

High-Resolution Remote Sensing Imagery for the Recognition of Traditional Villages

Mengchen Wang¹, Linshuhong Shen^{2*}

¹College of Geomatics and Urban Spatial Information, Beijing University of Civil Engineering and Architecture, Beijing 100044, China

²GeoScene, Beijing 100028, China

*Corresponding author: Linshuhong Shen, shen_linshuhong@yeah.net

Copyright: © 2024 Author(s). This is an open-access article distributed under the terms of the Creative Commons Attribution License (CC BY 4.0), permitting distribution and reproduction in any medium, provided the original work is cited.

Abstract: Traditional Chinese villages, vital carriers of traditional culture, have faced significant alterations due to urbanization in recent years, urgently necessitating artificial intelligence data updates. This study integrates high spatial resolution remote sensing imagery with deep learning techniques, proposing a novel method for identifying rooftops of traditional Chinese village buildings using high-definition remote sensing images. Using 0.54 m spatial resolution imagery of traditional village areas as the data source, this method analyzes the geometric and spectral image characteristics of village building rooftops. It constructs a deep learning feature sample library tailored to the target types. Employing a semantically enhanced version of the improved Mask R-CNN (Mask Region-based Convolutional Neural Network) for building recognition, the study conducts experiments on localized imagery from different regions. The results demonstrated that the modified Mask R-CNN effectively identifies traditional village building rooftops, achieving an of 0.7520 and an of 0.7400. It improves the current problem of misidentification and missed detection caused by feature heterogeneity. This method offers a viable and effective approach for industrialized data monitoring of traditional villages, contributing to their sustainable development.

Keywords: Traditional villages; Building rooftops; High spatial resolution remote sensing; Instance segmentation

Online publication: January 29, 2024

1. Introduction

Traditional villages in China are repositories of rural culture, historical heritage, and natural heritage, serving as living fossils and museums rich in cultural significance, often referred to as the DNA of the Chinese nation ^[1]. In recent years, traditional villages have been impacted by natural disasters and human destruction. Conventional manual preservation methods have become templated, data collection methods are invasive, and data storage is limited in scope. As a result, the number of original, ecological villages in China plummeted from 3.6 million to approximately 2.7 million in the first decade of the 21st century ^[2]. Recently, policies like the “Beijing Traditional Village Restoration Technical Guidelines” were introduced in Beijing, underscoring the importance of preserving traditional village architectural styles. By 2023, six batches of traditional villages,

totaling 8,155, have been certified and protected in China. Quantitative statistics on the area, number, and distribution of traditional sloped roofs are used to evaluate the protection and evolution of traditional village landscapes [3]. Common methods include interpreting unmanned aerial vehicle (UAV) remote sensing images manually to delineate the boundaries of different types of roofs. Increasingly, scholars are digitally preserving traditional architecture based on Geographic Information Systems (GIS), which includes digital management, modeling, remote sensing, and feature recognition of buildings.

