

Article

Not peer-reviewed version

Semantic Segmentation of High-Resolution Remote Sensing Images Based on RS³Mamba: An Investigation on the Extraction Algorithm of Rural Compound Utilization Status

[Xinyu Fang](#), [Zhenbo Liu](#)^{*}, [Suan Xie](#)

Posted Date: 21 July 2025

doi: 10.20944/preprints2025071607.v1

Keywords: deep learning; semantic segmentation; high-resolution remote sensing; multi-scale feature fusion; Mamba



Preprints.org is a free multidisciplinary platform providing preprint service that is dedicated to making early versions of research outputs permanently available and citable. Preprints posted at Preprints.org appear in Web of Science, Crossref, Google Scholar, Scilit, Europe PMC.

Copyright: This open access article is published under a Creative Commons CC BY 4.0 license, which permit the free download, distribution, and reuse, provided that the author and preprint are cited in any reuse.

Article

Semantic Segmentation of High-Resolution Remote Sensing Images Based on RS³Mamba: An Investigation on the Extraction Algorithm of Rural Compound Utilization Status

Xinyu Fang, Zhenbo Liu * and Suan Xie

School of Remote Sensing and Geomatics Engineering, Nanjing University of Information Science and Technology, Nanjing 210044, China

* Correspondence: ZBLiu@nuist.edu.cn

Abstract

In this study, we utilize Gaofen-2 satellite remote sensing images to optimize and enhance the extraction of feature information from rural compounds. This is achieved through the implementation of the RS³Mamba+ deep learning model and the construction of Mamba-assisted branching with the aid of multi-directional selective scanning (SS2D) and the STEM network framework. The primary objective of this approach is to capture the long-distance spatial dependence of the compounds in high-resolution remote sensing images. Additionally, it aims to minimize computational loss. The introduction of a multiscale attention feature fusion mechanism is an important development in this field. This new mechanism has been demonstrated to optimize feature extraction and fusion, enhance edge contour extraction accuracy in courtyards, and improve the recognition and differentiation ability of the courtyard and complex texture regions. The feature information of courtyard utilization status is finally extracted using empirical methods. A typical rural area in Weifang City, Shandong Province, is selected as the experimental sample area. The results show that the extraction accuracy reaches an average intersection and merger ratio (mIoU) of 79.64% and a Kappa coefficient of 0.7889. This improves the F1 score by at least 8.35% compared with models such as U-Net, ResNet, Transformer, and so on. Furthermore, the mIoU improves by 7.41%. The mIoU has been enhanced by 7.41%. The efficacy of the algorithm in suppressing false alarms triggered by shadows and intricate textures is noteworthy. It is a valuable instrument for the extraction of compounds from rural compounds by leveraging condition feature information.

Keywords: deep learning; semantic segmentation; high-resolution remote sensing; multi-scale feature fusion; Mamba

1. Introduction

The rapid development of China's economy and urbanization rate has led to a significant migration of rural populations to urban areas. This phenomenon, coupled with the aging of the rural population, has resulted in the "hollowing out" [1–3] of the countryside, a term used to describe the decline of rural areas as they lose population and economic activity. This process has led to a significant waste of land resources and has created a substantial obstacle to the effective allocation and use of resources in rural areas. As a result, a considerable number of compounds have become vacant or abandoned, further exacerbating these issues. This phenomenon not only results in a significant wastage of land resources but also presents a substantial impediment to the effective allocation and utilization of rural resources [4,5]. The ability to swiftly and precisely ascertain information regarding the utilization of rural compounds is of paramount importance for

comprehending regional land reserve resources, ensuring national food security, and effectively implementing the strategy of comprehensive rural revitalization.

The conventional approach of extracting compounds involves manual field surveys [6–8], a method that is both labor-intensive and inefficient. The advent of high-resolution remote sensing [9–11] satellite data has led to a marked increase in the utilization of remote sensing technology for the precise identification and extraction of feature information of rural compounds [12–14]. This has been particularly evident in the application of deep learning models such as UNet [15], DeepLab [15,16], Transformer [17–19], and analogous models, which have led to substantial enhancements in the performance of remote sensing semantic segmentation. Wang et al. [20] examined the impact of deep learning technology on traditional village landscape assessment and proposed an analysis framework based on pixel-level semantic segmentation algorithm and image feature extraction. The physical attributes and spatial features of village landscape images are extracted by convolutional neural networks, combined with image recognition techniques (e.g., HOG, SIFT algorithms) to realize the classification of architectural elements, and simulate the value perception logic of experts and the public. Zhao et al. [21] proposed a model for identifying hollow villages. This model integrates static remote sensing images, village views, and nighttime lighting (NTL) time series data. The extraction of static features from buildings is achieved through the utilization of ResNet18 in conjunction with the attention module, while the analysis of dynamic patterns in human activities within NTL data is facilitated by LSTM-FCN. The experiments demonstrate that NTL data plays a pivotal role in the identification of hollow villages, and remote sensing and view data can complement the specifics of the built environment, thereby providing a cross-scale solution for the monitoring of rural hollowing. In their study, Meng et al. [22] propose an automated classification methodology for rural building features, employing unmanned aerial vehicle (UAV) tilt photography and deep learning algorithms. A convolutional neural network, such as ResNet50, is employed to classify seven indicators, including building function, structure, and age. The recognition accuracy for the number of building layers achieves 99.5%, while the abandoned state demonstrates a recognition accuracy of 95.9%. The study establishes a standardized workflow from image acquisition to classification mapping, and it verifies the efficiency advantage of deep learning in a large-scale survey of rural buildings. This can significantly reduce the labor cost of traditional field surveys. Wang et al. [23] developed a technical framework integrating unmanned aerial vehicles (UAVs), deep learning, and machine learning algorithms to recognize the utilization state of rural compounds. The integrity of farmhouses is recognized by AlexNet (accuracy 94.68%), and the classification of compound residence/vacant/abandoned status is realized by combining with Adaboost algorithm (accuracy 0.933). The proposed method effectively overcomes the conventional limitations associated with utility data dependence, thereby providing a highly versatile quantitative instrument for the management of intensive rural land.

The existing semantic segmentation algorithms demonstrate efficacy in building extraction scenarios; however, due to the intricate environment of rural compound utilization, differentiating the use of residential compounds spectrally remains challenging. This field exhibits the following deficiencies: Firstly, the low contrast between the rural houses and the background, attributable to the roof material and vegetation cover, can result in blurring of the target boundary extraction. Secondly, the dispersed distribution of independent houses can impede the Transformer-class model's capacity to optimize computational efficiency and accuracy. Thirdly, there is an absence of specialized labeled datasets for rural residential compounds [24–26]. In this study, the authors employ domestic Gaofen-2 high spatial resolution remote sensing images to illustrate a spatial state model replacement of Transformer for long-range dependency modeling in typical rural areas of Shandong Province. This replacement enhances the sequence modeling capability and significantly reduces the computational loss. The replacement is based on RS³Mamba's two-branch framework. Concurrently, a manually annotated dataset is being constructed, with the relevant features being extracted to classify the compounds and for target boundary extraction. The fuzzy problem is

embedded with morphological operations to adjust the boundary shape, thereby ensuring the effective extraction of feature information regarding rural compound usage.

2. Data Processing and Algorithm Design

2.1. Study Area and Data Source

This study uses Mengjia Village in Shouguang City, Shandong Province (Figure 1) as the test area. Located in the middle of Shandong Province (118.79°E, 36.85°N), the village has a total area of about 0.4 square kilometers. Due to the exodus of the rural population, unused rural compounds have become increasingly prominent in the test area, generally showing the pattern of "vacant center — expanding periphery."



