Building Extraction Network based on High-resolution Remote Sensing Image

Han Li, Xian Guo*, Jie Jiang, Changyu Gong
Beijing University of Civil Engineering and Architecture, Beijing, China,
2108160224011@stu.bucea.edu.cn; guoxian@bucea.edu.cn; jiangjie@bucea.edu.cn; 2108160224014@stu.bucea.edu.cn

Keywords: Building extraction, Poly kernel inception network, Multi-scale feature fusion, High-resolution remote sensing images

Abstract

The segmentation of buildings from the background in high-resolution remote sensing images faces several challenges, including difficulties in extracting multi-scale information, insufficient capture of long-range contextual information, and the underutilization of multi-scale features. Existing methods often struggle to effectively capture features at different scales, which limits the segmentation accuracy. Furthermore, long-range contextual information is frequently overlooked, hindering model's ability in understanding the global structure of buildings. Additionally, balancing low-level details with high-level semantic information poses challenges in effectively fusing multi-scale features from high-resolution imagery. To address these issues, this paper proposes the Multi-Scale Multi-Kernel Building Extraction Network (MMAENet), which significantly enhances the capability to capture multi-scale features through the integration of Poly Kernel Inception Network (PKINet), and improves the capture of long-range contextual information. The Panoramic Feature Pyramid (PFP) structure is introduced to ensure the full integration of both high-level and low-level information. Performance evaluation on the WHU Aerial dataset demonstrates that the model achieves superior accuracy in building segmentation compared to Convnext, PSPNet, and Swin Transformer.

1. Introduction

Accurate building extraction from remote sensing images is critical for urban planning, urban evaluation, urban governance, and automated mapping (Y. Liu et al., 2022). The advancement of high spatial resolution imaging technologies has enhanced surface details, facilitating more refined and automated building extraction. However, the complexity of high-resolution imagery continues to present significant challenges in accurate building extraction.

Traditional building extraction methods primarily rely on spectral indices and classic machine learning algorithms to enhance extraction performance. Commonly used indices, such as the Normalized Difference Building Index (NDBI), Normalized Difference Vegetation Index (NDVI), Soil-Adjusted Vegetation Index (SAVI), Modified Normalized Difference Water Index (MNDWI), and Global Environmental Monitoring Index (GEMI), have been widely applied in building extraction from remote sensing imagery(Puttinaovarat and Horkaew, 2017). These indices effectively differentiate between various land cover types, such as buildings, vegetation, and water bodies. In addition, machine learning techniques such as Artificial Neural Networks (ANN), K-Nearest Neighbors (KNN), and Support Vector Machines (SVM)(Li et al., 2022) have also been employed to further enhance extraction accuracy. While these traditional methods have improved accuracy to some extent, they still face significant challenges in complex urban environments and multi-scale conditions in the high-resolution scenario, particularly in extracting buildings in dense urban areas or small-scale buildings.

Recent advances in deep learning (DL) t have substantially improved the extraction of small-scale, high-density buildings from high-resolution remote sensing imagery. In particular, the development of Convolutional Neural Networks (CNN)(Zhou et al., 2022) and Vision Transformers (ViT) (Dosovitskiy et al., 2020) has yielded significant progress in this area. Within the CNN framework, Li et al. (2019) proposed a U-Net-based semantic segmentation approach, leveraging the advantages of

CNNs in local feature extraction and hierarchical modeling to achieve efficient building extraction. Qiu et al. (2023) further refined this approach by introducing Refine-UNet, which enhances skip connections and employs depthwise separable convolutions to improve the capture of building details. Fan et al. (2023) incorporated residual modules (BasicBlock) and a spatially enhanced attention mechanism (SEAE) into the U-Net structure, further improving the model's fitting ability. However, CNNs are inherently limited in capturing global context. Hu et al. (2021) integrated a squeeze-and-excitation (SE) attention mechanism to dynamically emphasize critical regions, yet the constrained receptive fields of CNNs still hinder effective modeling of long-range spatial relationships. Conversely, ViTs leverage global self-attention mechanisms to enhance feature representation. Liu et al. (2021) proposed the Swin Transformer, which uses a hierarchical architecture and shifted window mechanism to process features at multiple scales, effectively balancing both local and global information. Similarly, Wang et al. (2022) utilized a dual-path structure in combination with linear multi-head attention to encode spatial details and capture global dependencies in high-resolution imagery. Nevertheless, Transformer architecture requires considerable computational resources and struggles to balance detail and global structure in large-scale scenarios.

