Li et al., 2019) have utilized recurrent neural network (RNN) (Yu et al., 2019) to forecast the top corners of buildings in a clockwise pattern as an illustration. Nevertheless, this method frequently encounters issues with the disappearance of corner vertices and irregular vertices. CNN-based techniques are presently the most widely used edge-based approaches. PolarMask (Xie et al., 2020), PolarMask++ (Xie et al., 2021), and LSNet (Duan et al., 2021) are efficient one-stage contour-based CNN methods that primarily focus on utilizing deep features of instance centers, resulting in the extraction of mostly coarse object contours.

Also the results require extensive post-processing: semantic segmentation does not have the capability to differentiate between separate buildings, while the bounding box generated by the instance segmentation method might encompass parts of neighboring buildings, leading to challenges in mask training. (Zorzi et al., 2022) offers an alternative approach with PolyWorld, a neural network that extracts building vertices from images and connects them to accurately create polygons. The graph neural network predicts the connection strength between each pair of vertices, while assignments are evaluated through the solution of a differentiable optimal transport problem. Additionally, the positions of vertices are optimized through the minimization of both the segmentation loss and the difference in polygon angles.

3. Method

This work outlines a thorough process for extracting buildings and enhances the accuracy of building area predictions by implementing boundary regularization techniques. To begin with, a convolutional neural network is employed to train a binary segmentation model. The neighborhood matrix is also utilized to arrange the predicted contours of buildings and enhance their geometry. The last stage involves converting the regularized building masks into polygons through the vectorization process, which allows them to be used in various applications. Figure 2 displays the layout of the algorithm.

3.1 Binary segmentation with U-netFormer

The initial stage of our technique involves binary segmentation of remote sensing images to identify buildings. To perform this task, the UNetFormer neural network was used.

The U-NetFormer(U-Transformer) (Petit et al., 2021) network complements the U-Nets (Ronneberger et al., 2015) with attention modules built from multi-head transformers. U-Transformer models long-distance contextual interactions and spatial dependencies using two types of attention modules (see Figure 3): multi-head self-attention (MHSA) and multi-head cross-attention (MHCA). Modules are designed to express a new representation of input data based on its own attention in the first case or attention given to higher-level functions in the second.

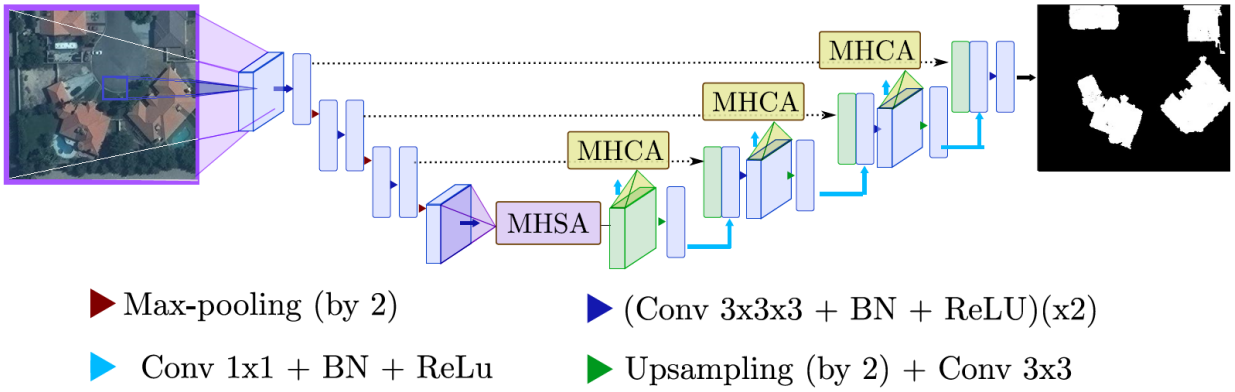

Figure 3. The structure of the convolutional neural network used for binary segmentation.

The MHSA module is designed to extract structural information from long-distance images. The main goal of MHSA is to connect each element of the superordinate feature map to each other, thereby providing access to a receptive field that includes the entire input image. Thus, the decision for one particular pixel can be influenced by any input pixel. The idea behind the MHCA module is to cut out unnecessary or noisy areas from skipping functions and highlight areas that are of significant interest to the application.

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

After generating predictions with trained models, a post-processing step is implemented to apply regularization in order to enhance the geometry and accuracy of the predicted building masks. Pixel-based classification often leads to rounded corners and closed-edge predictions, so implementing regularization is crucial for enhancing the accuracy of these predictions. Furthermore, following the binary segmentation process, the ”building” class may include additional pixels while certain crucial elements could potentially be omitted.

