Proceedings of CONF-MLA 2025 Symposium: Intelligent Systems and Automation: AI Models, IoT, and Robotic Algorithms
DOI: 10.54254/2755-2721/2025.LD29991

The Role of U-Net Variants in Semantic Segmentation of Remote Sensing Images: A Survey

Yiyang Liu

School of Computer Science and Technology, Wuhan University of Science and Technology, Wuhan, China
aayang1048596@outlook.com

Abstract. Semantic segmentation of high-resolution remote sensing imagery is pivotal for applications such as land-cover mapping, urban planning, and environmental monitoring. Since the introduction of U-Net, numerous variants have been proposed to address challenges unique to satellite data—namely, extreme class imbalance, small-object detection, and complex scene textures. This survey systematically reviews major U-Net extensions (including U-Net++, ResUNet-a, HCANet, CCT-Net, DIResUNet, CM-UNet, TransUNet, AER-UNet and U-KAN) and additional optimization techniques such as incremental learning. This study compares their architectural innovations—e.g., nested skip connections, residual or atrous blocks, multi-scale context modules, and attention mechanisms—and summarizes reported performance on standard benchmarks (ISPRS Vaihingen, Potsdam, GID, WHDLD, DeepGlobe, and GF-2). This work also identifies key factors that drive segmentation accuracy and discusses remaining challenges and promising directions for future research, including improved generalization, reduced annotation dependency, and better trade-offs between performance and computational efficiency.

Keywords: Deep Learning, U-Net, Remote Sensing, Semantic Segmentation

1. Introduction

Satellite imagery has become an indispensable source of data for earth observation, environmental monitoring, urban planning, and disaster management [1]. Among the many tasks involving satellite data, semantic segmentation of remote sensing (RS) images plays a crucial role, enabling detailed land cover mapping, building footprint extraction, and change detection [1]. However, RS images pose unique challenges for segmentation algorithms due to their high spatial resolution, complex textures, class imbalance, and varying illumination conditions [2,3].

In recent years, deep learning-based methods have revolutionized the field of semantic segmentation, achieving remarkable performance on natural and RS images alike. Among them, the U-Net architecture has emerged as a particularly effective and popular choice, owing to its encoder–decoder structure and ability to capture both global context and fine details [4,5]. Since its introduction, U-Net and its numerous variants have been widely adopted and further developed for satellite image segmentation tasks [6,7].

© 2026 The Authors. This is an open access article distributed under the terms of the Creative Commons Attribution License 4.0 (https://creativecommons.org/licenses/by/4.0/).

DOI: 10.54254/2755-2721/2025.LD29991

This survey aims to provide a comprehensive overview of the role of U-Net variants in RS image semantic segmentation, summarizing key advances, architectural innovations, and practical applications. Specifically, it reviews notable improvements such as nested skip connections, residual and attention mechanisms, multi-scale feature aggregation, and the incorporation of novel paradigms like the Kolmogorov–Arnold Network (U-KAN). The paper also discusses benchmark datasets, experimental comparisons, and optimization strategies, followed by an analysis of current challenges and future research directions outlined in the conclusion section.

2. Background

Semantic segmentation is a fundamental task in the field of computer vision (CV), aiming to perform pixel-level classification of images. Unlike object detection or image classification, which focuses on locating or categorizing entire objects or scenes, semantic segmentation assigns a specific class label to each individual pixel, thereby enabling a more detailed understanding of visual content. It partitions the image into coherent regions based on semantic similarity, with pixels sharing the same category annotated with the same label.

This fine-grained analysis allows semantic segmentation algorithms to simultaneously recognize, detect, and delineate visual elements within a scene, significantly enhancing the precision and comprehensiveness of image interpretation. Compared with traditional CV tasks, semantic segmentation provides richer spatial and structural information, which is essential for applications that require precise localization and contextual understanding.

Due to its ability to extract detailed scene information, semantic segmentation has demonstrated great potential across a wide range of applications. In autonomous driving, for example, it enables the system to distinguish between roads, vehicles, pedestrians, and obstacles, thereby facilitating safe navigation. In the context of remote sensing (RS), semantic segmentation plays a vital role in numerous tasks.