Despite these notable achievements in accuracy, robustness, and efficiency, challenges in building extraction persist. In particular, the extraction of multi-scale features remains problematic, and the modeling of long-range dependencies is still inadequate. To address these issues, this paper introduces the MMAENet, which enhances the capture of multi-scale features through PKINet (Cai et al., 2024). By extracting building features at different scales, MMAENet improves its ability to adapt to buildings of varying sizes, shapes, and textures. Additionally, the introduction of a Context Anchors Attention (CAA) module enables the model to capture long-range contextual information. To further optimize the use of multi-scale information extracted by PKINet, MMAENet incorporates a PFP structure (Kirillov et al., 2019), seamlessly integrating high-level semantic and low-level detail information, thereby improving the robustness of building extraction.

The main contributions of this paper are as follows:

1. Introduction of the PKINet Block within PKINet, which enhances the ability to capture multi-scale features of buildings.
2. Incorporation of the CAA module within the PKINet Block, enabling MMAENet to capture long-range contextual information.
3. Construction of the PFP structure, allowing MMAENet to fully integrate high-level semantic information with low-level details, thus balancing local features and global context to improve the robustness of building extraction.

2. Methodology

The overall architecture of MMAENet is presented in FIG. 1, consisting of a four-stage encoder and a PFP-based feature fusion structure. In the encoding phase, a Stem layer is employed to reduce the model's parameter count, thereby improving efficiency. The processed feature maps are then subjected to two distinct operations along the channel dimension following subsampling and convolution. One operation utilizes a simple feedforward network (FFN), while the other applies a Poly Kernel Inception (PKI) Block. Within the PKI Block, the PKI Module and CAA Module are integrated to fuse the feature maps produced by the two operations, and subsequently output the combined features. The PFP structure is then applied, enabling the comprehensive utilization and fusion of multi-scale information, thereby enhancing the model's ability to handle complex spatial dependencies and improve feature representation across varying scales.

