UNet-LSCNet: Integrating Dynamic Snake Convolution and Vision Transformer for Water Body Extraction with Complex Boundaries

Yukai Zhang , Xi Zhang , Zhenhua Wang , Wanwen He *

Posted Date: 20 April 2026

doi: 10.20944/preprints202604.1325.v1

Keywords: water body extraction; semantic segmentation; dynamic snake convolution; vision transformer; sentinel-2

 doi:10.20944/preprints202604.1325.v1

Article

UNet-LSCNet: Integrating Dynamic Snake Convolution and Vision Transformer for Water Body Extraction with Complex Boundaries

Yukai Zhang, Xi Zhang, Zhenhua Wang and Wanwen He *

College of Information Technology, Shanghai Ocean University, Shanghai 201306, China

* Correspondence: d250700181@st.shou.edu.cn

Abstract

Water body extraction using remote sensing is crucial for ecological environment monitoring and water resource management. Nevertheless, the irregular and complicated shapes of water bodies make it difficult to obtain fine-grained characterization and preserve structural consistency with current approaches. To overcome the shortcomings of fixed receptive fields and sampling schemes of traditional convolutional networks, this paper proposes UNet-LSCNet, an advanced architecture based on U-Net. The proposed model integrates dynamic snake convolutions (DSConv), the convolutional block attention module (CBAM), and a lightweight Vision Transformer (LaViT) to enable local adaptive geometric modeling and contextually enriched semantics representation. The experimental findings demonstrate that the suggested approach outperforms various widely applied models. Specifically, UNet-LSCNet achieves a mIoU of 95.67% and an F1-score of 96.32%, while maintaining a competitive inference speed of 4.18 frames per second (FPS). Furthermore, it exhibits greater stability in highly complex situations, such as slender meandering rivers and fragmented small-scale water bodies. Ablation experiments confirm the synergistic utility of each module, revealing that the proposed model enhances segmentation accuracy and morphological resilience without compromising inference efficiency.

Keywords: water body extraction; semantic segmentation; dynamic snake convolution; vision transformer; sentinel-2

1. Introduction

Water bodies constitute a crucial natural element within the Earth's surface system. Their spatial distribution patterns, temporal variation characteristics, and geometric morphology are vital for ecological and environmental assessments, water resource management, flood control and regional sustainable development [1]. With climate change and escalating human activity, the dynamic evolution of water body morphology exhibits heightened complexity and uncertainty, demanding more sophisticated and automated methods for acquiring water body information [2]. Consequently, developing a water body extraction method that balances accuracy, stability, and continuous operational capability holds significant scientific and practical value.

Traditional water body extraction methods can be broadly categorized into three types: threshold-based methods, traditional machine learning methods, and hybrid approaches. Threshold-based methods primarily utilize the spectral reflectance characteristics of water in specific bands. These include single-band thresholding [3] and multi-band mathematical operations, such as the Normalized Difference Water Index (NDWI) [4], Modified NDWI (MNDWI) [5], Automated Water Extraction Index (AWEI) [6], Background Difference Water Index (BDWI) [7], Sentinel Multi-Band Water Index (SMBWI) [8], Green-SWIR1 Reflectance Index (GSRI) [9], and the Normalized Difference Water Fraction Index (NDWFI) [10]. Traditional machine learning methods, e.g., Support Vector Machines (SVM) [11], Decision Trees (DT) [12], and Random Forests (RF) [13], formulate extraction

 doi:10.20944/preprints202604.1325.v1

as a pixel-wise classification problem. Hybrid methods employ strategies that integrate water features with multi-classifier ensembles to achieve high-precision extraction [14]. However, hybrid processing is often complex and uncertain. Overall, traditional methods rely heavily on expert knowledge and possess limited feature representation capabilities, making it difficult to capture deep semantic information and spatial relationships between pixels [15,16].

In recent years, with the rapid advancement of deep learning technologies, deep networks based on semantic segmentation have gradually become a core direction in remote sensing water body extraction research [17,18]. For instance, Wang et al. [19] proposed a multi-scale lake water extraction network to address the issue of large intra-class and small inter-class variances. Zhang et al. [20] designed a multi-feature extraction combination network to increase the diversity of water features while enhancing semantic information. Kang et al. [21] improved extraction accuracy by integrating multi-scale information extracted by a ResNet encoder with contextual information from a multi-kernel pooling unit. Furthermore, Liang et al. proposed MATLinkNet, incorporating a multi-scale pyramid structure to enhance the representation of multi-scale water targets [22]. Cao et al. [23] integrated multi-scale feature fusion into a U-Net architecture to achieve joint modeling of local details and global contexts, thereby strengthening boundary representations.

Despite these advancements, high-precision water body extraction still faces severe challenges. For optical imagery, major interfering factors include clouds, cloud shadows, terrain shadows, dark land surfaces, snow, and ice, which share similar spectral characteristics with water. To address these interference issues, researchers typically employ methods such as feature enhancement, interference suppression, and the use of auxiliary data [24,25]. Another critical challenge lies in the complex morphological boundaries and topological structures of lakes and rivers. To preserve the characteristics of tiny water bodies while enhancing land-water boundary information, models are often optimized through deep-shallow feature fusion, and loss function refinements [26,27].

Attention mechanisms have proven to be an effective strategy for highlighting water features while suppressing redundant background noise [28–31]. Among them, the CBAM [32] has proven to be a highly effective strategy for adaptively emphasizing key features. Simultaneously, the Transformer architecture has been increasingly introduced to model global contextual dependencies [33,34]. For instance, Kang et al. [35] proposed a model combining Convolutional Neural Networks (CNN) and Vision Transformers (ViT) to exploit both low-level spatial information and high-level semantic correlations. Wang et al. designed the MSFSwin network, combining an enhanced Swin Transformer with a feature fusion module to perceive water regions under complex terrains. To mitigate the high computational complexity of standard ViTs, lightweight local Transformer has been designed [36]. Specifically, the Local-Aggregation Vision Transformer (LaViT) achieves a balance between local context modeling and computational efficiency by performing self-attention computations within local windows [37,38].

Building upon this foundation, local geometric adaptation mechanisms are required to address the inherent issues of meandering, multi-scale, and irregular water body boundaries. DSConv [39] introduces learnable deformation offset parameters, enabling the convolution kernel's sampling position to adaptively adjust along target structures. Recent studies indicate that such deformable convolutions demonstrate strong boundary modeling capabilities in extraction tasks with distinct structural priors [40,41].

Motivated by these insights, this paper proposes the UNet-LSCNet, which integrates DSConv, CBAM, and LaViT. This model synergistically optimizes remote sensing water body extraction across three dimensions: adaptive modeling of local boundary structures, context-dependent modeling within local windows, and selective enhancement of spectral-spatial features. The proposed approach significantly improves the structural continuity and boundary precision of water bodies in complex scenarios while maintaining a favorable trade-off with computational efficiency. The contributions of this paper are as follows:

(1) DSConv is introduced into the U-Net architecture. By employing learnable sampling offsets, the convolution kernel adaptively adjusts along water body edges, overcoming the limitations of

 doi:10.20944/preprints202604.1325.v1

regular grid sampling. This effectively captures curved and narrow water bodies, and complex boundary structures, which improves the water body extraction accuracy.

