remote sensing

MDPI

Article

Improved Mask R-CNN for Rural Building Roof Type Recognition from UAV High-Resolution Images: A Case Study in Hunan Province, China

Yanjun Wang [1,2,3,*], Shaochun Li [1,2,3], Fei Teng [1,2,3], Yunhao Lin [1,2,3], Mengjie Wang [1,2,3] and Hengfan Cai [1,2,3]

1 Hunan Provincial Key Laboratory of Geo-Information Engineering in Surveying, Mapping and Remote Sensing, Hunan University of Science and Technology, Xiangtan 411201, China; lsc_gis@mail.hnust.edu.cn (S.L.); tengfei@mail.hnust.edu.cn (F.T.); linyunhao@mail.hnust.edu.cn (Y.L.); wangmengjie@mail.hnust.edu.cn (M.W.); chf@mail.hnust.edu.cn (H.C.)
2 National-Local Joint Engineering Laboratory of Geo-Spatial Information Technology, Hunan University of Science and Technology, Xiangtan 411201, China
3 School of Earth Sciences and Geospatial Information Engineering, Hunan University of Science and Technology, Xiangtan 411201, China
* Correspondence: wangyanjun@hnust.edu.cn; Tel./Fax: +86-731-5829-0092

Abstract: Accurate roof information of buildings can be obtained from UAV high-resolution images. The large-scale accurate recognition of roof types (such as gabled, flat, hipped, complex and mono-pitched roofs) of rural buildings is crucial for rural planning and construction. At present, most UAV high-resolution optical images only have red, green and blue (RGB) band information, which aggravates the problems of inter-class similarity and intra-class variability of image features. Furthermore, the different roof types of rural buildings are complex, spatially scattered, and easily covered by vegetation, which in turn leads to the low accuracy of roof type identification by existing methods. In response to the above problems, this paper proposes a method for identifying roof types of complex rural buildings based on visible high-resolution remote sensing images from UAVs. First, the fusion of deep learning networks with different visual features is investigated to analyze the effect of the different feature combinations of the visible difference vegetation index (VDVI) and Sobel edge detection features and UAV visible images on model recognition of rural building roof types. Secondly, an improved Mask R-CNN model is proposed to learn more complex features of different types of images of building roofs by using the ResNet152 feature extraction network with migration learning. After we obtained roof type recognition results in two test areas, we evaluated the accuracy of the results using the confusion matrix and obtained the following conclusions: (1) the model with RGB images incorporating Sobel edge detection features has the highest accuracy and enables the model to recognize more and more accurately the roof types of different morphological rural buildings, and the model recognition accuracy (Kappa coefficient (*KC*)) compared to that of RGB images is on average improved by 0.115; (2) compared with the original Mask R-CNN, U-Net, DeeplabV3 and PSPNet deep learning models, the improved Mask R-CNN model has the highest accuracy in recognizing the roof types of rural buildings, with *F1-score*, *KC* and *OA* averaging 0.777, 0.821 and 0.905, respectively. The method can obtain clear and accurate profiles and types of rural building roofs, and can be extended for green roof suitability evaluation, rooftop solar potential assessment, and other building roof surveys, management and planning.

Keywords: UAV high-resolution optical image; roof type recognition; VDVI; Sobel; improved Mask R-CNN; deep learning

Citation: Wang, Y.; Li, S.; Teng, F.; Lin, Y.; Wang, M.; 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. https://doi.org/10.3390/rs14020265

Academic Editor: Saeid Homayouni

Received: 21 November 2021
Accepted: 5 January 2022
Published: 7 January 2022

1. Introduction

The accurate identification of rural building roof types is significant in natural resource surveys [1], beautiful countryside planning and construction [2], detection of illegal

roofs [3], assessment of rooftop solar photovoltaic power generation potential [4,5] and disaster emergency management (e.g., detection of damaged rooftop areas after earthquakes and landslides) [6]. Compared with urban buildings, rural buildings have their own unique characteristics, which are mainly reflected in the following: firstly, the lack of unified planning and management leads to a chaotic building layout; secondly, the design of houses is mostly based on the construction experience of rural artisans, which makes the roof types of rural buildings complex and diverse [7]. However, the current building identification research mainly focuses on the extraction of large scale urban buildings [8], while less attention is paid to the more difficult and complex identification of multiple roof types of rural buildings. Therefore, there is an urgent need to develop methods for fine investigation and identification of rural roof types on a large scale.

Traditional building survey methods often require a lot of manpower and material resources for field mapping and surveying, which is a large workload and high cost, especially in rural areas [9]. With the launch of high-resolution remote sensing satellites (such as Worldview-2, GF-2, etc.), more and more scholars and mapping departments use high spatial resolution remote sensing images to extract building information [10,11]. Its spatial resolution reaching the sub-meter level allows buildings on remote sensing images to present richer detailed information such as internal structure, geometric contours and texture patterns, and the differences in geometric dimensions and texture features are the fundamental basis for identifying different categories of building roofs. However, due to the presence of a large number of shadows, features with similar spectral characteristics to buildings (such as roads, etc.) and intra-class hybrid image elements, the traditional remote sensing image classification methods based on pixel features cannot effectively, correctly and completely recognize different roof types [12]. Considering that the size of image elements in high spatial resolution remote sensing images reflecting the ground target is closer to the natural scene target taken on the ground, which is more in line with the human eye's perception compared with low and medium resolution images [13], there are many scholars using machine learning methods to extract different roof types, such as object-oriented classification, support vector machine classification (SVM) and random forest classification (RF) [14,15]. However, the performance of object-oriented methods depends mainly on the segmentation results of images, and the results of classifier methods such as SVM and RF often depend on the selection of a large set of valid samples, whose shallow structure makes the deeper information about buildings unavailable and not generalizable across images of different regions, which makes machine learning classification methods face great challenges in terms of reliability and generalizability in accurately identifying the roof types of buildings [16,17]. Although some studies have also combined LiDAR point cloud data and satellite image data using SVM and RF models to identify multiple roof types (e.g., flat, gabled, hipped, pyramidal and skillion roof types) to improve the accuracy of roof category identification by machine learning models [18], the high cost of acquiring LiDAR point cloud data prevents the effective achievement of the accurate identification of roof types on a large scale.

Similarly, large-scale high-resolution satellite remote sensing images have many shortcomings in roof type identification, for example, the long revisit period makes the time interval of different simultaneous data acquisition and processing longer, which makes the real-time update of roof database and disaster emergency monitoring impossible to be guaranteed [19]. In addition, compared with urban areas, rural areas are more likely to produce more cloudy weather, which makes the quality of satellite imaging lower and thus limits the accurate identification of roof types of complex buildings such as small areas in rural areas [20]. Low-altitude remote sensing technology, represented by UAV technology, can overcome the above shortcomings due to its advantages of high flexibility, high timeliness, low cost and not being restricted by geographic environment conditions, and it can provide centimeter-level ultra-high-resolution remote sensing images, which makes the spatial structure, surface texture features and edge feature information of the features on the images more clear [21]. However, the UAV remote sensing images with

significantly larger image data have higher requirements for classification methods than those of general high-resolution images [22].

With the rapid development of big data and high-performance computers [23], the application of deep learning technology in the field of automatic image recognition is expanding. Deep learning models, represented by Convolutional Neural Network (CNN), can automatically learn more complex abstract high-dimensional features from the low-level features of the input image [24], which obviously brings great advantages for acquiring complex spectral, geometric and texture features in ultra-high-resolution remote sensing images [25]. Based on this, many researchers are now using semantic segmentation frameworks (e.g., VGG-F [26], U-Net [27], SegNet [28,29], etc.) to identify building roof types in ultra-high-resolution remote sensing images. However, semantic segmentation suffers from the problems of difficulty in distinguishing different objects of the same range and easily connecting different building roof types at the edges [30], which is not conducive to the application and research of complex rural roof type recognition. In contrast to image semantic segmentation, instance segmentation can identify multiple objects of the same broad category as different individual entities and assign a pixel-level semantic category to each entity on this basis [31], which is ideal for the recognition of complex rural building roof types. Among the existing instance segmentation methods, Mask R-CNN has been proved to be a powerful and adaptable deep learning model in different domains [32] and consists of a combination of target detection and semantic segmentation techniques to segment objects into prediction frames by predicting the bounding boxes of target objects and finally outputting high precision vector segmentation results [33]. A large number of scholars have applied it to the recognition of buildings [34], for example, Stiller et al. [35] used fine-tuned Mask R-CNN to extract large-scale buildings in urban areas of Chile, while for more complex recognition of different building roof types Mask R-CNN is less applied at present. Most of the above studies, however, have focused on building extraction in urban areas, and less attention has been paid to the more difficult problem of identifying complex rural roof types. At the same time, the selection of deep learning feature extraction model also has a significant impact on the recognition accuracy of complex roof types [36]. As an important feature extraction structure of Mask R-CNN, the deep residual network [37] enables the model to extract more complex image features without decreasing the accuracy by increasing the number of residual convolution layers, and the traditional Mask R-CNN uses ResNet50 or ResNet101 as the feature extraction layer [38]. Whereas, for complex building roof type recognition in UAV remote sensing images with large data volume, the above deep residual networks are not able to extract very complex building roof type features at a deeper level [39], and ResNet152 [40], which is currently one of the best performers in classification, can solve the above problem and can be deployed in Mask R-CNN by migration learning.

Due to payload limitations [41], most UAV ultra-high-resolution remote sensing images only have red, green and blue (RGB) bands, making the inter-class similarity and intra-class variability of different features in the images more obvious [42]. Deep learning, while already the best method available in terms of automation and accuracy, has limitations in the recognition of low reflectance, features with similar spectral characteristics to buildings and complex building roof types, such as similar roof lawns and grasses, similar concrete roofs and floors, and dark gaps between different roof types [43]. It has been shown that adding more visual features to deep learning models can better address these problems [44]. Boonpook et al. [45] combined RGB data from UAV remote sensing imagery with the visible band vegetation index (VDVI) and digital surface model (DSM) to extract complex buildings using the SegNet deep learning method. The results show that the RGB combination with VDVI features can improve the separability of building areas from vegetation, the RGB combination with DSM features helps to separate buildings from ground objects, and the RGB combination with both features can identify small buildings that are low and obscured by vegetation, and the extraction results of each feature combination are higher than those of RGB only. However, the DSM data only contains

the height information of the ground objects and cannot distinguish the internal structure (e.g., surface texture features, etc.) of more complex different roof types. Zhang et al. [46] extracted buildings from high-resolution remote sensing images by fusing the Sobel edge detection algorithm and Mask R-CNN algorithm, and the results showed that using Sobel edge detection algorithm to segment building boundaries solved the problems of boundary texture extraction and object internal integrity in deep learning. The above studies show that adding VDVI feature bands can effectively distinguish buildings from green areas in UAV visible images and improve the inter-class similarity problem. What is more, adding Sobel edge detection features can clearly show the gradient, texture and boundary features of building roof surfaces, enhance the distinguishability of different building roof types and thus improve the intra-class variability problem.

In summary, to address the problems of extracting complex roof type features and easily confusing the building roofs of low reflectance with vegetation, roads and other objects of similar spectral features in existing methods, this paper proposes the improved Mask R-CNN method for rural building roof type recognition from UAV visible high-resolution remote sensing imagery. The improved Mask R-CNN model based on different feature combinations can fully extract the more complex features of different building roof types, effectively improve the differentiation between buildings and vegetation, as well as different building roof types, accelerate the convergence speed of the model and achieve a large range of high-precision recognition of building roof types in UAV visible remote sensing images. The main sections of this paper are organized as follows: Section 2 introduces the study areas and pre-processing of the experimental UAV image dataset. Section 3 describes the main methodological process of this study, including the visual feature extraction method of UAV visible images and the improvement and implementation of Mask R-CNN model. Section 4 shows the results of the rural building roof type recognition of this model. Section 5 mainly discusses the influences of different feature combinations and the number of ResNet layers on the training results, as well as the future probable improvement of this study. Finally, a summary of the conclusions of this study is given in Section 6.