Figure 1. Remote Sensing Images of the Study Area Taken by the GF-2 Satellite.

In this study, we used high-spatial-resolution GF-2 satellite remote sensing images to extract courtyard characteristics. GF-2 is China's first independently developed civilian optical remote sensing satellite with a spatial resolution better than 1 m. The width of a single camera is 23 km, and a dual-camera combination can reach up to 45 km. It can achieve a ground resolution of 0.81 m for panchromatic and 3.24 m for multispectral. The imaging time for the images selected for this study was October 16, 2023, with less than 10% cloud cover, and the images have been orthorectified.

2.2. Preprocessing

This study uses the remote sensing image processing software ENVI 5.3 to perform radiometric calibration and atmospheric correction on the multispectral bands of an image and relative radiometric correction on the panchromatic bands. Using the Red/Green/Blue bands for true-color synthesis and the GDAL library function to stretch and reconstruct the multispectral band correction onto the panchromatic bands produces an output true-color image. This image is then batch-cropped synchronously to a size of 512×512 pixels for model training.

Due to the high-density, patchy distribution of villages in the target area [27], the image is divided into several units. Areas with a concentrated distribution of compounds around the test area are selected to create a training set. The training set is about 100 times larger than the test set to extract the spatial distribution characteristics of unused compounds. To achieve the batch separation of

unused courtyards, we analyze fine features with sub-meter Google Earth images . We take vegetation coverage of courtyards as the core discriminative index and construct a decision rule by combining it with the damage status of buildings. To address the interference of shadows on feature extraction, we manually label non-vacant courtyards.

Figure 2 shows the technical route for data processing and model training.

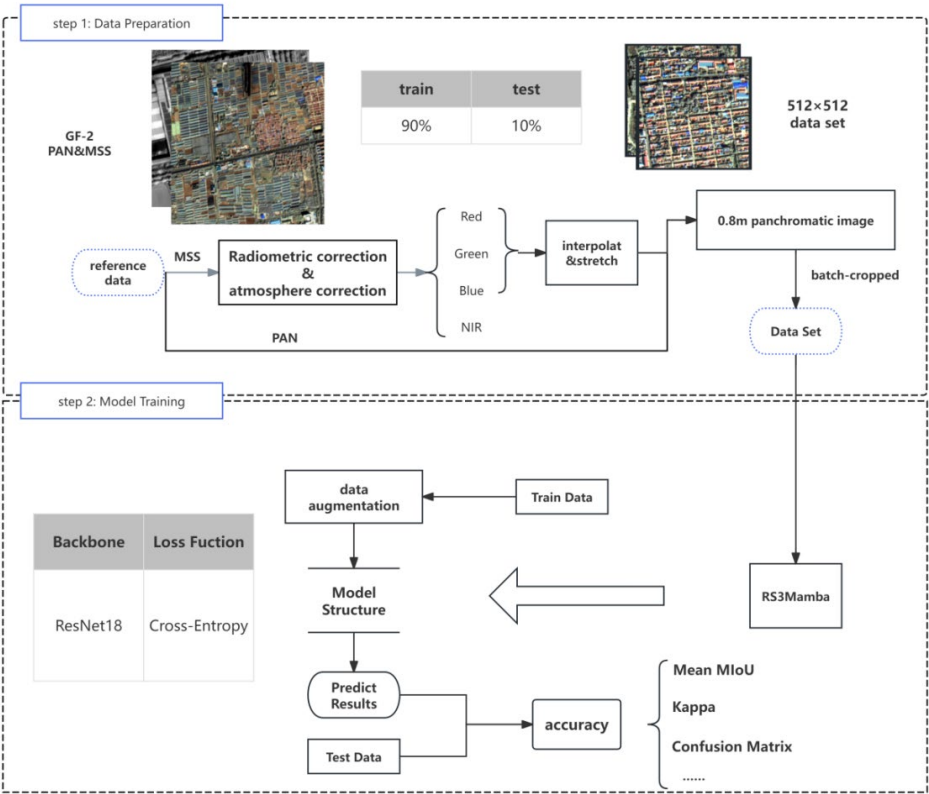


Figure 2. Workflow.

2.3. Algorithm

The RS³Mamba [28] model is based on the Mamba model and has a two-branch structure, as shown in Figure 3, containing an auxiliary branch and a main branch. The auxiliary branch is based on the VSS block and uses the Mamba model method to create long-range dependencies and provide global information. The main branch uses a convolutional neural network (ResNet) to learn local feature representations. The innovative multi-scale feature fusion module effectively combines global and local features, compensating for the lack of long-range modeling capability in traditional CNNs while avoiding the VSS model's unstable performance.

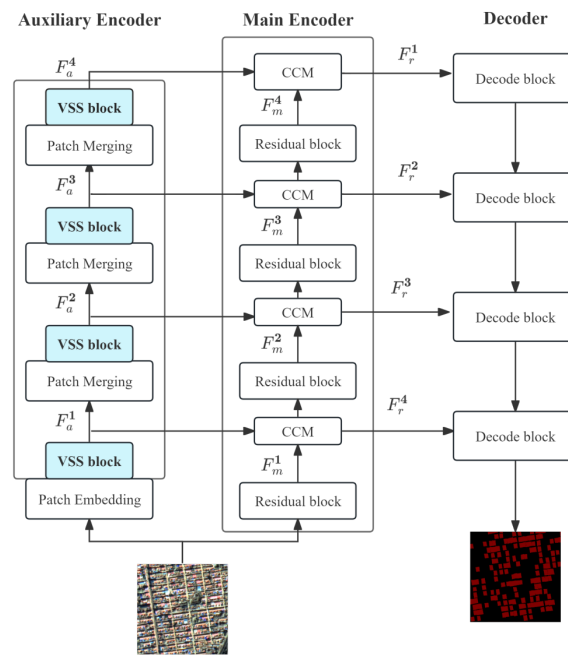


Figure 3. The overall architecture of RS3Mamba [28].

This study replaces the ReLU activation function used in lightweight, shallow networks with the SiLU activation function to construct nonlinear target features and alleviate the gradient descent problem in deep networks. The model's information loss rate is significantly reduced by improving the convolutional layer of the stem network and the MLP layer in Mamba's auxiliary branch. Additionally, the SS2D module's weight computation system is updated to create a multiscale attention fusion mechanism that optimizes and improves the model.

2.3.1. Auxiliary Branch Based on Mamba

The auxiliary branching based on the Mamba realizes its functionality primarily through the discrete state-space model of the VSS module in Figure 3. It uses an innovative multilayer perceptron for the VSS module's self-attentive output and utilizes a stem network to implement downsampling and connect the VSS module's layers in series.