(2) A lightweight LaViT module is incorporated into the bottleneck layer. This module models correlations between neighboring features through a local window self-attention mechanism, reusing aggregated representations to achieve effective integration of local contextual information without imposing an excessive computational burden.

(3) We conducted systematic experiments across diverse typical scenarios, such as large lakes, densely distributed ponds, and elongated rivers. The experimental results demonstrate that the proposed model exhibits superior accuracy and morphological stability compared to existing mainstream methods.

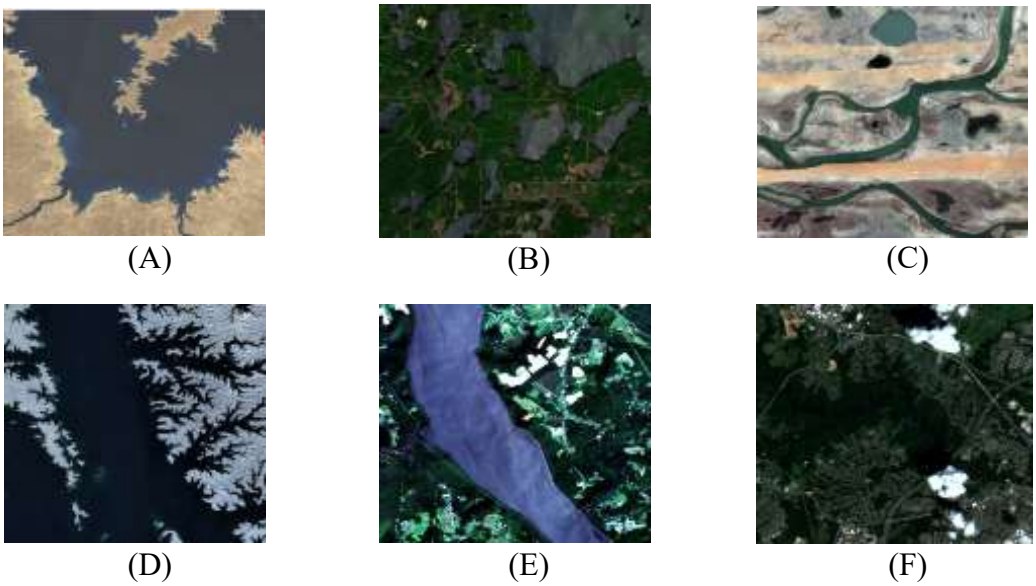
The remainder of this paper is organized as follows: Section 2 presents the dataset and the methodology of the proposed method; Section 3 elaborates on the experimental results and compares different methods. The role of each module is analyzed and discussed in Section 4. Finally, Section 5 concludes the paper.

2. Materials and Methods

2.1. Dataset

The Earth Surface Water Body Knowledge Base (ESWKB) dataset is used in this study [42]. The dataset can be publicly accessible through the following repository: https://github.com/xinluo2018/WatNet. There are 95 image scenes that use 31 Sentinel-2 satellite images, covering surface water bodies in different ecological settings such as mountains, flats, plateaus and urban-rural fringe areas. Given its extensive geographical coverage and diverse scenarios, this dataset is highly appropriate for surface water body extraction and segmentation research.

To assess the effectiveness of the proposed UNet-LSCNet model, the dataset has been divided into a training and testing subsets. Out of the 95 scenes available, 89 scenes were used to train models whereas the other six scenes were set aside as testing. The chosen testing images exhibit significant variation in water body morphology, encompassing different scales, geometric forms, channel widths and spatial distributions. The evaluation focused on challenging scenarios, ranging from large-scale continuous water bodies with broad borders to narrow, elongated watercourses with highly complex edges. Examples of the test scenarios are illustrated in Figure 1.



(A) (B) (C)

(D) (E) (F)

 doi:10.20944/preprints202604.1325.v1

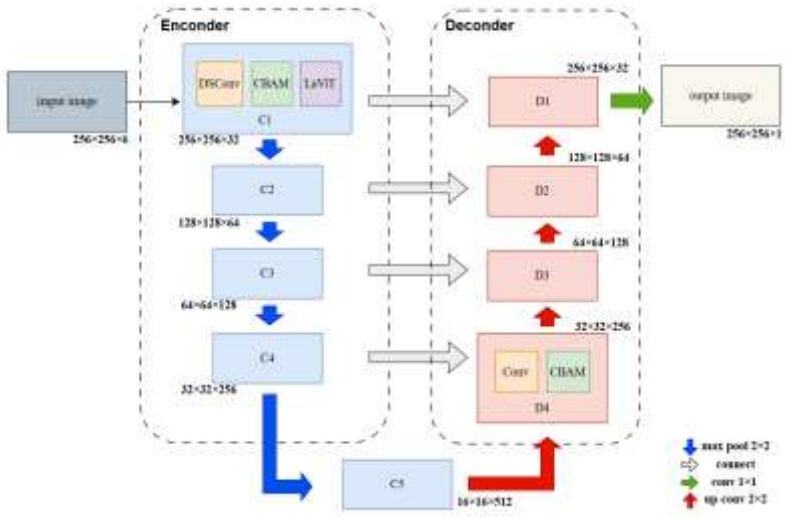
Figure 1. Illustration of the selected test images. The six kinds of images are: (A) large-scale continuous water bodies; (B) narrow, meandering river water bodies; (C) fragmented, small-scale water bodies; (D) high-density water networks with complicated borders; (E) major-channel-dominant river water bodies; (F) narrow, linear water bodies with complicated borders.

2.2. Overall Network Architecture

At present, various classic and new frameworks have been developed in the research area of semantic segmentation, and they have been used in the large-scale tasks of remote sensing image interpretation. Prominent examples include U-Net [43], DeepLabV3+ [44], and the ResNet series [45]. UNet is a symmetric encoder-decoder architecture and makes use of skip connections to perform multi-scale feature fusion, which is computationally efficient and stable when performing pixel-level segmentation tasks. DeepLabV3+ uses both dilated convolutions and the Atrous Spatial Pyramid Pooling (ASPP) module to improve segmentations in complicated scenes by means of multi-scale contextual fusion. ResNet efficiently addresses the vanishing gradient issue in deep networks through residual links, often being implemented as a high performance encoder structure. These convolutional neural networks, based on the encoder-decoder model, exhibit a positive balance between feature representation ability, generalization capability, and engineering workability. Consequently, they are widely applied to semantic segmentation work on surface water bodies.

Based on this background, this paper uses U-Net as a whole network structure, including the DSConv, CBAM attention mechanism, and the lightweight visual Transformer (LaViT) to formulate a new water body extraction network UNet-LSCNet. The model is efficient in terms of the encoder-decoder structure of U-Net with the benefit of end-to-end training and adds local geometric adaptive modeling, channel-spatial joint attention, and the mechanism of contextual improvement. The improvements have been shown to enhance the representation of complex water body shapes and small-scale objects.