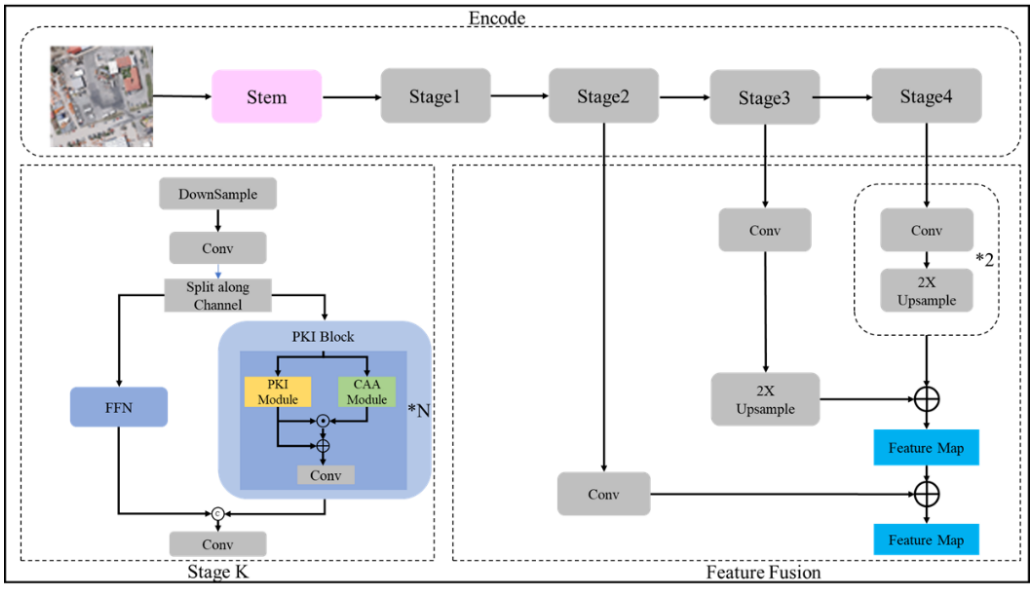

FIG. 1 General framework of the model

2.1 PKI Module

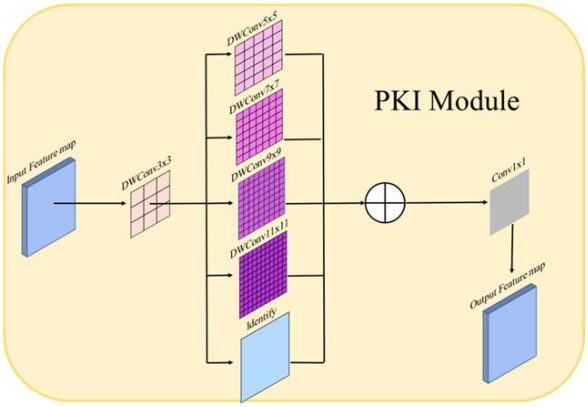

FIG. 2 PKI Module

The PKI Module is shown in FIG. 2. The PKI Module employs 3×3 depthwise convolution (DWConv) to extract local detail features of buildings, enhancing the representation of small buildings and edge information. A parallel structure is introduced with DWConv of varying receptive fields (5×5, 7×7, 9×9, 11×11) to capture multi-scale contextual information, ensuring the comprehensive representation of spatial features of buildings across different scales. Additionally, 1×1 convolutions are used to integrate features from different scales, capturing local contextual information within building features at each scale, thereby improving the completeness and accuracy of building extraction. The multi-scale calculation formula is shown in formula 1.

$$OutPut_i = \mathrm{DWConv}_{k^i \times k^i}(F), i = 1, ..., 4. \quad (1)$$

In formula 1, $\mathrm{DWConv}_{k^i \times k^i}$ is a depth-separable convolution operation of different sizes, $k^i \times k^i$ is available in sizes 5x5, 7x7, 9x9, and 11x11, F is the input feature map, $OutPut_i$ is the result of convolution operations of different sizes.

2.2 CAA Module

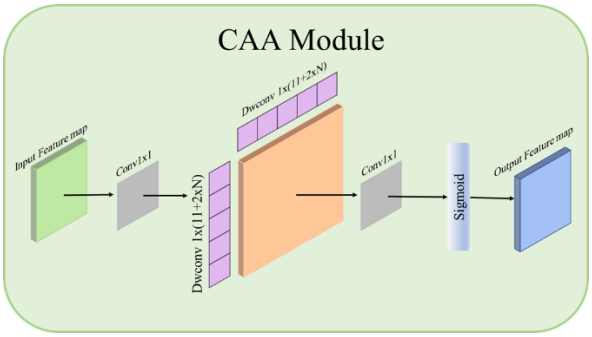

FIG. 3 CAA Module

The CAA Module is shown in FIG. 3. In the PKI Block, to capture long-range contextual information within building features, the CAA module is introduced to consider the dependencies between distant building and background pixels, while enhancing the central point features. The CAA module extracts local features through average pooling and convolutional layers, utilizing horizontal and vertical strip convolutions to improve the recognition of elongated building shapes. Simultaneously, the Sigmoid function generates attention weights, enabling the PKI Block to establish long-range pixel relationships in building high-resolution remote

sensing imagery. This mechanism further enhances the accuracy and robustness of building extraction. The bar convolution is shown in formula 2.

$$F_{\mathrm{w}} = \mathrm{DWConv}_{1\times k_b}(F_{\mathrm{pool}})$$
$$F_{\mathrm{h}} = \mathrm{DWConv}_{k_b\times 1}(F_{\mathrm{w}}) \quad (2)$$

In Formula 2, F_{pool} is the feature map after pool, $\mathrm{DWConv}_{1\times k_b}$ and $\mathrm{DWConv}_{k_b\times 1}$ are horizontal bar convolution and vertical bar convolution, where $K_b = 11 + 2\times N$, N is the number of PKI Blocks, F_w is the horizontal convolution result feature map, and F_h is the vertical convolution result feature map.

