RU-Net++: An automatic extraction method for Impervious Surface Area based on neural networks

FAN YU , XIAOKANG TU, LIN CAI, JIAYAO ZHANG, ZHENGXIN WANG

School of Surveying, Mapping and Urban Spatial Information, Beijing University of Civil Engineering and Architecture, Beijing,
yufan@bucea.edu.cn，1565779139@qq.com, 19838643380@163.com，2990016833@qq.com，2751085573@qq.com

Keywords: Remote Sensing,RU-Net++,Impervious Surface Area ,Attention Mechanism, Urban Planning

Abstract

Impervious Surface Area (ISA) is vital for urban planning, environmental monitoring, and water management. Traditional remote sensing methods struggle with complex urban landscapes, leading to accuracy limitations. To address this, we propose RU-Net++, a deep learning-based ISA extraction model integrating ResNet50 as the encoder with spatial, channel, and dual attention mechanisms. The decoder employs an Atrous Spatial Pyramid Pooling (ASPP) module and multiple refinement modules to enhance feature representation and edge restoration. Trained on GLC_FCS30D and GISA datasets, RU-Net++ outperforms traditional methods in IoU, F1 Score, and Overall Accuracy, offering a reliable tool for sustainable urban development and land-use management.

1. Research Contents

1.1 Introduction

In the fast-paced urbanization process of today’s world, profound changes have occurred in urban spatial structures and land use patterns. Impervious Surface Area (ISA) is a key indicator of urbanization and has become increasingly important for understanding ecological and environmental changes. ISA primarily consists of surfaces made from artificial materials, including roads, parking lots, sidewalks, rooftops, and other impervious surfaces in urban areas. These surfaces do not allow water to penetrate, unlike natural soils or vegetation, which significantly alters surface hydrological processes and ecological functions.

Geographical national condition monitoring is a key task and development direction for China’s mapping industry in the new era, tasked with accurately capturing changes in land use and spatial conditions. The distribution of impervious surfaces is one of the critical indicators for urban and regional ecological assessments. Accurate identification and quantification of impervious surfaces are of profound significance for urban and regional development planning and ecological evaluations. This directly affects the normal functioning of urban water cycles, heat island regulation, and other ecological functions, and is crucial for maintaining the stability and sustainability of regional ecosystems. A comprehensive understanding of ISA and its distribution allows for the precise application of strategies aimed at reducing impervious surfaces and mitigating their negative impacts on water resources and the environment in community planning, site design, and land use management. This not only optimizes urban spatial layouts but also helps alleviate the environmental issues caused by excessive impervious surface expansion during urbanization, such as urban flooding and water quality pollution, thus promoting sustainable urban development. The accurate quantification of impervious surfaces has become a key planning tool for urban land use development. As urban land use intensity continues to rise under the wave of urbanization, the area of impervious surfaces is steadily expanding. However, this unchecked expansion of impervious surfaces has led to numerous negative effects, such as the intensification of urban heat island effects, water quality degradation, and the destruction of natural habitats. The urban heat island effect raises city temperatures, reduces residents' comfort, and increases energy consumption; water quality degradation threatens drinking water safety and aquatic ecosystems; and the loss of natural habitats disrupts ecological balance, causing a decline in biodiversity. Therefore, scientifically and rationally quantifying impervious surfaces and implementing targeted measures to mitigate these adverse effects have become urgent tasks for sustainable urban development.

Remote sensing technology, with its wide coverage, high resolution, multi-temporal, and multispectral capabilities, has become a powerful tool for studying impervious surfaces. It can rapidly acquire extensive surface information, providing scientific data and technical support for urban planning, ecological conservation, and water resource management. Through remote sensing imagery, we can clearly observe the spatiotemporal variation characteristics of urban impervious surfaces, providing timely and accurate data support for decision-making, helping urban managers better address various environmental challenges arising from urbanization.

The extraction of impervious surfaces is the first and critical step in analyzing their evolution. In urban areas, impervious surfaces are mainly composed of buildings, roads, and other related structures. As urban construction progresses rapidly, urban areas are expanding, and buildings are emerging in various forms, both in height and layout. Roads, in particular, are intricately networked, making the texture information of impervious surfaces complex and diverse, which significantly increases the difficulty of achieving high extraction accuracy. The texture features of impervious surfaces vary significantly across different regions and land types, and are influenced by factors such as lighting conditions and shadow occlusion, making traditional extraction methods inadequate for achieving high-precision results.

Additionally, the proportion of impervious surfaces in urban areas is relatively small and their distribution is highly uneven. This characteristic makes it extremely difficult to obtain a large number of high-quality labeled samples, and labeled samples are crucial for training effective classification models. Given the limited labeled samples and the availability of only satellite data, achieving high-precision impervious surface classification

This contribution has been peer-reviewed.
https://doi.org/10.5194/isprs-archives-XLVIII-G-2025-1655-2025 | © Author(s) 2025. CC BY 4.0 License.

remains a major challenge. In this context, this paper innovatively uses neural networks as the model for extracting impervious surfaces. Neural networks, with their powerful learning and feature extraction capabilities, can automatically learn the complex feature patterns of impervious surfaces from limited sample data and apply them to large-scale remote sensing image classification. Through the carefully designed neural network architecture and optimization algorithms, this study aims to achieve high-precision impervious surface classification using the available satellite data, providing strong technical support for accurate monitoring and effective management of urban impervious surfaces, and contributing to related research and practices in sustainable urban development.

This study focuses on this challenging and highly relevant issue, aiming to explore an efficient and accurate impervious surface extraction method to address the environmental issues brought about by the rapid changes in impervious surfaces during urbanization, and to contribute to sustainable urban development and ecological environmental protection.

1.2 Impervious surface extraction methods

1.2.1 impervious surface extraction using the index-based method: The index-based method is an approach that distinguishes impervious surfaces from other land cover types by constructing specific spectral indices. These indices are designed based on the spectral characteristics of different land cover types, utilizing combinations and calculations of specific bands to highlight impervious surface features, thereby enabling their extraction. For example, the **Normalized Difference Impervious Surface Index (NDISI)** proposed by Xu H is one of the most commonly used indices. It leverages the ratio of the thermal infrared and near-infrared bands to enhance impervious surface information, improving its detectability in remote sensing imagery.

1.2.2

$$NDISI = \frac{TIRS - NIR}{TIRS + NIR}, \tag{1}$$

where TIRS = Thermal infrared band
NIR= Near infrared band

NDISI can highlight impervious surface information, but its accuracy can be further improved by combining it with other bands, such as the mid-infrared band and the Modified Normalized Difference Water Index (MNDWI).

1.2.3 impervious surface extraction based on multi-source remote sensing data fusion: The method of multi-source data fusion for impervious surface extraction is a technique that integrates the advantages of various remote sensing data sources. Through steps such as image fusion, feature optimization, and classifier training, it achieves high-precision impervious surface extraction. HUO Jiating, ZHAO Zhan, and others proposed an image fusion method based on band mapping and wavelet transformation, combining Sentinel-2 and GaoFen-2 imagery to obtain fused images that have both high spatial and spectral resolutions. These fused images contain rich spectral and spatial features, which enhance the ability to distinguish impervious surfaces from non-impervious surfaces in complex urban areas. The initial classification samples are automatically obtained using class information from the GlobeLand30 dataset. Based on the rich spectral information from the fused images, various vegetation indices, water body indices, and built-up area indices are constructed to optimize the initial classification samples. Finally, the optimized training samples are used to train classifiers with features such as spectral data and land cover indices, enabling the automatic and accurate extraction of urban impervious surfaces.

1.2.4 impervious surface extraction using machine learning and deep learning: Machine learning methods primarily include traditional algorithms such as Support Vector Machine (SVM) and Random Forest (RF). These methods learn the characteristic differences between impervious surfaces and other land cover types through training datasets, thereby enabling classification. For example, SVM distinguishes impervious surfaces from other land covers by finding the optimal separating hyperplane, while Random Forest improves classification accuracy and robustness by constructing multiple decision trees. These methods perform excellently when processing medium to low-resolution remote sensing images, effectively addressing the complex spectral features of land cover types.

In recent years, with the rapid development of artificial intelligence, remote sensing image classification has seen extensive advancements. Deep learning methods, particularly Convolutional Neural Networks (CNNs), have shown significant advantages in impervious surface extraction from high-resolution remote sensing images. CNNs automatically extract multi-level features from images, learning the unique texture and shape characteristics of impervious surfaces through the combination of convolutional layers, pooling layers, and fully connected layers. For instance, the U-Net architecture, commonly used for image segmentation, can effectively handle impervious surface extraction tasks in high-resolution remote sensing images through multi-scale feature fusion.

2. Research Methodology

2.1 Research Data and Experimental Platform

The data used in this study comes from two sources: the first is the world’s first 30-meter global land cover time-series dynamic remote sensing product (GLC_FCS30D) from 1985 to 2022, developed by the team of Professor Liu Liangyun at the Aerospace Information Research Institute, Chinese Academy of Sciences. The second is the 30-meter continuous global impervious surface dataset GISA 2.0 (1972–2019), developed by the Remote Sensing Information Engineering Institute, Wuhan University, under the guidance of Professor Huang Xin's research group. The former provides global-scale long-term temporal coverage to validate model generalization, while the latter optimizes local detail extraction through high-resolution

annotations. Together, these datasets complement each other, providing a reliable foundation for model training.
The experimental platform uses an Intel i5-13400F 16-core processor, equipped with 16.0 GB of memory and an NVIDIA GeForce RTX 3060 graphics card. The experiments were conducted on the Ubuntu 18.04 operating system, utilizing the Pytorch deep learning framework.

2.2 Experimental Data

2.2.1 Selection of Experimental Area： The experiment selected the cities of Beijing, Wuhan, Ili Kazakh Autonomous Prefecture, Urumqi, Baoding, and Langfang in China. The training, validation, and test sets were divided in a 7:2:1 ratio.

2.2.2 Product Reclassification： The impervious surface for 2020 from the GLC_FCS30 product was extracted using the BandMath tool in ENVI software, and the same method was applied to extract the impervious surface for 2019 from the GISA product.

2.2.3 Unified Coordinates: ensure that the labels correspond consistently with the images, both the GISA and GLC_FCS30 products were reprojected to the same coordinates as the corresponding GaoFen-1 imagery. This was done using ArcGIS, and the resolution was adjusted by resampling to 16 meters.

2.2.4 Data Fusion and Clipping： For the preprocessed data, pervious surfaces were assigned a value of 0, and impervious surfaces were assigned a value of 1. To address the issue of a small proportion of impervious surfaces, a sliding window approach was applied, where the data was cropped into 512m x 512m patches. Additionally, geometric transformations were applied to ensure a unified size for the impervious surface dataset.

2.3 U-Net Network

The study references the U-Net model architecture, a convolutional neural network (CNN) designed for image segmentation, initially proposed by Olaf Ronneberger et al. in 2015. Its main feature is the use of an encoder-decoder structure combined with skip connections to efficiently perform pixel-level image segmentation tasks. Skip connections involve directly adding input data to the output of a specific layer within the network. This design allows information to flow more freely and preserves both the detail and semantic information of the original input data, making it easier for information to propagate to later layers and preventing information loss. Skip connections are typically implemented through summation or concatenation operations.
The encoder extracts features through multiple 3×3 convolutions and ReLU activation functions, followed by 2×2 max pooling for downsampling. This process gradually reduces spatial resolution while increasing the number of channels. The decoder, on the other hand, performs upsampling through 2×2 transpose convolutions and integrates features from the corresponding encoder layers via skip connections to retain more detailed information. The final segmentation result is generated through a 1×1 convolution.

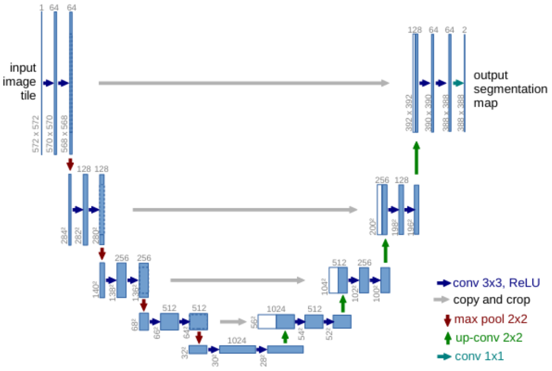


Figure 1. UNet architecture (Ronneberger et al., 2015).

Due to its skip connections and efficient feature learning, U-Net can achieve excellent segmentation results even with a relatively small amount of labeled data. It is particularly suitable for few-shot learning, where effective training is conducted with limited labeled data. This makes U-Net highly applicable in the study of impervious surfaces, where data annotation may be scarce.

2.4 ResNet Network

Typically, the more convolutional layers in a network, the better its performance. However, an excessive number of layers can exacerbate issues like vanishing and exploding gradients. To address these problems in deep learning, Kaiming He et al. proposed the Residual Neural Network (ResNet) in 2015. ResNet is a deep convolutional neural network (CNN) architecture that solves the degradation problem during deep network training by introducing "residual learning."
Residual learning works by bypassing certain intermediate layers and directly linking the activations of a layer to subsequent layers, thus creating a residual block. These residual blocks are stacked to form the ResNet. In ResNet, residual learning is achieved through the introduction of "shortcut connections" or "skip connections," which allow inputs from earlier layers in the network to be directly passed on to later layers.
As illustrated, H represents a hidden block, which is a module consisting of convolutional layers, activation layers, and batch normalization layers. The skip connections can be viewed as an identity mapping, enabling input data to be directly passed through the network.

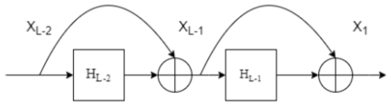


Figure 2. Residual block (He et al., 2016).

Instead of directly learning the target function H(x), ResNet learns the residual function between the input and output $F(x) = H(x) - x$, simplifying the optimization process. As a result, many subsequent methods are based on ResNet50 or ResNet101.
In this study, we reference the Bottleneck residual block of ResNet50. Each residual unit in ResNet50 contains three convolutional layers. First, a 1×1 convolution reduces the input channels to lower the computational load; then, a 3×3 convolution extracts spatial features; finally, a 1×1 convolution restores the number of channels. The input features are added to the convolutional output via identity mapping, effectively alleviating the gradient vanishing problem in deep networks. For example, in layer1, after passing through the Bottleneck structure, the output maintains 256 channels and the spatial

resolution remains unchanged, while the input feature map (256 channels) is processed.

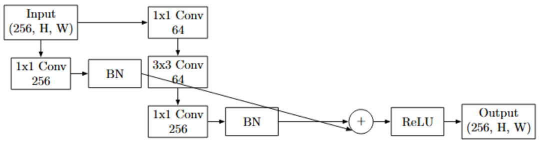


Figure 3. Bottleneck residual block of ResNet50.

2.5 RU-Net++ Network

2.5.1 Input Layer： A 7×7 convolutional kernel is used, providing a large receptive field (7×7 = 49 pixels) to capture large-scale features in remote sensing imagery. The stride is set to 2, resulting in a 50% downsampling rate to reduce the computational load while preserving key information. Padding is set to 3 to maintain edge integrity and prevent boundary information loss. The output channels are set to 64 to progressively expand features and prevent the explosion of shallow-layer information. The input layer consists of the following two steps: receiving a 512×512×3 RGB remote sensing image and applying a 7×7 convolutional kernel (stride 2, padding 3) to extract initial features.

2.5.2 Encoder： In terms of structural design, the ResNet50 architecture is used as the encoder. However, the original ResNet50 performs a large number of downsampling steps, which can lead to overly small feature maps. For example, a 512×512 input could be reduced to 16×16 after five downsampling steps, potentially losing too much detail. Therefore, the number of downsampling steps is reduced, and gradual upsampling is applied in the decoder to restore the size. As a result, the original ResNet50 structure is modified in three steps: (1) removing the original stage5 to avoid excessive downsampling (32×32 → 16×16); (2) adding spatial attention in Stage1; (3) adding channel attention in Stage4, and (4) inserting Atrous Spatial Pyramid Pooling (ASPP) after Stage4.

The spatial attention mechanism aims to guide the model to focus on important regions in the image. In Stage1 of the encoder, the input feature map is first processed using Global Average Pooling and Global Max Pooling, performing pooling operations across the channel dimension to generate two 2D feature maps. These feature maps are then concatenated along the channel dimension and passed through a 7×7 convolutional layer to generate the spatial attention weight map. Finally, this weight map is multiplied element-wise with the original feature map to emphasize features at key spatial locations. Introducing spatial attention in the early stages of the model helps it capture edge and shape features of impervious surfaces more effectively.

The channel attention mechanism focuses on different channels of the feature map to enhance the response to important features. In Stage4 of the encoder, the input feature map is first globally average pooled to obtain a global description of each channel. A Fully Connected Layer is then used to learn the weight coefficients for each channel. These weights are processed through an activation function and multiplied by the corresponding channels in the original feature map to adjust the response strength of each channel. Introducing channel attention in the deeper stages of the model helps capture high-level semantic information, improving the model's ability to recognize impervious surface features.

Atrous Spatial Pyramid Pooling (ASPP) is a module used to capture multi-scale contextual information, commonly found in semantic segmentation models like the DeepLab series. Its core design involves using convolutions with different dilation rates and global pooling in parallel, merging multi-scale features to enhance the model's ability to recognize objects of various sizes. Dilated convolutions (also known as atrous convolutions) increase the receptive field by inserting gaps between the convolutional kernel elements, and varying dilation rates control the size of the receptive field. This allows each branch of the module to capture context at different scales. The ASPP module is added after Stage5 and consists of five branches. Branch 1 uses a 1×1 convolution, branches 2 to 4 use 3×3 convolutions with dilation rates of 6, 12, and 18, respectively, and branch 5 performs global average pooling (GAP).

2.5.3 Decoder： In the decoder design of RU-Net++, specific attention and optimization modules are introduced at Levels 4 to 1 to address the need for recovering features at different resolutions, ensuring precise localization and detail recovery.

At Level 4, the ChannelGate module, combining both spatial attention and channel attention, is used. The spatial attention emphasizes the location of target regions in the input feature map, while the channel attention weights the importance of each channel. This allows high-level features to more accurately locate impervious surface targets during the upsampling process. Since Level 4 is in a higher layer of the decoder, where feature semantics are rich but spatial resolution is low, the ChannelGate module compensates for the lack of spatial information and enhances localization accuracy.

At Level 3, the Dual Attention module is introduced, which integrates both spatial and channel attention. These two attention mechanisms complement each other by weighting features from both global semantics and local details. This dual attention mechanism allows the intermediate features to more comprehensively represent target information, improving the model's ability to adapt to targets at different scales while reducing background interference.

At Level 2, the SpatialGate module is used, focusing on extracting and enhancing spatial information. At this stage, the feature map has a higher resolution, but it may also contain more noise and redundant information. SpatialGate calculates spatial attention weights to highlight key areas and edge details in the image, providing cleaner and more accurate feature inputs for final detail recovery.

Finally, at Level 1, the Boundary Enhancement Module (BEM) is introduced, specifically designed to address boundary blurring that may occur during upsampling. BEM utilizes local contextual information to refine and optimize the boundaries of the predicted result, ensuring that the segmentation result is clearer and more precise at the edges. This is especially beneficial for extracting impervious surface boundaries in complex scenarios.

2.5.4 Output Layer： The Sigmoid function independently maps the logit of each pixel to a value between 0 and 1, representing the probability of belonging to the positive class. On the other hand, the Softmax function ensures that the sum of probabilities for all categories equals 1 in multi-class classification. In binary classification, if two output channels are used, each channel represents the probability of the corresponding class, and their sum is also equal to 1.The table below provides a numerical comparison of the principles of the Sigmoid and Softmax functions.

Characteristic	Sigmoid	Softmax
Output range	Calculate probability independently for ea ch pixel∈[0,1]	Dual channel proba -bility∈[0,1]

formula	$\sigma(x)=\frac{1}{1+e^{-x}}$	$f(x_j)=\frac{e^{x_j}}{\sum_{k=1}^{K} e^{x_k}}$

Table 1. Comparison of Mathematical Principles between Sigmoid and Softmax

In the case of binary classification, both methods can theoretically be used, but there are some differences. For instance, when using Sigmoid, the model only requires one output channel, which makes the computation more efficient. However, it may face challenges with class imbalance. On the other hand, using Softmax with two channels allows each channel to independently learn features, providing more flexibility, but it requires more parameters.

In impervious surface extraction, there is often a class imbalance, where the non-impervious surface (such as vegetation and water bodies) dominates. In such cases, Softmax's mutual exclusivity can be more beneficial for the model in distinguishing between the two classes, especially when there is overlap between them. Additionally, when combined with cross-entropy loss, Softmax is generally more stable, as the loss calculation directly considers the probabilities of both classes. Finally, a 1×1 convolution is used to reduce the number of channels to 2.

2.5.5 Loss Function： Dice Loss is a metric that measures the similarity between the predicted results and the ground truth labels, and it is suitable for pixel-level segmentation tasks. The formula is as follows:

$$L_{Dice}=1-\frac{2\sum_{i=1}^{N} p_i g_i}{\sum_{i=1}^{N} p_i^2+\sum_{i=1}^{N} g_i^2}, \tag{2}$$

where p_i = Pixel values predicted by the model (between 0-1)

g_i = ground truth

N = Total number of pixels

Focal Loss is primarily used to address class imbalance and the insufficient learning of hard-to-classify samples. It is an improvement on BCE (Binary Cross-Entropy) loss, with a focus on difficult-to-classify samples. The formula is as follows:

$$L_{Focal}=-\sum_{i=1}^{N} \alpha_t (1-p_t)^{\gamma} \log(p_t), \tag{3}$$

where p_t = Probability of category, $p_t=p$ (the real category is 1) or $p_t=1-p$ (the real category is 0)

γ = The adjustment factor, usually set to 2

α_t = Category weight coefficient

RU-Net++ employs a hybrid loss function to optimize the accuracy of impervious surface extraction, addressing class imbalance and the uncertainty in boundary regions. Since the impervious surface areas in remote sensing images are typically small, using a single loss function such as Binary Cross-Entropy (BCE) or Dice Loss may lead to class bias. Therefore, we combine Dice Loss and Focal Loss, setting their weights to 0.7 and 0.3, respectively. The final loss function is as follows:

$$L_{total}=\alpha L_{Dice}+\beta L_{Focal}, \tag{4}$$

where α , β = Weighted coefficients of the loss function

Stage	Operation	Channels	Size
Input	7×7 Conv,s=2,pad=3	64	256×256
	MaxPool 3×3,s=2	64->256	128×128
Encoder	Stage1: 3×ResBlock +SpatialAttention	64->256	128×128
	Stage2: 4×ResBlock	128->512	64×64
	Stage3: 6×ResBlock +ChannelAttention	256->1024	32×32
	Stage4: 3×ResBlock	512->2048	16×16
ASPP	（1×1,3×3-d6, 3×3-d12, 3×3-d18,GAP）	2048->4096 ->2048	16×16
Decoder	Level4: Up2× +SpatialGate	2048->1024	32×32
	Level3: Up2×+ DualAttention	1024->512	64×64
	Level2: Up2×+ ChannelGate	512->256	128×128
	Level1: Up2×+ BEM	256->128	256×256
Output	1×1Conv +Soft-max	128->2	512×512

Table 2. RU-Net++ Model Structure.

2.6 Workflow Description

To provide a detailed description of the RU-Net++ workflow, the entire process structure is clearly illustrated in Figure 4.

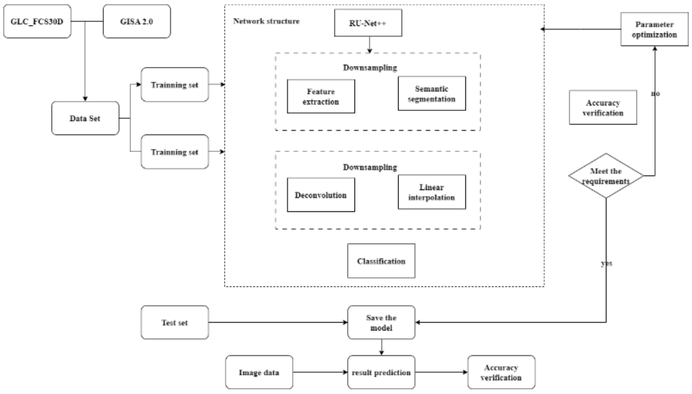


Figure 4.RU-Net++ Model workflow.

3. Result and Analysis

The RU-Net++ network has been demonstrated through experiments to have stronger feature extraction capabilities, effectively alleviate gradient vanishing, and improve classification accuracy through multi-scale feature fusion.

3.1 Remote sensing image accuracy metrics

To quantitatively analyze the model's segmentation accuracy, Intersection over Union (IoU), Kappa coefficient, and F1 score are used as evaluation metrics. Meanwhile, to assess the classification accuracy of remote sensing images, Overall Accuracy (OA), Kappa coefficient, and confusion matrix are used as evaluation metrics.

Specifically, IoU reflects the overlap between the segmented area and the ground truth segmentation area, the Kappa coefficient indicates the consistency between the ground truth and the classification result, the F1 score represents the model's accuracy and error, and OA indicates the proportion of correctly classified pixels. In general, the higher the values of these four metrics, the better the classification results.

3.1.1 IOU:

$$IOU = \frac{|A \cup B|}{|A \cap B|}, \tag{5}$$

where A = Pixel set of predicted results
B= The set of pixels with true values

3.1.2 Dice coefficient:

$$s = \frac{2|X \cap Y|}{|X|+|Y|}, \tag{6}$$

where $|X \cap Y|$ = Intersection between X and Y
$|X|+|Y|$ = The number of elements in X and Y

3.1.3 Kappa Coefficient:

$$Kappa = \frac{p0 - pe}{1 - pe}, \tag{7}$$

$$P_e = \frac{a1*b1 + a2*b2 + +an*bn}{1 - pe}, \tag{8}$$

where P_0= The proportion of correctly classified pixels in each category to the total number of pixels
a= Actual number of pixels for each category
b= Number of predicted pixels for each category
N= Total number of pixels

3.1.4 F1 Score:

$$F1 = 2*\frac{precision*recall}{precision + recall}, \tag{9}$$

where precision = The proportion of correctly classified pixels to the total predicted correct pixels
recall= The proportion of correctly classified pixels to the actual correct total number of pixels

3.1.5 Confusion Matrix : The Confusion Matrix is a commonly used tool to evaluate the performance of a classification model. It presents the relationship between the predicted results of the model and the true labels of each category in the form of a matrix. Specifically, the rows of the confusion matrix typically represent the true classes, while the columns represent the model's predicted classes. By examining the values in each element of the matrix, one can intuitively understand the number of correct and incorrect predictions for each category.

3.1.6 User's Accuracy（UA）： User's accuracy is used to describe the proportion of pixels that actually belong to a given class among those classified as that class in the classification results.

$$UA = \frac{nii}{\sum_{j=1}^{M} n_{ji}}, \tag{10}$$

where n_{ii} = The number of correctly classified pixels in category i
$\sum_{j=1}^{M} n_{ji}$ = The total number of pixels classified as the i th class in the confusion matrix (all pixels classified as the i th class)

3.1.7 Overall Accuracy（OA）： 。Overall accuracy (OA) represents the proportion of correctly classified pixels among all classified pixels in the dataset.

$$OA = \frac{\sum_{i=1}^{N} nii}{N_{total}}, \tag{11}$$

where n_{ii} = The number of correctly classified pixels in the i-th category of the confusion matrix
N_{total}= The total number of all pixels (sum of all elements in the confusion matrix

3.2 Accuracy Verification

3.2.1 Model Evaluation: Under the condition of maintaining a consistent dataset, ablation experiments were conducted by adding different modules. Exp-A introduces spatial attention in Stage 1 of the encoder, Exp-B incorporates channel attention in Stage 4 of the encoder, Exp-C integrates the ASPP module into the decoder, and Exp-D enhances skip connections by applying channel alignment and spatial attention.

Experimental Group	IOU(%)	Dice (%)	F1 Score	Kappa Coefficient
Baseline	72.3	82.1	72.3	0.685
Exp-A	74.1	83.5	75.1	0.712
Exp-B	73.6	83.0	74.6	0.703
Exp-C	75.2	84.3	76.8	0.728
Exp-D	74.8	84.0	79.5	0.721
RU-Net++	79.1	86.7	82.4	0.762

Table 3. Performance comparison of models on the test set

3.2.2 Validation Data: The validation dataset includes blue, green, red, and near-infrared bands, with a spatial resolution of 16 meters, a revisit cycle of 4 days, and a total coverage period of 41 days. Priority was given to clear-sky images, with cloud cover kept below 1%, and image acquisition was preferably conducted during the growing season. The validation images underwent orthorectification, geometric correction, radiometric calibration, atmospheric correction, and tiling, resulting in surface reflectance products. Each image covers an area of 100 km × 100 km and is projected using the UTM coordinate system.

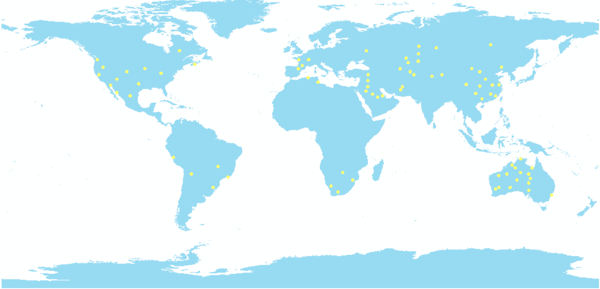


Figure 5. Location of verification area.

3.2.3 Validation Method: In ArcGIS, a total of 200 validation points were randomly generated across 78 selected images within the study areas. After determining the sample locations, experts familiar with the region conducted manual visual interpretation. Using Google Earth imagery from the same period, each sample point was cross-checked against its actual location to assess whether the land cover was impermeable. All sample points were individually verified. The final confusion matrix was derived through manual visual interpretation.

3.2.4 Accuracy Evaluation: To visualize the prediction results, Table 4 presents the classification outcomes in representative areas. By comparing the original remote sensing imagery with the classification results, the study validates the model's effectiveness and stability in real-world applications.

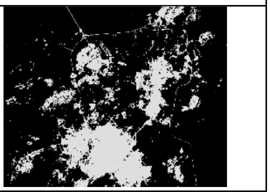
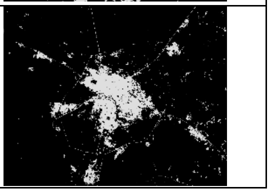
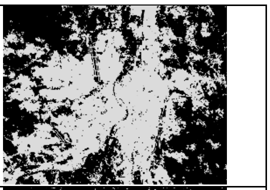
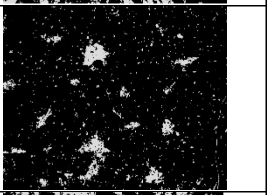
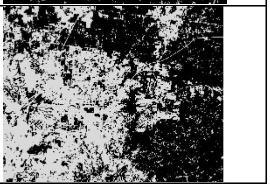
Serial Number	Oiginal image	Extract classification results of Impervious Surface
1		
2		
3		
4		
5		

Table 4. Extract Impervious Surface and compare it with the original image

To evaluate the classification performance of the proposed model in representative areas, random points were generated within each test region for visual interpretation. Based on the interpretation results, a confusion matrix was constructed to calculate three key metrics: User's Accuracy (UA), Overall Accuracy (OA), and Kappa Coefficient. Experimental results demonstrate that the model achieves high accuracy in distinguishing between impervious and pervious surfaces, confirming its robustness and effectiveness in real-world applications.

1	Pervious Surface	Impervious Surface	total	UA
Pervious Surface	188	2	190	0.989
Impervious Surface	0	10	10	1
total	188	12	200	0
UA	1	0.833	-	0.99
OA=0.99				
Kappa Coefficient=0.903				

2	Pervious Surface	Impervious Surface	total	UA
Pervious Surface	190	3	193	0.984
Impervious Surface	0	7	7	1
total	190	10	200	0
UA	1	0.7	-	0.985
OA=0.985				
Kappa Coefficient=0.815				

3	Pervious Surface	Impervious Surface	total	UA
Pervious Surface	189	2	191	0.989529
Impervious Surface	0	9	9	1
total	189	11	200	0
UA	1	0.818	-	0.99
OA=0.99				

Kappa Coefficient=0.894

4	Pervious Surface	Impervious Surface	total	UA
Pervious Surface	153	2	155	0.987
Impervious Surface	0	45	45	1
total	153	47	200	0
UA	1	0.957	-	0.99
OA=0.99				
Kappa Coefficient=0.971				

5	Pervious Surface	Impervious Surface	total	UA
Pervious Surface	166	2	168	0.988
Impervious Surface	0	32	32	1
total	166	34	200	0
UA	1	0.941	-	0.99
OA=0.99				
Kappa Coefficient=0.963				

Table 5. Confusion matrix of typical regional images

4. Discussion and conclusion

Although this study has made significant progress in impervious surface extraction, there is still room for further improvement. Future work will explore multiple perspectives to optimize model performance.

First, the current ablation experiments primarily analyze the contribution of each module within the model to overall performance. However, due to the lack of comparison with other mainstream models, it is challenging to comprehensively assess the model's relative advantages and limitations. In future work, we plan to introduce more classical and advanced models, such as DANet and Transformer, as benchmark comparisons. By using a unified dataset and evaluation metrics, we will conduct detailed performance comparisons under identical conditions to further validate the proposed model’s advantages and applicability in real-world scenarios.

Second, handling edge regions remains a major challenge, as impervious surfaces often have indistinct boundaries with surrounding land cover types such as vegetation and buildings. These boundary areas tend to lose fine-grained information, making accurate segmentation difficult.

Additionally, given that impervious surfaces may exhibit significant seasonal and climatic variations, future research will incorporate multi-temporal remote sensing data. By integrating temporal consistency constraints and employing change detection modules, we aim to enhance the model’s ability to capture dynamic variations over time.

By implementing these improvements, we expect to achieve significant enhancements in pixel-level segmentation accuracy and detail preservation while also strengthening the model’s stability and applicability. Ultimately, this will provide more reliable technical support for remote sensing applications, urban planning, and environmental monitoring.

References

LI Deren, LUO Hui, SHAO Zhenfeng. Review of Impervious Surface Mapping Using Remote Sensing Technology and Its Application. Geomatics and Information Science of Wuhan University, 2016, 41(5): 569-577,703.

Lu, D., Li, G., Kuang, W., & Moran, E. (2013). Methods to extract impervious surface areas from satellite images. *International Journal of Digital Earth*, *7*(2), 93–112. https://doi.org/10.1080/17538947.2013.866173

Xu H. Analysis of impervious surface and its impact on urban heat environment using the normalized difference impervious surface index (NDISI)[J]. Photogrammetric Engineering & Remote Sensing, 2010, 76(5): 557-565.

HUO Jiating, ZHAO Zhan, ZHU Xiuli. Extracting urban impervious area from multi-source image fusion data assisted by global land cover data[J]. Bulletin of Surveying and Mapping, 2024, 0(2): 19-25.

Ronneberger O, Fischer P, Brox T. U-net: Convolutional networks for biomedical image segmentation[C]//Medical image computing and computer-assisted intervention–MICCAI 2015: 18th international conference, Munich, Germany, October 5-9, 2015, proceedings, part III 18. Springer International Publishing, 2015: 234-241.

He K, Zhang X, Ren S, et al. Deep residual learning for image recognition[C]//Proceedings of the IEEE conference on computer vision and pattern recognition. 2016: 770-778.

Zhang X, Zhao T, Xu H, et al. GLC_FCS30D: The first global 30-m land-cover dynamic monitoring product with a fine classification system from 1985 to 2022 using dense time-series Landsat imagery and continuous change-detection method[J]. Earth System Science Data Discussions, 2023, 2023: 1-32.

Huang X, Song Y, Yang J, et al. Toward accurate mapping of 30-m time-series global impervious surface area (GISA)[J]. International Journal of Applied Earth Observation and Geoinformation, 2022, 109: 102787.

Appendix

GLC_FCS30D：*https://zenodo.org/records/8239305*
GISA:https://zenodo.org/record/5136330