2. Study Area and Data

2.1. Study Area

To test the performance of the proposed model, we used a UAV to obtain ultra-high-resolution remote sensing images covering Luxi County, Xiangxi Prefecture, Hunan Province, China. The selected study areas all contain relatively dense rural buildings, as shown in Figure 1. Luxi County is located in the northwestern part of Hunan Province, with mountainous terrain, a well-developed water system, high annual precipitation and obvious climatic differences in the region. It is an agricultural area with a predominantly ethnic minority population. The roofs of rural buildings in the selected area of this study are mainly flat and sloped, while there are numerous indistinguishable roof types, which pose great challenges to the task of building roof type identification. Although there are already open building datasets around the world (such as the WHU building dataset and the ISPRS Vaihingen dataset, etc.) that provide various patterns and styles of architectural landscapes [47], there are still fewer ultra-high-resolution remote sensing image building datasets proposed for rural areas in China, and the building patterns of rural areas in China are very different from urban areas; even urban and suburban buildings in western countries are very different, so in the process of studying the supervised learning method, some special representative rural buildings in the test area of this study can be considered to improve the generalizability of the model to identify the roof types of different styles of buildings. The total area of the seven test areas in this study is 62.34 km^2, of which the training area is 48.55 km^2 and the test area is 13.79 km^2.

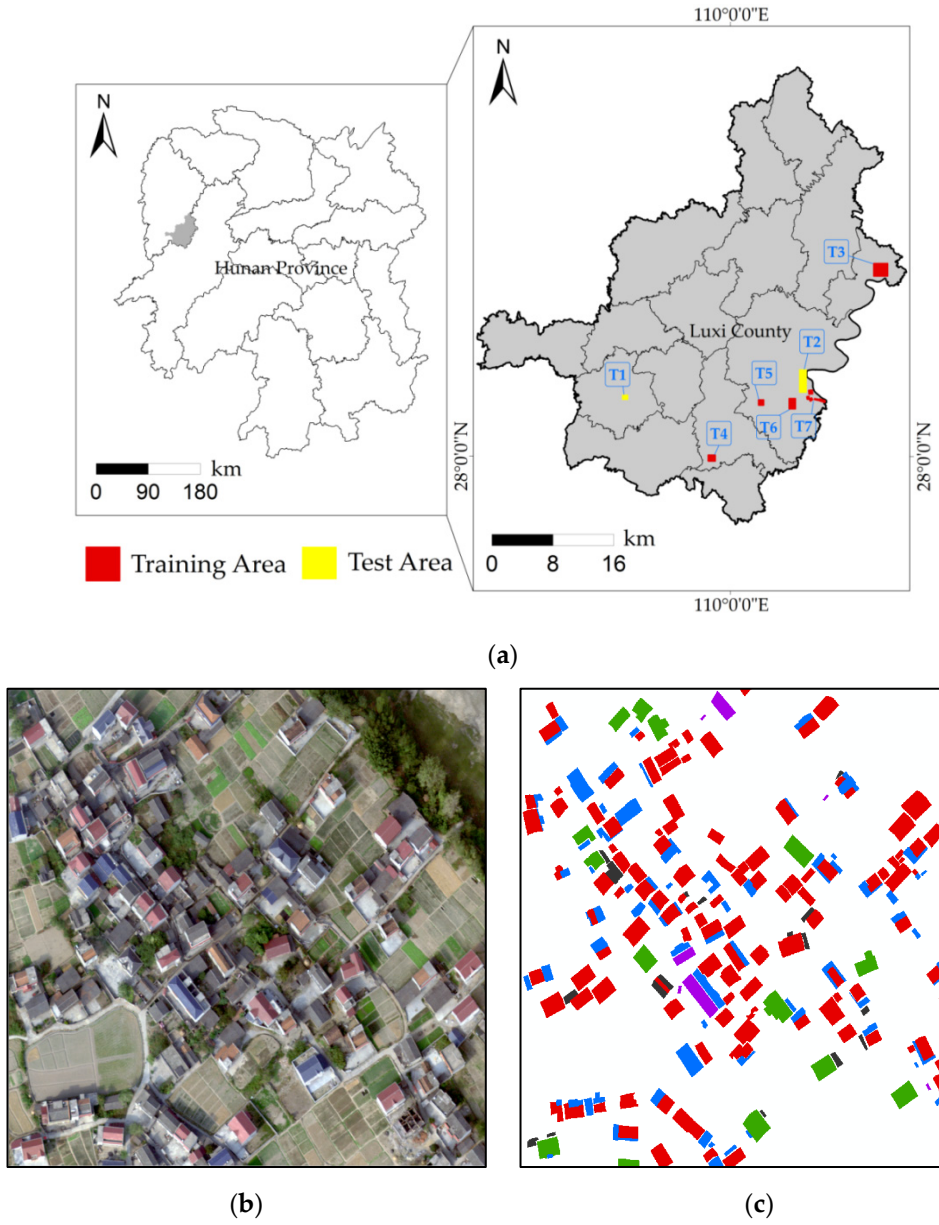

Figure 1. Study area. (**a**) Luxi County is located in Xiangxi Prefecture, Hunan Province, China; (**b**) Puxi village in Luxi County; and (**c**) roof type labels of rural buildings corresponding to Puxi village (red mask for gabled roof, blue mask for flat roof, purple mask for hipped roof, green mask for complex roof and black mask for mono-pitched roof).

2.2. Data Acquisition and Preprocessing

This study used a six-rotor UAV (KPM-28, Hunan Kunpeng Zhihui UAV Technology Co., LTD., Changsha, China) to acquire ultra-high-resolution true color aerial imagery of Luxi County, Hunan Province, in May 2018, and it had a wheelbase of 1.6 m, a payload weight of 8 kg, a cruise speed of 8 m/s, an endurance of about 60 min and was equipped with a SHARE-101S tilt photography camera. The SHARE-101S tilt photography camera consists of five complementary metal oxide semiconductor (CMOS) sensors (23.5 mm × 15.6 mm) with an effective pixel count of 24.3 megapixels, a tilt angle of 45°, a storage capacity of 320 G and a lens focal length of 35 mm × 4 and 25 mm for mapping. The UAV orthophoto data acquisition and processing process are carried out with Pix4D software, which mainly includes four steps: laying image control points, developing the flight plan, field UAV image acquisition and orthophoto generation. Luxi County is located

in a mountainous area with complex terrain. In order to ensure the final image accuracy of the survey area, it is necessary to lay image control points evenly in advance in areas with relatively flat terrain and clear feature points. Since the remote sensing images acquired by the UAV in each flight cover a small area, it is necessary to manually divide the survey area of Luxi County into several small areas before the flight. The flight design is carried out according to a ground resolution of 20 cm, with a heading overlap of 60% and a side overlap of 40%, and the average flight height is within the range of 200–250 m. The output image is in visible RGB mode. The images acquired by different sorties of UAVs are stitched together as a way to reduce the influence of weather and light on the images and to acquire image data of the whole county. After adding ground control points to the stitched images, an aerial triangulation leveling quality report is generated to meet production accuracy requirements [48]. Finally, after setting the CGCS2000 coordinate system, the UAV orthophoto can be generated. Zhang et al. [49] classified roofs into six categories (flat, gable, gambrel, half hip, hip and pyramid) based on roof edges. In this study, the types of roof samples in orthophotos obtained by UAVs were classified into five types: gabled, flat, hipped, complex and mono-pitched, based on the roof survey standards of local mapping departments and the surface texture and shape characteristics of roofs and the overall morphology of buildings in existing data sets. The typical roof types used in this paper are shown in Figure 2. Table 1 shows the number and percentage of training and test data for each type of roof.

(**a**) (**b**) (**c**) (**d**) (**e**)

Figure 2. There are different roof types according to the characteristics of building roof texture and shape. (**a**) Gabled; (**b**) flat; (**c**) hipped; (**d**) complex; and (**e**) mono-pitched.

Table 1. Number of training samples (60%/58%), number of testing samples (40%/42%) and total number of samples per roof category (number of pixels (million)/number of roofs).

Sample	Gabled	Flat	Hipped	Complex	Mono-Pitched
Training data	8.43/4252	8.74/3767	0.64/249	0.9/258	0.28/229
Test data	7.26/3111	4.04/2166	0.21/39	1/209	0.12/122
Total	15.69/7363	12.78/5933	0.85/288	1.9/467	0.4/351

3. Methods

This paper proposed an improved Mask R-CNN based on different visual feature combinations for the rural building roof type recognition of UAV visible high-resolution remote sensing images. All the processes are shown in Figure 3, which mainly include: (1) calculated VDVI spectral features and Sobel edge detection spatial features of UAV visible remote sensing images, and composed two visual features with RGB images into different feature combinations as the input dataset of the deep learning model; (2) the Mask R-CNN model based on ResNet migration learning were applied to train sample datasets with different feature combinations and to identify and evaluate the accuracy of rural building roof types in T1 and T2 test areas to achieve accurate identification of rural building roof types in UAV visible remote sensing images.

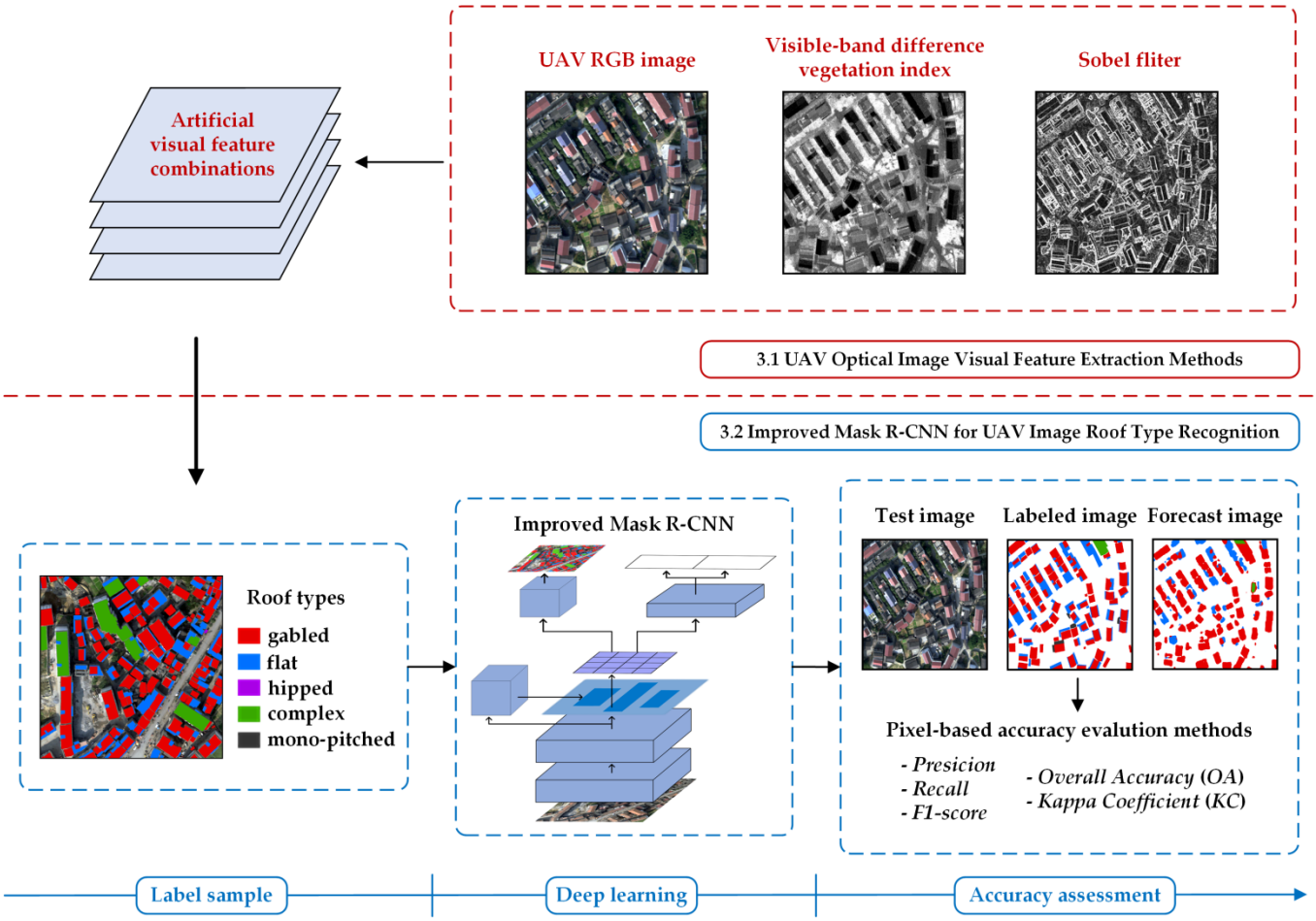

Figure 3. Improved Mask R-CNN for UAV high-resolution image rural building roof type recognition framework.