2.3 PFP Structure

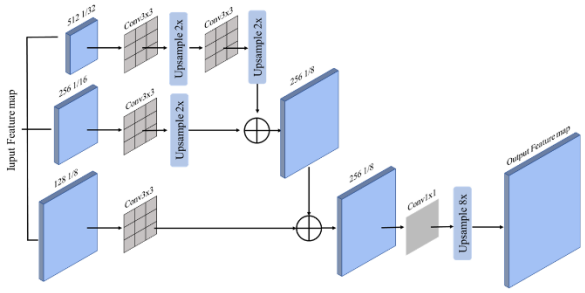

FIG. 4 PFP Structure

The PFP structure is illustrated in FIG. 4. To fully leverage the feature maps generated during the encoding stage, the PFP structure processes multi-scale feature maps. In the context of building high-resolution remote sensing imagery, to prevent the fusion of overly low-level fragmented texture information of buildings, the feature maps from the last three stages undergo an upsampling operation, restoring the feature maps to a quarter of their original size. Each upsampling stage consists of a 3×3 convolutional layer, a ReLU activation layer, and a twofold bilinear interpolation upsampling operation, ensuring that the building's detailed features are preserved. Finally, a 1×1 convolutional layer followed by an eightfold bilinear interpolation upsampling operation restores the image to its original size, thereby accurately reconstructing the building's high-resolution features.

3. Experiment

The performance of MMAENet was evaluated on the WHU aerial remote sensing dataset using IoU, Recall, F1-score, and Precision as accuracy metrics (Huang et al., 2024). The experiments were conducted on an NVIDIA GTX 4090D graphics card, utilizing pre-trained weights from 300 epochs on the ImageNet1k dataset. Training was carried out for 30 epochs with the support of the mmsegmentation toolbox, a batch size of 16, and the SGD optimizer, employing the PolyLR strategy for dynamic learning rate adjustment.

The configuration of the Encode phase is detailed in TABLE 1. The feature map dimensions were reduced to 1/2, 1/4, 1/8, 1/16, and 1/32 of the original image size. The number of PKI Blocks (N) in the overall framework varied across different stages, with values of 4, 12, 20, and 4. The number of output channels for each layer was 32, 64, 128, 256, and 512, respectively.

TABLE 1 configuration of Encode phase

	Feature Map Scale Size	Out Channels	PKI Block Num (N)
Stem	1/2×1/2	32	-
Stage 1	1/4×1/4	64	4
Stage 2	1/8×1/8	128	12
Stage 3	1/16×1/16	256	20
Stage 4	1/32×1/32	512	4

4. Analysis

The comparative test results are presented in FIG. 5, as well as TABLE 2 and 3. By comparing the performance of Convnext (Z. Liu et al., 2022), PSPNet (Zhao et al., 2017) and Swin Transformer (Liu et al., 2021) on WHU Aerial imagery dataset, our results show that MMAENet superior performance across all metrics, including IoU, Recall, F1-score and Precision. MMAENet outperforms Swin Transformer, PSPNet, and Convnext in both background and building segmentation.

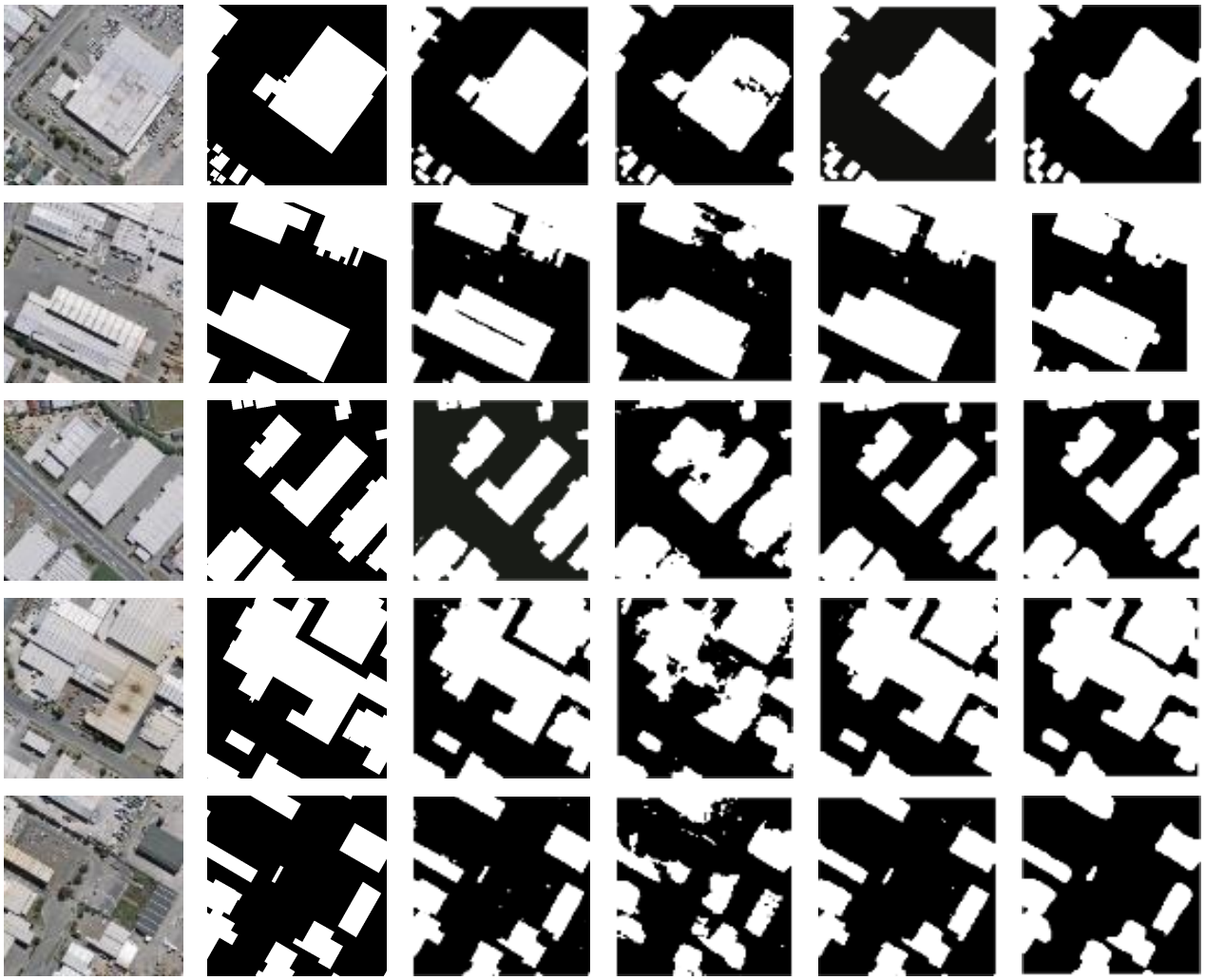

FIG. 5 Processing results of different models

TABLE 2 Different models are experimented in WHU Aerial imagery dataset

	mIoU(%)	mRecall(%)	mF1-score	mPrecision(%)
Convnext	79.34	90.85	87.71	85.18
PSPNet	85.91	94.80	92.09	89.79
Swin Transformer	90.15	96.47	94.66	93.03
MMAENet	**91.69**	**97.54**	**95.56**	**93.78**

TABLE 3 WHU Aerial imagery dataset category metrics

		Convnext	PSPNet	Swin Transformer	MMAENet
IoU(%)	Background	94.19	96.29	97.54	**97.94**
	Building	64.50	75.51	82.76	**85.44**
Recall(%)	background	95.86	97.21	98.18	**98.34**
	Building	85.85	92.39	94.76	**96.74**
F1-score	Background	97.01	98.11	98.76	**98.96**
	Building	78.42	86.06	90.56	**92.15**
Precision(%)	Background	98.19	99.03	99.34	**99.59**
	Building	72.17	80.55	86.73	**87.98**

MMAENet leverages a parallel multi-scale convolutional kernel mechanism to efficiently extract multi-scale building features from high-resolution remote sensing imagery. The experimental results demonstrate that MMAENet significantly outperforms the comparison models in terms of both building IoU and mIoU. The building IoU reaches 85.44%, outperforming Swin Transformer, PSPNet, and ConvNeXt by 2.68%, 0.93%, and 20.94%, respectively. The mIoU achieves 91.69%, with improvements of 1.54% and 5.78% over Swin Transformer and PSPNet, respectively. This enhancement in performance directly reflects the effectiveness of the PKI Module in extracting multi-scale features. By utilizing parallel multi-scale convolutional kernels, the model is capable of capturing both local details and global structures of buildings from high-resolution remote sensing imagery, thereby accommodating the significant scale variations of buildings in high-resolution remote sensing imagery.

Furthermore, MMAENet demonstrates enhanced capability in capturing long-range dependencies. In terms of Recall, MMAENet achieves a background recall rate of 98.34% and a building recall rate of 96.74%, surpassing Swin Transformer by 0.16% and 1.98%, respectively. This advantage highlights the ability of the CAA module to effectively model long-range semantic dependencies between buildings and the background in high-resolution remote sensing imagery, thus mitigating missegmentation issues caused by the predominance of local features. In dense building area segmentation tasks, the CAA module significantly improves the accuracy of boundary localization through dynamic weighting. Additionally, the mean recall (mRecall) reaches 97.54%, outpacing the comparison models and further solidifying the robustness of the CAA module in long-range information extraction.

With its multi-level feature fusion mechanism, the PFP structure effectively guarantees the optimal utilization of multi-scale features. The ablation experiments presented in TABLE 4 show that the PFP structure considerably enhances the model's efficiency in utilizing building features. Specifically, the mIoU, mRecall, mF1-score, and mPrecision metrics all show significant improvements over the PKINet baseline. MMAENet achieves mIoU, mRecall, mF1-score, and mPrecision values of 91.69%, 97.54%, 95.56%, and 93.78%, respectively. These results demonstrate that the PFP structure, by fusing shallow high-resolution features with deep semantic-rich features, strikes an effective balance between local detail representation and global semantic understanding. In small building segmentation tasks, the PFP structure retains edge details using shallow features while leveraging deep features to suppress background noise, thereby enabling high-precision segmentation.

TABLE 4 Ablation experiments on the WHU Aerial imagery dataset

	mIoU (%)	mRecall (%)	mF1-score	mPrecision (%)
PKINet	89.30	96.25	94.16	92.31
PKINet+ PFP	**91.69**	**97.54**	**95.56**	**93.78**

5. Conclusion

This paper presents an innovative and efficient building extraction network, MMAENet, developed through a comprehensive analysis of multi-scale features of buildings and the long-range dependencies between buildings and background in high-resolution remote sensing imagery. MMAENet employs a PKI module a PFP structure to fully exploit building features across various levels and scales, while a CAA module effectively captures long-range dependencies between buildings and background, thereby significantly enhancing the accuracy of building extraction. To validate the efficacy of MMAENet, comparative experiments were conducted on the WHU Aerial Image Dataset, benchmarking against mainstream networks such as ConvNeXt, PSPNet, and Swin Transformer. The experimental results demonstrate that MMAENet achieves IoU, Recall, F1-score, and Precision of 85.44%, 96.74%, 92.15%, and 87.98%, respectively, for building extraction from high-resolution remote sensing imagery. These metrics show improvements of 2.68%, 1.98%, 1.59%, and 1.25% over Swin Transformer, highlighting its remarkable ability to precisely segment buildings and background regions in high-resolution remote sensing images. Notably, MMAENet proves highly effective in handling challenges such as blurred building edges and strong background interference, particularly in dense urban areas and complex backgrounds. In the future, Zero-Shot learning will be incorporated to enable the model to generalize to the segmentation of different types of buildings.

References

Cai, X., Lai, Q., Wang, Y., Wang, W., Sun, Z., Yao, Y., 2024. Poly

Kernel Inception Network for Remote Sensing Detection, in: 2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). Presented at the 2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), IEEE, Seattle, WA, USA, pp. 27706–27716.

Dosovitskiy, A., Beyer, L., Kolesnikov, A., Weissenborn, D., Zhai, X., Unterthiner, T., Dehghani, M., Minderer, M., Heigold, G., Gelly, S., Uszkoreit, J., Houlsby, N., 2020. An Image is Worth 16x16 Words: Transformers for Image Recognition at Scale.

Fan, Z., Liu, Y., Xia, M., Hou, J., Yan, F., Zang, Q., 2023. ResAt-UNet: A U-Shaped Network Using ResNet and Attention Module for Image Segmentation of Urban Buildings. IEEE J. Sel. Top. Appl. Earth Observations Remote Sensing 16, 2094–2111.

Hu, Q., Zhen, L., Mao, Y., Zhou, X., Zhou, G., 2021. Automated building extraction using satellite remote sensing imagery. Automation in Construction 123, 103509.

Huang, H., Liu, J., Wang, R., 2024. Easy-Net: A Lightweight Building Extraction Network Based on Building Features. IEEE Trans. Geosci. Remote Sensing 62, 1–15.

Kirillov, A., Girshick, R., He, K., Dollár, P., 2019. Panoptic Feature Pyramid Networks, in 2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). Presented at the 2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), Long Beach, CA, USA, pp. 6392-6401.

Li, J., Huang, X., Tu, L., Zhang, T., Wang, L., 2022. A review of building detection from very high resolution optical remote sensing images. GIScience & Remote Sensing 59, 1199–1225.

Li, W., He, C., Fang, J., Zheng, J., Fu, H., Yu, L., 2019. Semantic Segmentation-Based Building Footprint Extraction Using Very High-Resolution Satellite Images and Multi-Source GIS Data. Remote Sensing 11, 403.

Liu, Y., Zhao, Z., Zhang, S., Huang, L., 2022. Multiregion Scale-Aware Network for Building Extraction From High-Resolution Remote Sensing Images. IEEE Trans. Geosci. Remote Sensing 60, 1–10.

Liu, Z., Lin, Y., Cao, Y., Hu, H., Wei, Y., Zhang, Z., Lin, S., Guo, B., 2021. Swin Transformer: Hierarchical Vision Transformer using Shifted Windows, in: 2021 IEEE/CVF International Conference on Computer Vision (ICCV). Presented at the 2021 IEEE/CVF International Conference on Computer Vision (ICCV), IEEE, Montreal, QC, Canada, pp. 9992–10002.

Liu, Z., Mao, H., Wu, C.-Y., Feichtenhofer, C., Darrell, T., Xie, S., 2022. A ConvNet for the 2020s, in: 2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). Presented at the 2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), IEEE, New Orleans, LA, USA, pp. 11966–11976.

Puttinaovarat, S., Horkaew, P., 2017. Urban areas extraction from multi sensor data based on machine learning and data fusion. Pattern Recognit. Image Anal. 27, 326–337.

Qiu, W., Gu, L., Gao, F., Jiang, T., 2023. Building Extraction From Very High-Resolution Remote Sensing Images Using Refine-UNet. IEEE Geosci. Remote Sensing Lett. 20, 1–5.

Wang, L., Fang, S., Meng, X., Li, R., 2022. Building Extraction With Vision Transformer. IEEE Trans. Geosci. Remote Sensing 60, 1–11.

Zhao, H., Shi, J., Qi, X., Wang, X., Jia, J., 2017. Pyramid Scene Parsing Network, in: 2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR). Presented at the 2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), IEEE, Honolulu, HI, pp. 6230–6239.

Zhou, Y., Chen, Z., Wang, B., Li, S., Liu, H., Xu, D., Ma, C., 2022. BOMSC-Net: Boundary Optimization and Multi-Scale Context Awareness Based Building Extraction From High-Resolution Remote Sensing Imagery. IEEE Trans. Geosci. Remote Sensing 60, 1–17.