As the identification of building boundaries relies on remote sensing images, it is assumed that the buildings depicted in the images do not intersect or overlap. In this situation, every building is represented by a closed limited set of pixels. By utilizing the neighborhood matrix, you are able to eliminate points that are not part of the ”building” class and also include any pixels that the classifier may have overlooked. Also note that, due to the specifics of remote sensing images, most of buildings in the images have straight edges, and the angles at the vertices are 90^0, i.e. the buildings are a composition of rectangles.

All points in the image are assigned values 1 and 0 depending on their membership in the “building” class (if the point is segmented, then the value is 1, if not - 0). The neighborhood matrix of a point is a 3x3 matrix filled with the corresponding values of the point itself in the center and its neighbors. Similarly, a 2nd order neighborhood matrix of a point is a 5x5 matrix filled with the corresponding values of the point in the center and its neighbors, etc.

Let us denote by K the sum of all elements in the matrix of the neighborhood of a certain pixel. K shows how many elements in the neighborhood of a pixel belong to the “building” class (including the membership of the point in question). Based on the results of studying various variations of neighborhood matrices, we can conclude that the smallest value of K for a point that belongs to a building is 4 (these are the vertices of the convex corners of the building, they have the fewest neighbors from the desired class - 3), and the values for points being the boundary of the building, vary from 5 to 8. Also note that for the internal points of buildings $K = 9$. Examples of neighborhood matrices for boundary points are shown in Figure 4. Therefore, all points for which $K \leq 3$ are considered incorrectly segmented and are removed from the “building” class.

$$\begin{bmatrix} 0 & 0 & 0 \\ 0 & 1 & 1 \\ 0 & 1 & 1 \end{bmatrix} \quad \begin{bmatrix} 1 & 1 & 0 \\ 1 & 1 & 0 \\ 0 & 0 & 0 \end{bmatrix} \quad \begin{bmatrix} 0 & 0 & 0 \\ 0 & 1 & 0 \\ 1 & 1 & 1 \end{bmatrix}$$

(a) (b) (c)

$$\begin{bmatrix} 1 & 1 & 1 \\ 0 & 1 & 1 \\ 0 & 0 & 1 \end{bmatrix} \quad \begin{bmatrix} 1 & 1 & 0 \\ 1 & 1 & 0 \\ 1 & 1 & 0 \end{bmatrix} \quad \begin{bmatrix} 0 & 0 & 0 \\ 1 & 1 & 0 \\ 1 & 1 & 1 \end{bmatrix}$$

(d) (e) (f)

Figure 4. Neighborhood matrices for boundary points. Matrices (a-c, f) show the main types of vertices of convex corners of buildings, and (d, e) - types of boundaries.

Figure 5. The result of the regularization process. First: input image. Middle: segmented image. Last: segmented image after the regularization process.

Also note that at the first stage of the algorithm, points with a value of $K = 8$ are added to the “building” class, but the neural network did not recognize them as representatives of the desired class. The above regularization steps are presented as an algorithm in Figure 6.

Algorithm 1: Regularization using neighborhood matrices

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 adding missing points to the set of points **B** ;
Procedure `Add(`$\boldsymbol{B}$`):`
  for *each point* x_j *not in class ”building”* $\boldsymbol{B}$ **do**
    Construct a neighborhood matrix A_j *for point* x_j
    if $K = 8$ **then**
      add x_j to **B** ;
    else
      skip;
  return B;
Search and removal unnecessary points from the set of points **B** ;

Procedure `Remove(`$\boldsymbol{B}$`):`
  while $|\boldsymbol{B}_i| < |\boldsymbol{B}_{i-1}|$ **do**
    for *each point* x_i *of class ”building”* $\boldsymbol{B}$ **do**
      Construct neighborhood matrix A_i *for point* x_i
      if $K \leq 3$ **then**
        delete x_i from **B** ;
      else
        skip;
    return $\mathbf{B}_i$;
  return B;
Search particularly important points to the set of points **B** ;

Procedure `Search(`$\boldsymbol{B}$`):`
  for *each point* x_i *of class ”building”* $\boldsymbol{B}$ **do**
    Construct a neighborhood matrix A_i *for point* x_i
    if $K = 4$ *and* $det(A_i) = 0$ *and* $(a_{ij} = a_{ji}$ *or* $a_{ij} = a_{4-j,4-i})$ **then**
      add x_i to vertices of **B** ;
    if $K = 6$ *and* $det(A_i) = 0$ *and* $A_i^2 = 2A_i$ **then**
      add x_i to boundary of **B** ;
    else
      skip;
  return B;

Figure 6. Algorithm for regularization and selection of particularly significant points.

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:

$$IoU = \frac{Intersection}{Union} = \frac{TP}{TP + FP + FN} \tag{3}$$

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

$$Precision = \frac{TP}{TP + FP} \tag{4}$$

$$Recall = \frac{TP}{TP + FN} \tag{5}$$

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

4.2 Experiment

Figure 7. Some images from the CrowdAI Mapping Challenge dataset.

The algorithm was trained on the open CrowdAI Mapping Challenge database (Mohanty et al., 2020), consisting of more than 280k satellite images for training and 60k images for testing. The training images were split into two portions: 80% of the images were utilized for training the algorithm, while

Figure 8. Experiment results.

the remaining 20% were set aside for validation. The training was conducted using CUDA 11.7 on an NVIDIA GeForce RTX 3070 GPU equipped with 8 GB of memory.

The meaning of the calculated measurements can be found in Table 1. The table includes the results of the leading methods on similar data in order to gauge the algorithm’s level of efficiency.

Method	AP	AR	IoU
Mask R-CNN	41.9	47.6	-
PolyMapper	55.7	62.1	-
PolyWorld	63.3	75.4	91.3
NM(our method)	64.1	75.1	91.2

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

5. Conclusion

Our study aimed to create a streamlined process for extracting building outlines, incorporating semantic segmentation, neighborhood matrix-based regularization, and vectorization into the workflow. Our findings indicate that regularization with a neighborhood matrix results high segmentation accuracy on average: AP (average precision) = 64.1 and AR (average recall) = 75.1. Regularization not only enhances predictions but also enhances the geometric shape of building outlines. Experiments demonstrated that the algorithm’s efficiency was on par with other top methods such as PolyWorld (Zorzi et al., 2022) for extracting building boundaries.

Additionally, there is an idea for our future research to move from using relationships between pixels to using relationships between objects and regions in images. These relationships will be presented in the form of a “scene graph”, the nodes of which will be objects, and the edges will be the relationships between them. For this purpose, graph convolutional networks will be used.

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

Chan, T., Vese, L., 2001. Active contours without edges. *IEEE Transactions on Image Processing*, 10(2), 266-277.

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.

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.

Duan, K., Xie, L., Qi, H., Bai, S., Huang, Q., Tian, Q., 2021. Location-Sensitive Visual Recognition with Cross-IOU Loss. *ArXiv*, abs/2104.04899. https://api.semanticscholar.org/CorpusID:233210422.

Emelyanov, A., Knyaz, V. A., Kniaz, V. V., 2024. Extracting building outlines based on convolutional neural networks using the property of linear connectivity. *The International Archives of the Photogrammetry, Remote Sensing and Spatial Information Sciences*, XLVIII-1-2024, 147–152. https://isprs-archives.copernicus.org/articles/XLVIII-1-2024/147/2024/.

Hatamizadeh, A., Sengupta, D., Terzopoulos, D., 2019. End-to-End Deep Convolutional Active Contours for Image Segmentation. *ArXiv*, abs/1909.13359. https://api.semanticscholar.org/CorpusID:203593984.

Huang, W., Tang, H., Xu, P., 2021. OEC-RNN: Object-Oriented Delineation of Rooftops With Edges and Corners Using the Recurrent Neural Network From the Aerial Images. *IEEE Transactions on Geoscience and Remote Sensing*, PP, 1-12.

Kass, M., Witkin, A., Terzopoulos, D., 1988. Snakes: Active contour models. *International journal of computer vision*, 1(4), 321–331.

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)*.

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.

Liu, Z., Tang, H., Huang, W., 2022. Building Outline Delineation From VHR Remote Sensing Images Using the Convolutional Recurrent Neural Network Embedded With Line Segment Information. *IEEE Transactions on Geoscience and Remote Sensing*, 60, 1-13.

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.

Marcos, D., Tuia, D., Kellenberger, B., Zhang, L., Bai, M., Liao, R., Urtasun, R., 2018. Learning Deep Structured Active Contours End-to-End. *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 8877-8885. https://api.semanticscholar.org/CorpusID:3939983.

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.

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.

Petit, O., Thome, N., Rambour, C., Soler, L., 2021. U-Net Transformer: Self and Cross Attention for Medical Image Segmentation. *ArXiv*, abs/2103.06104. https://api.semanticscholar.org/CorpusID:232170496.

Ronneberger, O., Fischer, P., Brox, T., 2015. U-Net: Convolutional Networks for Biomedical Image Segmentation. *ArXiv*, abs/1505.04597. https://api.semanticscholar.org/CorpusID:3719281.

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.

Xie, E., Sun, P., Song, X., Wang, W., Liu, X., Liang, D., Shen, C., Luo, P., 2020. Polarmask: Single shot instance segmentation with polar representation. *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 12190–12199.

Xie, E., Wang, W., Ding, M., Zhang, R., Luo, P., 2021. PolarMask++: Enhanced Polar Representation for Single-Shot Instance Segmentation and Beyond. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 44, 5385-5400. https://api.semanticscholar.org/CorpusID:233739844.

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.

Yu, Y., Si, X., Hu, C., Zhang, J., 2019. A Review of Recurrent Neural Networks: LSTM Cells and Network Architectures. *Neural Computation*, 31(7), 1235-1270. https://doi.org/10.1162/neco_a_01199.

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.