Remote sensing (RS) image segmentation has become an indispensable task in the field of earth observation, providing the foundation for a wide range of applications such as disaster assessment, crop yield estimation, land cover mapping, and monitoring of environmental changes. As RS technology and data acquisition capabilities continue to advance, segmentation methods have evolved to meet the increasing demand for higher accuracy, efficiency, and scalability. Traditional methods, while effective in certain scenarios, often struggle to cope with the high spatial resolution, spectral heterogeneity, and complex textures present in RS images [8,9].

Deep learning approaches, particularly convolutional neural networks (CNNs), have revolutionized RS image segmentation by enabling automatic feature extraction and end-to-end training. Among these, the U-Net architecture, first introduced in 2015, has attracted widespread attention due to its elegant encoder–decoder structure, skip connections, and fully convolutional design [4]. U-Net builds upon the concept of fully convolutional networks (FCNs) and enhances them with symmetric downsampling and upsampling paths, which facilitate precise localization while capturing contextual information [5]. Table 1 outlines important parts of the U-Net.

DOI: 10.54254/2755-2721/2025.LD29991

Table 1. Descriptions of components

Component	Description
Encoder	Responsible for extracting increasingly abstract and high-level features from the input image. Consists of convolutional layers to generate feature maps in different resolutions.
Decoder	Implemented by transposed convolutions or interpolation—combined with convolutional layers to refine the feature representations.
Skip Connections	Skip connections establish direct links between corresponding layers of the encoder and decoder at the same resolution level.

The principal innovation of U-Net resides in its upsampling pathway, which leverages a substantial number of feature channels to effectively restore and propagate contextual information to high-resolution representations. As depicted in Figure 1, the network exhibits a symmetric, U-shaped architecture. The left side constitutes the encoder (contracting path), adopting the typical design of convolutional neural networks. This path applies repeated convolution and pooling operations, where each pooling layer reduces the spatial resolution of the feature maps while simultaneously enriching their depth. Such a mechanism enables the extraction of progressively abstract and semantically meaningful features essential for segmentation.

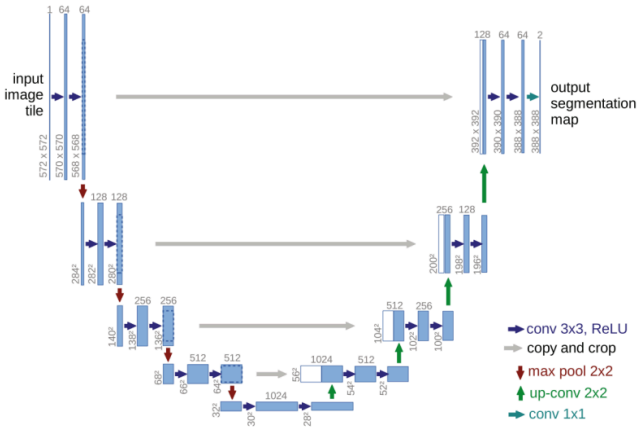

Figure 1. U-Net architecture

In contrast to standard CNNs, U-Net recovers the diminished spatial details through its decoder (expanding path) on the right, which incrementally upsamples the feature maps. To further improve segmentation fidelity, the U-Net integrates skip connections that directly link corresponding layers in the encoder and decoder. These connections function as shortcut pathways, fusing low-level, fine-grained features from the encoder with the higher-level decoder representations at matching resolutions. This strategy mitigates the information loss incurred during downsampling and facilitates more accurate and detailed segmentation outcomes. Figure 2 shows the number of publications about U-Net and remote sensing since 2018.

DOI: 10.54254/2755-2721/2025.LD29991

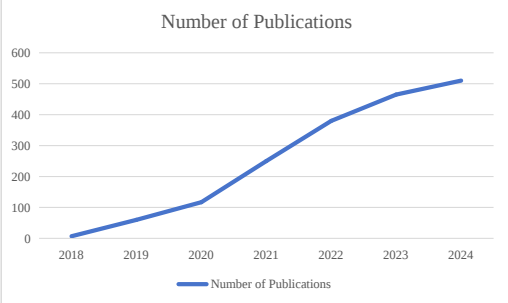

Figure 2. The number of U-Net remote sensing publications trend in last past years

U-Net remains a popular choice in various research domains because of its simple yet effective design, robustness, and versatility [10]. Its applications have extended beyond biomedical imaging to numerous RS tasks, including urban mapping, vegetation classification, and disaster monitoring. Over the years, many U-Net variants have been proposed to further improve its performance in RS scenarios, addressing challenges such as multi-scale object segmentation, boundary refinement, and long-range dependency modeling [6,7].

3. U-Net variants

U-Net is a suitable network for remote sensing image segmentation tasks. Since its simplicity and powerful scalability, the U-Net architecture has been modified into numerous variants. These modifications attempt to complement the shortcomings of the original U-Net in image segmentation tasks. This paper proposed several developments in the U-Net, which is used in remote sensing semantic segmentation.

U-Net++ [11] refines the standard U-Net through redesigned skip pathways and multi-scale deep supervision. In this structure, encoder outputs pass through dense convolution blocks whose depth depends on the pyramid level. Each layer connects to both the preceding convolution within the same block and upsampled outputs from lower-level blocks. Deep supervision applies auxiliary loss at various decoder stages rather than solely at the final output, accelerating convergence and improving results for small-scale targets and imbalanced datasets.

ResUNet-a [12] replaces the U-Net's conventional convolutions with enhanced residual units [13], maintaining stable gradients in deeper models. Within each residual block, parallel atrous convolutions [14,15] with different dilation rates expand the receptive field and capture spatial correlations across scales. A pyramid scene parsing pooling layer [16] further strengthens contextual understanding by incorporating broader scene information.

In HCANet [17], ResNet34 [18] serves as the encoder after removing its fully connected layers, forming a UNet–ResNet34 hybrid. The decoder upsamples low-resolution maps and fuses them with multi-scale features from the CASPP module. An extended CASPP+ module refines multi-scale feature extraction and aggregation for improved segmentation on high-resolution imagery.

CCT-Net [19] combines CNN-based local modeling with Transformer-based global context modeling [20]. The Local Adaptive Fusion Module (LAFM) and Coupled Attention Fusion Module (CAFM) integrate dual-branch features. During inference, Overlapping Sliding Window (OSW), Test-Time Augmentation (TTA), and post-processing (PP) steps help recover complete crop remote sensing maps.

DIResUNet [21] employs residual links to ease gradient flow and support deep feature learning. Atrous convolutions enlarge the receptive field without added computation, aiding recognition of

DOI: 10.54254/2755-2721/2025.LD29991

varying object scales. A pyramid scene parsing pooling layer boosts global awareness, while an adaptive Generalized Dice loss enhances convergence stability and generalization.

CM-UNet [22] applies channel attention in the encoder to emphasize critical shallow features. Residual connections bridge encoder and decoder paths for richer contextual use. Instead of typical upsampling, an improved sub-pixel convolution retains more details. A refined multi-feature fusion block preserves semantic information and reduces decoding loss, improving segmentation accuracy.

TransUNet [23] integrates Transformer components into both encoder and decoder, uniting self- and cross-attention for sequence-to-sequence segmentation. With a Transformer decoder and learnable queries, segmentation becomes a mask classification task. This design strategically embeds Transformers in the U-Net pipeline to address diverse segmentation challenges.

AER-UNet [24] enhances U-Net with attention mechanisms to suppress irrelevant background and highlight key regions via self-attention or attention gates. Residual blocks aid deep feature learning and avoid gradient issues. The model trains with an adaptive Adam optimizer (LR=0.001) that merges RMSprop, momentum, and SGD benefits, achieving a balance between fast convergence and stable learning while reducing overfitting risk.

U-KAN [25] introduces a Kolmogorov–Arnold Network (KAN)-based architecture into the U-Net framework. Unlike CNNs that rely on fixed kernel operations, KAN employs spline-based functions to model nonlinear mappings, enabling higher expressive capacity and smoother feature transformation. In U-KAN, both encoder and decoder blocks are constructed with KAN layers, preserving the hierarchical feature extraction ability of U-Net while enhancing adaptability to complex spatial patterns in remote sensing imagery. Furthermore, U-KAN integrates a lightweight attention refinement module to emphasize boundary information and suppress background noise. This design improves segmentation accuracy, particularly for small or irregularly shaped targets, while maintaining computational efficiency.

In addition to innovation on the Network, there are other optimization measures like incremental learning. Incremental learning offers a practical strategy for expanding semantic segmentation models to handle new classes over time without access to original labeled data. Tasar et al. [25] propose maintaining a frozen snapshot of the previous model as memory, and learning new classes by minimizing a composite loss: one term aligns old-class predictions to the frozen network (distillation), while another supervises new-class classification.

4. Dataset description and experimental discussion

This section introduces several benchmark datasets commonly used for semantic segmentation in remote sensing, with sample visualizations provided in Figure 3. Table 2 then compares the performance of various models on each dataset.

4.1. Dataset description

This section introduces a selection of widely used benchmark datasets for semantic segmentation in remote sensing. These datasets are derived from high-resolution satellite or aerial imagery captured by Earth observation sensors onboard satellites or aircraft. Each dataset offers distinct spatial resolutions, spectral characteristics, and land-cover annotation schemes, providing diverse evaluation scenarios for segmentation models.

DOI: 10.54254/2755-2721/2025.LD29991

4.1.1. ISPRS Vaihingen

The ISPRS Vaihingen dataset consists of very-high-resolution aerial imagery over a small German village [26]. It contains 33 ortho-rectified image patches (from a larger true orthophoto mosaic) with 9 cm ground sampling distance and an average size of about 2500×2000 pixels. Each image also includes a co-registered digital surface model (DSM). Of the 33 tiles, 16 are fully annotated and used for training, while the remaining 17 are held out for testing. Every pixel in the labelled tiles is classified into one of six land-cover categories: impervious surface (roads/concrete), building, low vegetation, tree, car, or clutter (undefined/background).

4.1.2. ISPRS Potsdam

The ISPRS Potsdam benchmark covers a 3.42 km^2 urban area subdivided into 38 equal-sized tiles [26]. Each tile is 6000×6000 pixels at 5 cm resolution. Every tile includes a four-band true orthophoto (near-infrared, red, green, and blue) and a co-registered one-band DSM, all on the same UTM/WGS84 grid. The orthophotos are provided in multiple channel combinations (e.g., IRRG, RGB, or 4-channel RGBIR) for convenience. A normalized DSM (nDSM) is also supplied, generated by ground filtering the DSM. This height-above-ground product was created automatically (without manual quality control) and may contain some errors. Ground truth labels (available for 24 of the 38 tiles) use the same six classes as Vaihingen: impervious surface, building, low vegetation, tree, car, and clutter. (The remaining 14 tiles are used as held-out test data in the standard benchmark.)

4.1.3. GID

The GID dataset is based on Gaofen-2 (GF-2) satellite imagery and covers large extents [27]. It comprises 150 high-resolution images (each 7200×6800 pixels) captured by GF-2, with a spatial resolution of about 0.8 m. Each image spans roughly 506 km^2 and contains four spectral bands (blue, green, red, and near-infrared). GID is divided into two parts: the GID-5 large-scale set and the GID-15 fine-grained set. In GID-5, each full-scene image is annotated at the pixel level with one of five broad land-cover classes (built-up, farmland, forest, meadow, or water). The dataset is split into 120 training images and 30 validation images. (The GID-15 subset further subdivides these into 15 more detailed classes, but we focus here on the five-class configuration.)

4.1.4. WHDLD

The WHDLD dataset consists of 4,940 small urban image chips over Wuhan, China [28]. These are RGB images of size 256×256 pixels with 2 m ground resolution, collected from Gaofen-1 and ZY-3 satellites. Each pixel is labeled into one of six categories: building, road, sidewalk, vegetation, bare soil (bare land), or water. This densely annotated urban dataset is used for semantic segmentation of city scenes.

4.1.5. DeepGlobe Land Cover

The DeepGlobe Land Cover dataset provides very-high-resolution satellite imagery (0.5 m GSD) for global land-cover classification [29]. It contains 1,146 RGB images, each of size 2448×2448 pixels. Every pixel mask is annotated with one of seven land-cover classes: urban land, agriculture,

DOI: 10.54254/2755-2721/2025.LD29991

rangeland, forest, water, barren land, or unknown. (In common usage, 803 of these images have labels for training and validation, while the rest form an unlabeled test set.)

4.1.6. GF-2

This dataset uses imagery from the Chinese Gaofen-2 (GF-2) satellite at 0.8 m resolution. Each image is 2000×2000 pixels. The raw GF-2 data are first pre-processed (for example, using ENVI software) to correct and mosaic the imagery. Ground truth labels are then created manually, often using MATLAB to paint different colors onto the images to indicate land-cover types. The specific classes depend on the study, but the annotation approach emphasizes diverse land-cover types distinguished by color.

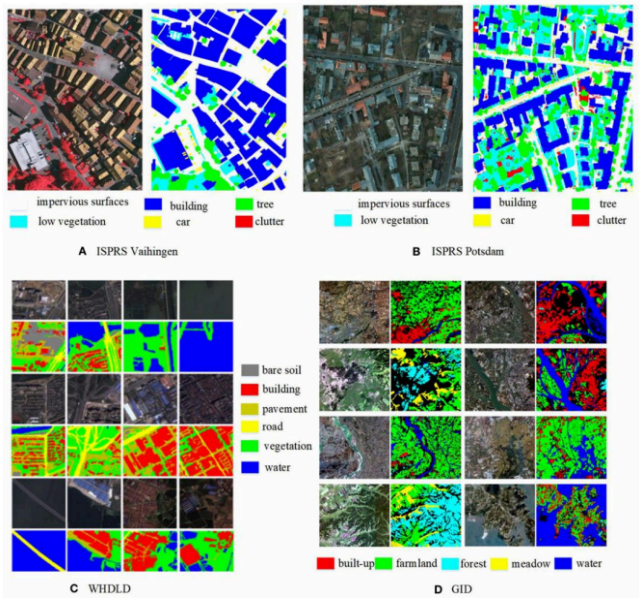

Figure 3. Visualization of the four common datasets [1]

Table 2. Comparison of different methods on ISPRS Potsdam and Vaihingen datasets

Models / Methods	ISPRS Potsdam			ISPRS Vaihingen		
	mF1 (%)	MIoU (%)	OA (%)	mF1 (%)	MIoU (%)	OA (%)
U-Net	84.08	72.10	87.05	-	-	-
U-Net++	-	72.36	88.23	-	-	-
ResUNet-a	92.9	-	91.5	-	-	-
HCANet	88.07	-	88.92	88.94	-	89.71
CM-UNet	90.4	87.41	-	90.22	88.12	-
Incremental Learning	-	-	84.25	-	-	87.44

4.2. Experimental comparison

Among the available benchmark datasets, ISPRS Vaihingen and ISPRS Potsdam are the most used for evaluating semantic segmentation methods in remote sensing. Table 2 summarizes the experimental results reported in related studies on these datasets, using mean F1 score (mF1), mean

DOI: 10.54254/2755-2721/2025.LD29991

Intersection over Union (mIoU), and Overall Accuracy (OA) as evaluation metrics. The U-KAN network is not included because of the lack of benchmark data on those datasets.

It is worth noting that the values cannot be used to compare the actual performance of the model. The reason for this is that the training set and test size of different models are different during experiments. However, the results still show us some useful information.

From the table, ResUNet-a performs prominently on Potsdam, with an mF1 of approximately 92.9% and an OA of 91.5%, indicating that its overall segmentation performance is superior to that of the traditional U-Net (with an mF1 of about 84% and an OA of 87%).

HCANet also achieves high accuracy (approximately 89% OA) on Vaihingen with stable indicators, demonstrating that its multi-scale scheme based on context fusion is robust in dense village environments.

CMUNet achieves mF1 scores in the 90–91 % range on both datasets, with Potsdam mIoU reaching 87.41 %, demonstrating a strong ability to distinguish medium- and high-frequency categories.

An incremental learning approach yields 87.44% OA on Vaihingen, showing promise for class-incremental adaptation even without intensive retraining.

4.3. Discussion

For model-specific strengths, we can discern the characteristics of these models from the data within Table 2. ResUNet-a integrates residual connections, alluring convolutions, and a multi-task structure, which has significantly improved the performance on small targets in Potsdam. HCANet enhances the ability to distinguish between buildings and vegetation in dense scenes by hierarchically fusing multi-scale contextual information through the CASPP module. CM-UNet leverages the channel attention mechanism to optimize the screening of shallow information and combines sub-pixel fusion to reduce information loss, which is an important reason for its high-precision performance.

From the aspect of dataset characteristics and model adaptability, Potsdam's finer spatial resolution (5 cm GSD) makes it particularly sensitive to small targets like cars and vegetation textures. ResUNeta and CMUNet excel here, benefiting from multi-scale modules and residual pathways. Besides, Vaihingen—with 9 cm resolution and a sparser, low-rise village layout—may favor methods like HCANet that exploit regional context and superpixel-level coherence.

5. Conclusion

This paper reviews the characteristics of several U-Net variants and their performance on common remote sensing semantic segmentation datasets from the angle of the DL framework and technology.

Despite the contributions of this study, several methodological limitations remain. In terms of remote sensing technology, the acquisition methods of remote sensing images were not described in detail, and only certain characteristics of the datasets were presented. With regard to deep learning, the technical strategies and architectural optimization methods employed by U-Net variants were not thoroughly explained or analyzed; rather, only their effects on model performance were discussed. Moreover, the number of reviewed studies on U-Net variants was relatively limited, and no large-scale comparative analysis was conducted.

Future work will aim to address these limitations. A more comprehensive investigation of deep learning techniques and U-Net architectures will be undertaken, with detailed explanations of the optimization strategies used in different variants. In addition, further study of remote sensing image

DOI: 10.54254/2755-2721/2025.LD29991

acquisition and classification will allow for a more elaborate description of dataset collection processes. Most importantly, future research will focus on integrating existing findings to design an efficient U-Net-based semantic segmentation framework.

References

[1] Lv, J., Shen, Q., Lv, M., Li, Y., Shi, L., & Zhang, P. (2023). Deep learning-based semantic segmentation of remote sensing images: a review. Frontiers in Ecology and Evolution, 11, 1201125.

[2] Ma, L., Liu, Y., Zhang, X., Ye, Y., Yin, G., & Johnson, B. A. (2019). Deep learning in remote sensing applications: A meta-analysis and review. ISPRS journal of photogrammetry and remote sensing, 152, 166-177.

[3] Zhu, X. X., Tuia, D., Mou, L., Xia, G. S., Zhang, L., Xu, F., & Fraundorfer, F. (2017). Deep learning in remote sensing: A comprehensive review and list of resources. IEEE geoscience and remote sensing magazine, 5(4), 8-36.

[4] Ronneberger, O., Fischer, P., & Brox, T. (2015, October). U-net: Convolutional networks for biomedical image segmentation. In International Conference on Medical image computing and computer-assisted intervention (pp. 234-241). Cham: Springer international publishing.

[5] Long, J., Shelhamer, E., & Darrell, T. (2015). Fully convolutional networks for semantic segmentation. In Proceedings of the IEEE conference on computer vision and pattern recognition (pp. 3431-3440).

[6] Yue, K., Yang, L., Li, R., Hu, W., Zhang, F., & Li, W. (2019). TreeUNet: Adaptive tree convolutional neural networks for subdecimeter aerial image segmentation. ISPRS Journal of Photogrammetry and Remote Sensing, 156, 1-13.

[7] Diakogiannis, F. I., Waldner, F., Caccetta, P., & Wu, C. (2020). ResUNet-a: A deep learning framework for semantic segmentation of remotely sensed data. ISPRS Journal of Photogrammetry and Remote Sensing, 162, 94-114.

[8] Lin, W., & Li, Y. (2020). Parallel Regional Segmentation Method of High-Resolution Remote Sensing Image Based on Minimum Spanning Tree. Remote Sensing, 12(5), 783. https: //doi.org/10.3390/rs12050783

[9] Fu, Z., Sun, Y., Fan, L., & Han, Y. (2018). Multiscale and Multifeature Segmentation of High-Spatial Resolution Remote Sensing Images Using Superpixels with Mutual Optimal Strategy. Remote Sensing, 10(8), 1289. https: //doi.org/10.3390/rs10081289

[10] Ramos, L. T., & Sappa, A. D. (2025). Leveraging U-Net and selective feature extraction for land cover classification using remote sensing imagery. Scientific Reports, 15(1), 784.

[11] Zhou, Z., Siddiquee, M. M. R., Tajbakhsh, N., & Liang, J. (2018). UNet++: A Nested U-Net Architecture for Medical Image Segmentation. Deep Learn Med Image Anal Multimodal Learn Clin Decis Support (2018), 11045, 3-11. https: //doi.org/10.1007/978-3-030-00889-5_1

[12] Diakogiannis, F. I., Waldner, F., Caccetta, P., & Wu, C. (2020). ResUNet-a: A deep learning framework for semantic segmentation of remotely sensed data. ISPRS Journal of Photogrammetry and Remote Sensing, 162, 94–114. https: //doi.org/10.1016/j.isprsjprs.2020.01.013

[13] He, K., Zhang, X., Ren, S., & Sun, J. (2016). Identity mappings in deep residual networks. In Lecture notes in computer science (pp. 630–645). https: //doi.org/10.1007/978-3-319-46493-0_38

[14] Chen, L. C., Papandreou, G., Kokkinos, I., Murphy, K., & Yuille, A. L. (2017). Deeplab: Semantic image segmentation with deep convolutional nets, atrous convolution, and fully connected crfs. IEEE transactions on pattern analysis and machine intelligence, 40(4), 834-848.

[15] Chen, L. C., Papandreou, G., Schroff, F., & Adam, H. (2017). Rethinking atrous convolution for semantic image segmentation. arXiv preprint arXiv: 1706.05587.

[16] Zhao, H., Shi, J., Qi, X., Wang, X., & Jia, J. (2017). Pyramid scene parsing network. In Proceedings of the IEEE conference on computer vision and pattern recognition (pp. 2881-2890).

[17] Bai, H., Cheng, J., Huang, X., Liu, S., & Deng, C. (2022). HCANet: A Hierarchical Context Aggregation Network for Semantic Segmentation of High-Resolution Remote Sensing Images. IEEE Geoscience and Remote Sensing Letters, 19, 1-5. https: //doi.org/10.1109/lgrs.2021.3063799

[18] He, K., Zhang, X., Ren, S., & Sun, J. (2016). Deep residual learning for image recognition. In Proceedings of the IEEE conference on computer vision and pattern recognition (pp. 770-778).

[19] Wang, H., Chen, X., Zhang, T., Xu, Z., & Li, J. (2022). CCTNet: Coupled CNN and Transformer Network for Crop Segmentation of Remote Sensing Images. Remote Sensing, 14(9). https: //doi.org/10.3390/rs14091956

[20] Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., ... & Polosukhin, I. (2017). Attention is all you need. Advances in neural information processing systems, 30.

[21] Priyanka, N, S., Lal, S., Nalini, J., Reddy, C. S., & Dell'Acqua, F. (2022). DIResUNet: Architecture for multiclass semantic segmentation of high resolution remote sensing imagery data. Applied Intelligence, 52(13), 15462-15482.

DOI: 10.54254/2755-2721/2025.LD29991

https: //doi.org/10.1007/s10489-022-03310-z
[22] Cui, M., Li, K., Chen, J., & Yu, W. (2023). CM-Unet: A Novel Remote Sensing Image Segmentation Method Based on Improved U-Net. IEEE Access, 11, 56994-57005. https: //doi.org/10.1109/access.2023.3282778
[23] Chen, J., Mei, J., Li, X., Lu, Y., Yu, Q., Wei, Q., Luo, X., Xie, Y., Adeli, E., Wang, Y., Lungren, M. P., Zhang, S., Xing, L., Lu, L., Yuille, A., & Zhou, Y. (2024). TransUNet: Rethinking the U-Net architecture design for medical image segmentation through the lens of transformers. Med Image Anal, 97, 103280. https: //doi.org/10.1016/j.media.2024.103280
[24] Jonnala, N. S., Siraaj, S., Prastuti, Y., Chinnababu, P., Praveen babu, B., Bansal, S., ... & Al-Mugren, K. S. (2025). AER U-Net: attention-enhanced multi-scale residual U-Net structure for water body segmentation using Sentinel-2 satellite images. Scientific Reports, 15(1), 16099.
[25] Li, C., Liu, X., Li, W., Wang, C., Liu, H., Liu, Y., Chen, Z., & Yuan, Y. (2025). U-KAN Makes Strong Backbone for Medical Image Segmentation and Generation. Proceedings of the AAAI Conference on Artificial Intelligence, 39(5), 4652-4660. https: //doi.org/10.1609/aaai.v39i5.32491
[26] Tasar, O., Tarabalka, Y., & Alliez, P. (2019). Incremental learning for semantic segmentation of large-scale remote sensing data. IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing, 12(9), 3524-3537.
[27] 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.
[28] Tong, X. Y., Xia, G. S., Lu, Q., Shen, H., Li, S., You, S., & Zhang, L. (2020). Land-cover classification with high-resolution remote sensing images using transferable deep models. Remote Sensing of Environment, 237, 111322.
[29] Shao, Z., Yang, K., & Zhou, W. (2018). Performance Evaluation of Single-Label and Multi-Label Remote Sensing Image Retrieval Using a Dense Labeling Dataset. Remote Sensing, 10(6), 964. https: //doi.org/10.3390/rs10060964
[30] Demir, I., Koperski, K., Lindenbaum, D., Pang, G., Huang, J., Basu, S., ... & Raskar, R. (2018). Deepglobe 2018: A challenge to parse the earth through satellite images. In Proceedings of the IEEE conference on computer vision and pattern recognition workshops (pp. 172-181).