3.1. UAV Optical Image Visual Feature Extraction Methods

The visible RGB bands and the high spatial resolution of the UAV remote sensing images make the inter-class similarity and intra-class variability of the features in the images significantly increase. The inter-class similarity will intensify the confusion of identifying different classes of features in the images, while the intra-class variability will make the subclasses of the same kind of features in the images present more complex image features, causing difficulties for the model to extract the features of the subclasses of the features [50]. Therefore, how to select UAV images with visual features added to the RGB band to enhance the difference between building roofs and other features, as well as highlighting building roof features to make different building roof types easier to identify, is the key to identifying building roof types in the countryside from UAV visible light high-resolution remote sensing images. The area of UAV remote sensing image data used in this paper is located in a rural area with dense vegetation, and there are many buildings obscured by dark vegetation, which can be easily misclassified as buildings. Therefore, it is very important to remove the influence of vegetation on buildings for building identification. The existing vegetation indices of visible UAV remote sensing images mainly include NGRDI [51], VDVI [52], EXG [53], etc. Among these methods, VDVI has proved to be the most effective in extracting green vegetation and can effectively distinguish vegetation from other features [54]. This study mainly classifies the types of rural building roofs based on the texture features of the roof surface, and to make the model extract the texture features of the roof surface more easily, image enhancement methods can be used, which start by improving the visual effect of the image and highlighting the

texture or boundary information of the image. The methods can usually be divided into two categories: spatial domain enhancement and transform domain enhancement [55]. The Sobel edge detection algorithm [56], as a commonly used spatial domain enhancement algorithm, is better for images with grayscale gradients and more noise, and it is sensitive to edges in both horizontal and vertical directions, which can reduce the blurring of image edges. It has been shown that this method is a more effective image edge enhancement algorithm than edge detection algorithms such as Canny [57] and Laplacian [58]. In this paper, two visual features, VDVI based on spectral visual features and the Sobel edge detection algorithm based on spatial visual features, are introduced and their applicability is compared with different combinations of features for UAV visible band images.

3.1.1. Calculation of Visible Difference Vegetation Index (VDVI)

In order to improve the separability of rural building roofs from vegetation and avoid misclassification due to the similarity of rural building roofs (e.g., roofs with green vegetation or roofs shaded by dark vegetation, etc.) and ground vegetation [20], this paper introduces the vegetation index as another spectral visual feature. Xu et al. introduced NDVI as a feature band to extract buildings from ultra-high-resolution color infrared remote sensing images [59], and the results showed that adding NDVI could further distinguish buildings from green areas. However, vegetation indices such as NDVI, which need to calculate multispectral information, cannot be applied to visible UAV high-resolution remote sensing images, so this paper uses a visible difference vegetation index (VDVI) with improved NDVI, which can use visible RGB band information to extract vegetation information. Ma et al. [60] demonstrated that the use of VDVI as an additional spectral feature for visible UAV remote sensing imagery can effectively reduce the interference of vegetation on building roof information extraction. VDVI can be calculated according to Equation (1), and its results range from -1 to 1.

$$\mathrm{VDVI} = \frac{2 \times \rho_{green} - \rho_{red} - \rho_{blue}}{2 \times \rho_{green} + \rho_{red} + \rho_{blue}} \tag{1}$$

where: ρ_{green}, ρ_{red} and ρ_{blue} denote the values of the visible green, red and blue bands of the UAV orthophoto, respectively. In this study, the VDVI calculation of the UAV orthophoto was done using the band operation of ENVI 5.3, and the grayscale image obtained after the index calculation was input into the training data as an additional visual feature together with the RGB image.

3.1.2. Calculation of Sobel Edge Detection Features

The Sobel edge detection algorithm uses a discrete differential operator to operate on the approximate gradient of image grayscale, and the larger the gradient is, the more likely it is an edge [61]. The deep learning network can learn the complex features of high-resolution remote sensing image features, but there are still deficiencies in the feature extraction of building roof types for very complex UAV ultra-high-resolution remote sensing images, mainly in the differences between different building roof types on the image edges and the integrity of the target [62]. To solve these problems, we combine the image features calculated by Sobel edge detection as additional spatial visual features with UAV RGB images to improve the discriminability of different building roof types on UAV visible images. The Sobel operator can smooth out the building boundaries in the filtered images, making the surface texture and shape features of different building roof types more prominent, while reducing the interference of background noise. The Sobel algorithm consists of two sets of 3×3 matrices, which are convolved along the x-axis, y-axis, from top to bottom and from left to right on the image, respectively, to obtain the horizontal and vertical luminance difference approximation; if $f(x, y)$ is the gray value of the (x, y) coordinate point on the image, S_x and S_y represent the gray value of the horizontal and

vertical edge detection of the image, respectively. The grayscale values are calculated as follows:

$$S_x = \begin{bmatrix} -1 & 0 & -1 \\ -2 & 0 & -2 \\ -1 & 0 & -1 \end{bmatrix} \times f(x,y) \quad (2)$$

$$S_y = \begin{bmatrix} +1 & +2 & -1 \\ 0 & 0 & 0 \\ -1 & -2 & -1 \end{bmatrix} \times f(x,y) \quad (3)$$

The approximate gradient M and the gradient direction θ of the grayscale at each pixel point of the image are calculated by combining the horizontal and vertical grayscale values of the point by applying the square root.

$$M = \sqrt{S_x{}^2 + S_y{}^2} \quad (4)$$

$$\theta = \arctan\left(\frac{S_y}{S_x}\right) \quad (5)$$

If the approximate gradient M is greater than a certain threshold, the point $(x,\ y)$ is considered an edge point. The Sobel edge detection feature of the UAV image in this study is calculated by an operation in ENVI5.3. After several experiments and comparisons, the calculated image enhancement parameters are set to linear 0–255 and the filter parameters are set to 18 sharpening degrees, which can make the image boundary more clearly displayed.

3.2. Improved Mask R-CNN for UAV Image Roof Type Recognition

This section describes the specific process of applying deep learning algorithms and theories, including an overview of the network architecture of Mask R-CNN based on ResNet152, the production and processing of sample sets, model implementation and training.

3.2.1. Migration Learning Deployment of ResNet152

When deep learning is used for building roof type recognition, the number of network layers of the neural network is crucial for the extraction of roof type features, which can extract higher-level abstract features such as texture, shape and color features when the pattern of the roof type features is not obvious. However, simply increasing the network depth can easily lead to gradient disappearance and network degradation problems, which make the image classification accuracy decrease rapidly after saturation [63]. These issues are addressed by ResNet [64], which can also reduce training errors while deepening the network by introducing identity mapping between layers. ResNet152 is one of the networks with deeper layers in the ResNet, which can effectively use the multilayer information of the network even though the number of layers is deeper, due to its lower complexity and better ability to extract features. Therefore, this study uses ResNet152 as the base network for complex feature extraction of roof types of buildings in the countryside of UAV images. Figure 4 shows the ResNet152 network structure, where from the first group of convolutional blocks up to the fifth group are residual modules. After inputting the image with an image size of 224 × 224, the final group output size is reduced to 7 × 7 by learning the features extracted from the training residual network, then the trained image is input to the average pooling layer to take the average, and finally, the softmax function of the fully connected layer is used to classify the image categories.

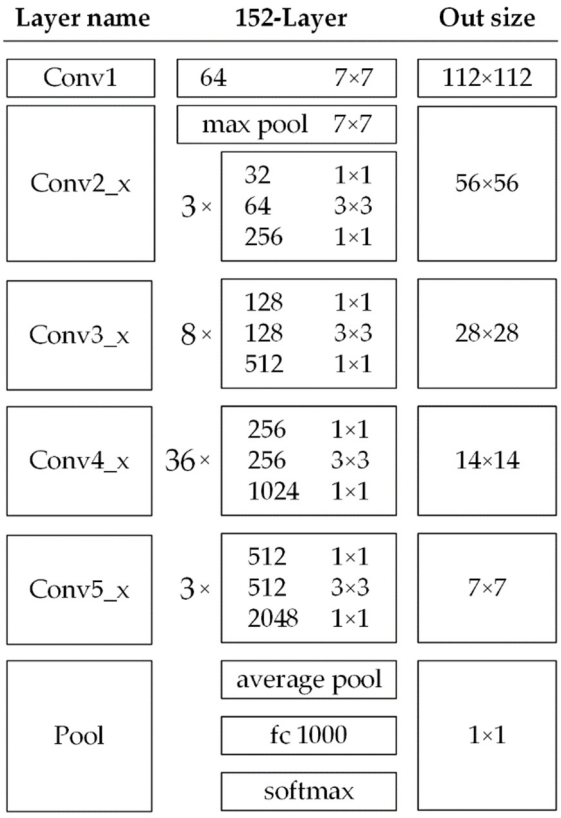

Figure 4. The network structure of ResNet152.

Transfer learning is a very effective method proposed to solve the problem of overfitting in the training process of neural network learning for small data volumes. It improves the efficiency and accuracy of small data classification problems by saving the feature parameters pre-trained in large datasets (such as ImageNet, etc.) and then applying them to the new target classification task to be solved, through the portability of feature model weights between different classification datasets. The two main common migration learning methods are feature migration and model migration [65]. In this study, we use model migration to migrate the ResNet152 pre-trained model, which is fully trained in the ImageNet dataset, to the feature extraction layer of Mask R-CNN, and then re-initialize the parameters of the last layer of the ResNet152 pre-trained model, while the other layers directly use the weight parameters of the pre-trained network and freeze them, and then use the rural building roof as a landmark. The model is then fine-tuned using the rural building roof type dataset to achieve optimal training of the building roof type recognition model.

3.2.2. Construction of Improved Mask R-CNN Model

Mask R-CNN is a widely used and efficient multi-task instance segmentation framework for integrated target detection and semantic segmentation, which is based on R-CNN [66], Fast R-CNN [67] and Faster R-CNN [68]. Mask R-CNN adds a branch using Full Convolutional Network (FCN) to Faster R-CNN to predict the segmentation mask, making it juxtaposed with the original bounding box layer and classification layer, and it can accurately detect the target class and location information in the image. In addition, Mask R-CNN uses region of interest (RoI) Align to optimize the spatial location misalignment problem caused by the RoI pooling layer, and by introducing a bilinear interpolation algorithm, each RoI is better aligned to the location of pixels on the original input image to achieve accurate pixel-level target segmentation. The network structure of Mask R-CNN used in this paper is shown in Figure 5, and the steps of building roof type recognition based on the improved Mask R-CNN are as follows:

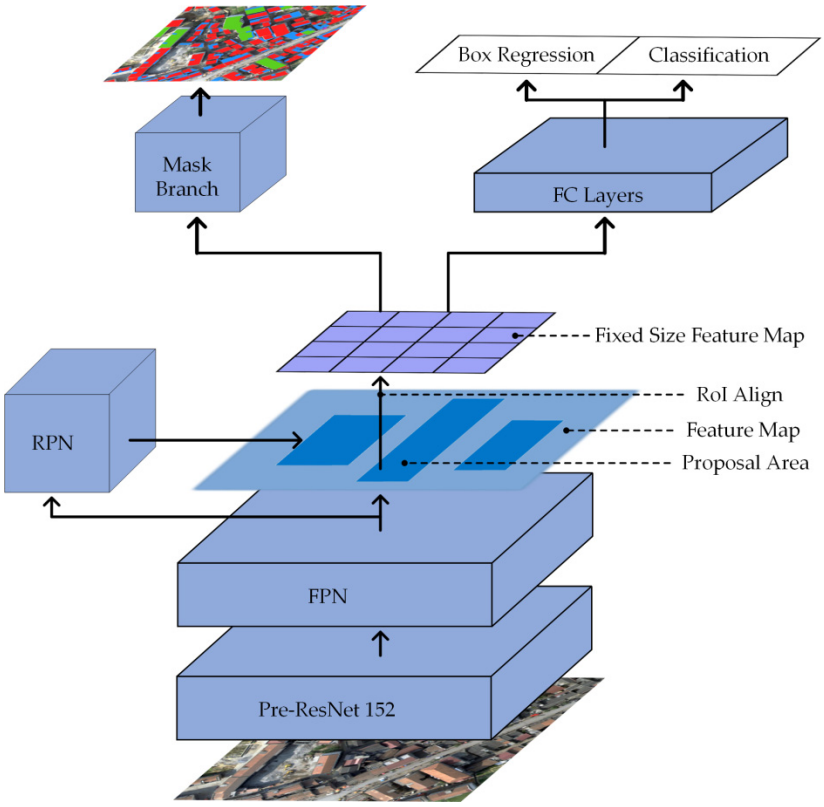

Figure 5. The network structure of improved Mask R-CNN.

(1) Input a pre-processed UAV remote sensing image of a specific size into the pre-trained ResNet152 network to obtain the corresponding feature maps.
(2) Assign a fixed number of RoIs to each point on the feature map, resulting in multiple RoIs.
(3) Transfer these candidate RoIs to the RPN network for binary classification (foreground and background) and fine-tuning of the location and size of the bounding box to obtain a more accurate bounding box for better fitting of the target. Simultaneously, filter out some of the candidate RoIs by using non-maximal value suppression.
(4) Run the RoI Align operation on the remaining RoIs, that is, first mapping the feature map's pixels to the original map, and then mapping the feature map to the fixed features.
(5) Finally, these RoIs are subjected to multi-category classification, bounding box regression, and mask generation by FCN in the sub-network.

3.2.3. Model Implementation and Training

(1) Software and hardware environment configuration: The computer used in this experiment is equipped with a 3.7 GHz octa-core Intel Core i9-10900K CPU, an 11 GB NVIDIA GeForce GTX 2080 Super graphics card, a 32 GB memory stick and Windows 10 as the operating system. The neural network design framework used in this paper is the Pytorch deep learning framework.
(2) Construction of a sample dataset of rural building roof types: We cropped the seven images and then calculated the spectral visual features and spatial visual features of the sample area images according to the method described in Section 3.1, and combined them with the original UAV visible band images for different features. We used ArcGIS Pro 2.8 to manually visually interpret the sample labeling of each representative roof type in these combined features and cross-checked it with multiple people to ensure the accuracy of the sample types, including the gabled type labeled as 1, flat type labeled as 2, hipped type labeled as 3, complex type labeled

as 4 and mono-pitched labeled as 5. The labeled images are converted to GeoTIFF format, which is used as the reference standard for training sample data and model accuracy verification of the deep learning model. Due to the hardware limitation in the training of the deep learning network model, the image needs to be segmented into several small pieces. Based on the random strategy [69], a 224 × 224 area is randomly intercepted from the manually labeled sample area as the input image for the training model.

(3) Data enhancement and sample data set assignment: To expand the training sample size of the UAV remote image dataset to avoid the model overfitting problem, we randomly select 50% of the images from the training dataset for data enhancement, and get 1.5 times the amount of image data as the original training data. These enhancement methods include rotate, crop, brightness enhancement, contrast enhancement, and scaling. In this paper, we use two regions, T1 and T2, as test regions, and other regions as training and validation regions, with the training sample set, validation sample set, and test sample set divided in the ratio of 5:1:4. It should be noted that because the number of hipped, complex and mono-pitched types of roofs on buildings in rural areas is small, which easily causes the unbalanced extraction of features from different roof types by the model, we conduct separate secondary training for the above three datasets of roof types with a small number of training samples, and the secondary training classification results are jointly processed with the full type training classification results as the final classification result output.

(4) Feature combination training mode setting: In this paper, different visual feature images are combined with UAV visible RGB band images as different feature combinations, and they are input to the model for training to analyze the performance of the model under several different feature combinations. The image features are divided into four different combinations as input layers: RGB, RGB + Sobel, RGB + VDVI and RGB + VDVI + Sobel.

(5) Model training parameters setting: After comparing the experimental results with several parameter selections, the improved Mask R-CNN deep learning network uses the average binary cross entropy as the loss function, which allows the generation of masks for each class, and there is no inter-class competition. The weight decay coefficient is 0.0001, the momentum coefficient is 0.9, the activation function is sigmoid, the batch size is set to 8, the epoch is set to 20, the initial learning rate is 0.001, and the optimization method uses the stochastic gradient descent (SGD) method, which can accelerate the convergence of the network.

3.2.4. Accuracy Evaluation Method

In this paper, the evaluation of the model includes two aspects: first, the accurate evaluation of the improved Mask R-CNN classification results in terms of their agreement with the true values; secondly, the feature applicability evaluation to determine the impact of different visual feature combinations on the accuracy of the recognition results of roof types of buildings in the countryside of UAV images. According to the combination of the true category and model classification category, the results can be classified into four cases: true positive (*TP*), false negative (*FN*), false positive (*FP*) and true negative (*TN*). The number of pixels correctly classified as positive samples is denoted by *TP*; the number of pixels correctly classified as negative samples is denoted by *FN*; the number of pixels with errors for negative samples is denoted by *FP*; and the number of pixels with errors for positive samples is denoted by *TN*. These values can be calculated using the pixel-based confusion matrix [70]. Based on the above calculation results, we use five accuracy evaluation methods, namely *Precision*, *Recall*, *F1-score*, Overall Accuracy (*OA*) and Kappa coefficient (*KC*), to check the overall prediction performance of the algorithm for different roof types. *Precision* is the ratio of the number of correctly classified positive samples to the number of all positive samples classified by the classifier. *Recall* is the ratio of the number of correctly classified positive samples to the number of all actual positive samples. In

practice, *Precision* sometimes contradicts *Recall* so we use the *F1-score* metric, which is the summed average of *Precision* and *Recall*. *OA* is the probability that the classified result is consistent with the actual type of the region on the ground. *KC* is used for consistency testing, which can be a better measure of classification accuracy. The specific formula for each accuracy evaluation index is as follows:

$$Precision = \frac{TP}{TP + FP} \tag{6}$$

$$Recall = \frac{TP}{TP + FN} \tag{7}$$

$$F1\text{-}score = 2 \times \frac{Precision \times Recall}{Precision + Recall} \tag{8}$$

$$OA = \frac{TP + TN}{TP + FN + FP + TN} \tag{9}$$

$$KC = \frac{N\sum_{i=1}^{r} x_{ii} - \sum_{i=1}^{r} (x_{i+}x_{+i})}{N^2 - \sum_{i=1}^{r} (x_{i+}x_{+i})} \tag{10}$$

In the equation of *KC*, r is the total number of categories in the confusion matrix; N is the total number of pixels used for accuracy evaluation; x_{ii} is the total number of pixels correctly extracted in the confusion matrix; x_{i+} and x_{+i} are the total number of pixels for each row and column of the confusion matrix, respectively.

4. Experimental Results

To verify the superiority of the improved Mask R-CNN model, the accuracy of the recognition results and the applicability of different combinations of visual features for the recognition of roof types of buildings in the countryside in UAV images were compared and evaluated. First, to investigate the applicability of different input feature combinations for the recognition of different roof types, we evaluate the effects of different feature combinations on the recognition results of single rural building roof types and the overall recognition results of roof types using the improved Mask R-CNN model, respectively, to verify the positive effects of different visual feature combinations in the recognition of complex roof types. Second, to evaluate the performance of the improved Mask R-CNN model, we trained the model on the feature combination images with the highest accuracy of roof type recognition and compared it with the original Mask R-CNN, U-Net [71], DeeplabV3 [72] and PSPNet [73] models, and we also verified the impact of different models on the roof type recognition results of single rural buildings and the overall roof type.

4.1. Accuracy Comparison of Roof Recognition with Different Feature Combinations Based on the Improved Mask R-CNN

4.1.1. Comparison of Accuracy of Roof Type Recognition Results of Single Rural Buildings with Different Feature Combinations

Spectral information and spatial information are significant features for remote sensing image classification and recognition. Based on the improved Mask R-CNN model, this paper compares the roof type recognition effects of two visual features, Sobel and VDVI, combined with UAV visible images and evaluates the impact of spectral- and spatial-based visual features on the recognition accuracy of complex building roof types. The improved Mask R-CNN model is used to conduct four sets of feature combination comparison tests: RGB, RGBS (RGB + Sobel), RGBV (RGB + VDVI) and RGBVS (RGB + VDVI + Sobel). RGB is the orthoimages in the visible band acquired by UAV, and VDVI and Sobel features are calculated from RGB images. Figure 6 shows the recognition results of single building roof types in the T1 and T2 test areas. From Figure 6d, it can be seen that the feature combination of RGBS has stronger sensitivity to the boundaries of single building roof categories, which can accurately outline the outlines of single buildings and

correctly separate the boundaries of adjacent roof types, while the recognition results of RGB band only have the problems of broken boundaries and incomplete extraction of internal information. From Figure 6d,e, it can be seen that the feature combination of RGBS and RGBV can identify flat roofs well, and in the case of vegetation distribution, the feature combination of RGBV is more advantageous than that of RGBS for distinguishing vegetation and buildings, while the recognition results of both RGBVS and RGB band only have some degree of under-recognition phenomenon.

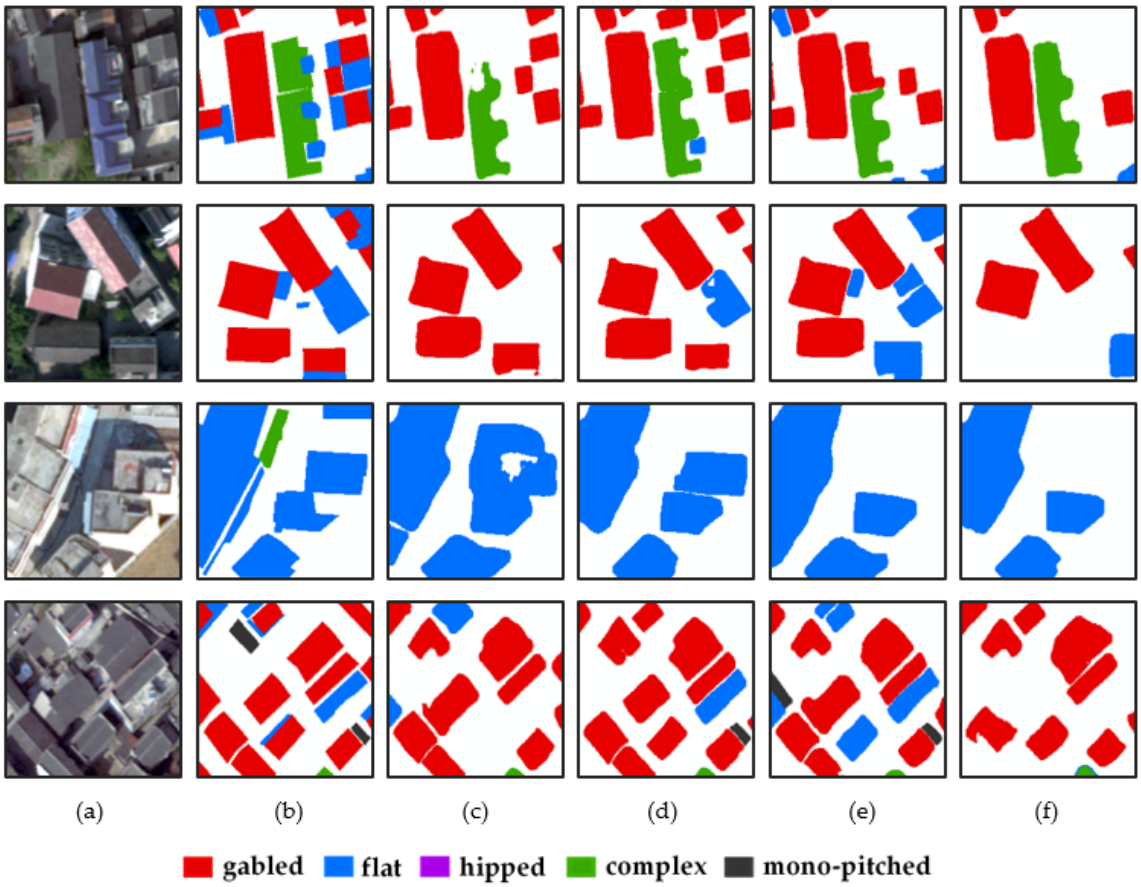

Figure 6. Roof type recognition results for single rural buildings in T1 and T2 regions are compared. (**a**) UAV image; (**b**) labeled image; (**c**) RGB; (**d**) RGB + Sobel; (**e**) RGB + VDVI; and (**f**) RGB + VDVI + Sobel.

The recognition accuracy of roof types of single rural buildings with different feature combinations is shown in Table 2. When RGB is combined with VDVI or with Sobel for features, the *Precision*, *Recall* and *F1-score* of each roof type are improved to some extent, among which the feature combination of RGBS is better. Specifically, the *F1-score* of RGBS in the T1 test area for each roof type improved by a minimum of 0.03 and a maximum of 0.18 compared to the test results for the RGB band only. In the T2 test area, RGBS shows the highest average *F1-score*, with superior recognition for gabled and flat roof types in particular and better recognition of other different complex roof types in the area. The RGBV feature combination is also slightly better than the RGB band-only recognition results, with an improved *F1-score* of at least 0.02 and at most 0.11, but the accuracy of RGBV is slightly lower than that of RGBS for the gabled and flat roof types. In addition, there is a certain degree of accuracy degradation in the recognition of each roof type by the combination of RGBVS features, which indicates that too many feature combinations may not necessarily improve the accuracy of feature recognition, but may lead to an overall decrease in accuracy.

Table 2. Comparison of recognition accuracy of single roof categories with different feature combinations.

Feature	Type	T1			T2		
		Precision	*Recall*	*F1-Score*	*Precision*	*Recall*	*F1-Score*
RGB	gabled	83.5%	76.4%	0.798	95.6%	93.9%	0.947
	flat	86.4%	98.1%	0.919	92.5%	90.0%	0.912
	hipped	47.5%	64.9%	0.549	73.8%	47.4%	0.577
	complex	90.9%	35.9%	0.515	63.7%	80.9%	0.713
	mono-pitched	77.8%	31.1%	0.444	56.3%	54.5%	0.554
RGB + Sobel	gabled	90.0%	82.4%	0.860	96.2%	92.1%	0.941
	flat	90.3%	98.0%	0.940	88.5%	91.6%	0.900
	hipped	56.3%	94.7%	0.706	69.7%	74.3%	0.719
	complex	97.7%	77.0%	0.861	67.8%	85.4%	0.756
	mono-pitched	37.5%	66.7%	0.480	54.2%	46.4%	0.500
RGB + VDVI	gabled	84.5%	81.0%	0.827	95.0%	81.3%	0.876
	flat	89.6%	98.2%	0.937	69.8%	94.7%	0.804
	hipped	65.5%	59.4%	0.623	67.3%	55.4%	0.608
	complex	74.3%	34.9%	0.475	63.6%	81.0%	0.713
	mono-pitched	93.3%	37.8%	0.538	53.1%	44.7%	0.485
RGB + VDVI + Sobel	gabled	86.0%	47.0%	0.608	93.1%	85.8%	0.893
	flat	80.6%	99.1%	0.889	62.9%	94.8%	0.756
	hipped	29.9%	92.1%	0.451	83.3%	37.5%	0.517
	complex	90.2%	44.0%	0.591	33.3%	1.0%	0.019
	mono-pitched	38.5%	12.5%	0.189	1.0%	1.0%	0.010

4.1.2. Comparison of the Overall Recognition Result Accuracy for Different Feature Combinations

The overall recognition results of roof types of rural buildings with different feature combinations using the improved Mask R-CNN are shown in Figures 7 and 8. It can be seen that the RGBS feature combinations in the T1 and T2 regions can identify more roof types and ensure the number of extracted roofs in the region, while all other feature combination methods have a considerable degree of missed extraction. From Figures 7c and 8c, it can be seen that the RGB band features by themselves can maintain high accuracy in extracting to different building roof types, but the RGB band features cannot accurately depict the gaps between different building roof types in dense building areas, and there are cases of misclassifying farmland plots into gabled roof types. The feature combination of RGBS improves this situation, and Figures 7d and 8d demonstrate the high performance of this feature combination in identifying medium-sized building roofs, extracting the shape of each building roof type well and separating them. However, in the case of insufficient Sobel feature detection, there are also some feature recognition errors, as shown in Figure 8d, which may not accurately identify the roof types at vegetation shading. While RGBV can identify the difference between each roof type and vegetation in this case, as shown in Figure 8e, the combination of RGBV features can improve the recognition of roof types that are heavily shaded by vegetation.

The overall recognition accuracies of roofs with different feature combinations are shown in Table 3. The results show that the feature combination of RGBS has the highest roof type recognition accuracy in both T1 and T2 test areas, with improvements of 0.105, 0.115 and 0.061, and 0.05, 0.115 and 0.075 over the *F1-score*, *KC* and *OA* of RGB band features, respectively. The roof type identification with the combination of RGBV features also has a higher *F1-score*, *KC* and *OA* than the RGB band features, improving by 0.023, 0.042 and 0.028, and 0.042, 0.097 and 0.078, respectively. In addition, the roof recognition accuracy of the RGBVS feature combination in both test areas is significantly lower than that of the RGB band features, indicating that too many feature inputs may instead hinder the model from extracting image features, resulting in low accuracy recognition. In contrast,

using the right combination of features can improve the recognition accuracy of roof types in UAV visible band images to a certain extent.

Table 3. Comparison of the overall recognition accuracy of different feature combinations.

Feature	T1					T2				
	Precision	*Recall*	*F1-Score*	*KC*	*OA*	*Precision*	*Recall*	*F1-Score*	*KC*	*OA*
RGB	77.2%	61.3%	0.683	0.716	0.842	69.8%	71.4%	0.706	0.696	0.832
RGB + Sobel	74.4%	83.8%	0.788	0.831	0.903	75.3%	78.0%	0.766	0.811	0.907
RGB + VDVI	81.4%	62.3%	0.706	0.758	0.870	76.4%	73.3%	0.748	0.793	0.910
RGB + VDVI + Sobel	65.0%	58.9%	0.618	0.565	0.791	54.7%	44.0%	0.488	0.626	0.827

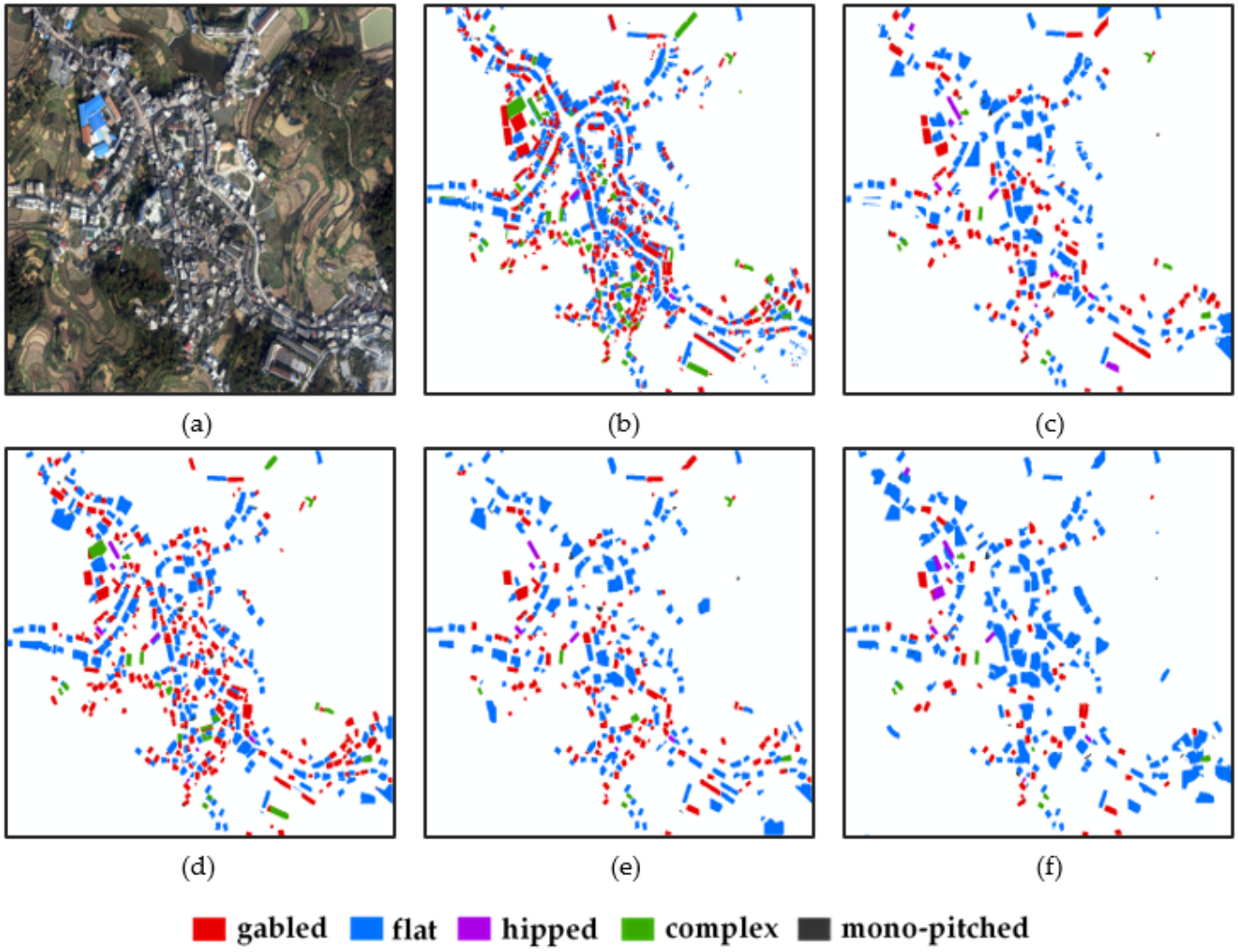

Figure 7. T1 test area identification results. (**a**) UAV image; (**b**) labeled image; (**c**) RGB; (**d**) RGB + Sobel; (**e**) RGB + VDVI; and (**f**) RGB + VDVI + Sobel.

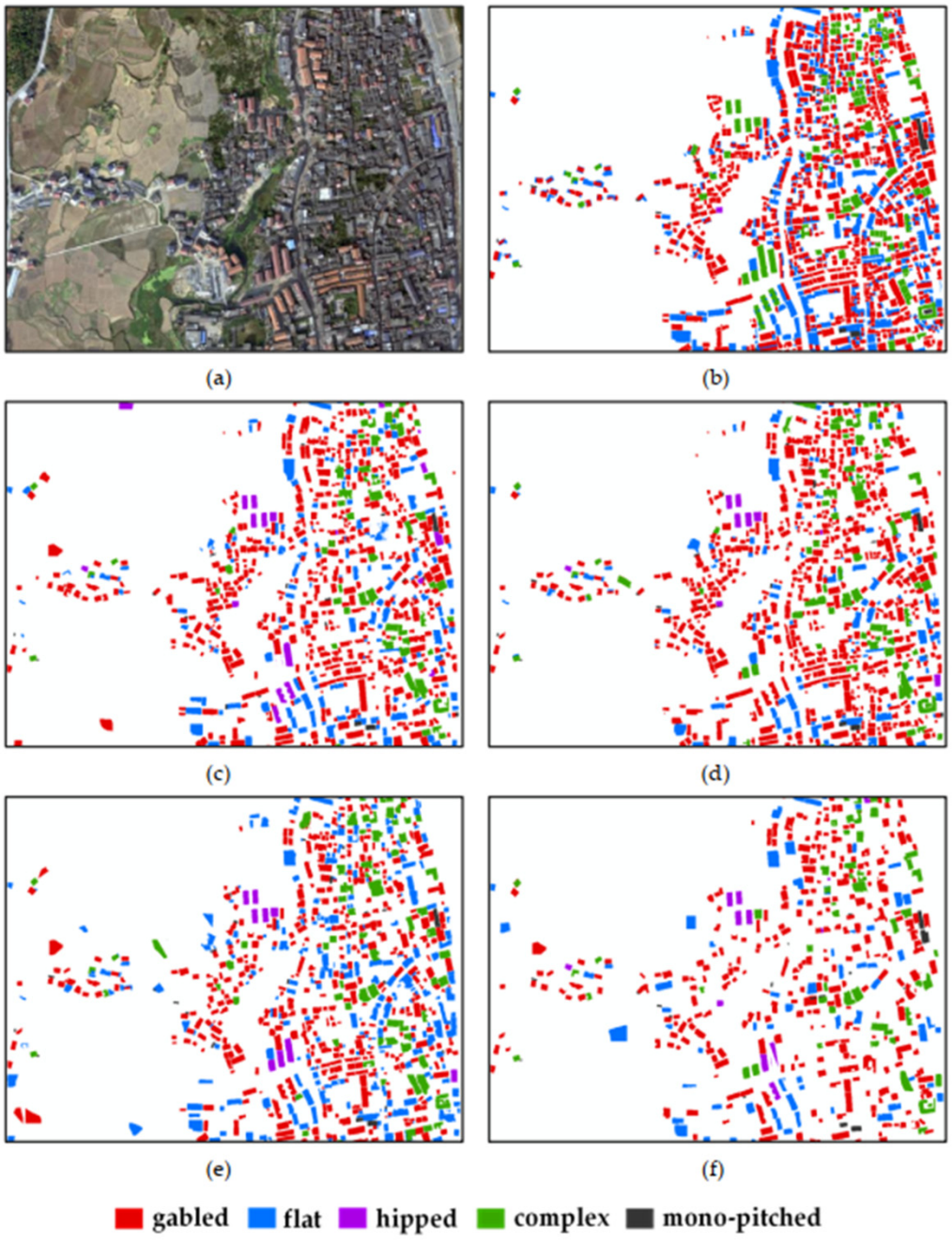

Figure 8. T2 test area identification results. (**a**) UAV image; (**b**) labeled image; (**c**) RGB; (**d**) RGB + Sobel; (**e**) RGB + VDVI; and (**f**) RGB + VDVI + Sobel.

4.2. Comparison of Roof Recognition Accuracy with Other Deep Learning Models

4.2.1. Comparison of Roof Type Recognition Results of Single Buildings with Different Deep Learning Models

The RGBS feature combination dataset with the highest accuracy for roof type recognition is input to the input layer of the improved Mask R-CNN model for training, and the performance of the improved Mask R-CNN model is evaluated by comparing the recognition results and accuracy with those of the original Mask R-CNN, U-Net, DeeplabV3 and PSPNet. The original Mask R-CNN uses ResNet50 as the feature extraction layer, which can also obtain good results in roof type recognition. U-Net has good applications in medical image recognition and has also achieved good results in remote sensing image building recognition. DeeplabV3 uses the ASPP module to mine convolutional features and image layer features at different scales, which has wide applications in high-resolution remote

sensing. DeeplabV3 has a wide range of applications in high-resolution remote sensing image classification [74]. PSPNet is able to aggregate global contextual information from different sub-region images and is suitable for image segmentation of buildings in different complex scenes. The evaluation metrics for the recognition results of roof types of single buildings with different deep learning models in T1 and T2 test areas are calculated and shown in Table 4. The mean *F1-score* of the improved Mask R-CNN is higher than the other models in the recognition of gabled, flat, hipped and complex types of roofs, indicating that it has a greater advantage in the recognition of different roof types. Although the recognition accuracy of the original Mask R-CNN model for different building roof types is not as high as that of the improved Mask R-CNN, its result accuracy is more stable and can also maintain a high recognition accuracy. On the other hand, the U-Net, DeeplabV3 and PSPNet models all show very low recognition accuracy on hipped, complex and mono-pitched types of roofs, indicating that there are still limitations in using only semantic segmentation networks for recognizing complex building roof types.

Table 4. Comparison of the accuracy of single building roof type recognition with other deep learning models.

Model	Type	T1			T2		
		Precision	*Recall*	*F1-Score*	*Precision*	*Recall*	*F1-Score*
Mask R-CNN	gabled	84.1%	92.3%	0.880	92.4%	87.2%	0.897
	flat	74.4%	71.1%	0.727	73.5%	92.6%	0.820
	hipped	40.0%	50.0%	0.444	61.2%	30.0%	0.403
	complex	90.9%	63.8%	0.750	69.7%	65.2%	0.674
	mono-pitched	65.4%	68.0%	0.667	66.7%	31.3%	0.426
U-Net	gabled	70.2%	93.0%	0.800	85.2%	95.1%	0.899
	flat	93.7%	86.5%	0.900	85.4%	79.6%	0.824
	hipped	66.7%	2.6%	0.050	51.9%	17.8%	0.265
	complex	81.0%	18.4%	0.300	45.1%	18.0%	0.257
	mono-pitched	20.0%	1.4%	0.026	100.0%	12.3%	0.219
DeeplabV3	gabled	76.8%	86.0%	0.811	86.9%	95.9%	0.912
	flat	95.4%	90.0%	0.926	89.0%	79.6%	0.840
	hipped	18.5%	41.9%	0.257	75.0%	8.2%	0.148
	complex	18.2%	4.4%	0.071	49.5%	33.9%	0.402
	mono-pitched	2.3%	3.9%	0.029	72.7%	16.0%	0.262
PSPNet	gabled	74.7%	90.0%	0.816	85.5%	95.9%	0.904
	flat	91.3%	92.2%	0.917	86.8%	86.9%	0.868
	hipped	34.2%	15.3%	0.211	44.3%	17.3%	0.249
	complex	40.6%	9.2%	0.150	48.1%	8.1%	0.139
	mono-pitched	16.7%	1.5%	0.028	20.0%	1.8%	0.033
Our Model	gabled	90.0%	82.4%	0.860	96.2%	92.1%	0.941
	flat	90.3%	98.0%	0.940	88.5%	91.6%	0.900
	hipped	56.3%	94.7%	0.706	69.7%	74.3%	0.719
	complex	97.7%	77.0%	0.861	67.8%	85.4%	0.756
	mono-pitched	37.5%	66.7%	0.480	54.2%	46.4%	0.500

4.2.2. Overall Recognition Accuracy of Roofs Compared to Other Deep Learning Models

The overall recognition accuracy of different deep learning models on the roof types of rural buildings is shown in Table 5, and the results show that the model proposed in this paper has higher evaluation indexes than other deep learning models in both T1 and T2 test areas, with *F1-score*, *KC* and *OA* improving, respectively, by 0.095, 0.125 and 0.076 on average over the original Mask R-CNN, 0.248, 0.178 and 0.082 on average over U-Net, 0.271, 0.164 and 0.082 on average over DeeplabV3, and 0.305, 0.151 and 0.07 on average over PSPNet. Although the original Mask R-CNN, U-Net, DeeplabV3 and PSPNet have more stable extraction results in both T1 and T2 test areas, there are more false identifications

and missed identifications for roof types with smaller sample sizes and more complex features, resulting in lower accuracy of roof type recognition, and thus these models have shortcomings in robustness and generalizability.

Table 5. Comparison of the overall recognition accuracy with other deep learning models.

Model	T1					T2				
	Precision	*Recall*	*F1-Score*	*KC*	*OA*	*Precision*	*Recall*	*F1-Score*	*KC*	*OA*
Mask R-CNN	71.0%	69.0%	0.700	0.688	0.807	72.7%	61.3%	0.665	0.705	0.851
U-Net	66.3%	40.4%	0.502	0.66	0.807	73.5%	44.6%	0.555	0.626	0.838
DeeplabV3	42.2%	45.2%	0.437	0.65	0.791	74.6%	46.7%	0.575	0.663	0.854
PSPNet	51.5%	41.6%	0.460	0.682	0.82	56.9%	42.0%	0.483	0.657	0.849
Our Model	74.4%	83.8%	0.788	0.831	0.903	75.3%	78.0%	0.766	0.811	0.907

5. Discussion

5.1. Sensitivity Analysis of Different Feature Combinations on the Training Results of the Improved Mask R-CNN

To verify the effect of different feature combinations on the training curves of the improved Mask R-CNN model used in this paper, we trained the improved Mask R-CNN model with 20 epochs of learning on sample datasets with different feature combinations and obtained the loss curves of the training and validation sets during the training process. As shown in Figure 9, in terms of training efficiency, the training curves of the RGBS feature combinations exhibit a faster convergence rate, which is 40% higher than other feature combinations under the same epochs, greatly improving the model training efficiency. In terms of model stability, the training curve of the RGBS feature combination has the least fluctuation and is highly stable, which can reduce the occurrence of overfitting problems. In terms of training accuracy, the training and validation loss values of the RGBS feature combination are closer to 0.5 than those of the other feature combinations. It can be seen that the combination of Sobel features and UAV RGB images is more conducive to improving the training efficiency, stability, and accuracy of the model.

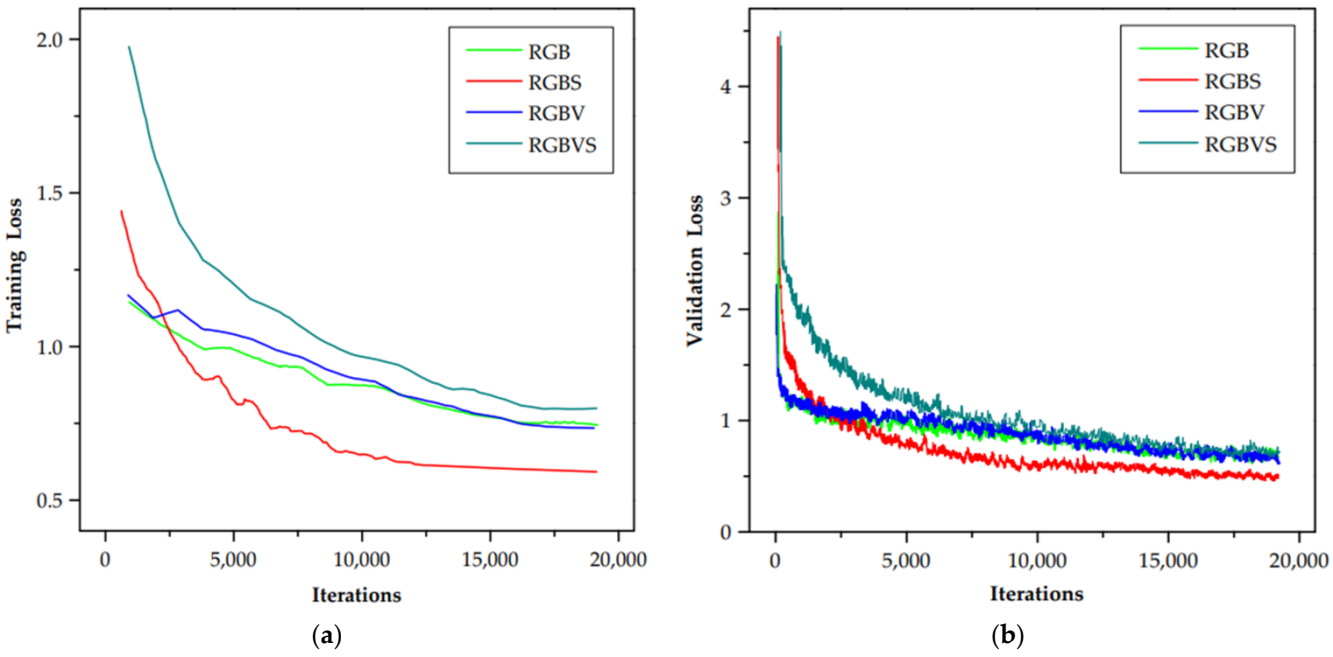

Figure 9. The effect of different feature combinations on the training results of the improved Mask R-CNN. (**a**) Training loss; and (**b**) validation loss.

5.2. Effect of Different Feature Extraction Layers of ResNet on the Accuracy of Results

Feature extraction is the key for deep learning models to maintain high accuracy in recognition results. To investigate the effect of using different typical layers of ResNet on the accuracy of Mask R-CNN models in recognizing complex rural building roof types, we used migration learning to deploy pre-trained ResNet18, ResNet34, ResNet50, ResNet101 and ResNet152 on the Mask R-CNN feature extraction layer and trained on the RGBS feature combination sample dataset with the highest recognition accuracy, and compared their recognition accuracy and efficiency of rural building roof types on T1 and T2 test areas. As shown in Table 6, using ResNet152 as the feature extraction layer of Mask R-CNN was able to obtain a much higher roof type recognition accuracy than ResNet18, ResNet34 and ResNet50. Although the recognition accuracy of the Mask R-CNN model based on ResNet101 is very close to that of ResNet152, it consumes much more training time than ResNet152, which may be due to the fact that ResNet152 has more residual blocks, which reduced the complexity of the model to extract features, thus improving the feature extraction capability and efficiency.

Table 6. Comparison of the roof type recognition accuracy of Mask R-CNN based on classical layers of different ResNet.

Model	T1					T2					*Training Time (min)*
	Precision	*Recall*	*F1-Score*	*KC*	*OA*	*Precision*	*Recall*	*F1-Score*	*KC*	*OA*	
ResNet18	39.6%	48.2%	0.435	0.443	0.596	77.0%	62.7%	0.691	0.588	0.733	187
ResNet34	53.4%	49.9%	0.516	0.512	0.633	57.2%	59.3%	0.583	0.615	0.752	198
ResNet50	71.0%	69.0%	0.700	0.688	0.807	72.7%	61.3%	0.665	0.705	0.851	211
ResNet101	69.8%	83.9%	0.762	0.808	0.884	77.1%	70.6%	0.737	0.779	0.913	301
ResNet152	74.4%	83.8%	0.788	0.831	0.903	75.3%	78.0%	0.766	0.811	0.907	220

5.3. Analysis of the Limitations of Roof Type Identification Methods for Complex Rural Buildings

With the wide application of UAV high-resolution remote sensing images, the accurate recognition of rural building roof types has gradually become possible. However, there are a large number of complex rural building roof types in UAV visible images and other features that are easily confused with building roof types, which poses a great challenge to the existing methods. Therefore, this paper proposes an improved Mask R-CNN model based on the combination of different visual features, which can effectively improve the recognition accuracy of complex rural building roof types in UAV visible images and provide a feasible reference solution for rural roof surveys. Specifically, the RGBS feature combination uses Sobel edge detection features to highlight the surface texture, shape and boundaries of different rustic building roof types, making it easier for the model to extract important features of complex building roof types, and the results show that the model based on this feature combination can completely and accurately segment the vector contours of different building roof types, clearly separating the gaps between buildings. The RGBV feature combination, on the other hand, uses VDVI features to highlight vegetation areas to distinguish them from buildings, reducing misclassification of dense areas of rural buildings, such as small building roofs covered by trees and green pads. In addition, we compare the accuracy of the improved Mask R-CNN model with other deep learning models for roof type recognition, and the results also show that the improved Mask R-CNN model exhibits the highest accuracy and the best robustness, with good performance in coping with different complex scenarios for recognizing roof types of rural buildings. However, the method proposed in this paper also has some false recognition and missed recognition for more complex building types and their morphologies (e.g., large buildings, irregularly shaped buildings, dark buildings, etc.), resulting in model recognition accuracy that is not very high, so we analyze the causes of these errors and their improvement methods.

(1) Uneven sample size across roof types

The uneven sample size of different roof types is an important reason for the error in identifying the roof types of complex buildings. In this paper, the roof types of buildings in UAV images are divided into five categories, but since rural buildings are far less numerous and dense than urban buildings, the sample size of these five categories of building roofs cannot be guaranteed to be evenly distributed, so the hipped, complex and mono-pitched roof type datasets with smaller sample sizes were trained separately and secondarily, and the results showed more accurate recognition than training the five categories together. We combined the results of training the full class dataset and the three class datasets with a smaller sample size to obtain the overall results of building roof type recognition. However, the recognition results and accuracy still did not reach a high level. This may be due to the unbalanced samples of the five categories of roof datasets, which makes the model pay more attention to the gabled and flat categories with large data volume, and the parameters in the network are optimized mainly based on the losses of these two categories, resulting in much lower test accuracy for the remaining categories; for example, rural buildings rarely have large irregular buildings with less training sample data, which reduces the model's ability to recognize these morphological building roof types (hipped and complex type roofs). Mono-pitched roof types with the same small sample size generally have more dark buildings and tend to cling to the sides of taller buildings, causing them to be obscured by shadows and other building walls, which also prevents the model from accurately extracting the features of this type of roof, resulting in low accuracy identification of mono-pitched roof types. In addition, the category with larger sample data does not mean that higher recognition accuracy can be obtained; the more samples of the category, the higher its recall rate will be, and therefore a certain accuracy will be lost accordingly [75]. For example, in the T2 test area, the recognition results of RGBV feature combinations have a higher *Recall* for flat type roofs (Table 2 and Figure 8e), but the *Precision* is 24.9% different, indicating a decrease in the *F1-score*. The recognition error problem caused by the imbalance between samples can be improved by increasing the sample size of complex roof types in other larger regions or by setting higher weights of network parameters for the categories with more complex features and smaller sample sizes.

(2) Limitations of different visual feature extraction methods

Different visual features also have limitations for feature extraction of building roof types for complex scenes. The Sobel edge detection algorithm in the RGBS feature combination used in this paper can only detect the edges in the horizontal and vertical directions of the image, which often has low detection accuracy for more complex scenes, and its detected image edges are coarse, which cannot precisely locate the location of the edge points and may generate additional background noise. Furthermore, while using VDVI in the RGBV feature combination distinguishes vegetation from buildings, it also eliminates building roof shapes and textures in densely built-up areas, resulting in a model that cannot effectively extract features of different building roof types. The Sobel algorithm can be improved to refine the detected edge features and improve its edge detection accuracy in complex scenes.

(3) Mask R-CNN structure problem

The structure of Mask R-CNN itself suffers from the problem of inadequate utilization of the features of each scale roof type. Although this paper uses the migration learning-based ResNet152 as the feature extraction layer of Mask R-CNN, there are still two problems with the structure of Mask R-CNN itself: first, the path between the highest-level features and the lowest-level features is too long, which easily leads to the loss of feature information transfer and cannot effectively utilize the lower-level features. Second, the feature mapping map input to the RPN network is only a map carrying information about itself and the higher-level features, which does not make full use of the feature information at each scale, resulting in lower detection accuracy [76]. These problems make the improved Mask R-CNN model unable to effectively utilize the extracted features of complex building

roof types at all scales, thus reducing the accuracy of the model in recognizing complex building roof types. Future research can improve the network structure of the Mask R-CNN model (e.g., FPN network) to shorten the path from low-level feature transfer to high-level mapping, reduce the feature information loss in the transfer process, and improve the feature utilization efficiency of the model, thus improving the performance of the model.

6. Conclusions

Rural areas in China account for nearly half of the Chinese population, and the survey of rural building roof types is of great significance for the planning and construction of beautiful villages in China. Aiming at the current problems that most of the UAV high-resolution remote sensing images only have a visible band, that the existing methods have difficulties extracting features of complex roof types, and that features with similar spectral features such as low reflection, obscured vegetation, and concrete roads are easily confused with building roof types, this paper proposes a method to identify rural building roof types in UAV visible images based on different combinations of visual features, and an improved Mask R-CNN deep learning model is used to improve the recognition accuracy of complex building roof types. VDVI features based on spectral vision and Sobel edge detection features based on spatial vision are combined with UAV visible images to form different feature datasets applied to a deep learning model for roof type recognition. We evaluate the recognition results of the models with four different feature combinations, RGB, RGB + Sobel, RGB + VDVI and RGB + VDVI + Sobel, and also compare the accuracy of the improved Mask R-CNN with the original Mask R-CNN, U-Net, DeeplabV3 and PSPNet deep learning models.

The results show that adding Sobel features or VDVI features to the UAV visible RGB images can improve the accuracy of the model in recognizing the roof types of rural buildings. Firstly, adding Sobel features to RGB images can identify the types and contours of different building roofs more clearly, especially in the dense building areas, and can show the gaps between different buildings well. Secondly, combining RGB images with VDVI features can effectively distinguish buildings and vegetation areas and improve the recognition accuracy of buildings obscured by vegetation. In contrast, when combining RGB images with VDVI and Sobel features together, the recognition accuracy of the model for roof types is reduced instead, indicating that too many feature combinations may not be beneficial to the recognition of building roof types. In addition, the *F1-score*, *KC* and *OA* of the improved Mask R-CNN rustic building roof type recognition results used in this paper are higher than those of other deep learning models, showing the highest accuracy and robustness in the test area.

Author Contributions: Conceptualization, Y.W. and S.L.; methodology, Y.W. and S.L.; software, Y.W. and S.L.; validation, Y.W. and S.L.; formal analysis, Y.W. and S.L.; investigation, Y.W. and S.L.; resources, Y.W. and S.L.; data curation, F.T. and H.C.; writing—original draft preparation, S.L., Y.L. and M.W.; writing—review and editing, Y.W. and S.L.; visualization, Y.W. and S.L.; supervision, Y.W. and S.L.; and project administration, Y.W. 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 (Nos. 41971423, 31972951, and 41771462), the Natural Science Foundation of Hunan Province (Nos. 2020JJ3020 and 2020JJ5164), the Science and Technology Program of Hunan Province (Nos. 2019RS2043 and 2019GK2132), and the Postgraduate Scientific Research Innovation Project of Hunan Province (No. CX20210991) and Open Fund of Hunan Provincial Key Laboratory of Geo-Information Engineering in Surveying, Mapping and Remote Sensing, Hunan University of Science and Technology (No. E22134).

Institutional Review Board Statement: Not applicable.

Informed Consent Statement: Not applicable.

Data Availability Statement: The data presented in this study are openly available in FigShare at https://doi.org/10.6084/m9.figshare.17696072.v1 (accessed on 20 November 2021).

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

References

1. Chen, D.; Loboda, T.V.; Silva, J.A.; Tonellato, M.R. Characterizing Small-Town Development Using Very High Resolution Imagery within Remote Rural Settings of Mozambique. *Remote Sens.* **2021**, *13*, 3385. [CrossRef]
2. Sun, L.; Tang, Y.; Zhang, L. Rural building detection in high-resolution imagery based on a two-stage CNN model. *IEEE Geosci. Remote Sens. Lett.* **2017**, *14*, 1998–2002. [CrossRef]
3. Varol, B.; Yılmaz, E.Ö.; Maktav, D.; Bayburt, S.; Gürdal, S. Detection of illegal constructions in urban cities: Comparing LIDAR data and stereo KOMPSAT-3 images with development plans. *Eur. J. Remote Sens.* **2019**, *52*, 335–344. [CrossRef]
4. Song, X.; Huang, Y.; Zhao, C.; Liu, Y.; Lu, Y.; Chang, Y.; Yang, J. An approach for estimating solar photovoltaic potential based on rooftop retrieval from remote sensing images. *Energies* **2018**, *11*, 3172. [CrossRef]
5. Tiwari, A.; Meir, I.A.; Karnieli, A. Object-based image procedures for assessing the solar energy photovoltaic potential of heterogeneous rooftops using airborne LiDAR and orthophoto. *Remote Sens.* **2020**, *12*, 223. [CrossRef]
6. Tu, J.; Sui, H.; Feng, W.; Sun, K.; Hua, L. Detection of damaged rooftop areas from high-resolution aerial images based on visual bag-of-words model. *IEEE Geosci. Remote Sens. Lett.* **2016**, *13*, 1817–1821. [CrossRef]
7. He, H.; Zhou, J.; Chen, M.; Chen, T.; Li, D.; Cheng, P. Building extraction from UAV images jointly using 6D-SLIC and multiscale Siamese convolutional networks. *Remote Sens.* **2019**, *11*, 1040. [CrossRef]
8. Guo, M.; Liu, H.; Xu, Y.; Huang, Y. Building extraction based on U-Net with an attention block and multiple losses. *Remote Sens.* **2020**, *12*, 1400. [CrossRef]
9. Benarchid, O.; Raissouni, N.; El Adib, S.; Abbous, A.; Azyat, A.; Achhab, N.B.; Lahraoua, M.; Chahboun, A. Building extraction using object-based classification and shadow information in very high resolution multispectral images, a case study: Tetuan, Morocco. *Can. J. Image Processing Comput. Vis.* **2013**, *4*, 1–8.
10. Schuegraf, P.; Bittner, K. Automatic building footprint extraction from multi-resolution remote sensing images using a hybrid FCN. *ISPRS Int. J. Geo-Inf.* **2019**, *8*, 191. [CrossRef]
11. Zhu, Q.; Li, Z.; Zhang, Y.; Guan, Q. Building extraction from high spatial resolution remote sensing images via multiscale-aware and segmentation-prior conditional random fields. *Remote Sens.* **2020**, *12*, 3983. [CrossRef]
12. Liao, C.; Hu, H.; Li, H.; Ge, X.; Chen, M.; Li, C.; Zhu, Q. Joint Learning of Contour and Structure for Boundary-Preserved Building Extraction. *Remote Sens.* **2021**, *13*, 1049. [CrossRef]
13. Nyandwi, E.; Koeva, M.; Kohli, D.; Bennett, R. Comparing human versus machine-driven cadastral boundary feature extraction. *Remote Sens.* **2019**, *11*, 1662. [CrossRef]
14. Chen, R.; Li, X.; Li, J. Object-based features for house detection from RGB high-resolution images. *Remote Sens.* **2018**, *10*, 451. [CrossRef]
15. Turker, M.; Koc-San, D. Building extraction from high-resolution optical spaceborne images using the integration of support vector machine (SVM) classification, Hough transformation and perceptual grouping. *Int. J. Appl. Earth Obs. Geoinf.* **2015**, *34*, 58–69. [CrossRef]
16. Li, W.; He, C.; Fang, J.; Zheng, J.; Fu, H.; Yu, L. Semantic segmentation-based building footprint extraction using very high-resolution satellite images and multi-source GIS data. *Remote Sens.* **2019**, *11*, 403. [CrossRef]
17. Zhang, C.; Jiao, J.-c.; Deng, Z.-l.; Cui, Y.-s. Individual Building Rooftop Segmentation from High-resolution Urban Single Multispectral Image Using Superpixels. *DEStech Trans. Comput. Sci. Eng.* **2019**, 188–193. [CrossRef]
18. Castagno, J.; Atkins, E. Roof shape classification from LiDAR and satellite image data fusion using supervised learning. *Sensors* **2018**, *18*, 3960. [CrossRef] [PubMed]
19. Tan, Y.; Wang, S.; Xu, B.; Zhang, J. An improved progressive morphological filter for UAV-based photogrammetric point clouds in river bank monitoring. *ISPRS J. Photogramm. Remote Sens.* **2018**, *146*, 421–429. [CrossRef]
20. Boonpook, W.; Tan, Y.; Ye, Y.; Torteeka, P.; Torsri, K.; Dong, S. A deep learning approach on building detection from unmanned aerial vehicle-based images in riverbank monitoring. *Sensors* **2018**, *18*, 3921. [CrossRef]
21. Shao, H.; Song, P.; Mu, B.; Tian, G.; Chen, Q.; He, R.; Kim, G. Assessing city-scale green roof development potential using Unmanned Aerial Vehicle (UAV) imagery. *Urban For. Urban Green.* **2021**, *57*, 126954. [CrossRef]
22. Liu, W.; Yang, M.; Xie, M.; Guo, Z.; Li, E.; Zhang, L.; Pei, T.; Wang, D. Accurate building extraction from fused DSM and UAV images using a chain fully convolutional neural network. *Remote Sens.* **2019**, *11*, 2912. [CrossRef]
23. Shao, Z.; Tang, P.; Wang, Z.; Saleem, N.; Yam, S.; Sommai, C. BRRNet: A fully convolutional neural network for automatic building extraction from high-resolution remote sensing images. *Remote Sens.* **2020**, *12*, 1050. [CrossRef]
24. Singh, P.; Verma, A.; Chaudhari, N.S. Deep convolutional neural network classifier for handwritten Devanagari character recognition. In *Information Systems Design and Intelligent Applications*; Springer: Berlin/Heidelberg, Germany, 2016; pp. 551–561.
25. Yang, H.; Wu, P.; Yao, X.; Wu, Y.; Wang, B.; Xu, Y. Building extraction in very high resolution imagery by dense-attention networks. *Remote Sens.* **2018**, *10*, 1768. [CrossRef]
26. Alidoost, F.; Arefi, H. A CNN-based approach for automatic building detection and recognition of roof types using a single aerial image. *PFG—J. Photogram. Remote Sens. Geoinfor. Sci.* **2018**, *86*, 235–248. [CrossRef]

27. Ronneberger, O.; Fischer, P.; Brox, T. U-net: Convolutional networks for biomedical image segmentation. In Proceedings of the International Conference on Medical Image Computing and Computer-Assisted Intervention, Munich, Germany, 5–9 October 2015; pp. 234–241.
28. Badrinarayanan, V.; Kendall, A.; Cipolla, R. Segnet: A deep convolutional encoder-decoder architecture for image segmentation. *IEEE Trans. Pattern Anal. Mach. Intell.* **2017**, *39*, 2481–2495. [CrossRef]
29. Li, E.; Xia, J.; Du, P.; Lin, C.; Samat, A. Integrating multilayer features of convolutional neural networks for remote sensing scene classification. *IEEE Trans. Geosci. Remote Sens.* **2017**, *55*, 5653–5665. [CrossRef]
30. Arnab, A.; Zheng, S.; Jayasumana, S.; Romera-Paredes, B.; Larsson, M.; Kirillov, A.; Savchynskyy, B.; Rother, C.; Kahl, F.; Torr, P.H. Conditional random fields meet deep neural networks for semantic segmentation: Combining probabilistic graphical models with deep learning for structured prediction. *IEEE Signal Processing Mag.* **2018**, *35*, 37–52. [CrossRef]
31. Pan, Z.; Xu, J.; Guo, Y.; Hu, Y.; Wang, G. Deep learning segmentation and classification for urban village using a worldview satellite image based on U-Net. *Remote Sens.* **2020**, *12*, 1574. [CrossRef]
32. Wu, T.; Hu, Y.; Peng, L.; Chen, R. Improved Anchor-Free Instance Segmentation for Building Extraction from High-Resolution Remote Sensing Images. *Remote Sens.* **2020**, *12*, 2910. [CrossRef]
33. Zhao, K.; Kang, J.; Jung, J.; Sohn, G. Building extraction from satellite images using mask R-CNN with building boundary regularization. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops, Salt Lake City, UT, USA, 18–22 June 2018; pp. 247–251.
34. Ji, C.; Tang, H. Number of Building Stories Estimation from Monocular Satellite Image Using a Modified Mask R-CNN. *Remote Sens.* **2020**, *12*, 3833. [CrossRef]
35. Stiller, D.; Stark, T.; Wurm, M.; Dech, S.; Taubenböck, H. Large-scale building extraction in very high-resolution aerial imagery using Mask R-CNN. In Proceedings of the 2019 Joint Urban Remote Sensing Event (JURSE), Vannes, France, 22–24 May 2019; pp. 1–4.
36. Chen, M.; Wu, J.; Liu, L.; Zhao, W.; Tian, F.; Shen, Q.; Zhao, B.; Du, R. DR-Net: An improved network for building extraction from high resolution remote sensing image. *Remote Sens.* **2021**, *13*, 294. [CrossRef]
37. Zhong, Z.; Li, J.; Ma, L.; Jiang, H.; Zhao, H. Deep residual networks for hyperspectral image classification. In Proceedings of the 2017 IEEE International Geoscience and Remote Sensing Symposium (IGARSS), Fort Worth, TX, USA, 23–28 July 2017; pp. 1824–1827.
38. Hu, Y.; Guo, F. Building Extraction Using Mask Scoring R-CNN Network. In Proceedings of the 3rd International Conference on Computer Science and Application Engineering, Sanya, China, 22–24 October 2019; pp. 1–5.
39. Yang, F.; Li, W.; Hu, H.; Li, W.; Wang, P. Multi-scale feature integrated attention-based rotation network for object detection in VHR aerial images. *Sensors* **2020**, *20*, 1686. [CrossRef]
40. Kumar, A.; Abhishek, K.; Kumar Singh, A.; Nerurkar, P.; Chandane, M.; Bhirud, S.; Patel, D.; Busnel, Y. Multilabel classification of remote sensed satellite imagery. *Trans. Emerg. Telecommun. Technol.* **2021**, *4*, 118–133. [CrossRef]
41. Zhuo, X.; Fraundorfer, F.; Kurz, F.; Reinartz, P. Optimization of OpenStreetMap building footprints based on semantic information of oblique UAV images. *Remote Sens.* **2018**, *10*, 624. [CrossRef]
42. Li, J.; Cai, X.; Qi, J. AMFNet: An attention-based multi-level feature fusion network for ground objects extraction from mining area's UAV-based RGB images and digital surface model. *J. Appl. Remote Sens.* **2021**, *15*, 036506. [CrossRef]
43. Sun, W.; Wang, R. Fully convolutional networks for semantic segmentation of very high resolution remotely sensed images combined with DSM. *IEEE Geosci. Remote Sens. Lett.* **2018**, *15*, 474–478. [CrossRef]
44. Marmanis, D.; Schindler, K.; Wegner, J.D.; Galliani, S.; Datcu, M.; Stilla, U. Classification with an edge: Improving semantic image segmentation with boundary detection. *ISPRS J. Photogramm. Remote Sens.* **2018**, *135*, 158–172. [CrossRef]
45. Boonpook, W.; Tan, Y.; Xu, B. Deep learning-based multi-feature semantic segmentation in building extraction from images of UAV photogrammetry. *Int. J. Remote Sens.* **2021**, *42*, 1–19. [CrossRef]
46. Zhang, L.; Wu, J.; Fan, Y.; Gao, H.; Shao, Y. An efficient building extraction method from high spatial resolution remote sensing images based on improved mask R-CNN. *Sensors* **2020**, *20*, 1465. [CrossRef]
47. Ji, S.; Wei, S.; Lu, M. Fully convolutional networks for multisource building extraction from an open aerial and satellite imagery data set. *IEEE Trans. Geosci. Remote Sens.* **2018**, *57*, 574–586. [CrossRef]
48. Li, W.; Li, Y.; Gong, J.; Feng, Q.; Zhou, J.; Sun, J.; Shi, C.; Hu, W. Urban Water Extraction with UAV High-Resolution Remote Sensing Data Based on an Improved U-Net Model. *Remote Sens.* **2021**, *13*, 3165. [CrossRef]
49. Zhang, X.; Fu, Y.; Zang, A.; Sigal, L.; Agam, G. Learning classifiers from synthetic data using a multichannel autoencoder. *arXiv* **2015**, arXiv:1503.03163.
50. Yan, G.; Li, L.; Coy, A.; Mu, X.; Chen, S.; Xie, D.; Zhang, W.; Shen, Q.; Zhou, H. Improving the estimation of fractional vegetation cover from UAV RGB imagery by colour unmixing. *ISPRS J. Photogramm. Remote Sens.* **2019**, *158*, 23–34. [CrossRef]
51. Jannoura, R.; Brinkmann, K.; Uteau, D.; Bruns, C.; Joergensen, R.G. Monitoring of crop biomass using true colour aerial photographs taken from a remote controlled hexacopter. *Biosyst. Eng.* **2015**, *129*, 341–351. [CrossRef]
52. Xiaoqin, W.; Miaomiao, W.; Shaoqiang, W.; Yundong, W. Extraction of vegetation information from visible unmanned aerial vehicle images. *Trans. Chin. Soc. Agric. Eng.* **2015**, *31*, 152–159.

53. Zhang, Y.; Zhang, F.; Shakhsheer, Y.; Silver, J.D.; Klinefelter, A.; Nagaraju, M.; Boley, J.; Pandey, J.; Shrivastava, A.; Carlson, E.J. A batteryless 19μW MICS/ISM-Band energy harvesting body sensor node SoC for ExG applications. *IEEE J. Solid-State Circuits* **2012**, *48*, 199–213. [CrossRef]
54. Yuan, H.; Liu, Z.; Cai, Y.; Zhao, B. Research on vegetation information extraction from visible UAV remote sensing images. In Proceedings of the 2018 Fifth International Workshop on Earth Observation and Remote Sensing Applications (EORSA), Xi'an, China, 18–20 June 2018; pp. 1–5.
55. Huang, Y.-h.; Chen, D.-w. Image fuzzy enhancement algorithm based on contourlet transform domain. *Multimed. Tools Appl.* **2020**, *79*, 35017–35032. [CrossRef]
56. Vincent, O.R.; Folorunso, O. A descriptive algorithm for sobel image edge detection. In Proceedings of the Informing Science & IT Education Conference (InSITE), Macon, GA, USA, 12–15 June 2009; pp. 97–107.
57. Ding, L.; Goshtasby, A. On the Canny edge detector. *Pattern Recognit.* **2001**, *34*, 721–725. [CrossRef]
58. Burt, P.J.; Adelson, E.H. The Laplacian pyramid as a compact image code. In *Readings in Computer Vision*; Elsevier: Amsterdam, The Netherlands, 1987; pp. 671–679.
59. Xu, Y.; Wu, L.; Xie, Z.; Chen, