remote sensing

MDPI

Article

Traditional Village Building Extraction Based on Improved Mask R-CNN: A Case Study of Beijing, China

Wenke Wang [1], Yang Shi [1,2,*], Jie Zhang [1,3], Lujin Hu [2,4], Shuo Li [1], Ding He [1,2] and Fei Liu [2,4]

[1] School of Architecture and Urban Planning, Beijing University of Civil Engineering and Architecture, Beijing 100044, China
[2] Research Center for Urban Big Data Applications, Beijing University of Civil Engineering and Architecture, Beijing 100044, China
[3] School of Architecture, Tsinghua University, Beijing 100084, China
[4] School of Geomatics and Urban Spatial Informatics, Beijing University of Civil Engineering and Architecture, Beijing 100044, China
* Correspondence: shiyang@bucea.edu.cn

Abstract: As an essential material carrier of cultural heritage, the accurate identification and effective monitoring of buildings in traditional Chinese villages are of great significance to the sustainable development of villages. However, along with rapid urbanization in recent years, many towns have experienced problems such as private construction, hollowing out, and land abuse, destroying the traditional appearance of villages. This study combines deep learning technology and UAV remote sensing to propose a high-precision extraction method for conventional village architecture. Firstly, this study constructs the first sample database of traditional village architecture based on UAV remote sensing orthophotos of eight representative villages in Beijing, combined with fine classification; secondly, in the face of the diversity and complexity of the built environment in traditional villages, we use the Mask R-CNN instance segmentation model as the basis and Path Aggregate Feature Pyramid Network (PAFPN) and Atlas Space Pyramid Pool (ASPP) as the main strategies to enhance the backbone model for multi-scale feature extraction and fusion, using data increment and migration learning as auxiliary means to overcome the shortage of labeled data. The results showed that some categories could achieve more than 91% accuracy, with average precision, recall, F1-score, and Intersection over Union (IoU) values reaching 71.3% (+7.8%), 81.9% (+4.6%), 75.7% (+6.0%), and 69.4% (+8.5%), respectively. The application practice in Hexi village shows that the method has good generalization ability and robustness, and has good application prospects for future traditional village conservation.

Keywords: traditional village building; improved Mask R-CNN; UAV remote sensing images; PAFPN; ASPP; multi-scale feature extraction and fusion

Citation: Wang, W.; Shi, Y.; Zhang, J.; Hu, L.; Li, S.; He, D.; Liu, F. Traditional Village Building Extraction Based on Improved Mask R-CNN: A Case Study of Beijing, China. *Remote Sens.* **2023**, *15*, 2616. https://doi.org/10.3390/rs15102616

Academic Editors: Dimitrios Skarlatos and Andreas Georgopoulos

Received: 7 April 2023
Revised: 7 May 2023
Accepted: 15 May 2023
Published: 18 May 2023

1. Introduction

As an essential material carrier of cultural heritage, traditional villages provide significant resources for us to explore historical stories and folk customs of different times and regions [1,2]. They are rural settlements formed spontaneously during the long-term interaction between people and nature [3], and they are also the non-renewable cultural resources of Chinese civilization [4]. However, along with rapid urbanization, the contradiction between people's pursuit of a better living environment and preserving traditional villages has become increasingly prominent. Many traditional villages have experienced severe problems such as hollowing out, land abuse, and demolition of the old to build new ones, leading to the decay of village buildings and a large amount of cultural heritage facing extinction [5,6]. Strengthening the protection and development of traditional villages is essential to China's "rural revitalization strategy" [7,8]. However, due to the large number and wide distribution of traditional Chinese villages and the highly complex

built environment, it is challenging to achieve rapid and high-precision acquisition of building information in traditional villages by relying on existing methods. Therefore, it is urgent to explore a technique that can automatically and accurately extract the buildings of traditional Chinese villages.

In the past, scholars mainly evaluated traditional village building through manual surveys by experts, such as combining historical image maps [9], field observation and recording [10,11], and in-depth interviews with residents [12,13]. Although much architectural information has been obtained, disadvantages include labor and material resource consumption, reliance on subjective judgment, and susceptibility to environmental interference. In recent years, deep learning models have made breakthroughs in computer vision with their robust data fitting and feature representation capabilities. They are widely used in remote sensing image information extraction [14–20], bringing a new research perspective for efficiently evaluating traditional village buildings. Compared with traditional methods such as the feature detection method [21–24], region segmentation method [25–30], and auxiliary information combination method [31–36], remote sensing information extraction based on deep learning can spatially model adjacent pixels and obtain higher quality results in processing numerous vision tasks. Xiong et al. [14] proposed a detection model for traditional Chinese houses—Hakka Weirong Houses (HWHs)—based on ResNet50 and YOLO v2 and combined with multi-metric evaluation proved that the model has high accuracy and excellent performance; Liu et al. [15] drew the advantages of U-Net and ResNet and proposed the SSNet deep residual learning sequence semantic segmentation model, and proved the superiority of the model through experiments. Compared with the semantic and target detection models, the instance segmentation model can simultaneously achieve detection, positioning, and segmentation and extract richer feature information. Some scholars introduced the Mask R-CNN instance segmentation model [37], such as Chen et al. [16], who combined the Mask R-CNN model with a transfer learning technique to achieve high precision building area estimation; Li, Wang et al. [17,18] combined the Mask R-CNN model with data augmentation technique to achieve new and old buildings in the Chinese countryside, Chinese rural building roofs with high accuracy; in addition, Zhan et al. [19] earned high accuracy extraction of residential buildings based on the Mask R-CNN model and by improving the feature pyramid network. Tejeswari et al., achieved high-precision urban building extraction based on the Mask R-CNN model combined with a large dataset automatically generated by Google API, demonstrating that sufficient data samples are the key point limiting the accuracy of contemporary deep learning models [20].

In summary, some studies have shown the excellent performance of instance segmentation models in some specific object extractions. However, these objects often have clear boundaries, similar morphology and size, fewer classes, and apparent background features on remote sensing images. Traditional villages have a long history, and various factors such as geographic location, economic and cultural factors, and human-made factors have led to a diverse and complex overall built environment. As an important carrier for shaping the image of Chinese architecture and characterizing the cultural connotation, as shown in Figure 1, roofs include small green tile roofs inherited from ancient times, red tile roofs, small green tile roofs with cement roofs after the founding of the country, and multi-colored plastic steel roofs and resin roofs of the beginning of this century, and many other types [38]. At the same time, multi-functional needs and the size of the house lots lead to high heterogeneity of form and size within the group. The above multi-scale, multi-type, and multi-temporal roofs enhance the complexity of the built environment and pose a significant challenge to the existing building extraction models.

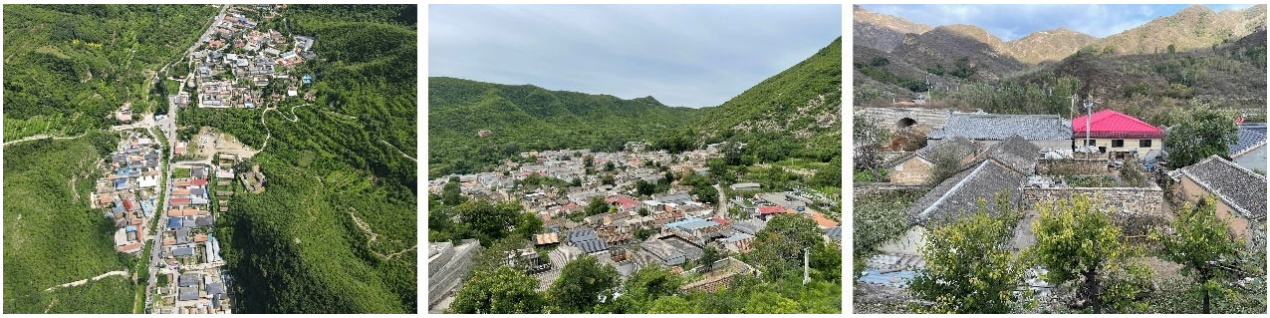


Figure 1. Aerial view of traditional village buildings.

In recent studies, multi-scale feature extraction and fusion have effectively coped with complex environments [39–51]. Chen et al., incorporated "dilated convolution" into the DeepLabv2 [39] model to increase the perceptual field by inserting "holes" in the filter to achieve accurate recognition of targets in complex environments; The method was later further improved by several researchers in DeepLabV3 [40], ResUNet++ [41], and PSPNet [42], which proposed the Atrous Spatial Pyramid Pooling (ASPP) module to further obtain multi-scale information and enhance the model's recognition and detection of objects in the context of complex environments. Additionally, Li et al., demonstrated the effectiveness of the ASPP module for the building extraction task [43]. Furthermore, multi-scale feature fusion based on Feature Pyramid Networks (FPN) [44] is widely used to improve recognition performance in complex environments. This model is based on a top-down architecture to construct high-level semantic feature maps at all scales and predict feature maps at different scales. However, Liu et al., found that the unidirectional information propagation mechanism in FPN led to the underutilization of the underlying features and proposed the Path Aggregation Feature Pyramid Network (PAFPN) [45], which enhances the feature pyramid by shortening the information path using the precise localization signal at the bottom and establishes a bottom-up path enhancement that strengthens the contextual relationship between the multilayer features. The superior performance of the model has been proven in numerous other detection tasks, such as ground aircraft [46], tunnel surface defects [47], traffic obstacles [48], and other complex environments for target extraction. Still, its potential for building extraction tasks has yet to be effectively explored. The bottom features have been applied in some early models [49–51], but it is needed to sufficiently extract and fuse the multi-scale elements to enhance the whole feature hierarchy.

With the above challenges, this paper proposes a high-precision automatic extraction method for traditional village buildings by combining UAV remote sensing orthophoto and improved Mask R-CNN. First, we construct the first traditional Chinese village building roof orthophoto dataset, which consists of fine-grained labels of building roofs in eight historical villages in Beijing. Second, we use the Mask R-CNN instance segmentation model as the basis; Path Aggregation Feature Pyramid Network (PAFPN) and Atrous Spatial Pyramid Pooling (ASPP) as the main strategies to enhance the backbone model for multi-scale feature extraction and fusion; and data augmentation and transfer learning as auxiliary means to overcome the shortage of labeled data. After comparison experiments and ablation experiments combined with common evaluation indexes of computer vision, the superiority of the performance of this model is verified. Finally, Hexi village is selected as the validation object. The extraction results are visualized in GIS, and the complete workflow is to prove our method's excellent robustness and generalization ability. This paper is the first time a fast extraction method has been developed in China for traditional village buildings.

This paper is organized as follows: Section 2 describes the study area; Section 3 describes in detail the traditional village building extraction method proposed in this paper, which includes data acquisition and preprocessing, and improved Mask R-CNN model architecture; Section 4 conducts detailed experiments to verify the performance of the

model; Section 5 discusses various aspects of the research results, and proposes limitations and improvement directions; finally, the paper concludes.

2. Study Area

As the capital of five major dynasties (Liao, Jin, Yuan, Ming, and Qing) and China's political center, Beijing has many traditional villages of unparalleled historical, cultural, and social value. With a long history, most of these villages originated in the Ming Dynasty (1948 to 1990 A.D.). They made remarkable contributions to the defense of the ancient capital as the early homes of the Great Wall defenders and are essential for transmitting ancient military culture.

However, along with rapid urbanization, the contradiction between people's pursuit of a better living environment and the preservation of traditional villages has become increasingly prominent, with many traditional villages experiencing severe problems such as hollowing out, land abuse, demolition of the old and construction of the new, and the destruction of a large number of traditional buildings, directly leading to the disappearance of cultural heritage. In recent years, Beijing has successively issued policy documents such as the Beijing Urban Master Plan (2016–2035) and the Technical Guidelines for Repairing Traditional Villages in Beijing, emphasizing the importance of protecting the architectural style of traditional villages.

In this study, nine representative historical villages in Beijing (shown in Figure 2) were selected as the source of data acquisition (including five traditional villages and four non-traditional villages), among which eight towns, namely Baimaguan, Fengjiayu, Jijiaying, Shangyu, Xiaokou, Xiaying, Xituogu, and Yaoqiaoyu, were used as training sets, and Hexi village was used for application practice. These villages have many traditional buildings with small gray tile roofs and modern facilities with various roof forms, which are typical of the classic village style in North China.

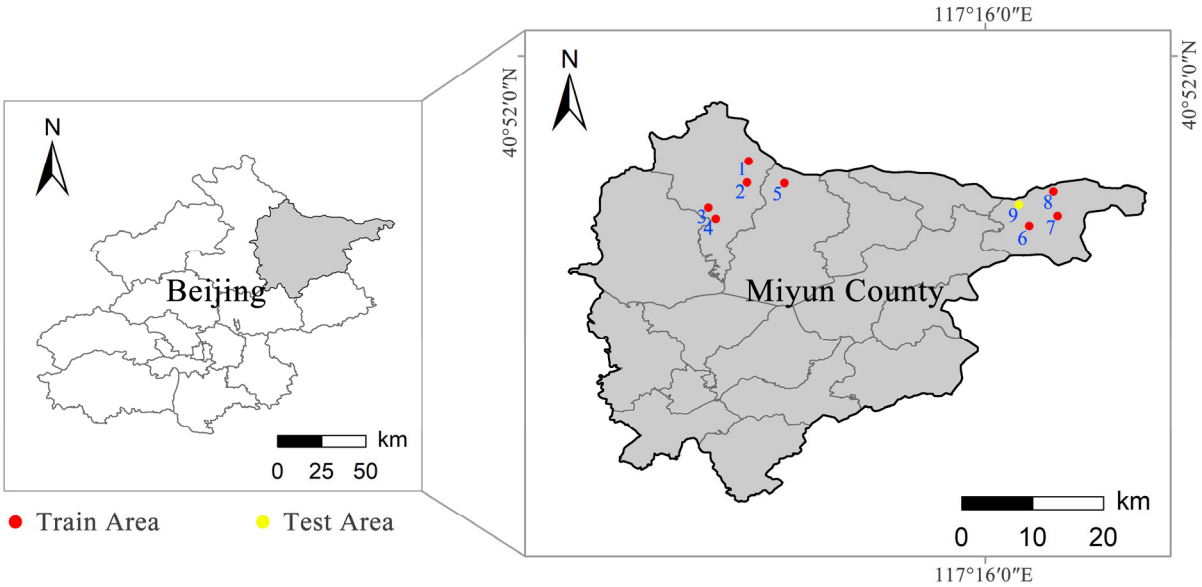


Figure 2. Study area and village location. 1: Baimaguan village; 2: Xiaying village; 3: Shangyu village; 4: Fengjiayu village; 5: Xituogu village; 6: Jijiaying Village; 7: Yaoqiaoyu Village; 8: Xiaokou Village; 9: Hexi village.

3. Materials and Methods

This study aims to propose a method for automatically extracting traditional village buildings in northern China by combining remote sensing images to automatically identify existing and potential traditional villages and promote the continuation of traditional village cultural heritage. Figure 3 illustrates the whole workflow and main steps, including four steps: data acquisition, data preprocessing, experimental analysis, and application

practice. Data acquisition includes low-altitude remote sensing data acquisition by UAV, and super-resolution reconstruction of UAV orthophotos to obtain the sample database. Data preprocessing includes steps such as roof classification, data annotation, expert cross-checking, format conversion, data enhancement, etc. The experimental analysis includes comparison and ablation experiments of various existing advanced models. The application practice consists of three parts: data acquisition and preprocessing, building extraction, and visualization.

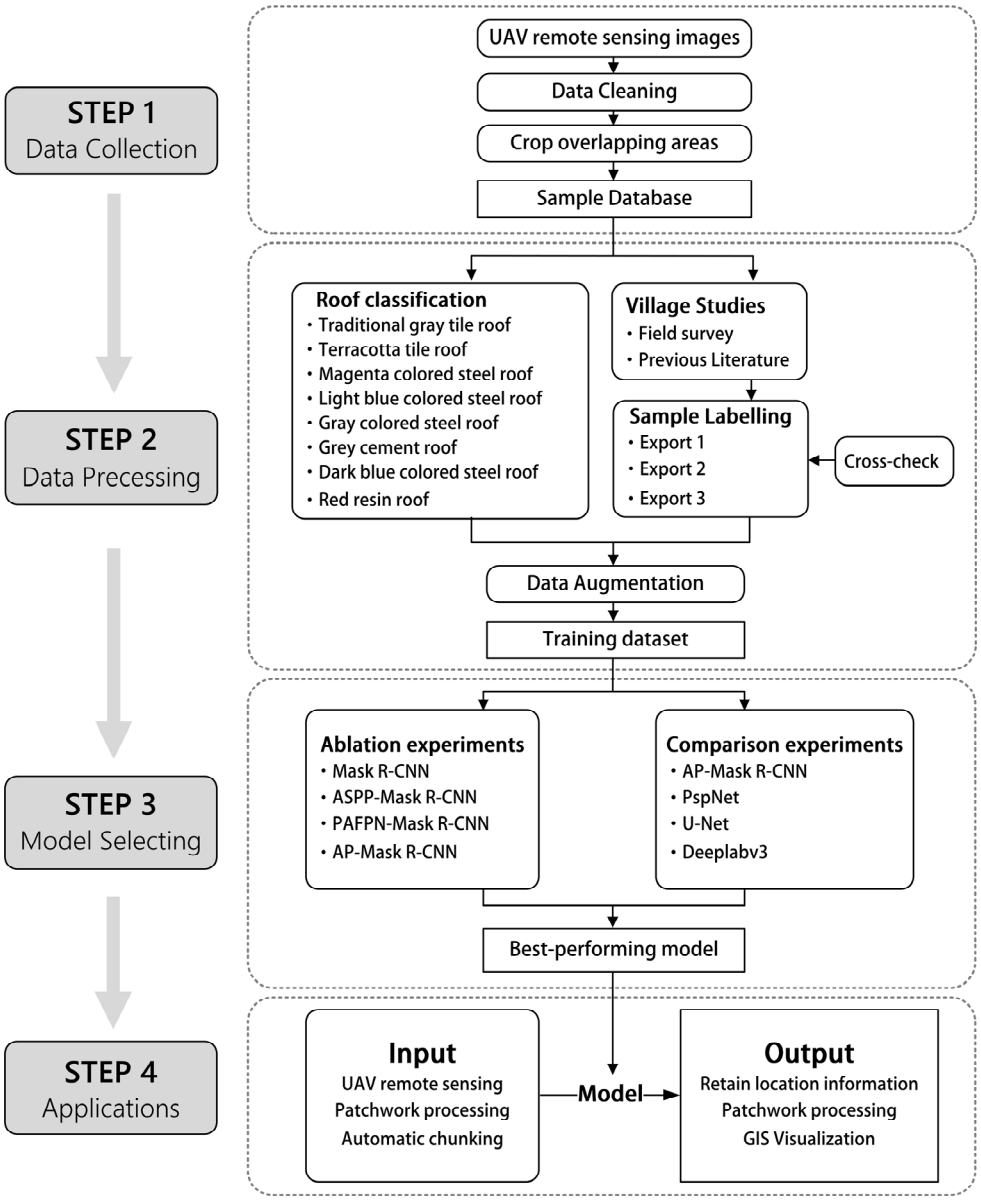


Figure 3. Overall technical route.

3.1. Data Acquisition and Preprocessing

3.1.1. Data Acquisition

This study uses an uncrewed aerial vehicle (DJI M300; specific parameters are shown in Table 1) to collect low-altitude remote sensing data. In recent years, UAV data acquisition has been rapidly developed with the advantages of portability, low cost, and high operability. It is widely used in commercial and scientific fields and heritage conservation. In addition, numerous studies based on UAV low-altitude remote sensing images have shown that extracting multi-scale targets in complex ground environments has more advantages than satellites [23,52]. During the acquisition process, the effects of sun height, wind speed,

and weather conditions on the quality of the acquired data were taken into account, and we selected a clear and breezy day (between 9:00 a.m. and 3:00 p.m.) for the acquisition. The specific flight plan parameters are shown in Table 2. The acquired image data were stitched together using Context Capture software, and the generated super-resolution orthophotos are shown in Figure 4.

Table 1. Description of data acquisition equipment.

Technical Indicators		Parameters
Aircraft	Brand	DJI
	Model	M300RTK
	Maximum flight time	55 min
	Protection rating	IP45
	RTK position accuracy	1 cm + 1 ppm (horizontal); 1.5 cm + 1 ppm (vertical)
Camera	Camera model	DEFAULTQ
	Aperture	f/0
	Exposure time	1/500 s
	ISO speed	ISO-200
	Focal length	36 mm

Table 2. Flight plan parameters.

Technical Indicators	Parameters
Aviation high	100 m
Return altitude	200 m
Flight mode	RTK
Speed	8 m/s
Overlap	80%

Figure 4. UAV orthophoto of the village building training set.

3.1.2. Data Preprocessing

The performance of the deep learning model depends mainly on the quality of the data provided to the model. Due to the lack of traditional village road building roof datasets in China, this study divided them into eight categories based on roof form, material, and color

through field survey and the divergent characteristics of traditional Chinese buildings [52], including one type of traditional roof (gray small green tile roofs) and seven categories of non-traditional roofs (terra cotta tile roofs, magenta-colored steel roofs, light blue-colored steel roofs, gray-colored steel roofs, gray cement roofs, red resin roof, dark blue-colored steel roof), as shown in Figure 5. After that, the raw data acquired by the UAV are preprocessed by manual cleaning, and the cropped images are labeled with the help of LabelMe labeling software and converted to Coco dataset format using Python programming. Finally, new synthetic data are obtained by combining data augmentation techniques to expand the dataset. It has been demonstrated that combining different image enhancement techniques (e.g., flip, rotation, and grayscale) can enhance the detection performance of deep learning methods [53,54]. In this paper, we use five data augmentation methods, namely color gamut transform, image expansion, random cropping, random mirroring, and scale dithering, for data augmentation, and the complete data preprocessing flow is shown in Figure 6.

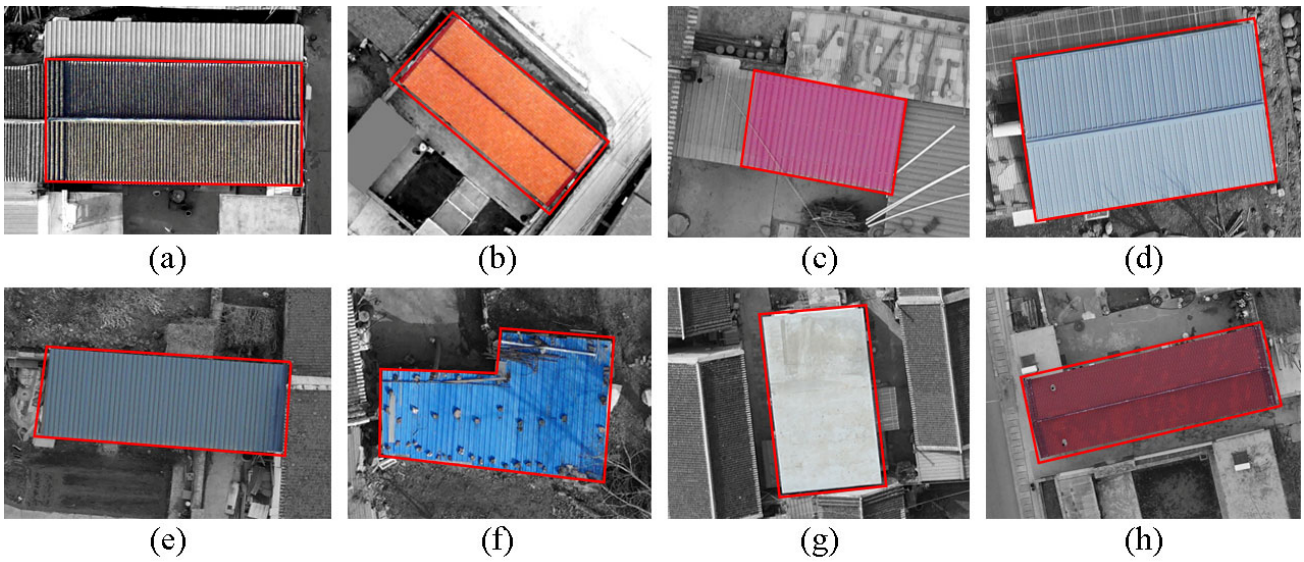


Figure 5. Traditional village building roof classification. (**a**) TGTR: traditional gray tile roof; (**b**) TTR: terracotta tile roof; (**c**) MCSR: magenta-colored steel roof; (**d**) LBCSR: light blue-colored steel roof; (**e**) GCSR: gray-colored steel roof; (**f**) DBCSR: dark blue-colored steel roof; (**g**) GCR: gray cement roof; (**h**) RRR: red resin roof.

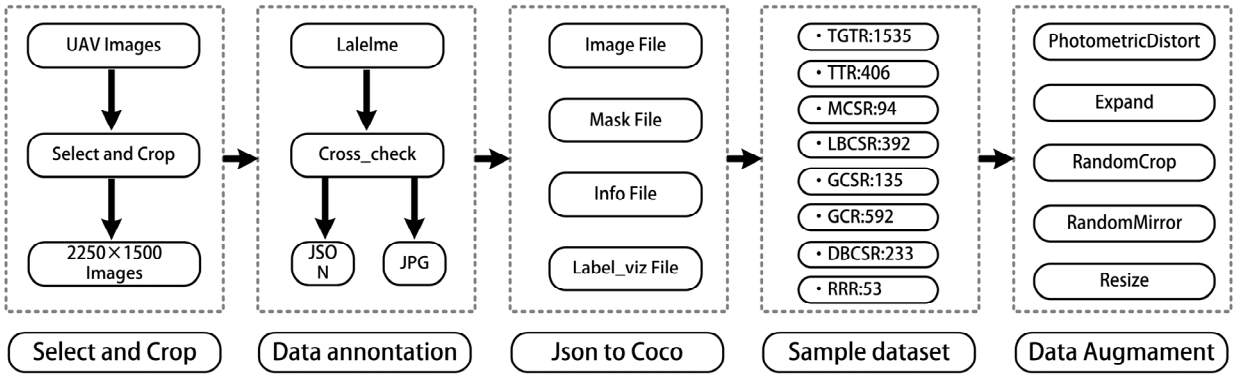


Figure 6. Data preprocessing process.

3.2. Improved Mask R-CNN Model Architecture

For the complex architectural environment of multi-scale, multi-category, and multi-temporal in traditional villages, this study takes the Mask R-CNN instance segmentation model [37] as the basic framework, improves the multi-scale feature extraction and fusion capability of the model by incorporating Path Aggregation Feature Pyramid Network

(PAFPN) and Atrous Spatial Pyramid Pooling (ASPP) with a slight increase in computation, and constructs a high-precision traditional village architectural extraction model AP_Mask R-CNN. The core architecture of the AP_Mask R-CNN model is shown in Figure 7, which consists of three parts: a backbone model comprised of the Resnet50 model and path aggregation feature pyramid network (PAFPN) with atrous spatial pyramid pooling (ASPP), a region proposal network (RPN), and a fully convolutional network (FCN) model including mask branches. In addition, this model uses a transfer learning strategy during training to transfer knowledge and share knowledge structures from a large amount of labeled data in the Imagenet dataset to facilitate learning tasks in this domain.

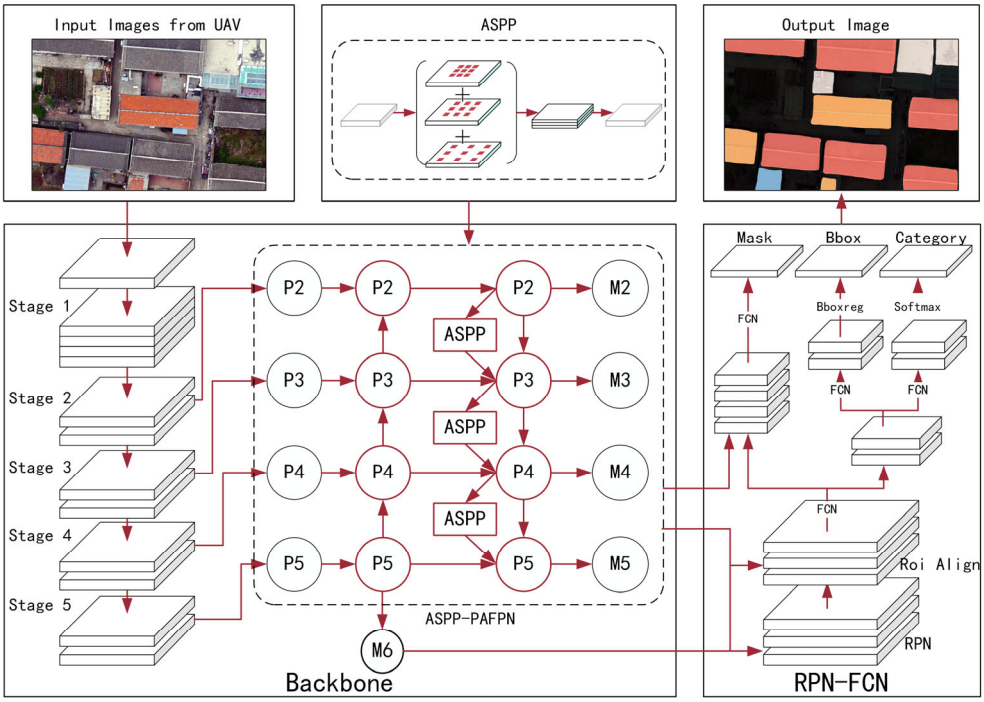


Figure 7. AP_Mask R-CNN model architecture.

3.2.1. ResNet-50

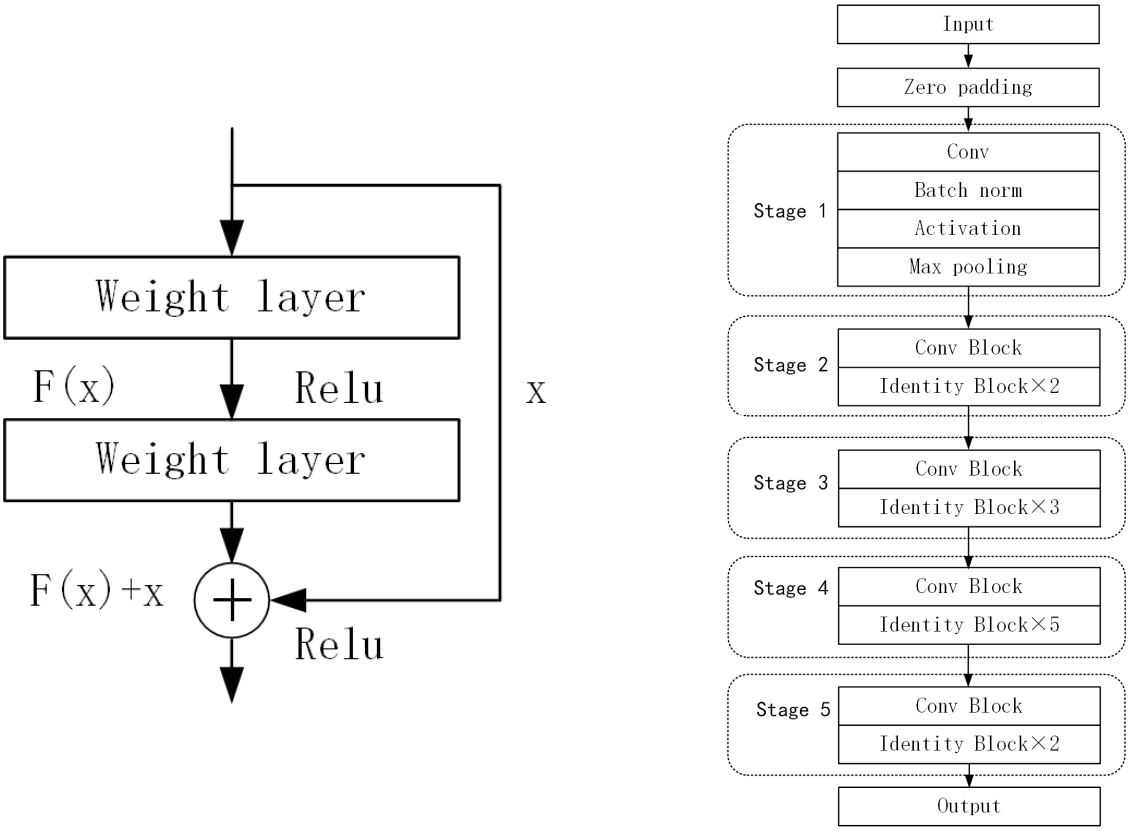
ResNet, proposed by He et al. [55] in 2016, is mainly used to extract features from input images. The residual mechanism (Figure 8a) significantly improves the convergence and accuracy of deep learning and has thus achieved wide application in significant models. One of the main principles of the residual mechanism is as follows:

$$y = F(x, \{W_i\}) + x \quad (1)$$

In Equation (1), x and y are the input and output vectors of the layer, and F(x, {W_i}) is the residual mapping to be learned. If the dimensions of x and F are not equal, a linear projection W_s can be used to match the measurements, as shown in Equation (2):

$$y = F(x, \{W_i\}) + W_s x \quad (2)$$

The experimental results of Kaiming He show that the residual model can effectively alleviate the gradient disappearance problem caused by more profound training of the model without increasing the model parameters, thus improving the detection performance of the model. In this study, for the data characteristics and underlying structure of the traditional village building roof dataset and to balance the accuracy and training efficiency of the model, we chose the 50-layer Resnet (Figure 8b) as the feature extraction model.



(**a**) (**b**)

Figure 8. Model architecture diagram. (**a**) Residual structure; (**b**) ResNet-50.

3.2.2. ASPP-PAFPN

The purpose of constructing ASPP-PAFPN is to develop a feature fusion mechanism with advanced semantics using the pyramidal hierarchy of the convolutional neural model to enhance the extraction and fusion of features of multi-scale buildings. The initial FPN is a single-path feature fusion, which has the problems of a single sensory field and insufficient feature utilization [45]. Its model structure is shown in Figure 9a. Given the traditional village's complex built environment and the FPN model's limitations, this study proposes an ASPP-PAFPN convergent path feature pyramid extraction model based on the original FPN by extending the fusion path and updating the convolution module in two ways.

First, multi-scale context fusion can enhance the recognition performance of the model for buildings in complex environments. We propose the path aggregation feature pyramid network (PAFPN) based on the original single path feature fusion, which fully uses the information relationship between different scale feature maps. The original FPN enhances the model's performance by adding top-down feature fusion paths to the backbone model to promote the mutual fusion of high-resolution information from low-level and high-level features. Borrowing from the idea of PANet, a reverse feature fusion mechanism is superimposed on the FPN to enhance the feature information of the image so that the whole model can obtain better detection results, and its model structure is shown in Figure 9b.

Secondly, this study uses a combination of ASPP module and global average pooling to replace the original 3×3 convolutional layers (stride = 2) to enhance feature extraction in the complex environment of the model by improving the model perception field.

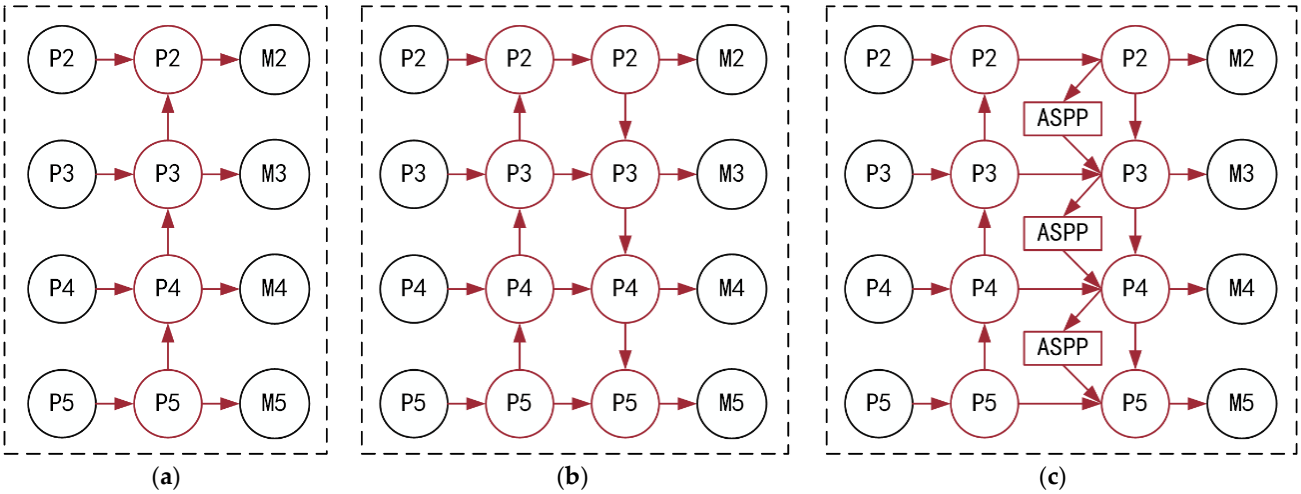


Figure 9. PAFPN improvement methods. (**a**) FPN; (**b**) PAFPN; (**c**) ASPP-PAFPN.

ASPP (Atrous Spatial Pyramid Pooling) is an improved framework of SPP (Spatial Pyramid Pooling) [56], whose most significant advantage is the use of dilated convolution instead of general convolution. With this improvement, computational resource consumption can be reduced while obtaining sufficient sensory fields. Moreover, to keep spatial invariance and continuity context information for larger objects, we combine global average pooling and ASPP modules to replace the original general convolution (stride = 2). The model structure of the fused ASPP is shown in Figure 9c. Pi' is first generated by the ASPP module after expanding the feeling field. Then, the spatial dimension of the feature map is reduced by half by the global average pooling, and then a parallel strategy is performed to fuse Pi' and Pi+1' to form a new feature graph.

In this paper, the original features are input to each of the three dilated convolutions, as shown in Figure 10. Each dilated convolution unit consists of one dilated convolution module. Among them, the dilated rates of the three dilated convolutions are 2, 4, and 8, and the convolution kernel size is 3 × 3. Then, the output feature map of ASPP is obtained from the output of the dilated convolution unit. Finally, the feature maps are merged using a parallel strategy, and the combined results are fed into the 1 × 1 convolution module to obtain the high-level features, as shown in Equation (3).

$$o[i] = \sum_{l=1}^{L} x[i + r \cdot l] f[l] \tag{3}$$

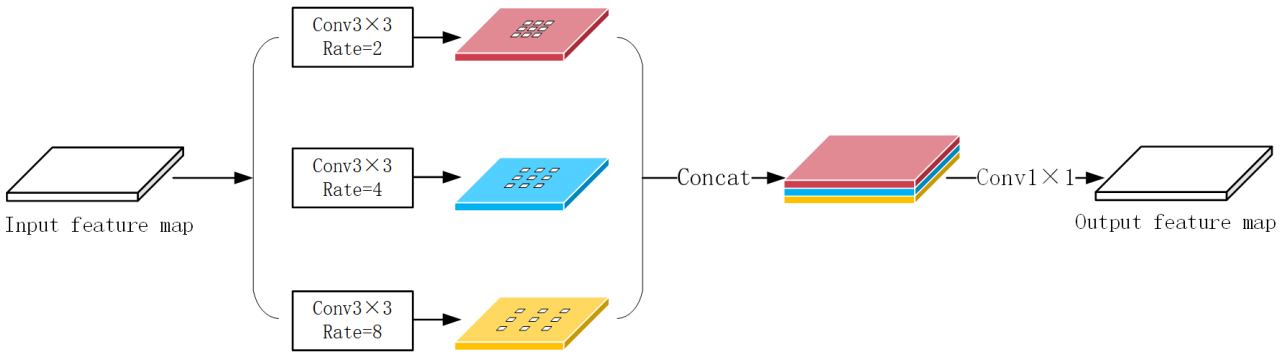


Figure 10. ASPP modular architecture.

Equation (3) where x[i] is the input signal, f[l] is a filter of length l, and r is the dilated rate. Note that when r = 1 is a standard convolution. In two-dimensional convolution operations, dilated convolution can be considered by inserting "holes" in the convolution

filter (inserting zeros between two adjacent pixels). By defining the dilated rate r, we can modify the size of the perceptual field without changing the filter size.

The introduction of this structure enables multi-scale sampling of the feature map using the dilated convolution with different sampling rates, extending the perceptual capability of the convolution kernel, avoiding the loss of image detail features, and enhancing the adaptability to complex targets.

3.2.3. RPN-FCN

The RPN-FCN model is mainly responsible for generating the multi-scale feature maps generated by the backbone model and selecting the best-suggested frames for positioning, classification, and segmentation. First, the feature maps extracted from the backbone model are transported to the RPN, which generates the ROI containing the building roof at each point on the feature map. To cope with the multiple scales and orientations of the building, we designed five scales of 32 × 32, 64 × 64, 128 × 128, 256 × 256, and 512 × 512 and three aspect ratios of 1:2, 1:1, and 2:1 for ROI. Thus, 15 ROIs are generated for each point on the feature map. After that, the ROIs are subjected to classification and regression operations to output the class and boundary coordinates of the region of interest. The boundary coordinates of the ROIs are used to adjust the bounding box to fit the area where the building is located. The generated ROI and the corresponding feature maps are input into ROIAlign. RoIAlign is used to adjust the size of the anchor box, and the extracted features are aligned with the input with the help of a bilinear interpolation algorithm, which improves the boundary box positioning and pixel segmentation precision. Then, the ROI classifier and the bounding box regressor are used to generate a specific class of ROI (i.e., a detailed classification of roofs in traditional villages) and its associated bounding box. Finally, the fully convolutional layer generates ROI segmentation masks using the positive regions selected by the ROI classifier.

3.2.4. Loss Calculation

The loss function represents the difference between the predicted result and the labeled actual value and is the basis for optimizing the model parameters. First, this model inputs the feature maps output by ROIAlign to the fully connected and convolutional layers, respectively. The fully connected layer is used for classification and bounding box regression, and the fully convolutional layer is used for building instance segmentation. Classification is done by passing the output of the fully linked layer through the softmax layer, and instance segmentation is achieved by convolution and deconvolution. In this study, the model is trained using a three-part joint loss function of classification, bounding box regression, and masked branching, and the loss is calculated as shown in Equations (4)–(7):

$$L = L_{cls} + L_{bbox} + L_{mask} \tag{4}$$

$$L_{cls} = \sum_{i} -\log[p_i^* p_i + (1 - p_i^*)(1 - p_i)] \tag{5}$$

$$L_{bbox} = \frac{1}{N_{reg}} \sum_{i} p_i^* R(t_i - t_i^*) \tag{6}$$

$$L_{mask} = -\frac{1}{m^2} \sum_{1 \le i,j \le m} \left[y_{ij}^* \log y_{ij} + \left(1 - y_{ij}^*\right) \log\left(1 - y_{ij}\right) \right] \tag{7}$$

where L_{cls} is the classification loss; L_{bbox} is the bounding box regression loss, L_{mask} is the mask loss; p_i and p_i^* are the prediction probability and ground truth of anchor point i, respectively; N_{reg} is the number of pixels in the feature map; t_i and t_i^* are the prediction coordinates and ground truth coordinates; and $R(\cdot)$ is the smoothing L1 function. For the improved Mask R-CNN, the mask branch has an output of size m^2 for each ROI. In Equation (7), y_{ij}^* is the ground truth of the coordinate point (i, j) in the $m \times m$ region, and y_{ij} is the predicted value.

3.2.5. Evaluation Metrics

This study evaluated model performance using precision, recall, F1-score metrics, and Intersection over Union (IoU) values. Precision is the sample size ratio correctly classified to all predicted sample sizes. Recall estimates the percentage of a correctly classified sample size among all labeled sample sizes. F1-score is the summed average of precision and recall, and the higher the F1-score value, the better the result. The Intersection over Union (IoU) value is the ratio of the intersection of the predicted building pixels to the actual building pixels. It is commonly used for evaluating semantic segmentation tasks. In this study, we separately evaluate the overall and single-category precision to analyze the effect of different models on detection accuracy. Equations (8)–(11) show the specific index calculation formula.

$$\text{precision} = \text{TP}/(\text{TP} + \text{FP}) \times 100\% \tag{8}$$

$$\text{recall} = \text{TP}/(\text{TP} + \text{FN}) \times 100\% \tag{9}$$

$$\text{F}_1 = 2 \times \text{precision} \cdot \text{recall}/(\text{precision} \times \text{recall}) \tag{10}$$

$$\text{IoU} = \text{TP}/(\text{TP} + \text{FN} + \text{FP}) \times 100\% \tag{11}$$

where TP is a true positive, which represents the accurately identified building roof; FP is a false positive, which is defined as the detected non-roof area, as the roof area in the image; while FN remains as a false negative, which shows the actual roof is not detected by the applied method.

4. Results

To verify the AP_Mask R-CNN model's precision and practicality, we designed three sets of experiments in this section: an ablation experiment, a comparison experiment, and application practice. Among them, the ablation experiment and comparison experiment are based on the evaluation indexes; the application practice is based on the UAV remote sensing images of traditional villages as the test objects. The qualitative analysis is conducted based on the extraction results. The software environment used for all experiments is shown in Table 3.

Table 3. Description of the working environment.

Working Environment		Versions
Software	Pycharm	2022.3.1
	Anaconda	3
	Computer System	Window10 64-bit
	Cuda	11.3
Hardware	Processor	I9-9900k
	GPU	NVIDIA 2080 super (8 g)
Frame	Tensorflow	2.2
	Python	3.7

4.1. Comparison Experiments

The comparison experiments are mainly used to investigate whether this model has advantages over the traditional village building extraction task. This section compares it with three advanced semantic segmentation models for quantitative comparison and comparative analysis of recognition results. Among the models used are DeepLabv3 [40], PspNet [42], and U-Net [57]. In these semantic segmentation models, pixel accuracy is used as a metric. The contrast models were all built on the open-source code of bubbliiiing (https://github.com/bubbliiiing, accessed on 25 October 2022) for U-Net, PspNet, and DeepLabv3.

The learning rate was initially set to 0.01 for all comparison models. The learning rate size was adjusted periodically during training using an SGD (stochastic gradient descent) optimization strategy combined with a cosine annealing strategy. In addition, due to the imbalance in the number of pixels between different categories in the dataset, all models use the focal loss function instead of the original cross-entropy loss function. Through the training of 150 epochs, the overall evaluation metrics of different models are shown in Table 4. First, AP_Mask R-CNN outperforms the other models in terms of overall metrics. Regarding category-specific metrics, the accuracy change curves are similar for all models, verifying the effect of the quality of different categories of data on their accuracy. Secondly, the F1-score and the IoU of AP_Mask R-CNN are higher than other models in recognizing terracotta tile roofs, light blue-colored steel roofs, gray-colored steel roofs, gray cement roofs, and dark blue-colored steel roof types, indicating that it has a more significant advantage in recognizing different roof types. Especially for dark blue-colored steel roofs, U-Net, DeeplabV3, and PSPNet models show very low recognition accuracy, which indicates that there are still limitations in using only semantic segmentation models to recognize multi-scale, multi-type, and multi-temporal roofs.

Table 4. Evaluation results of metrics for comparative experiments.

Model	Metric	Overall	TGTR	TTB	MCSR	LBCSR	GCSR	GCR	DBCSR	RRR
AP_Mask R-CNN	Precision	0.713	0.905	0.985	0.475	0.65	0.545	0.675	0.897	0.574
	Recall	0.819	0.918	0.996	0.639	0.809	0.582	0.791	0.991	0.77
	F1-Score	0.757	0.912	0.991	0.545	0.721	0.563	0.728	0.942	0.658
	IoU	0.694	0.873	0.956	0.462	0.633	0.528	0.65	0.89	0.561
PspNet	Precision	0.64	0.79	0.93	0.63	0.53	0.38	0.52	0.71	0.63
	Recall	0.66	0.79	0.9	0.53	0.68	0.59	0.37	0.61	0.81
	F1-Score	0.65	0.79	0.915	0.576	0.596	0.462	0.432	0.656	0.709
	IoU	0.496	0.657	0.845	0.418	0.435	0.302	0.274	0.481	0.556
Deeplabv3	Precision	0.418	0.64	0.78	0.33	0.37	0.2	0.38	0.29	0.35
	Recall	0.46	0.76	0.75	0.47	0.43	0.44	0.17	0.37	0.29
	F1-Score	0.438	0.695	0.765	0.388	0.398	0.275	0.235	0.325	0.317
	IoU	0.298	0.547	0.623	0.245	0.258	0.164	0.146	0.202	0.198
U-Net	Precision	0.663	0.86	0.89	0.77	0.54	0.36	0.61	0.61	0.67
	Recall	0.751	0.88	0.95	0.68	0.67	0.56	0.67	0.83	0.78
	F1-Score	0.704	0.87	0.919	0.722	0.598	0.44	0.639	0.704	0.721
	IoU	0.601	0.825	0.901	0.603	0.454	0.312	0.515	0.586	0.608

4.2. Ablation Experiments

In this study, ablation experiments were conducted to verify the effectiveness of each module, where the models used were (1) Baseline Mask R-CNN; (2) P_Mask R-CNN (incorporated into PAFPN); (3) A_MaskR-CNN (incorporated into ASPP); and (4) AP_Mask R-CNNN (incorporated into ASPP-PAFPN). The overall evaluation metrics of all models were derived by training 150 epochs, and the change in accuracy values and the training loss were calculated.

The overall evaluation results of each model, the numerical changes of the metrics under different module combinations, and training loss curves are shown in Tables 5 and 6, and Figure 11. From the overall metrics, the precision of all three models under different module combinations has improved to some extent. The average precision of Mask R-CNN incorporated into ASPP-PAFPN has been enhanced by 7.8%, the recall has improved by 4.6%, the F1-score has improved by 6.0%, and the IoU has improved by 8.5%; the precision of Mask R-CNN incorporating only ASPP has been enhanced by 2.1%, the recall has improved by 3.4%, the F1-score improved by 2.7%, and the IoU has improved by 3.1%; and Mask R-CNN incorporating only PAFPN improved by 4.9% in precision, 3.1% in the recall, 4.2% in the F1-score, and 2.3% in the IoU. In summary, ASPP-PAFPN brings better results than PAFPN or ASPP. Regarding category-specific metrics, the accuracy changes are mainly

concentrated in the categories with low accuracy. Some types with precision higher than 90% show slight improvements, such as traditional gray tile roofs and terracotta tile roofs. In terms of improvement strategies, incorporating only the ASPP module, the recognition accuracy improvement is more significant for dark blue-colored steel roofs (+9.0%) and magenta-colored steel roofs (+7.7%). Using only PAFPN, the gain is substantial for red resin roofs (+15.7%) and dark blue-colored steel roofs (+10.9%), followed by gray cement roofs (+4.2%) and light blue-colored steel roofs (+3.6%). When incorporating both ASPP and PAFPN, red resin roofs (+15.7%), dark blue-colored steel roofs (+13.9%), magenta-colored steel roofs (+6.9%), light blue-colored steel (+5.3%), gray cement roofs (+4.6%), gray-colored steel roofs (+ 3.2%), terra cotta tile roofs (+2.1%), and traditional gray tile roofs (1.2%), in order of decreasing F1-score. From the training loss curves, the convergence speed improvement effect after adding ASPP and PAFPN modules is not particularly obvious. However, it can still prove the advantage of AP-Mask R-CNN in loss convergence. The above results show that the ASPP and PAFPN modules can improve the model's performance. ASPP focuses on dark blue and magenta-colored steel roofs with more diverse scales and forms. PAFPN is more sensitive to red resin roofs, light blue-colored steel roofs, and gray cement roofs, which are easily confused with the background, with a smaller sample size. Combining the two improvements in this study can retain more accuracy advantages for the model.

Table 5. Results of metrics evaluation of ablation experiments.

Model	Metric	Overall	TGTR	TTB	MCSR	LBCSR	GCSR	GCR	DBCSR	RRR
AP_Mask R-CNN	Precision	0.713	0.905	0.985	0.475	0.65	0.545	0.675	0.897	0.574
	Recall	0.819	0.918	0.996	0.639	0.809	0.582	0.791	0.991	0.77
	F1- Score	0.757	0.912	0.991	0.545	0.721	0.563	0.728	0.942	0.658
	IoU	0.694	0.873	0.956	0.462	0.633	0.528	0.65	0.89	0.561
A_Mask R-CNN	Precision	0.656	0.901	0.951	0.493	0.569	0.496	0.617	0.828	0.396
	Recall	0.807	0.913	1	0.629	0.806	0.603	0.781	0.97	0.74
	F1-Score	0.724	0.907	0.98	0.553	0.667	0.544	0.689	0.893	0.515
	IoU	0.64	0.893	0.912	0.481	0.543	0.492	0.605	0.81	0.386
P_Mask R-CNN	Precision	0.684	0.897	0.966	0.422	0.619	0.476	0.654	0.861	0.581
	Recall	0.804	0.913	1	0.581	0.817	0.572	0.81	0.97	0.76
	F1-Score	0.739	0.905	0.988	0.489	0.704	0.52	0.724	0.912	0.658
	IoU	0.632	0.828	0.932	0.383	0.571	0.406	0.592	0.827	0.513
Mask R-CNN	Precision	0.635	0.892	0.941	0.436	0.579	0.504	0.611	0.738	0.376
	Recall	0.773	0.909	1	0.524	0.789	0.562	0.771	0.88	0.75
	F1-Score	0.697	0.9	0.97	0.476	0.668	0.531	0.682	0.803	0.501
	IoU	0.609	0.874	0.916	0.41	0.553	0.476	0.571	0.696	0.373

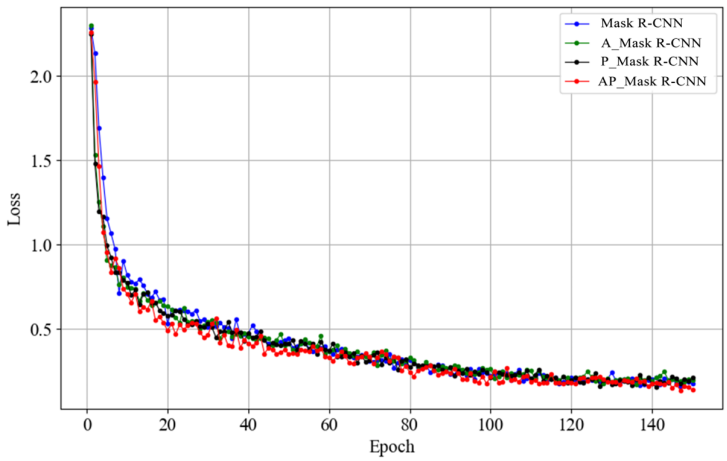


Figure 11. Loss curves of four models during training.

Table 6. Variation of metrics values for ablation experiments.

Model	Metric	Overall	TGTR	TTB	MCSR	LBCSR	GCSR	GCR	DBCSR	RRR
AP_Mask R-CNN	Precision	0.078	0.013	0.044	0.039	0.071	0.041	0.064	0.159	0.198
	Recall	0.046	0.009	−0.004	0.115	0.02	0.02	0.02	0.111	0.02
	F1- Score	0.06	0.012	0.021	0.069	0.053	0.032	0.046	0.139	0.157
	IoU	0.085	−0.001	0.04	0.052	0.08	0.052	0.085	0.194	0.182
A_Mask R-CNN	Precision	0.021	0.009	0.01	0.057	−0.01	−0.008	0.006	0.09	0.02
	Recall	0.034	0.004	0	0.105	0.017	0.041	0.01	0.09	−0.01
	F1-Score	0.027	0.007	0.01	0.077	−0.001	0.013	0.007	0.09	0.014
	IoU	0.031	0.019	−0.004	0.071	−0.01	0.016	0.034	0.114	0.013
P_Mask R-CNN	Precision	0.049	0.005	0.025	−0.014	0.04	−0.028	0.043	0.123	0.205
	Recall	0.031	0.004	0	0.057	0.028	0.01	0.039	0.09	0.01
	F1-Score	0.042	0.005	0.018	0.013	0.036	−0.011	0.042	0.109	0.157
	IoU	0.023	−0.046	0.016	−0.027	0.018	−0.07	0.021	0.131	0.14
Mask R-CNN	Precision	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000
	Recall	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000
	F1-Score	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000
	IoU	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000

In this study, to illustrate the impact of each module on the detection, eight representative samples were selected from the test sample data set based on the principles of category combination, target scale size, roof density, and roof orientation, and the prediction was performed based on the above model. After that, the prediction results are shown together with the original images and ground truth marker data for comparison and analysis, as shown in Figure 12a–f, respectively. In addition, to highlight the differences between ground truth and extraction results, different colored boxes are drawn. The red boxes in the extraction results indicate the wrong extraction, and the green boxes indicate an entirely or partially missing extraction. By analyzing the variations of the boxes, the effect of different modules on different types of data can be derived.

From the detection results of different models, it can be seen that the complex environment of the village influences the baseline model. The modeling of some categories of features needs to be more apparent. There are more false detections, missed detections, and serious jittering of detection edges, mainly in the small and medium-scale targets of gray cement roofs, red resin roofs, magenta-colored steel roofs, and gray-colored steel roofs. After incorporating ASPP, the detection rate of the model increases for small-scale gray cement roofs and decreases for red resin roofs; after incorporating PAFPN, the false detection of red resin roofs and magenta-colored steel roofs is improved, and the detected roof shapes are more regular. Therefore, this study incorporates the aggregated path feature fusion method (PAFPN) and feature pyramid pooling (ASPP) into the Mask R-CNN model to successfully avoid the loss of underlying feature information, strengthen the contextual relationship between multilayer features, improve the feature extraction capability of the Mask R-CNN backbone model, and better respond to the complex multi-scale, multi-type, and multi-temporal building roof environment.

4.3. Application Practices

This section selects a traditional village, Hexi village, for application practice to test this method's generalization ability and transferability. The village has a long history and covers an area of 4.7 square kilometers with various types and scales of buildings. The selected village is of double-layer significance for the performance verification of the AP_Mask R-CNN model and for preserving the village building.

Figure 12. Building extraction results of different models for eight samples: (**a**) origin images; (**b**) Ground Truth; (**c**) AP-Mask R-CNN; (**d**) A-Mask R-CNN; (**e**) P-Mask R-CNN; (**f**) Mask R-CNN.

The workflow includes six steps: UAV image data acquisition, orthophoto reconstruction, chunking, model detection, stitching, and visualization. After the data acquisition of the selected area by UAV is completed, a super-resolution orthophoto is generated with the help of ContextCapture software. Since the deep learning vision model has a specific limitation on the size of single image data, it is necessary to chunk the remote sensing image and retain the original projection coordinate system and geographic location. The chunked image data are recognized using the trained AP_Mask R-CNN model, and then the model output is stitched together with Python and visualized in GIS.

The visualization results of Hexi village and the original images are shown in Figures 13 and 14. Compared with the comparison and ablation experiments, the overall detection effect of Hexi village achieves a better state, except for the slightly decreased detection precision of terra cotta tile roofs, red resin roofs, and gray cement roofs, the detection rate and accuracy of the other types, such as traditional gray tile roofs and dark

blue-colored steel roofs, have been greatly improved, which indicates that the proposed method has better generalization ability and transferability.

Figure 13. UAV orthophoto of Hexi village.

Figure 14. Building extraction visualization results for Hexi village.

5. Discussion

5.1. Advantages Brought by AP_Mask R-CNN Model

AP_Mask R-CNN, as an instance segmentation model, can perform various tasks such as target classification, target detection, and semantic segmentation and has high performance while extracting more comprehensive and integrated information. Secondly, traditional villages' multi-scale, multi-type, and multi-temporal building roofs are prone to information loss in the backbone model feature extraction. They are easy to miss in the detection results. Numerous applications have demonstrated that Atrous Spatial Pyramid Pooling (ASPP) can improve the ability of the model to obtain complex features and enhance the recognition and detection of multi-scale objects by the model [39–41]. In this paper, the general convolution of the feature fusion model is replaced with an ASPP module, and feature extraction is performed simultaneously using convolution kernels with three dilation rates. The ablation experiments show that introducing ASPP extends the convolution kernel's perceptual capability, enhances the model's sensitivity to multi-scale foreground targets, and improves the F1-score by nearly 3%.

In addition, in the actual recognition process, making full use of the bottom features can improve the model generalization ability, but usually, the bottom features also occupy more computational resources. In this case, feature fusion mechanisms with advanced

semantics are needed to enhance feature mining for multi-scale, multi-type, and multi-temporal building roofs. In this paper, we refer to the idea of PANet [44,45] to construct a path aggregation feature pyramid fusion mechanism. The 4.2% improvement of the F1 score in the experimental results indicates that the model narrows the gap between the exact and predicted values of the markers and reflects the model's attention to the category features with a smaller sample size and the easily confused category features. In addition, this paper uses data augmentation and migration learning to reduce the need for data volume and improve the recognition performance of the model.

5.2. Limitations of Automatic Extraction of Traditional Village Building

The method proposed in this paper effectively improves the performance of extracting buildings from traditional villages in low-altitude UAV remote sensing images. It achieves high-accuracy extraction of traditional village buildings. However, there are still problems, such as low accuracy for a few roof types (e.g., blue-colored steel roofs, red resin roofs, etc.) and a single area to which the model is applicable. We analyze the reasons and improvement methods. First, the uneven sample size of different roof types is an important reason for the low average accuracy. This paper divides the building roofs in traditional villages into eight categories. Still, the sample size needs to be more balanced because different roof types cannot be guaranteed to be evenly distributed in the villages. Therefore, this paper improves the model performance by enhancing the backbone model, data augmentation, migration learning, and other strategies. The result effectively improves the accuracy of building recognition with a small sample size. However, there is still room for improvement in recognition results and accuracy. Therefore, the data imbalance between different types of roofs can be bridged by higher quality data sources, replacement of loss functions, etc., in future work. In addition, the data enhancement strategy based on generative adversarial networks [58] is an effective measure for some of the smaller sample size categories.

In addition, this study applies UAV remote sensing images and computer vision to the building extraction of traditional Chinese villages for the first time. It achieves the accurate identification of different roofs in traditional villages. However, because the dataset is derived from typical historical villages in Beijing, this model only supports building extraction for villages in northern China. In the subsequent research, the datasets should be produced by selecting village images from different regions and training the model separately for other residential zoning characteristics to realize the excavation of potential traditional villages in China on a larger scale and to promote the development of the Chinese traditional village investigation business.

6. Conclusions

In this study, an innovative workflow for automatically extracting traditional village buildings based on UAV images is established to promote the conservation and development of traditional Chinese villages. The method is based on deep learning and UAV remote sensing to automatically classify, locate, and segment different building roofs in traditional villages. For this purpose, the authors collected orthophoto datasets of the roofs of typical traditional village buildings in Beijing and finely annotated them with good classification. Secondly, the improved Mask R-CNN model performs well in extracting roof features in the homemade dataset under the premise of the complex architectural environment with multiple scales, categories, and temporalities in traditional villages and accurately extracting traditional village architectural information. Faced with the architectural survey task in many villages in China at present, efficient and accurate extraction of buildings is an effective means to cope with the problem.

Although the current framework needs to be more mature, it has shown considerable potential for the automatic extraction of traditional village buildings in terms of precision and practicality. Accuracy is the first focus of this study, with combining ASPP and PAFPN to improve the backbone model for complex feature extraction and fusion as the primary strategy and transfer learning and data augmentation techniques as auxiliary strategies to

enhance the performance of the model for automatic building segmentation, localization, and classification. The results of the ablation experiments show that precision, recall, and F1-score have significant advantages, indicating that the improved method in this study improves the stability of the training process and the effectiveness of the final model for building extraction. Practicality, another focus of this study, is mainly ensured by selecting the study area and following strict classification criteria. Starting from this goal, firstly, eight historical villages with different geographical features, village scales, and building types were selected as the study area in Beijing, which combined good weather conditions and flight techniques to lay down the reliability of the source of the building roof dataset. Secondly, methods such as cross-validation support the classification of building roofs to ensure that the classification and labeling errors in the dataset input to the model are minimized. Finally, we chose the traditional Chinese village of Hexi for our application practice. Its extraction results verify the robustness and generalization ability of the model proposed in this study, indicating that the model has some migration effect in evaluating other traditional villages in North China.

Author Contributions: Conceptualization, Y.S. and J.Z.; methodology, W.W. and Y.S.; software, D.H.; validation, W.W. and Y.S.; formal analysis, W.W. and Y.S.; investigation, Y.S. and F.L.; resources, S.L., Y.S. and W.W.; data curation, W.W.; writing—original draft preparation, W.W.; writing—review and editing, W.W., S.L. and L.H.; visualization, W.W.; supervision, Y.S. and J.Z.; project administration, Y.S. and J.Z.; funding acquisition, Y.S. and J.Z. All authors have read and agreed to the published version of the manuscript.

Funding: This research was funded by the National Natural Science Foundation of China Key Projects (grant number 51938002), the National Natural Science Foundation of China (grant number 52178029), and the Soft Science Project of the Ministry of Housing and Construction of China (grant number 2020-R-022).

Data Availability Statement: The data presented in this study are available on request from the corresponding author.

Acknowledgments: The authors would like to thank the editor and the anonymous reviewers who provided insightful comments on improving this article.

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

References

1. Ghosh, M. Traditional folk art community and urban transformation: The case of the artists' village at Kalighat, India. *J. Archit. Plan. Res.* **2019**, *36*, 70–89.
2. Xu, Q.; Wang, J. Recognition of values of traditional villages in Southwest China for sustainable development: A case study of Liufang village. *Sustainability* **2021**, *13*, 21. [CrossRef]
3. Liu, Y. On the protection dilemma of traditional Chinese villages: A case study of Xisuguazi Tibetan village. In Proceedings of the Euro-Asian Conference on Corporate Social Responsibility (CSR) and Environmental Management—Tourism, Society and Education Session (Part III), Tianjin, China, 23–24 November 2018; pp. 45–52.
4. Xie, X.B.; Li, X.J. The formation and transformation of "Cultural Matrix" in traditional village. *J. Hunan Univ. Soc. Sci. Ed.* **2019**, *33*, 8.
5. 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*, 19. [CrossRef] [PubMed]
6. Lu, Y.; Ahmad, Y. Heritage protection perspective of sustainable development of traditional villages in Guangxi, China. *Sustainability* **2023**, *15*, 23. [CrossRef]
7. Liu, T.; Liu, P.; Wang, L. The protection and tourism development path of ancient villages and old towns under the background of new-type urbanization: A case study of old town of Xuanzhou in Hunan province. *Geogr. Res.* **2019**, *38*, 133–145.
8. Xia, M. The rural revitalization strategy-strategies for cultural heritage and transformation of traditional village Zhu Jiayu, Zhangqiu. In Proceedings of the International Workshop on Advances in Social Sciences (IWASS), Hong Kong, China, 12–13 December 2018; pp. 1192–1195.
9. Yu, Y. Landscape transition of historic villages in Southwest China. *Front. Archit. Res.* **2013**, *2*, 234–242. [CrossRef]
10. Olczak, B.; Wilkosz-Mamcarczyk, M.; Prus, B.; Hodor, K.; Dixon-Gough, R. Application of the building cohesion method in spatial planning to shape patterns of the development in a suburban historical landscape of a 'village within Kraków'. *Land Use Policy* **2022**, *114*, 105997. [CrossRef]

11. Fu, J.; Zhou, J.; Deng, Y. Heritage values of ancient vernacular residences in traditional villages in Western Hunan, China: Spatial patterns and influencing factors. *Build. Environ.* **2021**, *188*, 107473. [CrossRef]
12. Wang, N.; Fang, M.; Beauchamp, M.; Jia, Z.; Zhou, Z. An indigenous knowledge-based sustainable landscape for mountain villages: The Jiabang rice terraces of Guizhou, China. *Habitat Int.* **2021**, *111*, 102360. [CrossRef]
13. Song, L.A.; Gl, B.; Ming, X.C. The linguistic landscape in rural destinations: A case study of Hongcun Village in China. *Tour. Manag.* **2020**, *77*, 104005.
14. Xiong, Y.; Chen, Q.; Zhu, M.; Zhang, Y.; Huang, K. Accurate detection of historical buildings using aerial photographs and deep transfer learning. In Proceedings of the IGARSS 2020—2020 IEEE International Geoscience and Remote Sensing Symposium, Waikoloa, HI, USA, 26 September–2 October 2020.
15. Liu, J.; Wang, S.; Hou, X.; Song, W. A deep residual learning serial segmentation network for extracting buildings from remote sensing imagery. *Int. J. Remote Sens.* **2020**, *41*, 5573–5587. [CrossRef]
16. Chen, J.; Wang, G.; Luo, L.; Gong, W.; Cheng, Z. Building area estimation in drone aerial images based on mask R-CNN. *IEEE Geosci. Remote Sens. Lett.* **2020**, *18*, 891–894. [CrossRef]
17. Li, Y.; Xu, W.; Chen, H.; Jiang, J.; Li, X. A novel framework based on mask R-CNN and histogram thresholding for scalable segmentation of new and old rural buildings. *Remote Sens.* **2021**, *13*, 1070. [CrossRef]
18. Wang, Y.; Li, S.; Teng, F.; Cai, H. Improved mask R-CNN for rural building roof type recognition from UAV high-resolution images: A case study in Hunan province, China. *Remote Sens.* **2022**, *14*, 265. [CrossRef]
19. Zhan, Y.; Liu, W.; Maruyama, Y. Damaged building extraction using modified mask R-CNN model using post-event aerial images of the 2016 Kumamoto earthquake. *Remote Sens.* **2022**, *14*, 1002. [CrossRef]
20. Tejeswari, B.; Sharma, S.K.; Kumar, M.V.R.; Gupta, K. Building footprint extraction from space-borne imagery using deep neural networks. *Int. Arch. Photogramm. Remote Sens. Spat. Inf. Sci.* **2022**, 641–647. [CrossRef]
21. Katartzis, A.; Sahli, H.; Nyssen, E.; Cornelis, J. Detection of buildings from a single airborne image using a Markov random field model. In Proceedings of the IEEE 2001 Geoscience and Remote Sensing Symposium (IGARSS), Sydney, NSW, Australia, 9–13 July 2001.
22. Simonetto, E.; Oriot, H.; Garello, R. Rectangular building extraction from stereoscopic airborne Radar images. *IEEE Trans. Geosci. Remote Sens.* **2005**, *43*, 2386–2395. [CrossRef]
23. Jung, C.R.; Schramm, R. Rectangle detection based on a windowed hough transform. In Proceedings of the Brazilian Symposium on Computer Graphics & Image Processing, Curitiba, Brazil, 20 October 2004.
24. Ma, Y.; Gu, X.D.; Wang, Y.Y. Feature fusion method for edge detection of color images. *J. Syst. Eng. Electron.* **2009**, *20*, 394–399.
25. Zhang, Z.Z.; Zhang, Y.J. Building extraction from airborne laser point cloud using NDVI constrained watershed algorithm. *Acta Opt. Sin.* **2016**, *36*, 1028002.
26. Zhou, S.L.; Liang, D.; Wang, H.; Kong, J. Remote sensing image segmentation approach based on quarter-tree and graph cut. *Comput. Eng.* **2010**, *36*, 3.
27. Wei, D.Q. Research on Building Extraction Technology on High Resolution Remote Sensing Images. Ph.D. Thesis, PLA Information Engineering University, Zhengzhou, China, 2013.
28. Iyer, B.; Macleod, M.D. Multi-scale region segmentation of images using nonlinear methods. In Proceedings of the Visual Communications and Image Processing, Perth, Australia, 30 May 2000.
29. Zguira, A.; Doggaz, N.; Zagrouba, E. Region-based objective evaluation of polygonal mesh segmentation methods. In Proceedings of the VISAPP 2011—Sixth International Conference on Computer Vision Theory and Applications, Algarve, Portugal, 5–7 March 2011.
30. Hui, Z.; Fritts, J.E.; Goldman, S.A. A fast texture feature extraction method for region-based image segmentation. In Proceedings of the SPIE—The International Society for Optical Engineering, San Jose, CA, USA, 18–20 January 2005; p. 5685.
31. Huang, X.; Zhang, L. Morphological building/shadow index for building extraction from high-resolution imagery over urban areas. *IEEE J. Sel. Top. Appl. Earth Obs. Remote Sens.* **2012**, *5*, 161–172. [CrossRef]
32. Gao, X.; Wang, M.; Yang, Y.; Li, G. Building extraction from RGB VHR images using shifted shadow algorithm. *IEEE Access* **2018**, *6*, 22034–22045. [CrossRef]
33. Maruyama, Y.; Tashiro, A.; Yamazaki, F. Use of digital surface model constructed from digital aerial images to detect collapsed buildings during earthquake. *Procedia Eng.* **2011**, *14*, 552–558. [CrossRef]
34. Tournaire, O.; Bredif, M.; Boldo, D.; Durupt, M. An efficient stochastic approach for building footprint extraction from digital elevation models. *ISPRS J. Photogramm. Remote Sens.* **2010**, *65*, 317–327. [CrossRef]
35. Saeid, P.; Meisam, A. Building extraction from fused LiDAR and hyperspectral data using Random Forest Algorithm. *Geomatica* **2017**, *71*, 185–193.
36. Ferro, A.; Brunner, D.; Bruzzone, L. Automatic detection and reconstruction of building radar footprints from single VHR SAR images. *IEEE Trans. Geosci. Remote Sens.* **2013**, *51*, 935–952. [CrossRef]
37. He, K.; Gkioxari, G.; Dollár, P.; Girshick, R.B. Mask R-CNN. *IEEE Trans. Pattern Anal. Mach. Intell.* **2017**, *42*, 386–397. [CrossRef]
38. Yuan, Y. *Study on the Inheritance and Update of Residence in Jijiaying Village*; Beijing University of Civil Engineering and Architecture: Beijing, China.

39. Chen, L.C.; Papandreou, G.; Kokkinos, I.; Murphy, K.; Yuille, A.L. DeepLab: Semantic image segmentation with deep convolutional nets, atrous convolution, and fully connected CRFs. *IEEE Trans. Pattern Anal. Mach. Intell.* **2018**, *40*, 834–848. [CrossRef]
40. Chen, L.C.; Papandreou, G.; Schroff, F.; Adam, H. Rethinking atrous convolution for semantic image segmentation. *arXiv* **2017**, arXiv:1706.05587.
41. Jha, D.; Smedsrud, P.H.; Riegler, M.A.; Johansen, D. Simulamet. ResUNet++: An advanced architecture for medical image segmentation. *arXiv* **2019**, arXiv:1911.07067.
42. Zhao, H.; Shi, J.; Qi, X.; Wang, X.; Jia, J. Pyramid scene parsing network. In Proceedings of the 2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Honolulu, HI, USA, 21–26 July 2017; pp. 6230–6239. [CrossRef]
43. Li, W.; Sun, K.; Zhao, H.; Li, W.; Wei, J.; Gao, S. Extracting buildings from high-resolution remote sensing images by deep ConvNets equipped with structural-cue-guided feature alignment. *Int. J. Appl. Earth Obs. Geoinf.* **2022**, *113*, 102970. [CrossRef]
44. Lin, T.Y.; Dollar, P.; Girshick, R.; He, K.; Hariharan, B.; Belongie, S. Feature pyramid networks for object detection. *IEEE Comput. Soc.* **2017**, 2117–2125.
45. Liu, S.; Qi, L.; Qin, H.; Shi, J.; Jia, J. Path aggregation network for instance segmentation. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Salt Lake City, UT, USA, 18–23 June 2018; pp. 8759–8768.
46. Ge, J.; Wang, C.; Zhang, B.; Xu, C.; Wen, X. Azimuth-sensitive object detection of high-resolution SAR images in complex scenes by using a spatial orientation attention enhancement network. *Remote Sens.* **2022**, *14*, 2198. [CrossRef]
47. Yingying, X.; Li, D.; Xie, Q.; Wu, Q.; Wang, J. Automatic defect detection and segmentation of tunnel surface using modified Mask R-CNN. *Measurement* **2021**, *178*, 109316.
48. He, D.; Qiu, Y.; Miao, J.; Zou, Z.; Li, K.; Ren, C.; Shen, G. Improved mask R-CNN for obstacle detection of rail transit. *Measurement* **2022**, *190*, 110728. [CrossRef]
49. Fu, C.Y.; Liu, W.; Ranga, A.; Tyagi, A.; Berg, A.C. DSSD: Deconvolutional Single Shot Detector. *arXiv* **2017**, arXiv:1701.06659.
50. Kong, T.; Yao, A.; Chen, Y.; Sun, F. HyperNet: Towards accurate region proposal generation and joint object detection. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Las Vegas, NV, USA, 27–30 June 2016; pp. 845–853.
51. Noh, H.; Hong, S.; Han, B. Learning deconvolution network for semantic segmentation. In Proceedings of the IEEE International Conference on Computer Vision, Las Vegas, NV, USA, 27–30 June 2016.
52. Wang, D.G.; Lv, Q.Y.; Wu, Y.F.; Fan, Z.Q. The characteristic of regional differentiation and impact mechanism of architecture style of traditional residence. *J. Nat. Resour.* **2019**, *34*, 1864.
53. Monna, F.; Rolland, T.; Denaire, A.; Navarro, N.; Granjon, L.; Barbé, R.; Chateau-Smith, C. Deep learning to detect built cultural heritage from satellite imagery—Spatial distribution and size of vernacular houses in Sumba, Indonesia. *J. Cult. Herit.* **2021**, *52*, 171–183. [CrossRef]
54. Shorten, C.; Khoshgoftaar, T.M. A survey on image data augmentation for deep learning. *J. Big Data* **2019**, *6*, 48. [CrossRef]
55. He, K.; Zhang, X.; Ren, S.; Sun, J. Deep residual learning for image recognition. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Las Vegas, NV, USA, 27–30 June 2016.
56. He, K.; Zhang, X.; Ren, S.; Sun, J. Spatial pyramid pooling in deep convolutional networks for visual recognition. In Proceedings of the European Conference on Computer Vision, Zurich, Switzerland, 6–12 September 2014.
57. Ronneberger, O.; Fischer, P.; Brox, T. U-Net: Convolutional networks for biomedical image segmentation. In *Proceedings of the Medical Image Computing and Computer-Assisted Intervention–MICCAI 2015: 18th International Conference, Munich, Germany, 5–9 October 2015*; Springer International Publishing: Cham, Switzerland, 2015; pp. 234–241.
58. Goodfellow, I.J.; Pouget-Abadie, J.; Mirza, M.; Xu, B.; Warde-Farley, D.; Ozair, S.; Courville, A.C.; Bengio, Y. Generative Adversarial Nets. In Proceedings of the NIPS, Montreal, QC, Canada, 8–13 December 2014.