A state space model (SSM) maps an input sequence to an output sequence through a set of state equations. Mathematically, this process can be expressed as state and output equations using linear ordinary differential equations (ODEs), respectively:

$$\begin{aligned} h'(t) &= Ah(t) + Bx(t) \\ y(t) &= Ch(t) \end{aligned} \quad (1)$$

Where $A \in \mathbb{R}^{N \times N}$ is the state transfer matrix, $B \in \mathbb{R}^{N \times 1}$ is the projection matrix, and $C \in \mathbb{R}^{1 \times N}$ is the projection matrix, $h(t)$ denotes the hidden state, and $x(t), y(t)$ represent the input and output sequences, respectively. The parameters A and B are continuous, so they cannot be directly mapped to the mainstream framework and are difficult to realize in digital systems. Therefore, this study discretizes the SSM using zero-order hold [29], transforming the continuous parameters A and B into discrete state and input matrices, \bar{A} and \bar{B} , respectively:

$$\begin{cases} \bar{A} = e^{A \cdot \Delta} \\ \bar{B} = \int_0^\Delta e^{As} ds \cdot B \approx B \cdot \Delta (\text{当 } \Delta \text{ 较小时}) \end{cases} \quad (2)$$

In particular, the time scale parameters Δ and projection matrices B and C are dynamically generated by linearly projecting the input feature map x_k (linear layer) to adaptively model spatial dependencies in different regions:

$$\Delta, B, C = \text{Linear}(x_k) \quad (3)$$

The discretized equation of state is then obtained as follows:

$$\begin{cases} h_k = \bar{A}h_{k-1} + \bar{B}x_k \\ y_k = Ch_k \end{cases} \quad (4)$$

The RSMamba model uses SS2D to capture the spatial global dependence of remote sensing images. The feature map is expanded into a one-dimensional sequence in four directions and processed separately for SSM:

$$x_v = \text{expand}(x, v), \quad v \in \{1, 2, 3, 4\} \quad (5)$$

Perform discrete SSM for each direction v :

$$\bar{x}_v = \text{SSM}(x_v) = \sum_{k=1}^L \left(\bar{A}_v^{(k)} h_{k-1,v} + \bar{B}_v^{(k)} x_{v,k} \right) \quad (6)$$

$x_v \in \mathbb{R}^{L \times D}$, where $L = H \times W$ denotes the sequence length and D is the feature dimension. The outputs from the four directions are then combined into the global feature F via weighted summation: notes the sequence length and D is the feature dimension. The outputs from the four directions are then combined into the global feature F via weighted summation.

$$F = \sum_{v=1}^4 w_v \cdot \text{Reshape}(y_v) \quad (7)$$

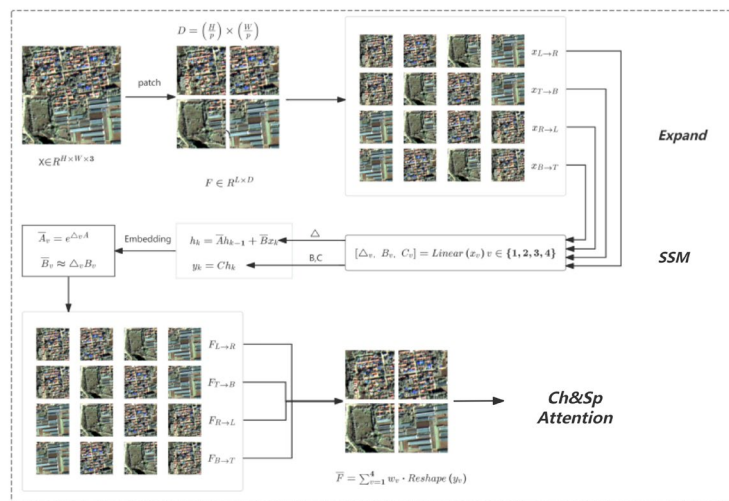


Figure 4. SS2D Dependency Analysis Schematic Diagram.

The final extension yields the following complete equation of state for Mamba. In this equation, $\bar{A}_v^{(k)}, \bar{B}_v^{(k)}, C_v^{(k)}$ is generated from input $x_{k,v}$ via linear projection to achieve input-dependent global modeling.

$$\begin{cases} h_{k,v} = \bar{A}_v^{(k)} h_{k-1,v} + \bar{B}_v^{(k)} x_{k,v} \\ y_{k,v} = C_v^{(k)} h_{k,v} \\ \bar{x} = \sum_{v=1}^4 WeightedSum(y_{1:L,v}) \end{cases} \quad (8)$$

In accordance with the above state equations, each layer of VSS modules obtains long-range dependencies and is connected to each other through a stem network. This setup efficiently preprocesses and extracts features from remote sensing images through improved convolution, normalization, and activation operations.

First, this study transforms a single-layer convolution (7x7) into a two-layer convolution structure to capture richer image patterns through progressive feature extraction. The first 3x3 convolution layer focuses on extracting base edge and texture features. It increases the number of channels from three to 32 and uses the BatchNorm normalization function after the first convolution layer. This accelerates model convergence by exploiting batch statistical information. The second 3x3 convolution layer further integrates and refines these features. It increases the number of channels to 48. This provides better-quality input for subsequent modules and significantly reduces information loss.

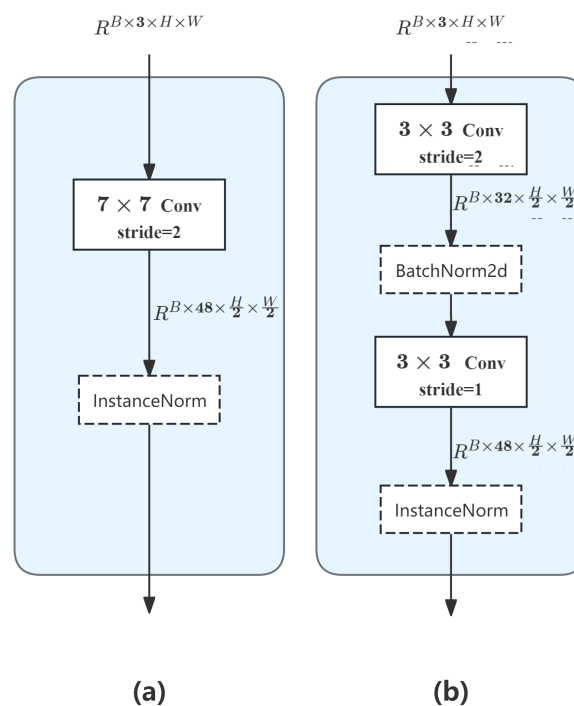


Figure 5. A comparison of the stem network innovation. (a) shows the original model's stem network framework, and (b) shows the RS³Mamba+ model's network framework.

In the MLP module [30] with self-attentive output after VSS, standard convolution is decomposed into depth convolution and point-by-point convolution using depth-separable convolution. This drastically reduces the number of parameters to about one-ninth of standard convolution and captures spatially localized information using a 3×3 depth convolution combined with a 1×1 point-by-point convolution to achieve inter-channel information fusion.

2.3.2. Multiscale Attention Feature Fusion Module

In dual-branch decoder feature fusion, the architecture integrates spatial, channel, and gate attention mechanisms to construct a feature fusion weight calculation system. This system is cascaded

in the channel-first, then spatial order during the multiscale attention feature fusion process and introduces an implicit gate attention mechanism to optimize the fusion process. Specifically, the gate attention module dynamically generates a weight matrix based on the spatial context of the input features when the long-range dependent features output from the VSS block and the local convolutional features derived from ResNet interact with each other through the multiscale feature fusion module. This matrix carries out the element-by-element fusion of the dual-channel features with adaptive modulation to effectively inhibit background noise interference with the boundary features of the house.

During the decoder stage, the gate attention mechanism adjusts the fusion ratio of features at different levels nonlinearly through the activation function. Transforming the feature fusion process into a learnable attention allocation problem enhances the model's ability to discriminate low-contrast targets (e.g., vegetation-covered vacant houses) in remote sensing images. This approach avoids the dimensionality explosion caused by traditional feature splicing.

During the post-processing stage of prediction, the architecture uses a gating mechanism to filter out meaningless fragmented spots. It then constructs rectangles with variable side lengths to predict the shapes and boundaries of labels. Finally, it creates dynamic filtering rules based on the connected domain area and morphological operations to further improve the spatial consistency of the segmentation results.

The model uses the gate attention mechanism multiple times to dynamically model semantic associations within input sequences via triad mapping of query, key, and value with gating logic. The core idea is to map the input features into three different vector spaces (Q, K, and V). After calculating the similarity between the query and key, we generate attention weights and then weight and sum the value based on these weights to capture long-range dependencies within the sequence. The technical details [31] are as follows:

Specify the image input as $X \in \mathbb{R}^{N \times L \times D}$, where N denotes the batch size, L is the sequence length, and D is the feature dimension:

$$\begin{aligned} Q &= XW^Q \in \mathbb{R}^{N \times L \times d_k} \\ K &= XW^K \in \mathbb{R}^{N \times L \times d_k} \\ V &= XW^V \in \mathbb{R}^{N \times L \times d_k} \end{aligned} \quad (9)$$

Where W^Q, W^K, W^V are learnable mapping matrices, d_k is the dimension of Q, K, and V, and num_heads denotes the number of branches of parallel computation that are satisfied:

$$d_k = D / \text{num_heads} \quad (10)$$

Backpropagate the gradient through the training data and update the weight matrix so that:

- W^Q extracts features related to the current semantic basic unit to be attended to
 - W^K extracts features related to how other basic units are attended to
 - W^V extracts the actual semantic content delivered
- Backpropagation adjusts the different mapping matrices. Calculate the similarity of each basic unit with W^Q and all units with W^K , then weight the aggregated W^V to obtain the attention weights:

$$\text{Attention}(Q, K, V) = \text{Softmax}\left(\frac{QK^T}{\sqrt{d_k}}\right)V \quad (11)$$

Where $\sqrt{d_k}$ is a scale factor that stabilizes the gradient. For each semantic basic unit in the input X, the Q vector interacts with the K vectors of the other semantic units to determine the attention score. The final output is a weighted sum of the V vectors, where the weights are determined by the attention score.

Ultimately, the gating unit is introduced into the attention mechanism to dynamically adjust the weights. The formula is as follows:

$$\begin{aligned} G &= \sigma(XW^G) \\ \text{Gated_Attention} &= G \odot \text{Attention}(Q, K, V) \end{aligned} \quad (12)$$

Where $G \in \mathbb{R}^{N \times L \times 1}$, the gating weight is defined, and \odot denotes element-wise multiplication. This enhances the expression of key semantics by suppressing the attentional response in regions of low relevance. Key technical steps of the model, such as two-branch feature fusion and predictive model post-processing, can be realized on this basis.

2.3.3. Model Enhancement

Together, optimization algorithms for neural networks, loss function design, and predictive post-processing strategies form the core link for optimizing the performance of semantic segmentation models.

1. Loss Function

This study introduces the cross-entropy loss function to minimize the relative entropy between the actual and expected outputs. It can accurately measure the difference between the model's predicted distribution and the true distribution by calculating the KL dispersion of two discrete probability distributions. The function receives the model's predictions, true labels, and weights as inputs. The mathematical expression for the cross-entropy loss is:

$$L(\theta) = H(p, q) = - \sum_x p(x) \log q(x) \quad (13)$$

The probability of the true distribution is denoted by $p(x)$, and the probability of the model's predictive distribution is denoted by $q(x)$.

2. Optimizer

The optimizer adjusts the model parameters to accelerate convergence based on the gradient calculated by the loss function. After obtaining the loss value, the gradient of the loss with respect to the model parameters is calculated using backpropagation. In this study, the Adam W optimizer [32] updates the loss function parameters according to the gradient descent rule to minimize prediction error in the next iteration. The parameter update formula is as follows:

First, the gradient with weight decay is computed.

$$g_t' = g_t + \lambda \cdot \theta_{t-1} \quad (14)$$

Then, the first-order moment estimate is updated accordingly.

$$m_t = \beta_1 \cdot m_{t-1} + (1 - \beta_1) \cdot g_t' \quad (15)$$

Update the second-order moment estimates, i.e., the exponential moving average of the gradient squares.

$$v_t = \beta_2 \cdot v_{t-1} + (1 - \beta_2) \cdot (g_t')^2 \quad (16)$$

Correcting the bias of the first- and second-order moment estimates, respectively.

$$\hat{m}_t = \frac{m_t}{1 - \beta_1^t} \quad (17)$$

$$\hat{v}_t = \frac{v_t}{1 - \beta_2^t} \quad (18)$$

The final update parameters are as follows:

$$\theta_t = \theta_{t-1} - \eta \cdot \frac{\hat{m}_t}{\sqrt{\hat{v}_t} + \epsilon} \quad (19)$$

In this model, the parameters are denoted by θ_t , the gradient by g_t , the time step by t , the learning rate by η , the weight decay coefficient by λ , the exponential decay rate of the moment estimation by β_1 and β_2 , and the numerical stability minor constant by ϵ (set to 10^{-8} by default).

3. Predictive reprocessing

In this study, morphological operations are introduced to assist prediction. Firstly, a connectivity domain analysis is performed on the predicted images, and small targets are removed by setting an area threshold to remove obvious noise. Subsequently, closed operations are used to fill internal holes and connect neighboring fractured regions to restore the complete morphology of the target. Finally, open operations are used to clean up the edge noise and smooth the boundaries to enhance the clarity of the target contour. The gate attention mechanism is further superimposed in the prediction post-processing to construct rectangles with variable side lengths to predict the label shapes and boundaries. The dynamically generated spatial attention weights are used to accurately recognize and suppress the meaningless broken spots. When combined with the connectivity domain area threshold filtering, this mechanism can significantly reduce the occurrence of small-size error prediction and construct dynamic filtering rules based on the connectivity domain area and morphological operations. These rules improve the spatial consistency of the segmentation results and strengthen the connectivity and integrity of the target area. This makes the mechanism especially suitable for fragmented courtyard fine segmentation scenarios.

3. Feature Extraction Process and Implementation

3.1. Dataset Labeling

In this study, the rural compound utilization status dataset is initially constructed through visual interpretation. The general features of vacant compounds are extracted by manually extracting vacant compounds and combining them with labeled multispectral information. Finally, the data are manually labeled to adjust the scale constraints.

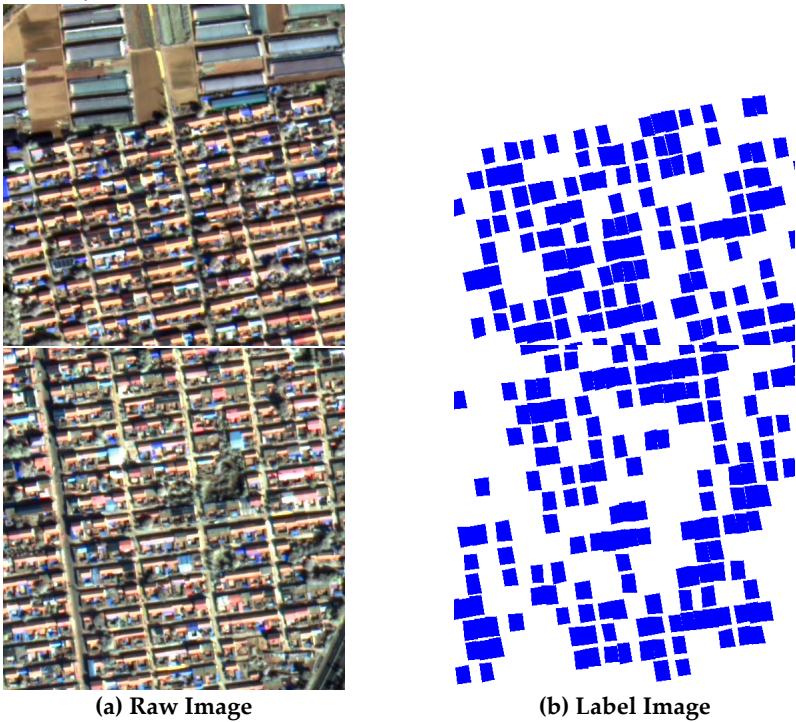


Figure 6. Example of cut image and manual labeling.

The study utilizes high-resolution image data to analyze visual features such as roof condition, spatial distribution, and vegetation condition to assist in extracting compound use conditions. The non-village portion of the preprocessed GF-2 image (see Figure 3) is filtered, and approximately 14,000 labels are selected to construct a manually labeled dataset. Random cropping is then applied to the image to enhance the diversity of samples and improve the robustness of the model to the target location. Flipping is employed in a seemingly random manner to simulate a mirrored scene. This is done to enhance the model's invariance to the target direction, improve the utilization of features, and introduce smooth noise through Gaussian fuzzy operations. The purpose of this is to

enhance the model's adaptability to blurred images. The model is enhanced by MixUp and CutMix data, employing linear interpolation to blend different samples and their labels, thereby enhancing the model's comprehension of inter-class boundaries. Additionally, random cropping and replacement of image regions is utilized to compel the model to prioritize local features.

3.2. Parameter

During the configuration process, this study adjusts the hyperparameters and evaluates the performance of the model in detecting housing utilization in the study area. The present study was conducted within a Linux operating system, utilizing a particular environment that comprised a 1 TB hard disk and a GPU. Nvidia 4090, CPU: The Intel Core i7 13700KF, equipped with CUDA: V11.2, boasts 128 GB of RAM. The RS³Mamba+ network was implemented in Python 3.8, while the deep learning framework was Pytorch. The base learning rate was set to 1×10^{-2} , weight decay was set to 5×10^{-4} , and batch size was set to 10. The total training period is set to 50. The learning rate is tuned by a multi-step learning rate scheduler. The loss function was defined as an amalgamation of cross-entropy loss and an ignore labeling mechanism.

3.3. Model Performance Evaluation

The performance of RS³Mamba+ on the dataset is evaluated by the F1 score, the mean intersection and merger ratio (mIoU), and the Kappa coefficient, calculated based on the aggregated confusion matrix. The following calculations have been performed:

$$P = \frac{1}{N} \sum_{k=1}^N \frac{TP_k}{TP_k + FP_k} \quad (20)$$

$$R = \frac{1}{N} \sum_{k=1}^N \frac{TP_k}{TP_k + FN_k} \quad (21)$$

$$F1 = 2 \times \frac{P \times R}{P + R} \quad (22)$$

$$mIoU = \frac{1}{N} \sum_{k=1}^N \frac{TP_k}{TP_k + FP_k + FN_k} \quad (23)$$

$$Kappa = \frac{1}{N} \sum_{k=1}^N \frac{P \times R - (P + R - 1)}{(1 - P) \times R + P} \quad (24)$$

其中 In this study, TP_k , FP_k , TN_k , FN_k represent true cases, false positive cases, true negative cases, and false negative cases, respectively, for a specific object indexing category k . P , R , $F1$, $mIoU$, $Kappa$ represent the precision rate, the recall rate, the F1 score, the average intersection and merger ratio, and the Kappa coefficient, respectively.

3.4. Realization of Results

The present study employs the RS³Mamba+ network model to train the manually labeled sample set, as illustrated in Figure 2. Figure 7 presents a trend plot of the number of iterative rounds (epoch) versus the mIoU and loss function during model training. The horizontal axis indicates the number of iterative rounds, while the vertical axes represent the corresponding loss function and performance parameters, respectively. The total number of epochs in the training process is 50, with each epoch comprising 1,000 training iterations. These iterations are based on the weights from the previous iteration, and they are performed through multiple rounds of iteration.

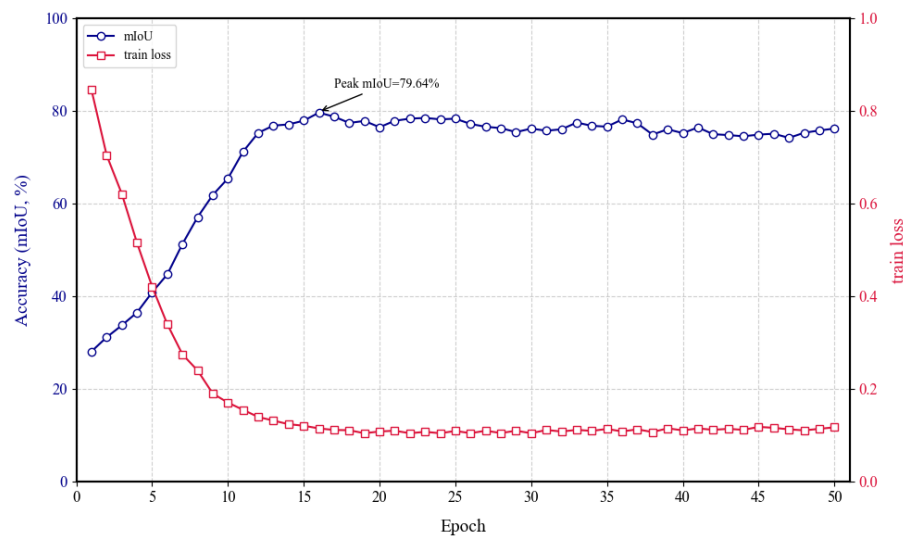


Figure 7. Trend of Loss and Accuracy in RS³Mamba+.

The parameter variations of the RS³Mamba+ model as a whole reflect the model's progression from feature learning to overfitting risk exposure. The model commences with a substantial initial loss, rapidly descending below 0.2 within 10 epochs. This observation signifies that the model has been trained with exceptional efficiency. The efficacy of the dual-branch co-training approach is evident in the rapid escalation of the mIoU to its maximum of 79.64% by the 16th epoch, signifying an optimal alignment between the model and the dataset. The primary branch enhances local detail extraction, while the auxiliary branch models long-range dependencies through the SS2D module. This module combines with the attention mechanism to focus on key regions, thereby facilitating rapid feature learning and fusion. This combination demonstrates the robustness of RS³Mamba+.

A comprehensive investigation was conducted to ascertain the courtyard area of Mengjia Village in Weifang City, Shandong Province. This investigation was conducted using ENVI software statistics obtained through high-precision multi-source visual interpretation (see Figure 1). The investigation revealed that the total courtyard area of Mengjia Village is 345,386 m², of which 140,469 m² is utilized regularly. The remaining 155,381 m² has been extracted using deep learning algorithms. Preliminary calculations indicate that the utilization rate of the courtyard in Mengjia Village is 40.67%, while the utilization rate of the courtyard as a result of deep learning extraction is 44.99%. The accuracy rate of the deep learning method in extracting the utilized compounds is 90.40%, mF1 is 0.8851, mean mIoU is 79.64%, and the Kappa coefficient is 0.7889. These findings indicate that the deep learning method is highly efficient in recognition and is evidently superior to the general statistical method in terms of recognition accuracy.

Table 2. Confusion matrix based on RS³Mamba+.

User\Reference Class	Background	Building	Sum
Background	5038452	177017	5215469
Building	150224	578513	728737
Sum	5188676	755530	

4. Discussion

4.1. Comparison of Base Algorithm Accuracy

The outcomes of the comparative experiments on the manually labeled dataset are presented in Table 3. The experimental results indicate that the accuracy of a single model network is suboptimal in terms of extracting features from rural buildings. The RS³Mamba+ model implemented in this study attains the optimal mIoU (0.7964) and Kappa coefficient (0.7889), which are a minimum of

8.35% and 7.41% more effective than the single-branch network mF1 and mIoU, respectively. A comparison of RS³Mamba+ with Transformer, which also incorporates long-range dependency modeling, reveals that RS³Mamba+ enhances the mF1 and mIoU scores by 10.78% and 11.8%, respectively. This observation substantiates the model's superior capacity to integrate global and local feature information, thereby demonstrating its efficacy in feature fusion.

Table 3. Segmentation results of different models on the dataset. The accuracy of each category is presented in the format of mF1/IoU(%).

Method	Building	Background	mF1	mIoU
U-Net	0.6316/0.4615	0.9046/0.8259	0.7680	0.6437
ResNet	0.7020/0.5558	0.8905/0.8888	0.7962	0.7223
Transformer	0.6605/0.5102	0.8869/0.8478	0.7737	0.6784
RS ³ Mamba+(Ours)	0.8131/0.7626	0.9772/0.9055	0.8815	0.7964

4.2. Local Visualization Analysis

As illustrated in Figure 7, the RS³Mamba+ model demonstrates efficacy in the domain of house area extraction. A comparative analysis of the U-Net, ResNet, and Transformer models reveals that RS³Mamba+ demonstrates optimal performance in terms of detail capture and boundary definition. The initial two columns illustrate the prediction accuracy of the clustered areas of houses, and the model in this study exhibits exceptional performance in addressing fragmented spots. The final column demonstrates the prediction effect of independent houses, with the red box highlighting its discriminative ability. In the semantic segmentation task, the predictive quality of the models exhibits significant variability. U-Net demonstrates successful building recognition in certain regions; however, mis-segmentation and miss-detection occur in the peripheral regions. ResNet enhances edge recognition; nevertheless, there are still erroneous edges and minor miss-detection regions, which impairs the discriminative power, particularly in damaged areas and boundaries. The transformer demonstrates proficiency in the domain of region recognition; however, in the presence of intricate region edges, the clarity is inadequate, resulting in mis-segmentation. The residual neural network (ResNet) and transformer exhibit high accuracy in the recognition of tree-related structures. Nevertheless, the accurate classification of houses shaded by trees poses a significant challenge due to the interference of background noise and the absence of comprehensive modeling of the house compound.

The RS³Mamba+ demonstrates proficiency in recognizing and segmenting dispersed or small-scale buildings, and it is effective in dealing with meaningless fragmented spots, which is a difficult achievement for other models. Furthermore, the model exhibits reduced propensity for misclassifying agricultural areas as residential zones when confronted with intricate backgrounds and noise, thereby demonstrating its efficacy in handling complex data structures. Additionally, it evinces notable robustness in independent house prediction and determination, substantiating its effectiveness and superiority in the task of building refinement classification.

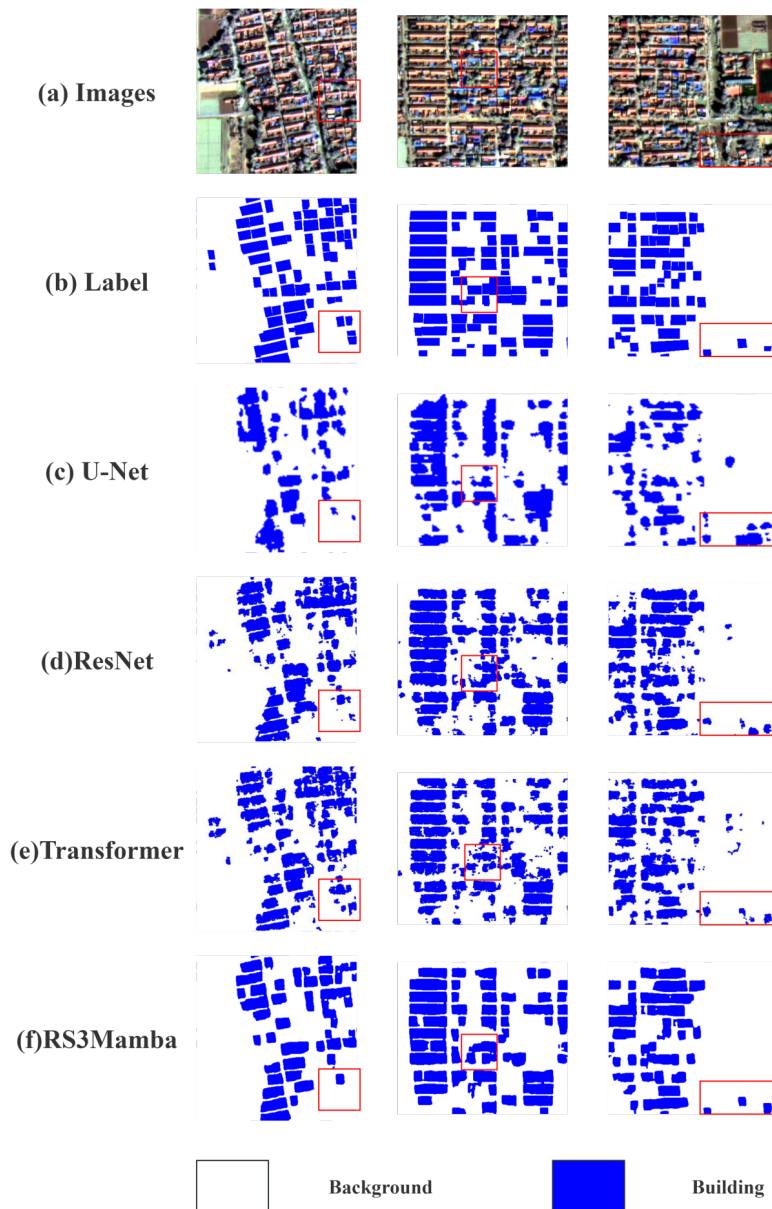


Figure 8. Visualization results of different models on the dataset.

As illustrated in Figure 9, the predicted labels for the building refinement extraction task for remotely sensed imagery are presented. In this figure, the areas designated for buildings are indicated in blue, while the remaining areas encompass farmland, bare soil, and primary roadways within villages that cannot be rejected by cropping. The figure illustrates the distribution of buildings within a complex background, encompassing both dense and sparse regions. It is evident that the delineations of the structures are distinct and there is an absence of superfluous fragmentation that would compromise the statistical precision. This observation suggests that the RS³Mamba+ prediction results possess a high degree of boundary identification precision and are capable of differentiating between structures and non-structural domains. This capacity is more conducive to subsequent statistical and analytical endeavors in comparison to the boundaries of the remaining single-model network. The RS³Mamba+ model's efficacy in preserving the continuity and consistency of building areas is indicative of its superior identification capabilities.



Figure 9. Visualization results of RS³Mamba+ on the dataset.

4.3. Ablation Experiments

The backbone of this study is ResNet18, which quantifies the enhancement effect of different functional modules on the model respectively, and Fig. 9 shows the experimental results of the ablation experiment. (b) Segmentation using only ResNet18, the segmentation results are fragmented, and there are more misclassification and meaningless fragmented spots in the building area, which is difficult to accurately outline the building, and thus has a low mIoU (0.6953) and Kappa (0.6851). (c) Introducing Mamba assisted branching on the basis of backbone, compared with (b), the segmentation effect is improved, the integrity of the building area is enhanced, the misclassification is reduced while the meaningless fragmented spots are significantly eliminated, so that the prediction target tends to be regularized, and the enhancement of mIoU and Kappa of 8.09% and 8.88% are realized. (d) Introducing gate processing on the basis of backbone, focusing on reducing misjudgment and improving the recognition rate of small-sized buildings, realizing the improvement of mIoU and Kappa 4.02% and 4%. The specific parameters of each model for the ablation experiment are shown in Table 4.

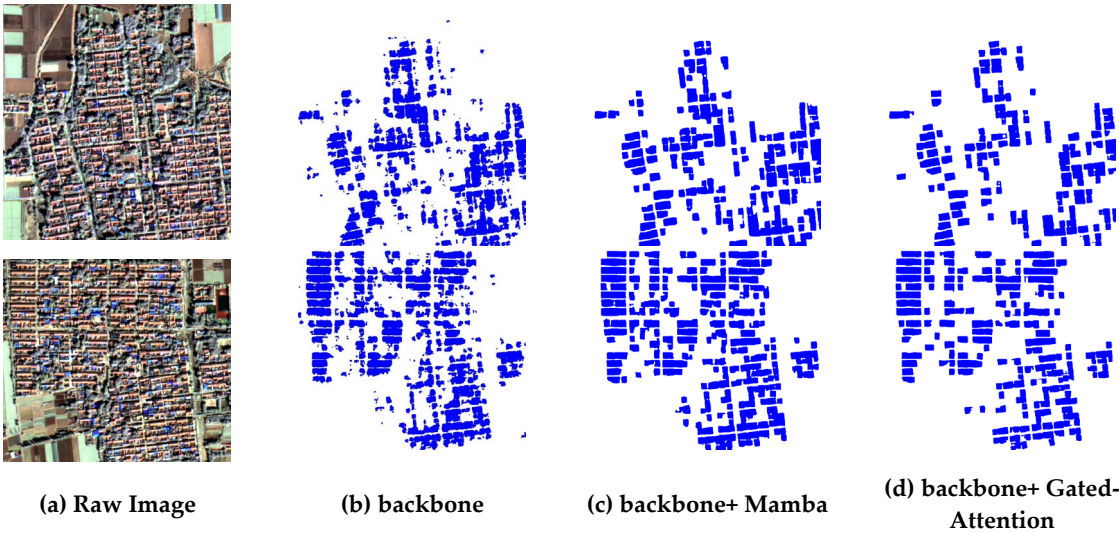


Figure 10. Results of ablation experiments.

Table 4. Segmentation results of Ablation Study on the dataset. The accuracy of each category is presented in the format of mF1/IoU(%).

Model Name	Building IoU	Background IoU	mF1	mIoU	Kappa
Backbone	0.626	0.7646	0.7124	0.6953	0.6851
Backbone+ Mamba	0.7251	0.8272	0.8692	0.7762	0.7739
Backbone+ Gated-Attention	0.6797	0.7912	0.8125	0.7355	0.7251
RS³Mamba+(Ours)	0.7626	0.9055	0.8815	0.7964	0.7889

5. Conclusions

In this paper, RS³Mamba+ is demonstrated to be an enhanced semantic segmentation framework based on the state space model (SSM). It exhibits a reduction in computational complexity while capturing long-distance spatial dependencies of remote sensing images through a two-branch synergistic architecture with the multi-directional selective scanning (SS2D) mechanism. It optimizes feature fusion by combining the gate-attention mechanism and strengthens the edge sharpness through morphology post-processing. The result is an efficient balance between global contextual modeling and local detail preservation. The findings indicate that the RS³Mamba+deep learning extraction method demonstrates a high degree of accuracy (90.40%) and attains an average intersection and merger ratio (mIoU) of 79.64% and a Kappa coefficient of 0.7889 in GF-2 satellite images. This method effectively facilitates the extraction of information concerning the utilization status of rural compounds. The high recall rate of the model for small edifices may be accompanied by misdetection, especially in low-contrast scenes susceptible to interference by vegetation shadows. In the future, the model may be combined with unmanned aerial vehicles and multi-temporal, high-resolution satellite imagery to provide richer texture details and annotation precision. Furthermore, with the advent of algorithmic lightweighting technology, the RS³ Mamba+ framework is anticipated to enhance extraction precision, facilitating the extraction of information on the utilization status of rural compounds. This advancement is expected to augment the technical support for rural revitalization and regional development.

Author Contributions: validation, X.F., S.X.; formal analysis, X.F., S.X.; investigation, S.X.; resources, X.F., Z.L.; data curation, Z.L.; writing—original draft preparation, X.F.; writing—review and editing, Z.L.; visualization, X.F.; supervision, Z.L.; funding acquisition, Z.L. All authors have read and agreed to the published version of the manuscript.

Funding: This research was funded by The Major Project of High Resolution Earth Observation System, Grant No. “30-Y60B01-9003-22/23”.

Conflicts of Interest: The authors declare no conflicts of interest.

References

1. Liu, C.; Xu, M. Characteristics and Influencing Factors on the Hollowing of Traditional Villages—Taking 2645 Villages from the Chinese Traditional Village Catalogue (Batch 5) as an Example. *Int. J. Environ. Res. Public. Health* **2021**, *18*, 12759.
2. Smith, G. The Hollow State: Rural Governance in China. *China Q.* **2010**, *203*, 601–618.
3. Comprehensive Evaluation of Rural Courtyard Utilization Efficiency: A Case Study in Shandong Province, Eastern China | Journal of Mountain Science Available online: <https://link.springer.com/article/10.1007/s11629-019-5824-x> (accessed on 16 July 2025).
4. Characteristics and Influencing Factors on the Hollowing of Traditional Villages—Taking 2645 Villages from the Chinese Traditional Village Catalogue (Batch 5) as an Example Available online: <https://www.mdpi.com/1660-4601/18/23/12759> (accessed on 16 July 2025).
5. Liu, Y.; Liu, Y.; Chen, Y.; Long, H. The Process and Driving Forces of Rural Hollowing in China under Rapid Urbanization. *J. Geogr. Sci.* **2010**, *20*, 876–888.
6. Sun, H.; Liu, Y.; Xu, K. Hollow Villages and Rural Restructuring in Major Rural Regions of China: A Case Study of Yucheng City, Shandong Province. *Chin. Geogr. Sci.* **2011**, *21*, 354–363, doi:10.1007/s11769-011-0474-0.
7. Li, S. The Renewal of Hollow Village A Research and Proposal of Erhai Area in YunNan, China 2014.
8. Sun, H.; Liu, Y.; Xu, K. Hollow Villages and Rural Restructuring in Major Rural Regions of China: A Case Study of Yucheng City, Shandong Province. *Chin. Geogr. Sci.* **2011**, *21*, 354–363, doi:10.1007/s11769-011-0474-0.
9. Chen, Y.; Wang, Y.; Xiong, S.; Lu, X.; Zhu, X.X.; Mou, L. Integrating Detailed Features and Global Contexts for Semantic Segmentation in Ultra-High-Resolution Remote Sensing Images. *IEEE Trans. Geosci. Remote Sens.* **2024**.
10. Zeng, Q.; Zhou, J.; Tao, J.; Chen, L.; Niu, X.; Zhang, Y. Multiscale Global Context Network for Semantic Segmentation of High-Resolution Remote Sensing Images. *IEEE Trans. Geosci. Remote Sens.* **2024**, *62*, 1–13, doi:10.1109/TGRS.2024.3393489.
11. Wu, Z.; Li, J.; Wang, Y.; Hu, Z.; Molinier, M. Self-Attentive Generative Adversarial Network for Cloud Detection in High Resolution Remote Sensing Images. *IEEE Geosci. Remote Sens. Lett.* **2020**, *17*, 1792–1796, doi:10.1109/LGRS.2019.2955071.
12. Fan, R.; Wang, L.; Feng, R.; Zhu, Y. Attention Based Residual Network for High-Resolution Remote Sensing Imagery Scene Classification. In Proceedings of the IGARSS 2019 - 2019 IEEE International Geoscience and Remote Sensing Symposium; IEEE: Yokohama, Japan, July 2019; pp. 1346–1349.
13. Ji, H.; Li, X.; Wei, X.; Liu, W.; Zhang, L.; Wang, L. Mapping 10-m Resolution Rural Settlements Using Multi-Source Remote Sensing Datasets with the Google Earth Engine Platform. *Remote Sens.* **2020**, *12*, 2832, doi:10.3390/rs12172832.
14. Li Zheng. Research on Key Technology for Acquiring Building Information of Hollow Village Based on UAV High Resolution Imagery [D]. Southwest Jiaotong University, 2018. in Chinese (accessed on 16 July 2025).

15. Chiu, W.-T.; Lin, C.-H.; Jhu, C.-L.; Lin, C.; Chen, Y.-C.; Huang, M.-J.; Liu, W.-M. Semantic Segmentation of Lotus Leaves in UAV Aerial Images via U-Net and DeepLab-Based Networks. In Proceedings of the 2020 International Computer Symposium (ICS); IEEE: Tainan, Taiwan, December 2020; pp. 535–540.
16. Qian, Z.; Cao, Y.; Shi, Z.; Qiu, L.; Shi, C. A Semantic Segmentation Method for Remote Sensing Images Based on Deeplab V3. In Proceedings of the 2021 2nd International Conference on Big Data & Artificial Intelligence & Software Engineering (ICBASE); IEEE: Zhuhai, China, September 2021; pp. 396–400.
17. Zhang, R.; Zhang, Q.; Zhang, G. LSRFormer: Efficient Transformer Supply Convolutional Neural Networks With Global Information for Aerial Image Segmentation. *IEEE Trans. Geosci. Remote Sens.* **2024**, *62*, 1–13, doi:10.1109/TGRS.2024.3366709.
18. He, X.; Zhou, Y.; Zhao, J.; Zhang, D.; Yao, R.; Xue, Y. Swin Transformer Embedding UNet for Remote Sensing Image Semantic Segmentation. *IEEE Trans. Geosci. Remote Sens.* **2022**, *60*, 1–15, doi:10.1109/TGRS.2022.3144165.
19. Wu, H.; Huang, P.; Zhang, M.; Tang, W.; Yu, X. CMTFNet: CNN and Multiscale Transformer Fusion Network for Remote-Sensing Image Semantic Segmentation. *IEEE Trans. Geosci. Remote Sens.* **2023**, *61*, 1–12, doi:10.1109/TGRS.2023.3314641.
20. Wang, T.; Chen, J.; Liu, L.; Guo, L. A Review: How Deep Learning Technology Impacts the Evaluation of Traditional Village Landscapes. *Buildings* **2023**, *13*, 525, doi:10.3390/buildings13020525.
21. Zhao, H.; Li, X.; Gu, Y.; Deng, W.; Huang, Y.; Zhou, S. Integrating Time-Series Nighttime Light Data With Static Remote Sensing and Village View Images for Hollow Villages Identification. *IEEE J. Sel. Top. Appl. Earth Obs. Remote Sens.* **2025**, *18*, 9151–9165, doi:10.1109/JSTARS.2025.3548420.
22. Automatic Classification of Rural Building Characteristics Using Deep Learning Methods on Oblique Photography | Building Simulation Available online: <https://link.springer.com/article/10.1007/s12273-021-0872-x> (accessed on 28 June 2025).
23. Wang, M.; Xu, W.; Cao, G.; Liu, T. Identification of Rural Courtyards' Utilization Status Using Deep Learning and Machine Learning Methods on Unmanned Aerial Vehicle Images in North China. *Build. Simul.* **2024**, *17*, 799–818, doi:10.1007/s12273-023-1099-9.
24. Huang, X.; Liu, H.; Zhang, L. Spatiotemporal Detection and Analysis of Urban Villages in Mega City Regions of China Using High-Resolution Remotely Sensed Imagery. *IEEE Trans. Geosci. Remote Sens.* **2015**, *53*, 3639–3657, doi:10.1109/TGRS.2014.2380779.
25. Zhan, Z.; Zhang, X.; Liu, Y.; Sun, X.; Pang, C.; Zhao, C. Vegetation Land Use/Land Cover Extraction From High-Resolution Satellite Images Based on Adaptive Context Inference. *IEEE Access* **2020**, *8*, 21036–21051, doi:10.1109/ACCESS.2020.2969812.
26. A Novel Framework Based on Mask R-CNN and Histogram Thresholding for Scalable Segmentation of New and Old Rural Buildings Available online: <https://www.mdpi.com/2072-4292/13/6/1070> (accessed on 16 July 2025).
27. Land Use Transitions and Their Dynamic Mechanism: The Case of the Huang-Huai-Hai Plain | Journal of Geographical Sciences Available online: <https://link.springer.com/article/10.1007/s11442-016-1283-2> (accessed on 16 July 2025).
28. Ma, X.; Zhang, X.; Pun, M.-O. RS³ Mamba: Visual State Space Model for Remote Sensing Image Semantic Segmentation. *IEEE Geosci. Remote Sens. Lett.* **2024**, *21*, 1–5, doi:10.1109/LGRS.2024.3414293.
29. Chen, K.; Chen, B.; Liu, C.; Li, W.; Zou, Z.; Shi, Z. RSMamba: Remote Sensing Image Classification with State Space Model 2024.
30. Shawky, O.A.; Hagag, A.; El-Dahshan, E.-S.A.; Ismail, M.A. Remote Sensing Image Scene Classification Using CNN-MLP with Data Augmentation. *Optik* **2020**, *221*, 165356, doi:10.1016/j.ijleo.2020.165356.

31. Gated Recurrent Multiattention Network for VHR Remote Sensing Image Classification Available online: <https://ieeexplore.ieee.org/abstract/document/9495118> (accessed on 3 July 2025).
32. Comparison between Adam, AdaMax and Adam W Optimizers to Implement a Weather Forecast Based on Neural Networks for the Andean City of Quito Available online: <https://ieeexplore.ieee.org/abstract/document/9590681> (accessed on 3 July 2025).

Disclaimer/Publisher's Note: The statements, opinions and data contained in all publications are solely those of the individual author(s) and contributor(s) and not of MDPI and/or the editor(s). MDPI and/or the editor(s) disclaim responsibility for any injury to people or property resulting from any ideas, methods, instructions or products referred to in the content.