The International Archives of the Photogrammetry, Remote Sensing and Spatial Information Sciences, Volume XLVIII-4/W18-2025
Symposium on GeoSpatial Technologies: Visions and Horizons 2025 (GeoVisions2025), 8–10 October 2025, Çanakkale, Türkiye

Advancing Building Footprint Extraction with Multi-Stage Regularization Techniques

Anton Emelyanov[1,2], Vladimir Knyaz[1,2], Vladimir Kniaz[1,2], Nikita Borisov[1], Victor Aleksandrov[2]

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

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

Abstract

This paper introduces an automated building extraction method combining CNN segmentation with multi-stage regularization. We address urban mapping challenges including boundary inaccuracies and topological errors through: (1) neighborhood matrix processing for local refinement, (2) spectral graph optimization (SBO) for global consistency, and (3) curvature-adaptive contour refinement (ACR) to preserve geometric features. The pipeline converts initial segmentations into precise polygons through hierarchical processing. Experiments show performance matching state-of-the-art methods like PolyWorld, with superior handling of complex geometries. Key innovations include integrated local-global artifact removal and topology-preserving regularization. The curvature-adaptive approach maintains critical architectural features while eliminating noise. Particularly effective for high-resolution imagery, our solution improves geometric fidelity in urban mapping applications. The framework demonstrates robust performance for 3D city modeling and GIS tasks, overcoming common segmentation limitations. Results confirm accurate building outline extraction from satellite/aerial data, advancing automated urban feature mapping.

1. Introduction

Cutting-edge research in remote sensing focuses on creating automated algorithms capable of matching the accuracy of manual methods for delineating building footprints. Although challenges such as imperfect image quality, diverse architectural styles, and cluttered backgrounds persist, advancing these algorithms is essential for applications like urban monitoring, 3D city reconstruction, disaster response, and population analysis.

For decades, aerial imagery has served as a fundamental tool for building identification and facilitating vector map production (Paparoditis et al., 1998, Persson et al., 2005, Yang et al., 2018). The field has witnessed significant progress in building detection accuracy through modern remote sensing innovations (Li et al., 2019, Chen et al., 2020, Šanca et al., 2023), driven primarily by the adoption of deep learning architectures like convolutional neural networks (CNNs) (LeCun et al., 1989) and fully convolutional networks (FCNs) (Long et al., 2014), combined with enhanced datasets and computing power. However, fully automated generation of precise building vector maps from aerial photos remains unachievable in most urban environments. This limitation stems partly from inherent constraints in current deep learning approaches, which face difficulties with obscured roof structures (e.g., under vegetation or shadows) (Chen et al., 2019) and lack robust generalization across diverse geographical contexts (Maggiori et al., 2017). Another often-overlooked issue involves minor detection errors along building edges, where even small omissions or false positives can lead to distorted polygonal shapes during vectorization. The core difficulty lies in precisely reconstructing building polygons to create accurate vector representations suitable for various applications.

This paper presents an algorithm that automatically extracts building outlines through a combination of binary semantic segmentation, regularization, and vectorization techniques. The framework's key contribution involves augmenting conventional regularization approaches - including the established neighborhood matrix technique and Spectral Boundary Optimization (SBO) - with our newly developed Adaptive Contour Refinement (ACR) algorithm. This synergistic combination addresses distinct aspects of boundary refinement: the neighborhood matrix ensures local spatial consistency, SBO maintains global topological coherence, ACR preserves geometrically significant features. The complete regularization pipeline produces a refined segmentation mask that subsequently undergoes polygonal vectorization, generating accurate geospatial representations of building outlines suitable for GIS applications and urban planning. In summary, the main contributions of this paper are as follows:

Figure 1. Example of extracting a building boundary.

- We investigate how the developed regularization methods can improve the quality and accuracy of binary segmentation.

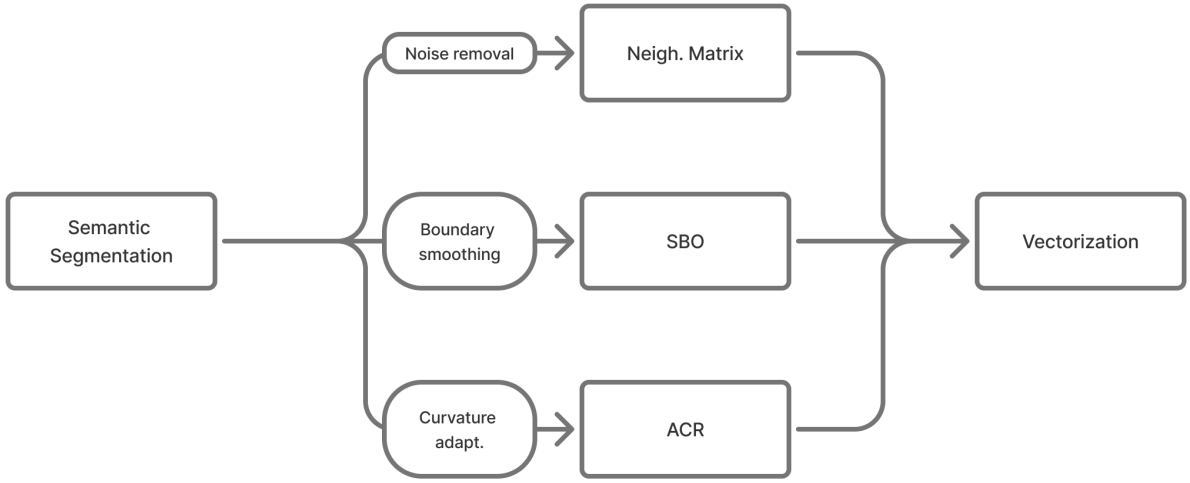

Figure 2. The structure of the proposed algorithm, integrating neighborhood matrix regularization, Spectral Boundary Optimization, and Adaptive Contour Refinement.

- We employ the CrowdAI dataset (Mohanty et al., 2020), a prevalent benchmark in building vectorization studies, to evaluate results and conduct comparative analysis with existing boundary extraction methods.

2. Related work

Contemporary building extraction methodologies primarily employ two dominant techniques: 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). While these approaches operate at a fine-grained pixel level, they face inherent limitations. Insufficient global context can lead to prediction gaps in building footprints, while inadequate local detail resolution may cause smaller structures to be overlooked.

To address these challenges, researchers have developed several innovative solutions. (Wei et al., 2019) introduced a multi-scale aggregation FCN architecture that integrates building features across different scales to enhance prediction accuracy. Another comprehensive framework (Šanca et al., 2023) combines binary semantic segmentation with subsequent regularization and vectorization processes, demonstrating its effectiveness through application to a novel building dataset and introducing an original vectorization methodology. (Knyaz et al., 2020) proposed an advanced masking technique specifically designed for segmenting repetitive architectural elements, achieving an 11% performance improvement in segmentation tasks.

Recent advances in building extraction have introduced innovative instance segmentation approaches. (Zhao et al., 2020) developed a multi-stage instance segmentation model that specifically examines how detection quality affects mask preservation. Their framework combines object detection with segmentation refinement to improve both boundary precision and geometric accuracy of building footprints. In a related development, (Emelyanov et al., 2024) presented an automated pipeline for building outline extraction that integrates instance segmentation with subsequent regularization and vectorization. This approach distinguishes itself through a novel regularization technique that employs principles of linear connectivity and point set convexity, offering improved results compared to conventional methods.

Recent research has explored instance segmentation through contour regression approaches (Huang et al., 2021, Liu et al., 2021), where the task involves predicting polygon vertex coordinates (essentially locating a shape's corner points). Traditional techniques employ active contour models (Kass et al., 1988, Chan and Vese, 2001) that derive object boundaries by optimizing hand-crafted energy functions. Modern approaches have improved robustness by integrating CNNs with active contour models (Marcos et al., 2018, Hatamizadeh et al., 2019).

Current methodologies increasingly adopt unified deep learning frameworks for contour extraction. Some studies (Liu et al., 2022, Li et al., 2019) have implemented recurrent neural networks (RNNs) (Yu et al., 2019) to sequentially predict building roof corners in clockwise order, though these methods often struggle with vertex disappearance and irregular point distributions.

Among contemporary edge-based solutions, CNN architectures dominate the field. Notable examples include PolarMask (Xie et al., 2020), PolarMask++ (Xie et al., 2021), and LSNet (Duan et al., 2021) - efficient single-stage systems that leverage deep features from instance centers. While computationally effective, these methods typically produce only approximate object contours.

Current approaches face significant post-processing challenges: semantic segmentation cannot distinguish between adjacent buildings, while instance segmentation may produce bounding boxes that incorrectly include portions of nearby structures, complicating mask generation. To address these limitations, (Zorzi et al., 2022) proposed PolyWorld - an innovative neural network that directly predicts building vertices from imagery and constructs precise polygons by establishing connections between them. The framework employs a graph neural network to estimate connection probabilities between vertex pairs, with final assignments determined through a differentiable optimal transport formulation. Furthermore, vertex positions are refined by jointly optimizing segmentation accuracy and polygon angle consistency.

3. Method

This study introduces an optimized pipeline for automated building footprint extraction, focusing on precision enhancement through innovative boundary regularization strategies. Our methodology implements a multi-stage processing chain that sequentially executes: (1) binary semantic segmentation using a deep convolutional neural network, (2) comprehensive boundary refinement through integrated regularization techniques, and (3) geometric vectorization of the refined outputs. The algorithm's structure, integrating these regularization techniques, is illustrated in Figure 2.

3.1 Semantic segmentation with U-NetFormer

Our methodology employs a sophisticated deep learning approach for building footprint detection from remote sensing imagery, beginning with binary segmentation using the enhanced UNetFormer architecture. This advanced neural network extends the capabilities of conventional U-Net frameworks by integrating powerful transformer-based attention mechanisms, enabling more comprehensive analysis of spatial relationships across multiple scales.

The UNetFormer (U-Transformer) (Petit et al., 2021) architecture fundamentally enhances the original U-Net design (Ronneberger et al., 2015) through its innovative attention modules, which operate in complementary fashion. The Multi-Head Self-Attention (MHSA) mechanism provides global contextual understanding by establishing long-range dependencies across the entire feature map, effectively connecting each pixel to all others in the image. This creates a fully comprehensive receptive field that allows segmentation decisions at any location to incorporate relevant information from distant regions of the input.

Working in concert with MHSA, the Multi-Head Cross-Attention (MHCA) module performs intelligent feature selection, dynamically filtering out irrelevant or noisy information in skip connections while precisely highlighting the most salient regions for building detection. This dual-attention system generates optimized feature representations through two distinct but complementary pathways: MHSA handles intrinsic feature relationships within the data, while MHCA strategically incorporates higher-level contextual information to focus processing on the most diagnostically valuable image regions.

The model was optimized using the Adam algorithm with Binary Cross-Entropy with Logits (BCEWithLogitsLoss) as the objective function. This loss metric quantifies the discrepancy between the network's predictions and the ground truth annotations through the following formulation:

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

where N represents the batch size, x_i denotes the ground truth binary mask for the i-th sample, y_i corresponds to the model's raw output logits for sample i, σ indicates the sigmoid activation function.

The sigmoid function, characterized by its distinctive S-shaped curve, transforms the logit values into probabilistic outputs. Specifically, we employ the standard logistic function defined as:

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

This activation function maps the unbounded logit outputs to the [0, 1] range, enabling probabilistic interpretation of the model's predictions while maintaining differentiability for backpropagation.

3.2 Application of the developed regularization method

3.2.1 Neighborhood Matrix-Based Initial Regularization The first stage of our regularization pipeline employs a local spatial consistency approach to correct pixel-level classification errors. For each pixel $p_{i,j}$ classified as "building" (class label 1), we construct a 3×3 neighborhood matrix $N(p_{i,j})$ encompassing the central pixel and its eight immediate neighbors. This matrix serves as a local context window, allowing us to:

- **Detect misclassified pixels:** A pixel is flagged as potentially misclassified if its label contradicts the majority of its neighbors (e.g., an isolated "building" pixel surrounded by "non-building" pixels).
- **Apply probabilistic correction:** The final label of $p_{i,j}$ is reassigned based on a weighted vote of its neighbors, where weights are inversely proportional to their Euclidean distance from $p_{i,j}$. This step effectively removes salt-and-pepper noise while preserving legitimate small-scale structures.

3.2.2 Spectral Boundary Optimization (SBO) for Global Consistency To enforce topological coherence across building segments, we model the segmentation output as an undirected graph $G = (V, E)$, where Nodes V correspond to pixels labeled as "building", Edges E connect spatially adjacent pixels (8-neighborhood), with weights w_{ij} defined by a Gaussian affinity kernel:

$$w_{ij} = \exp\left(-\frac{|p_i - p_j|_2^2}{2\sigma_d^2} - \frac{|I_i - I_j|_2^2}{2\sigma_I^2}\right), \quad (3)$$

where σ_d and σ_I control the sensitivity to spatial distance and intensity variation, respectively.

The graph Laplacian $L = D - W$ (where D is the degree matrix and W the weighted adjacency matrix) encodes the global structure of building regions. We solve the spectral optimization problem:

$$\min_f \left(f^T L f + \lambda |f - y|^2\right), \quad (4)$$

where f is the regularized label field, y the initial segmentation, and λ a trade-off parameter. This step eliminates fragmented regions and smooths irregular boundaries while respecting image edges.

3.2.3 Adaptive Contour Refinement (ACR) for Geometric Precision The final stage refines building boundaries by explicit curvature adaptation. For each extracted boundary contour $\mathcal{C}$:

Figure 3. The result of the regularization process. First: input image. Middle: segmented image. Last: segmented image after the regularization process.

- **Compute local curvature** κ_i at each point i using a derivative-based estimator:

$$\kappa_i = \frac{|x'_i y''_i - y'_i x''_i|}{(x'^2_i + y'^2_i)^{3/2}}, \tag{5}$$

where derivatives are approximated via central finite differences (e.g., $x' = \frac{x_{i+1} - x_{i-1}}{2}$).

- **Adjust smoothing strength** with a curvature-adaptive kernel:

$$h_i = \exp\left(-\frac{\kappa_i^2}{2\sigma_c^2}\right), \quad \sigma_c = \text{curvature tolerance}. \tag{6}$$

High-curvature regions (e.g., corners with $\kappa_i > \sigma_c$) undergo weak smoothing, preserving sharp features, while low-curvature segments (straight edges) are aggressively regularized.

- **Resample the contour** using B-spline interpolation to ensure uniform point spacing, critical for high-quality vectorization. The final contour is reconstructed as:

$$\mathcal{C}\text{refined} = \sum ih_i \cdot K(|\mathcal{C}_i - \mathcal{C}|) * \mathcal{C}, \tag{7}$$

where K is a Gaussian smoothing kernel.

4. Results

4.1 Evaluation metrics

Following the methodology of (Zorzi et al., 2022), we employ several standard metrics to assess model performance.

The Intersection-over-Union (IoU), also known as the Jaccard index, quantifies segmentation accuracy by measuring the overlap between predicted and ground truth masks. It is calculated as:

$$IoU = \frac{Intersection}{Union} = \frac{TP}{TP + FP + FN} \tag{8}$$

Additionally, we compute precision and recall metrics to derive Average Precision (AP) and Average Recall (AR) scores:

$$Precision = \frac{TP}{TP + FP} \tag{9}$$

$$Recall = \frac{TP}{TP + FN} \tag{10}$$

where TP (True Positives) represents correctly identified building pixels, FP (False Positives) indicates non-building pixels misclassified as buildings, FN (False Negatives) denotes building pixels missed by the prediction.

4.2 Experiment

Figure 4. Some images from the CrowdAI Mapping Challenge dataset.

The model was trained using the publicly available CrowdAI Mapping Challenge dataset (Mohanty et al., 2020), which contains over 280,000 satellite images for training and an additional 60,000 for testing. We adopted an 80-20 split for the training data, allocating 80% for model training and reserving 20% for validation purposes. All training procedures were implemented using CUDA 11.7 on an NVIDIA GeForce RTX 3070 GPU with 8GB VRAM.

For performance evaluation, Table 1 presents the quantitative results of our method alongside comparative benchmarks from state-of-the-art approaches. This comparative analysis enables assessment of our algorithm's relative performance against existing solutions on comparable data.

5. Conclusion

This study has presented a comprehensive end-to-end workflow for automated building footprint extraction, combining binary semantic segmentation with advanced multi-stage regularization and vectorization techniques. Our experimental results demonstrate that the proposed method achieves competitive

Figure 5. Experiment results.

Method	AP	AR	IoU
Mask R-CNN	41.9	47.6	-
PolyMapper	55.7	62.1	-
PolyWorld	63.3	75.4	91.3
Neigh. Matrix	64.1	75.1	91.2
Our method	69.8	80.3	92.0

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

performance with state-of-the-art approaches like PolyWorld (Zorzi et al., 2022), while offering several distinct advantages through its novel hierarchical regularization framework.

The key strengths of our methodology include:a multi-scale regularization approach, topological integrity maintenance through graph-based processing that prevents over-smoothing of complex building structures and adaptive geometric handling that intelligently preserves critical features like sharp corners while eliminating noise.

The framework has proven effective for processing high-resolution satellite imagery, where it successfully addresses common segmentation artifacts (jagged edges, spurious pixels) that typically degrade the quality of downstream applications including 3D urban modeling and GIS analysis. The method's robustness stems from its balanced integration of local and global processing, combining the precision of pixel-level operations with the structural awareness of graph-based and curvature-adaptive techniques.

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.