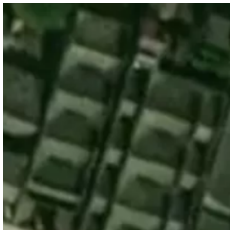

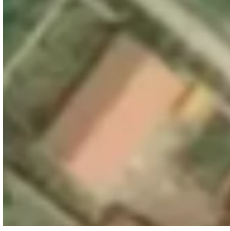

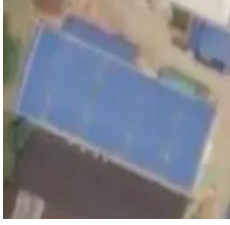
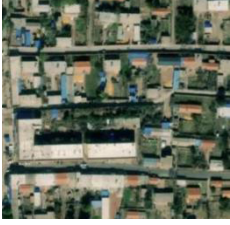
Technologies like deep learning have enhanced the efficiency of identifying and interpreting architectural features. He *et al.* [4] proposed Mask R-CNN (Mask Region-based Convolutional Neural Network), combining the Faster R-CNN object detection framework with a branch for predicting object masks, enabling efficient object detection and pixel-level instance segmentation. Li *et al.* [5] introduced Histogram Thresholding Mask R-CNN (HTMask R-CNN), a convolutional neural network based on histogram threshold masking regions, using dynamic grayscale thresholds inferred from the results of dual-object instance segmentation tasks with scarce training data, to extract rural building roofs. Zhan *et al.* [6] developed an instance segmentation model based on Mask R-CNN, feature pyramid networks, and online hard example mining, enhancing the precision of building extraction. Wang *et al.* [7] combined the Path Aggregation Feature Pyramid Network and the Atlas Spatial Pyramid Pool with ResNet-50 as the backbone network for the Mask R-CNN model, training on eight different color-scale sample sets specifically constructed for Beijing's traditional villages. However, due to the heterogeneity, complexity, and similarity of building roofs to other land features, there are still inaccuracies in target recognition. OpenAI [8] introduced the CLIP model, which connects text and images through large-scale contrastive learning, training both image and text representations simultaneously. This model understands a wide range of images and texts, exhibiting strong cross-modal capabilities suitable for various visual tasks.

This study integrates high spatial resolution remote sensing imagery with multi-scale deep learning techniques to propose a method for identifying building rooftops in traditional Chinese villages using high-definition remote sensing images. Based on the characteristic features of traditional Chinese village landscapes, this method categorizes building rooftops into traditional, ancient-style, and modern architectural styles, thus creating a spectralized sample set that focuses on the spectral, geometric, and semantic granularity features of these rooftops. In this paper, the core feature extraction component of the CLIP model is embedded into the Mask R-CNN framework, resulting in a semantically enhanced and improved Mask R-CNN model. This innovative approach not only leverages the strengths of advanced deep learning and remote sensing technologies but also provides a nuanced and effective tool for the identification and categorization of architectural styles, contributing significantly to the field of cultural heritage preservation and architectural analysis.

2. Analysis of traditional village building rooftop features









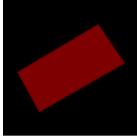

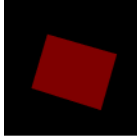









An in-depth analysis of the image targets of building rooftops in high-resolution remote sensing images of villages selected for the Chinese Traditional Village Directory is conducted. As shown in **Table 1**, the rooftop types of traditional villages are categorized into three categories: traditional architecture, ancient-style architecture, and modern architecture, based on the spectral, geometric, and semantic scale features of the target images.

Table 1. Types of roofs in traditional village buildings

Type	Image (partial)	Image (whole)	Visual interpretation features
Traditional architecture			Traditional architectural roofs are mostly composed of gray tiles with a hard and sloping shape, with a few pavilion-style roofs. They exhibit uniform spectral color characteristics, appearing in light gray and dark gray shades. In terms of geometric features, they predominantly exhibit a rectangular shape. The houses are evenly and densely distributed.
Ancient-style architecture			Ancient-style architecture clusters are predominantly composed of clay tile roofs and resin sloping roofs. The spectral color characteristics consist mainly of light gray, dark gray, orange, and purple, with occasional instances of blue and red. The geometric shape features are consistent, primarily rectangular, and the distribution of houses is dispersed.
Modern architecture			In modern architectural clusters, the roofs predominantly consist of colored steel sloping roofs, flat concrete roofs, and cement flat roofs. The spectral color characteristics often include light gray, blue, and orange. The geometric shapes are primarily rectangular and L-shaped, with a few instances of combination shapes. The houses are distributed in an orderly yet scattered manner.

Based on the interpretation of high-resolution remote sensing images of traditional village rooftops, an analysis of spectral heterogeneity features and geometric heterogeneity features of building rooftops is conducted. As shown in **Table 2**, spectral heterogeneity features can be categorized into six types: deep gray, light gray, red, orange, blue, and purple. Geometric heterogeneity features can be categorized into four types: rectangular, L-shaped, U-shaped, and combination.

Table 2. Examples of roof types in traditional village buildings

Types	Deep gray	Light gray	Red	Orange	Blue	Purple
Figure						
Mask						
Types	Rectangle	L-shape	U-shape	Combination		
Figure						
Mask						

3. Methods for identifying traditional village building rooftops

3.1. Overview of the improved Mask R-CNN model

In response to the need for traditional village building roof recognition in China, by leveraging the strong cross-modal learning capability of the CLIP^[8] model, this study enhances the feature extraction and fusion of the Mask R-CNN model^[4], complemented by data augmentation and transfer learning techniques. As illustrated in **Figure 1**, the improved Mask R-CNN network model primarily utilizes the Mask R-CNN instance segmentation network model as its underlying infrastructure. Based on the multi-scale representation capabilities of the Feature Pyramid Network (FPN)^[9] and combined with the multimodal interpretability of the CLIP model, it refines the feature extraction mechanism. This creates a semantically enhanced deep convolutional neural network aimed at improving the precision in identifying building rooftops. As illustrated in **Figure 1**, the overall architecture of the improved Mask R-CNN network model is divided into three parts. The first part is the enhanced shared convolutional backbone network, which integrates the visual encoder and the text encoder from CLIP, enabling deep feature extraction and semantic understanding. The second part consists of the Region Proposal Network (RPN), which, based on the feature maps extracted from the previous part, inputs into the RPN to generate region candidates (Proposals). It then refines the candidates for the region of interest (RoI). The generated target candidate boxes are input into the RoIAlign layer, which ensures better alignment of the features extracted from each RoI with the original image's RoI area and converts their dimensions into a fixed value for the subsequent fully connected layer. The third part comprises three parallel branches for classification, bounding box regression, and mask generation. The model employs an FPN network for feature fusion, enhancing its ability to recognize multi-scale targets. The optimizer is configured with AdamW, along with specific learning rate and weight decay strategies, to facilitate efficient network training and optimization.

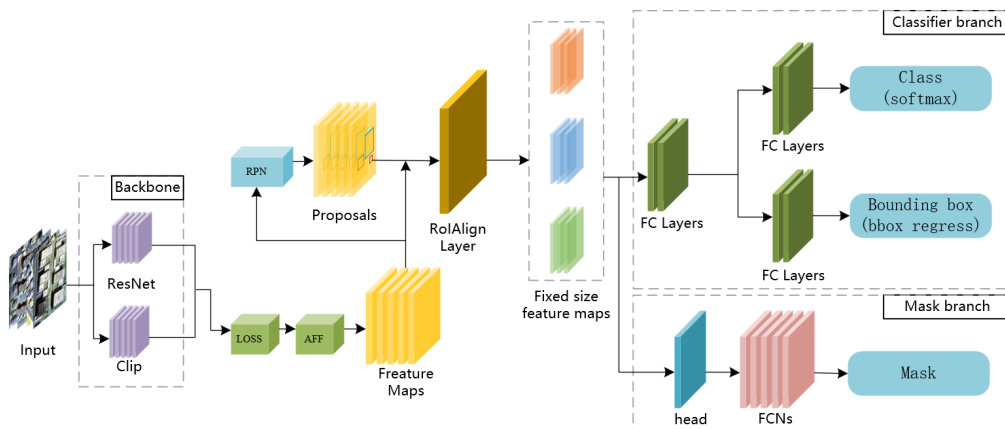


Figure 1. General architecture of improved Mask R-CNN network model

3.2. Experimental data

In this study, we have selected 553 traditional Chinese villages distributed across 30 provinces and municipalities throughout China as our experimental areas. These villages were chosen from the village listings in the Chinese Traditional Village Directory published by various government departments up to the year 2023.

In the experiment, high spatial resolution remote sensing imagery data with a spatial resolution of 0.54 meters was utilized. To convert this imagery data into a format suitable for deep neural networks, it underwent further data preprocessing. Considering that village buildings are small in volume and often scattered, the images were cropped into segments of 512×512 pixels. This segmentation was done to select images containing buildings with representative features as the basic units for model recognition training. To preserve

the detail of image features, the remote sensing images were subjected to 2% linear stretching, converting the image depth from 16-bit to 8-bit, thereby enhancing the visual contrast of the images. Additionally, Gaussian filtering was employed to enhance the multi-scale feature representation of the images.

This paper combines the scale granularity characteristics of traditional Chinese village building roofs to collect sample data that matches roof target features, thereby constructing a spectralized sample set adapted to these features. Post-preprocessing, high-resolution remote sensing images were used for sample annotation with the open-source tool LabelMe. This involved collecting category, feature, and location information of building roofs along target contours and extracting samples of land reclamation features. By parsing the sample JSON files, attributes and annotation information of roof targets were generated, constructing a COCO (Common Objects in Context)^[10] sample dataset for training the improved Mask R-CNN model.

3.3. Methods for identifying roofs of buildings in traditional villages

The sample library set of building roof target samples from high-resolution remote sensing imagery is randomly divided into training dataset, validation dataset, and test dataset in the ratio of 4:1:2, as shown in **Table 3**.

Table 3. Dataset statistics

	Training	Test	Validation
Sample size	3056	750	1500

Firstly, this study trained the improved Mask R-CNN model and obtained the initial parameters based on the training dataset, and measured the accuracy over the pass validation dataset. Secondly, a series of parameterized recognition models were constructed by assigning different values to the model parameters. Based on the results of the accuracy assessment of each parameterized model on the validation dataset, the depth of the network and the structure of the network backbone feature extraction were adjusted and optimized, and the optimal network structure and corresponding hyperparameters were selected. The network structure and parameters used in this paper are detailed in **Table 4**, and the specific values of the relevant hyperparameters are shown in **Table 5**. Finally, experiments on the recognition of roof targets on buildings in traditional Chinese villages are carried out on the test dataset to validate the effectiveness and accuracy of the proposed method.

Table 4. Convolutional neural network structure

Layer name	Conv1	Conv2_x	Conv3_x	Conv4_x	Conv5_x
50-layer					
Output-size	112 112	56 56	2828	1414	77

Table 5. Hyperparameter table

Parametric	Parameter type	Parameter value
GPU_COUNT	Number of GPUs used	1
IMAGES_PER_GPU	Amount of images trained on each GPU	1
BACKBONE	Backbone network	CLIPResNet+FPN
IN_CHANNELS	Number of input channels	
BACKBONE_STRIDES	Downsampling multiplier based on backbone network	

Table 5 (Continue)

Parametric	Parameter type	Parameter value
OPTIMIZER	Optimizer	AdamW
NUM_CLASSES	Number of categories in the training dataset (including background)	2
8*STRIDES	Anchor frame sizes corresponding to each feature map p2–p6	
RATIOS	3 scales corresponding to each size of anchor frame	
iou_threshold	Threshold for non-maximal suppression	0.7
max_per_img	Number of Rols per image input for later classification and masking networks	1000
DETECTION_MAX_INSTANCES	Maximum number of instances included in the final detection	100
Batch Size	Amount of images per training	2
Epochs	Number of iterations	12
LEARNING_RATE	Learning rate	0.0002
Optimizer_betas	Learning momentum	0.9
WEIGHT_DECAY	Weight decay	0.0001

3.4. Results

The improved Mask R-CNN model is used for building roof target recognition in the traditional Chinese village area, and the part of the image containing the target mask is obtained as shown in **Figure 2**.

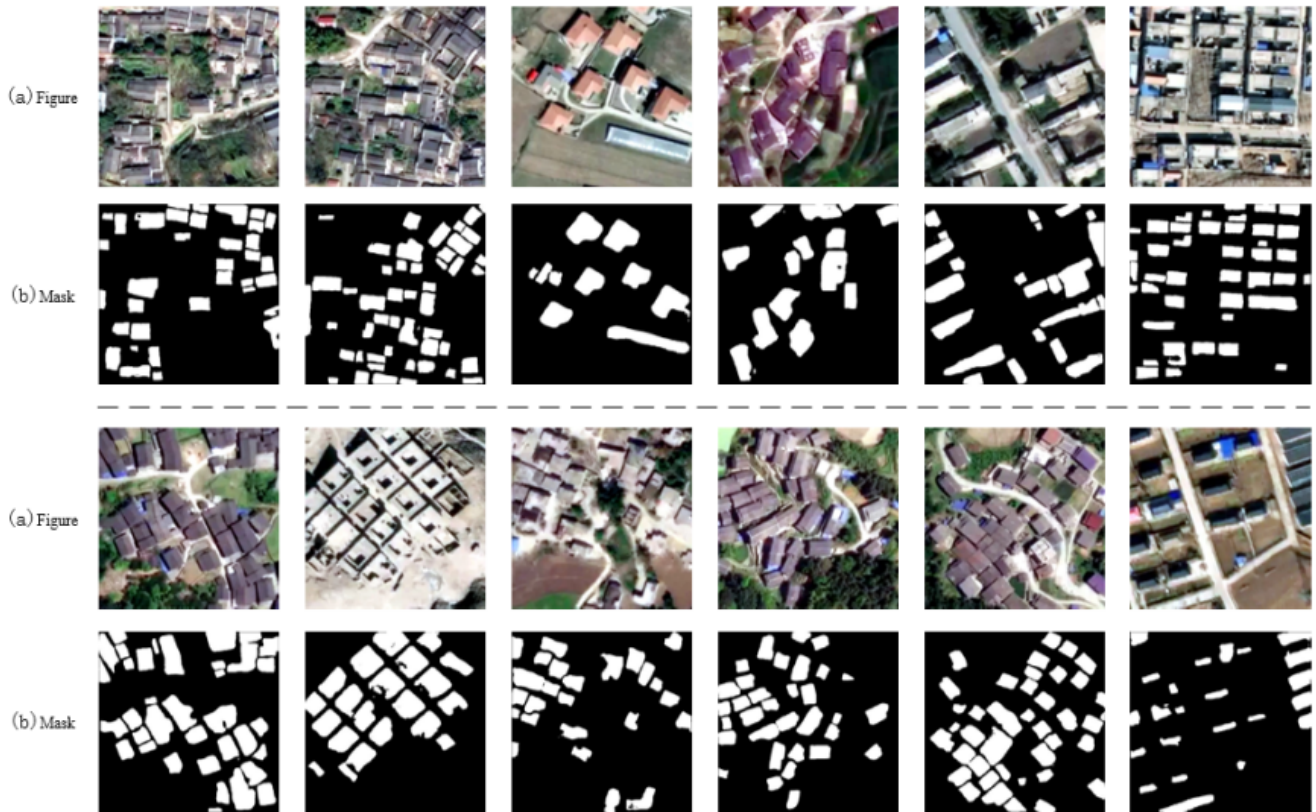


Figure 2. Experimental area target identification results

In accordance with the characteristics of the COCO ^[10] dataset, the performance of the improved Mask R-CNN model and the accuracy of its object detection results were evaluated using the metric of mean Average Precision (mAP). mAP is employed to objectively assess the effectiveness of object detection and instance segmentation methods. A prediction is considered a true positive only when the predicted mask's Intersection over Union (IoU) with the corresponding true annotation exceeds a threshold T and the predicted category matches the true category. This study utilizes for evaluation. The results of the accuracy assessment are presented in **Table 6**.

Table 6. Accuracy evaluation of test results

mAP						
Build	0.6830	0.8800	0.7520	0.6880	0.8820	0.7400

In the accuracy evaluation results, and reach 0.7520 and 0.7400, respectively, which proves that the improved Mask R-CNN model can effectively identify the roof targets of traditional buildings in China with high-resolution remote sensing, and it has good performance and model generalisation ability.

3.5. Analysis and discussion

The feasibility and effectiveness of the improved deep recognition model were validated through experiments involving the identification of building roof targets in high-resolution remote sensing images of traditional Chinese villages. Utilizing the enhanced Mask R-CNN network model, the precision of the bounding box for building roof recognition, denoted as , was 0.7520, and the accuracy of the mask, denoted as , was 0.7400. The precise prediction of reclaimed land demonstrates the model's capability in accurately recognizing and locating targets.

As shown in **Figure 3**, a comparison between the identified building roof masks and the manually quality-checked vectorized roof patches reveals occurrences of missed detections in traditional buildings. This issue arises due to the irregular layout, dense distribution, and indistinct roof edge contours of traditional village building clusters, leading to missed detections in the recognition results. In contrast, the roofs of buildings in antique-style and modern clusters, characterized by distinct spectral, geometric, and semantic features, are accurately identified with more complete edges. The vector data generated in subsequent steps, after undergoing a manual quality assurance process, meets the standards required for engineering applications.



Figure 3. Comparison of identification results in the experimental area

4. Conclusion

In this paper, we propose a method for target recognition of building roofs in traditional Chinese villages using an improved Mask R-CNN deep neural network recognition model. Additionally, leveraging high spatial resolution remote sensing imagery data and based on the analysis of spectral, geometric, and semantic characteristics of traditional, antique-style, and modern buildings, a sample set geared towards deep learning instance segmentation was constructed. Subsequently, building roof target recognition was achieved using the improved Mask R-CNN model, with a bounding box accuracy of 0.7520 and a mask accuracy of 0.7400, demonstrating the engineering feasibility and accuracy of this method for high-resolution remote sensing recognition of building roofs in traditional villages. This recognition method, by integrating the language-image joint understanding ability of CLIP with the precise instance segmentation advantage of Mask R-CNN, enhances the model's semantic understanding of image content, thereby improving segmentation accuracy and robustness in complex scenarios. It addresses issues of target misidentification and omissions caused by the complexity of building roof features, offering a more effective approach for the automated interpretation of building roofs in traditional Chinese villages.

However, there are areas in this study that require improvement. The roof target recognition results for traditional building clusters generated by the recognition model still exhibit missed detections. Therefore, future improvements should focus on combining self-attention mechanisms with multi-scale feature extraction methods in the feature extraction part of the deep learning neural network. Simultaneously, further enrichment and refinement of the sample set's spectral lineage are necessary to construct a more optimized spectralized building roof feature sample set, aiming to enhance the accuracy and efficiency of target recognition.

Disclosure statement

The authors declare no conflict of interest.

Author contributions

Conceptualization: Mengchen Wang

Investigation: Linshuhong Shen

Formal analysis: Mengchen Wang

Writing: Mengchen Wang

References

- [1] Nie XY, Zhang Y, Sun LS, et al., 2015, Types and Value Recognition of Traditional Villages: A Case Study of Traditional Villages in Shijiazhuang, Hebei. *Planner*, 2015(S2): 5.
- [2] Chen J, 2013, Reduction of 900,000 Natural Villages in 10 Years: Traditional Villages in China are "Dialing 120," *People's Daily*, June 5, 2013.
- [3] Liu XF, Tu ZZ, 2017, Quantitative Analysis Method for the Evolution of Traditional Village Scenery. *Journal of Huaqiao University: Natural Science Edition*, 38(6): 7.
- [4] He K, Gkioxari G, Dollár P, et al., 2017, Mask R-CNN. *Proceedings of the IEEE International Conference on Computer Vision*, 2961–2969.
- [5] Li Y, Xu W, Chen H, et al., 2021, A Novel Framework Based on Mask R-CNN and Histogram Thresholding for

Scalable Segmentation of New and Old Rural Buildings. *Remote Sensing*, 13(6): 1070.

- [6] Zhan Y, Liu W, Maruyama Y, 2022, Damaged Building Extraction Using Modified Mask R-CNN Model Using Post-Event Aerial Images of the 2016 Kumamoto Earthquake. *Remote Sensing*, 14(4): 1002.
- [7] Wang W, Shi Y, Zhang J, et al., 2023, Traditional Village Building Extraction Based on Improved Mask R-CNN: A Case Study of Beijing, China. *Remote Sensing*, 15(10): 2616.
- [8] Radford A, Kim JW, Hallacy C, et al., 2021, Learning Transferable Visual Models from Natural Language Supervision. *International Conference on Machine Learning*, PMLR 2021, 8748–8763.
- [9] Lin TY, Dollár P, Girshick R, et al., 2017, Feature Pyramid Networks for Object Detection. *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2117–2125.
- [10] Lin TY, Maire M, Belongie S, et al., 2014, Microsoft COCO: Common Objects in Context. *Computer Vision–ECCV 2014: 13th European Conference, Zurich, Switzerland, September 6–12, 2014, Proceedings, Part V 13*. Springer International Publishing, Cham, 740–755.

Publisher's note

Bio-Byword Scientific Publishing remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.