Figure 2 illustrates the overall structural design of the proposed network. UNet-LSCNet follows a classical symmetric U-Net encoder-decoder model, which consists of an encoder, bottleneck layer and a decoder. In the process of the features being extracted, the synergy between the DSConv and the CBAM module helps improve the performance of the model in terms of the irregular water body edges, as well as other important aspects. At the same time, LaViT module is presented as a measure of reducing computation costs due to the ability to calculate attention less often and reuse learnt attention relationships. It provides the alignment to features and representational capacity but addresses the problem of excessive deep attention. It accepts six-band Sentinel-2 remote sensing images as inputs and delivers a binary segmentation result based on water bodies. It goes through optimization and performance monitoring during training using various pixel-level evaluation and structural-level evaluation indicators.



 doi:10.20944/preprints202604.1325.v1

Figure 2. The proposed UNet-LSCNet architecture.

2.3. DSConv module

As shown in Figure 3, DSConv differs significantly from regular convolution. It was introduced by Qi et al. [39], to address the drawbacks of conventional convolution using a fixed grid sampling when modeling irregular target geometries. By introducing learnable dynamic offsets to the convolutional kernels, DSConv allows sampling points to align with the target structure along a continuous, smooth curve. In contrast to the hard spatial sampling properties of typical convolutions, DSConv dynamically optimizes the receptive fields depending on local structure. This can greatly improve the representation of irregular geometric shapes with minimal change to the form of convolution being calculated.

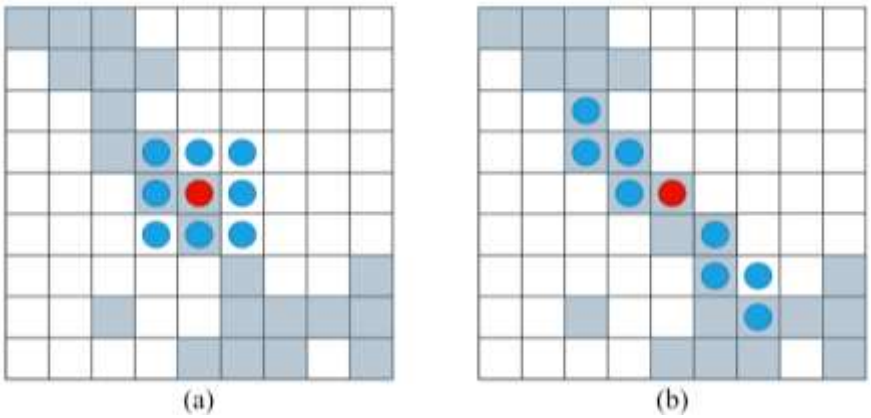


Figure 3. The Difference between Standard Convolution and DSConv [39]. (a) Standard Convolution; (b) DSConv.

Water boundaries often exhibit nature-made complex and winding shape, especially at the confluence of tributaries, narrow river banks, and lake shores. High-curvature regions change abruptly, making them difficult to accurately represent using ordinary convolutions, which frequently causes edge blurring or structural gaps. . To overcome these difficulties, DSConv is integrated into both the encoder and decoder of the proposed network. This allows the convolution kernel to undergo slight, continuous deformations along the water body edge, thereby improving the model's capacity to delineate complex water edges and fine-scale structures.

Specifically, in connection with an entry feature map, referred to as X, the dynamic snake convolution first generates the spatial offsets through an offset prediction branch.

These offsets dynamically alter the sampling locations for subsequent feature aggregation. The offset Δ is calculated as [39]:

$$\Delta = \tanh\left(f_{\text{offset}}(X)\right) \quad (1)$$

The hyperbolic tangent (tanh) function is employed to constrain the magnitude of the offsets, thereby ensuring stable model convergence during training. Subsequently, the calculated offset is added to the original feature coordinates [39]:

$$X' = X + \Delta \quad (2)$$

A standard convolution is then applied to the deformed features X′. This adaptive manner without imposing excessive computational costs which leads to increasing sensitivity to water-boundary-curvature and low-contrast variations in the network.

 doi:10.20944/preprints202604.1325.v1

2.4. CBAM Attention Module

The configuration of the CBAM segment is represented in Figure 4. Two sub components in series are included in this model: Channel Attention and Spatial Attention. First, Average and Maximum pooling are applied to extract the global context by Channel Attention. A shared Multi-Layer Perceptron (MLP) is then utilized to generate the channel attention map M_c [32].

$$M_c = \sigma\,(\,\mathrm{MLP}\,(\,\mathrm{AvgPool}\,(\,X\,)) + \mathrm{MLP}\,(\,\mathrm{MaxPool}\,(\,X\,)))\tag{3}$$

Where σ denotes the sigmoid activation function. The input features are then multiplied element-wise by this channel attention map. In the subsequent Spatial Attention step, maximum pooling and average pooling are applied to the channel-weighted features. The results are then concatenated, each one being passed through a 7 x 7 convolution to create the resulting spatial attention map [32].

$$M_s = \sigma\,(f^{7\times7}([\mathrm{AvgPool}_c(\,X\,))\,,\,\mathrm{MaxPool}_c(\,X\,)]))\tag{4}$$

Finally, the refined output feature X″ is obtained through element-wise multiplication [32]:

$$\mathrm{X}\text{‘’} = X \odot M_c \odot M_s\tag{5}$$

By incorporating the CBAM model, the network can concentrate on important spectral characteristics of water surfaces and their respective areas, thereby increasing the precision of recognizing small and fragmented water bodies [32].

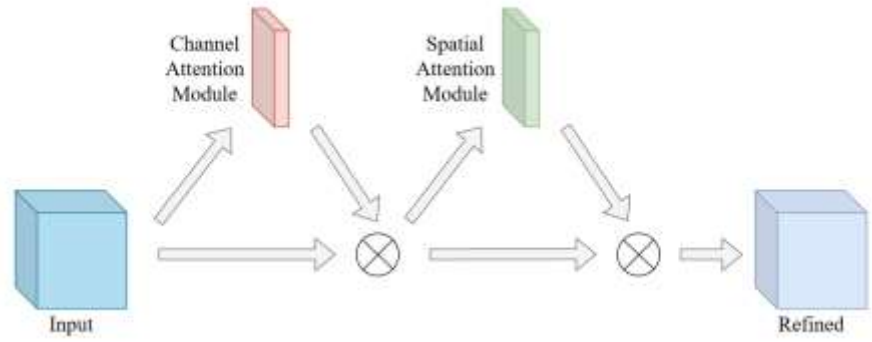


Figure 4. The network structure of the CBAM module [32].

2.5. The LaViT Module

Figure 5 illustrates the construction of the LaViT module. The proposed module has been designed to address the problem of high computational overhead and attention redundancy of layer-wise self-attention calculations in Vision Transformers according to the idea by Zhang et al [36]. Its core mechanism includes minimizing attention operations at each step and reusing previously learned attention relationships based on an attention transformation mechanism. This provides an efficient way to improve total computational performance without degrading feature representation ability.

 doi:10.20944/preprints202604.1325.v1

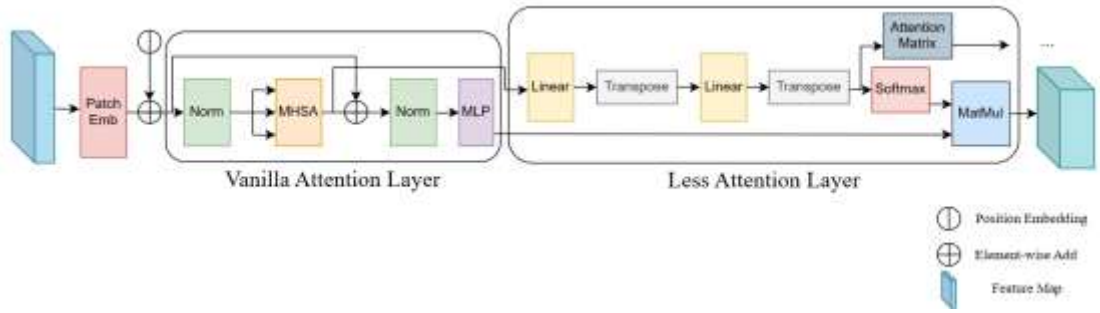


Figure 5. Design of the LaViT module [36].

LaViT takes a hybrid design approach that combines both local convolutional embedding and windowed self-attention, with the goal of enhancing the ability of the network to learn dependencies between features inside a local window at minimal computational cost. First, the input features *B* that go to the bottleneck layer are locally embedded via typical 3 x 3 convolutions, which enhances local structural information and generates the initial feature representation that may be used in transformer-based steps. The local embedding process is defined as [36]:

$$B_c = f_{conv}(B) \tag{6}$$

Here, f_{conv} denotes the standard convolution operation, which provides a stable neighborhood prior to the self-attention process in order to increase the locality of the attention calculations.

Subsequently, the embedded features are divided into several local windows of constant size, every window contains the multi-head self-attention mechanism. Assume that *X* represent the feature sequence within a particular window; the general form of window self-attention is [36]:

$$\mathrm{MSA}(X) = \mathrm{Concat}(\mathrm{head}_1, \ldots, \mathrm{head}_M) W^O \tag{7}$$

where each head is computed as [36]:

$$head_i = Softmax\left(\frac{Q_i K_i^T}{\sqrt{d}}\right) V_i$$

$$Q_i = XW_i^Q, K_i = XW_i^K, V_i = XW_i^V$$

In order to reduce the possible problem of saturation in attention computations in deep networks, LaViT repeats the use of aggregated attention features representations across adjacent layers. It allows the further layers to slowly incorporate more extensive contextual details at each stage of the process, eliminating the need to repeat the entire self-attention computation at every step, thus increasing the stability of feature representations and keeping the local models [36].

$$g = \mathrm{GAP}(B_c) \tag{8}$$

$$X_m = X_{MSA} + W_g g \tag{9}$$

Here is the train-able weight matrix which modifies the effect intensity of global tokens over the features in the window.

Here, W_g denotes the trainable weight matrix, which is used to modulate the influence of the global token g on the local features within the window; X_{MSA} denotes the window self-attention output applied to the local convolutional embeddings, which is used to capture contextual relationships within each window.

Once local convolutional embeddings, windowed self-attention and global tokens are integrated, the result is the final output of LaViT which is combined with the input features via a residual link in order to enhance the ability of training [36]:

 doi:10.20944/preprints202604.1325.v1

$$B^{'} = B + \mathrm{LaViT}(B) \tag{10}$$

The residual design ensures that the Transformer architecture acts as an additional enhancement module of the existing feature distribution [36].

2.6. Loss Function

The sparse pixel-level cross-entropy loss is employed in the water body classification task to ensure high pixel-level prediction accuracy and sufficient inter-class discriminability. The input image $X(X \in R^{H\times W\times C})$and the ground truth labels ($Y\in\{0,1\}^{H\times W}$) are given and the model output is in terms of probabilities of each class P. The loss function is defined as the average cross-entropy across all pixels [46].

$$\mathcal{L}_{CE} = -\frac{1}{H\times W}\sum_{\mathrm{h}=1}^{H}\sum_{\omega=1}^{W}\sum_{c=0}^{1} 1\{Y_{\mathrm{h},\omega}=c\}\log P_{\mathrm{h},\omega,c} \tag{11}$$

Where the indicator function 1 {} is equal to 1 as the pixel label is equal to category c, and it is equal to 0 in case of all other categories.

2.7. Evaluation Metrics

To quantitatively evaluate the effectiveness of the proposed UNet-LSCNet, six evaluation metrics are employed: Accuracy, Recall, Precision, F1-score, Kappa and Mean Intersection over Union (mIoU). Accuracy measures the overall proportion of correctly classified pixels. Precision is the proportion of the pixels which are supposed to be water and are actually water and this will be used to assess reliability of predictions. Recall represents the proportion of actual water pixels successfully identified by the model. The F1-score is the harmonic mean of Precision and Recall, providing a balanced measure of the model's detection performance [47]:

$$\mathrm{Precision} = \frac{TP}{TP+FP} \tag{12}$$

$$\mathrm{Recall} = \frac{TP}{TP+FN} \tag{13}$$

$$\mathrm{F1\_score} = \frac{2\,Precision Recall}{Precision+Recall} \tag{14}$$

TP is the number of pixels that were correctly identified as water.

FP is the number of background pixels that were incorrectly identified as water.

FN is the number of water pixels that were incorrectly identified as background.

TN means the count of pixels identified as background. We compute the accuracy rate in this study as follows:

$$\mathrm{Accuracy} = \frac{TP+TN}{TP+TN+FP+FN} \tag{15}$$

The mIoU was also calculated to have a good understanding of how the model performed in segmenting various classes of objects [50]:

$$\mathrm{mIoU} = \frac{1}{C}\sum_{i=1}^{C}\frac{TP_i}{TP_i+FP_i+FN_i} \tag{16}$$

Where C is the total number of categories. In this binary classification task C=2. During training, these metrics are computed iteratively at the end of each epoch using customized TensorFlow metric classes, allowing for real-time performance tracking.

 doi:10.20944/preprints202604.1325.v1

3. Results

3.1. Experimental Environment and Training Model

The model training and assessment were conducted in a standardized experimental environment. The main internal components of the computer are Intel Core i7-13700H CPU, 64 GB of memory, and NVIDIA GeForce RTX 4060 with 8 GB of GDDR6 memory and support of CUDA 11.0 that allows accelerating deep learning. TensorFlow 2.12.0 and Keras were selected to perform deep learning and Python 3.10 used as the primary language. The environment was developed with PyCharm 2025.1. All images were resized to 512x512 pixels. The models were trained for 200 epochs with a batch size of 16, using an initial learning rate of 0.002.

The image composed of 6 bands based on Sentinel-2 data and their respective single-channel labels had been preprocessed by the process of path matching and pairing to have proper alignment. The next step was to divide the dataset randomly into three sets: training set, validation set, and test set. Care was taken so as not to overlap between training and test sets but maintain the same overall distribution of data. In preprocessing, the images have been converted into a floating point representation, to accommodate the process of feature learning in the network, followed by compressing label information into single channel binary images to be consistent with water body extraction task purpose.

3.2. Methods of Comparison

After the creation of three data subsets, a sample pipeline that can be read directly by the network was built. The pipeline simplified the processes of image loading, format change, and resizing, serving as the source of data fed into the model using batch processing. In validation, a constant amount of image blocks of the same size were randomly cropped out of the original data to increase the effectiveness and resilience of the assessment process. A group of data augmentation methods was used in training to make the model more robust to spatial rotations and flipping transformations in remote sensing images.

It consisted of random horizontal flipping, vertical flipping, and directional changes due to right-angle rotations. Such augmentations were successful in increasing the morphological variability of training examples so that the network could support steady segmentation under more complicated space patterns.

To comprehensively evaluate the performance of the proposed UNet-LSCNet, this study selected four representative semantic segmentation networks as baselines: SegNet [49], DeepLabV3+ [50], HRNet [51], and PSPNet [52]. These models were chosen because they represent four distinct and classic architectures in remote sensing image interpretation: encoder-decoder, dilated convolutions, multi-resolution feature fusion, and pyramid pooling, respectively, and are widely deployed in waterbody segmentation tasks. Their fundamental principles are described as follows:

SegNet: It is a typical encoder-decoder structure that uses max-pooling indices from the encoding phase to perform non-linear upsampling in the decoding phase. In contrast to conventional deconvolution-based reconstruction, this index-based upsampling is more efficient in preserving edge contour information while decreasing model parameters, thus providing consistent results in water body segmentation task.

DeepLabV3+: This network integrates dilated convolution with the ASPP, enabling the simultaneous capture of local textures and large-scale contextual information. Moreover, the model uses a small decoder structure to promote detailed boundary recovery, making it highly robust in challenging remote sensing backgrounds.

HRNet: The emphasis is on keeping a high-resolution branch during the course of feature extraction instead of down sampling to lower resolution and reconstruction as in conventional methods and multi stage feature exchange with other branches of different scales. The architecture maintains small scale information of space without compromising large scale receptive areas

 doi:10.20944/preprints202604.1325.v1

providing strong benefits to accurately locate edges of an object. Since remote sensing images have highly fragmented and complex edges of water bodies, HRNet can be used for comparison purpose.

PSPNet: The core of PSPNet is its pyramid pooling module that permits global context modelling of feature maps at various scales. The model excels at understanding structural relationships between scenes and distinguishing between visually homogeneous regions, making it highly effective for separating water bodies from backgrounds.

3.3. Analysis of Results

The quantitative assessment results of each method on the test dataset are given in Table 1. Figure 6 presents a qualitative comparison across six typical water body scenarios (A-F), displaying the original images, ground truth label, along with predictions of each model. The generalization capacity and robustness of all the methods in complex aquatic environments are best assessed through the joint consideration of both quantitative metrics and qualitative visual findings.

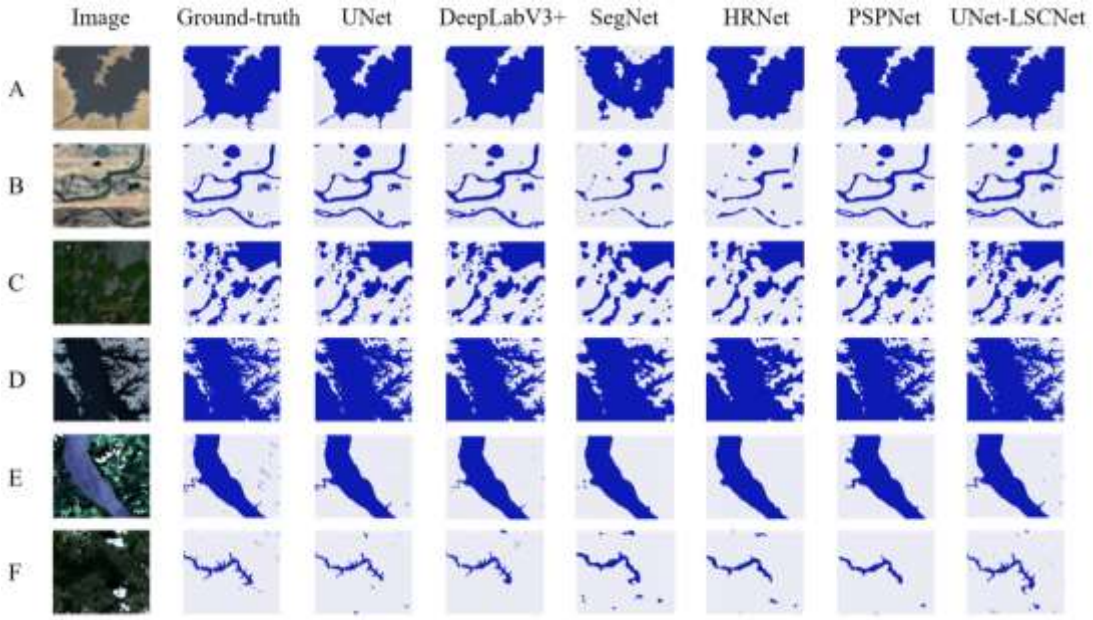


Figure 6. Comparison between six sets of test images. The six image types are: (A) Large-scale uninterrupted water bodies; (B) Slender, meandering river waters bodies; (C) Small-scale fragmented water bodies; (D) Water networks with high density and intricate borders; (E) Water bodies in which main channels prevail over tributaries; (F) Narrow linear water bodies with intricate borders.

According to Table 1, the proposed UNet-LSCNet outperforms all comparative networks across all evaluation metrics. It achieves an Accuracy of 98.70%, Precision of 97.41%, mIoU of 95.67%, and F1-score of 96.32%. Specifically, the mIoU and F1-score improve by 1.89% and 0.63% over the baseline U-Net, indicating that the approach performs better in regional overlap and overall segmentation quality.

Table 1. PERFORMANCE EVALUATION METRICS OF VARIOUS METHODS.

Method	Accuracy (%)	Precision (%)	mIoU (%)	Recall (%)	F1-score (%)	FPS(images/s)
UNet	98.07	95.27	93.78	**96.12**	95.69	4.32
DeepLabV3+	96.88	95.10	92.11	91.20	93.11	**4.68**
SegNet	96.82	86.75	90.32	87.01	86.88	1.28
HRNet	96.59	96.05	92.06	90.20	93.03	0.98

 doi:10.20944/preprints202604.1325.v1

PSPNet	97.56	86.60	91.15	94.50	90.48	3.16
UNet-LSCNet	**98.70**	**97.41**	**95.67**	95.26	**96.32**	4.18

In the case of large-scale continuous water bodies with relatively homogeneous shorelines, as is depicted by Figure 6(A), all the algorithms extract the central parts of water bodies with high completeness even though they experienced different levels of boundary misalignment in localized, complex shoreline areas. While U-Net, HRNet and PSPNet struggled with boundary accuracy in the locally irregular peripheries; and DeepLabV3+ and SegNet, on the other hand, tend to oversmooth boundaries, UNet-LSCNet demonstrated superior contour maintenance and local detail rendering. Its predictions were consistent with ground truth labels, corroborating its high Accuracy and mIoU.

For slender, meandering rivers (Figure 6(B)), the requirements for continuity modeling and boundary perception are significantly higher. In some narrow parts of a river, SegNet and DeepLabV3+ fail to maintain channel connectivity, while PSPNet generated blurred borders along curves, overestimating local river widths. Although both U-Net and HRNet captured primary structures, discontinuities remain in sharp bends and width variation regions. UNet-LSCNet proved effective in ensuring continuity and morphological constancy of rivers with a recall of 95.26%, which is higher than DeepLabV3+ (91.20%) and SegNet (87.01%), thus significantly reducing missed detections of linear water bodies.

In scenarios with small-scale fragmented water bodies (Figure 6(C)), which are easily confused with background features, SegNet and PSPNet demonstrate significant failures, with Precision of 86.75% and 86.60%, respectively. Despite U-Net recognizing most water areas, it erroneously connected adjacent distinct patches. UNet-LSCNet successfully preserved the integrity and independence of small water patches, outperforming HRNet by 3.61% in mIoU, indicating a superior capability to distinguish fragmented water body.

Within high-density complex water networks (Figure 6(D)), the PSPNet and SegNet tend to be vulnerable to background interference, leading to false positives in non-water zones. Although U-Net and HRNet can isolate major water structures they have difficulties in differentiating between adjacent water edges. Conversely, UNet-LSCNet delivered highly complete results with the lowest false detection rate, achieving an F1-score 0.63% higher than U-Net, indicating robustness against complex background clutter.

In river systems dominated by major channels (Figure 6(E)), UNet-LSCNet maintained the structural integrity of the main channels while providing precise estimations for small-scale tributaries. In contrast to the missed detections in minor tributary areas observed in DeepLabV3+ and HRNet, the proposed UNet-LSCNet demonstrated robust feature representation capabilities when handling multi-scale water body architectures.

Finally, for narrow linear water bodies with complex boundaries (Figure 6(F)), UNet-LSCNet effectively preserves structural continuity and captures fine boundary details, even in highly curved and irregular regions. In contrast, DeepLabV3+ and SegNet tend to miss thin segments or produce over-smoothed results, while U-Net and HRNet still exhibit local discontinuities. These results demonstrate that UNet-LSCNet has stronger capability in modeling slender structures and maintaining topological consistency.

Furthermore, in terms of computational efficiency, the proposed UNet-LSCNet demonstrates competitive inference speed. As shown in Table 1, UNet-LSCNet achieves an inference speed of 4.18 frames per second (FPS), which is equivalent to the baseline U-Net (4.32 FPS) and DeepLabV3+ (4.68 FPS), and significantly faster than SegNet (1.28 FPS) and HRNet (0.98 FPS). This minimal drop in speed confirms that the integration of DSConv, CBAM, and the lightweight LaViT module improves segmentation accuracy without imposing computational burden. Consequently, the proposed model achieves a superior balance between high-precision and operational efficiency.

Figure 7 visually compares the Accuracy, Precision, Recall, mIoU, and F1-score across all tested models. The comprehensive results confirm that UNet-LSCNet exhibits the most stable and superior performance across diverse scenarios.

 doi:10.20944/preprints202604.1325.v1

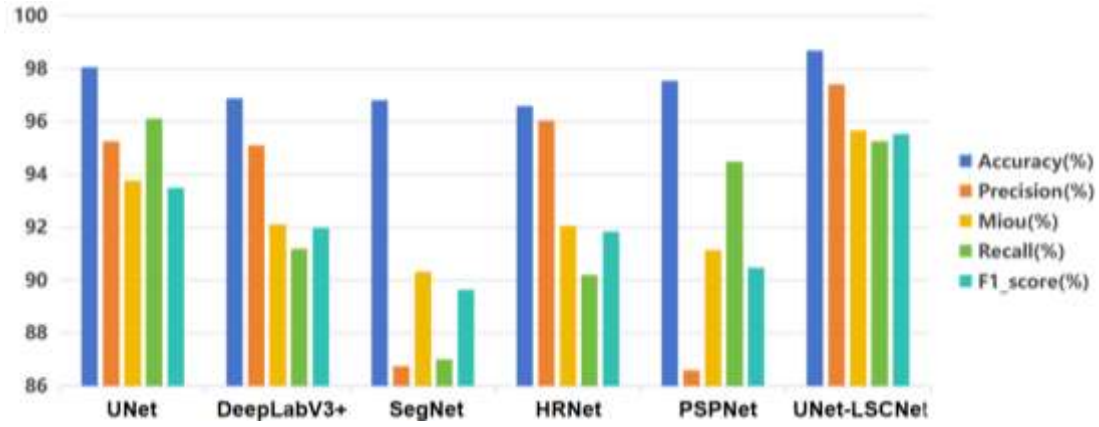


Figure 7. Comparison of Metrics across Different Models.

To summarize, the proposed UNet-LSCNet exhibits better segmentation stability and generalization ability across various shapes of water bodies, especially excelling in elongated rivers, dense networks, and multi-scale composite scenarios. The integration of deep and shallow features, coupled with enhanced local structural perception, enables the model to effectively address challenging remote sensing water body extraction tasks.

4. Discussion

To systematically evaluate the contribution of each key module, a series of ablation experiments were designed using U-Net as the baseline. The experiment incrementally incorporated the DSConv, the CBAM, and the LaViT. By analyzing changes in evaluation metrics, the contributions of each module were comprehensively analyzed. The results are summarized in Table 2.

Table 2. ABLATION STUDY.

Model	Setting				Evaluation			
	UNet	DSConv	CBAM	LaViT	Precision (%)	mIoU (%)	Recall (%)	F1-score (%)
1	√				95.27	93.78	96.12	95.69
2	√	√			94.36	94.14	95.62	94.98
3	√	√	√		96.37	95.77	96.47	96.42
4	√			√	94.05	93.06	94.20	94.12
5	√	√	√	√	97.41	95.67	95.26	96.32

The DSConv module was designed to enhance the network's adaptive modelling capability for complex, irregular water body boundaries. Model 2 incorporates DSConv into the U-Net backbone. As shown in Table 2, Model 2 achieves a Recall of 95.62% and an F1-score of 94.98%, demonstrating an improvement in overall water body detection capability compared to the baseline U-Net, which relies on standard convolutions alone. This demonstrates that DSConv can mitigate false negatives, particularly in regions containing small water bodies or meandering river sections. However, lacking explicit attention constraints, Model 2 still exhibits false positives in complex background areas, resulting in relatively limited improvements in Precision and mIoU metrics.

To further enhance the model's discrimination capability, the CBAM attention module was introduced to Model 2, forming Model 3. CBAM jointly utilizes channel attention and spatial attention, guiding the network to adaptively focus on discriminative water bodies during feature learning while suppressing background noise interference. Quantitative results show that Model 3's

 doi:10.20944/preprints202604.1325.v1

Precision improved to 96.37%, mIoU reached 95.77%, and F1-score rose to 96.42%, all significantly outperforming Model 2. This underscores the role of CBAM in reducing false positives and improving the consistency between prediction and the ground truth. Particularly in areas where water and land exhibit similar spectral characteristics, the attention mechanism assists the network in more accurately delineating water boundaries, thereby improving the overall segmentation quality.

To evaluate the contribution of the LaViT module, Model 4 incorporates the LaViT into the U-Net backbone without other components. As shown in Table 2, this configuration achieves 94.05% Precision, 93.06% mIoU, 94.20% Recall, and 94.12% F1-score. Notably, when deployed in isolation, the performance of LaViT does not surpass the baseline U-Net. This phenomenon can be attributed to the inherent characteristics of Vision Transformers, that without the local geometric constraints provided by specialized convolutions or explicit feature selection, self-attention mechanisms applied in complex scenes may lead to over-smoothing of features. This finding suggests that in high-resolution water body extraction tasks, local Transformers are better suited as complementary components to convolutional features, rather than functioning independently as the primary feature extraction mechanism.

Model 5 simultaneously integrates DSConv, CBAM attention modules, and LaViT modules to form the complete UNet-LSCNet network architecture. It can be observed that this model achieves optimal results across all evaluation metrics, with Precision reaching 97.41%, mIoU increased to 95.67%, Recall reaching 95.26%, and F1-score attaining 96.32%. Compared to models incorporating only a single or partial modules, multi-module collaboration significantly enhances the model's overall performance in water body extraction tasks. These findings demonstrate the complementarity among the three modules: DSConv provides locally adaptive boundary geometry modelling, the CBAM ensures precise feature selection and background noise suppression, and the LaViT models contextual feature dependencies within local windows. Their combined application enables the network to achieve finer boundary localization and more stable predictions while preserving the structural continuity of diverse water bodies. In summary, the ablation experiments confirm that the three introduced modules are complementary, enabling the network to handle complex water body extraction tasks.

5. Conclusions

This study addresses the challenge of extracting multi-scale water bodies characterized by highly variable boundaries. To this end, we proposed UNet-LSCNet, a novel architecture that integrates local geometric feature learning with global semantic modeling to achieve boundary-sensitive water body extraction from remote sensing imagery. Built upon the efficient encoder-decoder framework, the model integrates DSConv, CBAM, and LaViT to simultaneously capture the complex boundaries and enhance the feature representation of water bodies.

Extensive experimental evaluations across diverse scales and morphology water bodies demonstrated that the proposed UNet-LSCNet outperforms existing semantic segmentation techniques. It exhibits superior capabilities in ensuring water body completeness, refining boundary definitions, and detecting fragmented, small-scale water bodies. Ablation studies further confirmed the contributions of each module: DSConv overcomes the expressive limitations of standard convolutions along complex boundaries; the attention mechanism significantly enhances regional feature selectivity and noise suppression; and the local window context modeled by LaViT greatly improves the continuity of large-scale water distribution patterns.

Altogether, the proposed architecture delivers precise water body segmentation results, while maintaining a favorable trade-off with computational inference efficiency. This demonstrates that UNet-LSCNet is a robust and practical method for automated water body extraction tasks in the field of remote sensing.

 doi:10.20944/preprints202604.1325.v1

References

1. Pickens, A.H.; Hansen, M.C.; Hancher, M.; Stehman, S.V.; Tyukavina, A.; Potapov, P.; Marroquin, B.; Sherani, Z. Mapping and sampling global inland water dynamics using Landsat data. Remote Sens. Environ. 2020, 243, 111792.
2. Zhu, X.X.; Tuia, D.; Mou, L.; Xia, G.-S.; Zhang, L.; Xu, F.; Fraundorfer, F. Deep learning in remote sensing: A comprehensive review and list of resources. IEEE Geosci. Remote Sens. Mag. 2021, 9, 8–36.
3. Tong, X.; Luo, X.; Liu, S. An approach for flood monitoring by the combined use of Landsat 8 optical imagery and COSMO-SkyMed radar imagery. ISPRS J. Photogramm. Remote Sens. 2018, 136, 144–153.
4. Du, Y.; Zhang, Y.; Ling, F.; Wang, Q.; Li, W.; Li, X. Water bodies mapping from Sentinel-2 imagery with modified normalized difference water index at 10-m spatial resolution. Remote Sens. 2021, 13, 142.
5. Rokni, K.; Ahmad, A.; Selamat, A.; Hazini, S. Water feature extraction and change detection using multitemporal Landsat imagery. Remote Sens. 2014, 6(5), 4173–4189.
6. Liu, H.; Hu, H.; Liu, X.; Jiang, H.; Liu W.; Yin, J.; Yin, W. A comparison of different water indices and band downscaling methods for water bodies mapping from Sentinel-2 imagery at 10-m resolution. Water 2022, 14(17), 2696.
7. Li, L.; Su, H.; Du, Q.; Wu, T. A novel surface water index using local background information for long-term and large-scale Landsat images. ISPRS J. Photogramm. Remote Sens. 2021, 172, 59–78.
8. Su, Z.; Xiang, L.; Steffen, H.; Jia, L.; Deng, F.; Gao, P. A new and robust index for water body extraction from Sentinel-2 imagery. Remote Sens. 2024, 16(15), 2749.
9. Abdulkareem, H.; Al-Hadithi, M.; Mozihim, R. A new water body extract index utilizing Sentinel-2 satellite imagery: A case study of Bahr Al Najaf, Iraq. IOP Conf. Ser. Earth Environ. Sci. 2025, 1545, 012016.
10. Cai, Y.; Shi, Q.; Liu, X. Spatiotemporal mapping of surface water using Landsat images and spectral mixture analysis on Google Earth Engine. J. Remote Sens. 2024, 4, 0117.
11. Shalev-Shwartz, S.; Ben-David, S. Support vector machines and kernel methods. In Understanding Machine Learning: From Theory to Algorithms; Cambridge Univ. Press: Cambridge, UK, 2023; pp. 403–445.
12. Criminisi, A.; Shotton, J.; Konukoglu, E. Decision forests: A unified framework for classification, regression, density estimation, manifold learning and semi-supervised learning. Found. Trends Comput. Graph. Vis. 2022, 7, 81–227.
13. Islam, K.I.; Elias, E.; Carroll, K.C.; Brown, C. Exploring random forest machine learning and remote sensing data for streamflow prediction. Remote Sens. 2023, 15, 3999.
14. Chen, F.; Chen, X.; de Voorde, T.; Roberts, D.; Jiang, H.; Xu, W. Open water detection in urban environments using high spatial resolution remote sensing imagery. Remote Sens. Environ. 2020, 242, 111706.
15. Ma, L.; Liu, Y.; Zhang, X.; Ye, Y.; Yin, G.; Johnson, B.A. Deep learning in remote sensing applications: A meta-analysis and review. ISPRS J. Photogramm. Remote Sens. 2021, 152, 166–177.
16. Sun, Z.; Di, L.; Fang, H.; Burgess, A. Random forest based land cover classification using remote sensing data. Remote Sens. 2022, 14, 1357.
17. Dong, Z.; Wang, G.; Amankwah, S.; Wei, X.; Hu, Y.; Feng, A. Monitoring the summer flooding in the Poyang Lake area of China in 2020 based on Sentinel-1 data and multiple convolutional neural networks. Int. J. Appl. Earth Obs. Geoinf. 2021, 102, 102400.
18. Nie, P.; Cheng, X.; Song, Z.; Ma, M.; Wang, T.; Meng, L. Rethinking BiSeNet: A lightweight network for urban water extraction. IEEE Trans. Geosci. Remote Sens. 2023, 61, 4203910.
19. Wang, Z.; Gao, X.; Zhang, Y.; Zhao, G. MSLWENet: A novel deep learning network for lake water body extraction of Google remote sensing images. Remote Sens. 2020, 12, 4140.
20. Zhang, Z.; Lu, M.; Ji, S.; Yu, H.; Nie, C. Rich CNN features for water-body segmentation from very high resolution aerial and satellite imagery. Remote Sens. 2021, 13, 1912.
21. Kang, J.; Guan, H.; Peng, D.; Chen, Z. Multi-scale context extractor network for water-body extraction from high-resolution optical remotely sensed images. Int. J. Appl. Earth Obs. Geoinf. 2021, 103, 102499.
22. Liang, X.; Xiao, H.; Chen, L.; Fan, X. Multi-scale attention network for automatic water body detection in SAR images. High-Resol. Remote Sens. 2024, 5, 100–112.

 doi:10.20944/preprints202604.1325.v1

23. Cao, Y.; Hassan, M.; Elmahdy, A. Surface water mapping from remote sensing imagery using an improved U-Net with multi-scale information and attention mechanism. Int. J. Appl. Earth Obs. Geoinf. 2025, 76, 102–115.
24. Xu, Y.; Lin, J.; Zhao, J.; Zhu, X. New method improves extraction accuracy of lake water bodies in Central Asia. J. Hydrol. 2021, 603, 127180.
25. Liu, B.; Du, S.; Bai, L.; Ouyang, S.; Wang, H.; Zhang, X. Water extraction from optical high-resolution remote sensing imagery: A multi-scale feature extraction network with contrastive learning. GIScience & Remote Sensing 2023, 60(1), 1–18.
26. Shi, W.; Sui, H. An effective superpixel-based graph convolutional network for small waterbody extraction from remotely sensed imagery. Int. J. Appl. Earth Obs. Geoinf. 2022, 109, 102777.
27. Zhong, H.; Sun, Q.; Sun, H.; Jia, R. NT-Net: A semantic segmentation network for extracting lake water bodies from optical remote sensing images based on Transformer. IEEE Trans. Geosci. Remote Sens. 2022, 60, 1–13.
28. Li, Z.; Wegner, J.D.; Lucchi, A. Topological feature extraction for water segmentation. Remote Sens. Environ. 2020, 237, 111540.
29. Mou, L.; Zhu, X.X. Learning to pay attention on spectral domain. ISPRS J. Photogramm. Remote Sens. 2018, 144, 1–12.
30. Guo, H.; Wang, J.; Zhang, X. Attention-based U-Net for water extraction. Remote Sens. 2021, 13, 470.
31. Yu, X.; Zhang, Y.; Liu, X. WaterHRNet: High-resolution network for water body extraction. Remote Sens. 2022, 14, 2186.
32. Woo, S.; Park, J.; Lee, J.-Y.; Kweon, I.S. CBAM: Convolutional block attention module. In Proceedings of the European Conference on Computer Vision (ECCV), Munich, Germany, 2018.
33. Dosovitskiy, A.; Beyer, L.; Kolesnikov, A.; et al. An image is worth 16 × 16 words: Transformers for image recognition at scale. In Proceedings of the International Conference on Learning Representations (ICLR), Virtual Event, 2021.
34. Yuan, Y.; Chen, X.; Wang, J. Transformers in vision: A survey. IEEE Trans. Pattern Anal. Mach. Intell. 2022, 44, 8289–8314.
35. Kang, J.; Guan, H.; Ma, L.; Wang, L.; Xu, Z.; Li, J. WaterFormer: A coupled transformer and CNN network for waterbody detection in optical remotely sensed imagery. ISPRS J. Photogramm. Remote Sens. 2023, 206, 222–241.
36. Zhang, S.; Liu, H.; Lin, S.; He, K. You only need less attention at each stage in vision transformers. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), Seattle, WA, USA, 2024; pp. 6057–6066.
37. Han, K.; Wang, Y.; Zhang, H.; et al. Vision transformer with local aggregation. In Advances in Neural Information Processing Systems (NeurIPS), 2021.
38. Li, Y.; Zhang, K.; Wang, Y. Local aggregation transformer for dense prediction. IEEE Trans. Geosci. Remote Sens. 2022, 60, 1–13.
39. Qi, Y.; He, Y.; Qi, X.; Zhang, Y.; Yang, G. DSConv based on topological geometric constraints for tubular structure segmentation. In Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2023.
40. Ding L.; Tang H.; Bruzzone L. LANet: Local attention embedding to improve the semantic segmentation of remote sensing images. IEEE Trans. Geosci. Remote Sens., 59(1): 426–435, 2021.
41. Wang, L.; Liu, Y.; Zhang, S.; Yan, J.; Tao, P. Structure-Aware convolution for 3D point cloud classification and segmentation.
Remote Sens. 2020, 12(4), 634.
42. Luo, L.X.; Tong, X.H.; Hu, Z.W. An applicable and automatic method for earth surface water mapping based on multispectral images. Int. J. Appl. Earth Obs. Geoinf. 2021, 103, 102472.
43. Feng, W.; Sui, H.; Tu, J.; Huang, W. Water body extraction from high-resolution remote sensing imagery using an improved U-Net network. Remote Sens. 2022, 14, 2396.
44. Li, Y.; Zhang, H.; Shen, Q.; Li, X. DeepLabV3+ based semantic segmentation for high-resolution remote sensing images. Sensors 2023, 23, 2187.

 doi:10.20944/preprints202604.1325.v1

45. Wang, S.; Chen, Y.; Li, X.; Zhang, B. Deep residual network based semantic segmentation for remote sensing images. Remote Sens. 2021, 13, 2056.
46. Chen, F.; Liu, H.; Zeng, Z.; Zhou, X.; Tan, X. BES-Net: Boundary enhancing semantic context network for high-resolution image semantic segmentation.Remote Sens. 2022, 14(7), 1638.
47. Zhang, C.; Li, W.; Travis, D. Gated shape CNNs for semantic segmentation. IEEE Trans. Pattern Anal. Mach. Intell. 2019, 43, 1095–1109.
48. Yuan, Q.; Wei, Y.; Meng, X.; Shen, H.; Zhang, L. A multiscale and multitask deep learning framework for remote sensing image semantic segmentation. IEEE Geosci. Remote Sens. Mag. 2021, 9, 12–27.
49. Sarker, M.M.K.; Song, M.; Ullah, A.; Li, J. A modified SegNet for semantic segmentation of remote sensing images. Remote Sens. 2022, 14, 1023.
50. Wang, Y.; Yang, L.; Liu, X.; Zhang, H.; Li, J. An improved semantic segmentation algorithm for high-resolution remote sensing images based on DeepLabV3+. Sci. Rep. 2024, 14, 9716.
51. Jin, Y.; Liu, X.; Huang, X. EMR-HRNet: A multi-scale feature fusion network for landslide segmentation from remote sensing images. Sensors 2024, 24, 3677.
52. Li, Y.; Huang, X.; Zhang, Z.; Chen, Y. Deep learning-based surface water mapping from remote sensing imagery: A comprehensive review and perspective. ISPRS J. Photogramm. Remote Sens. 2023, 202, 160–182.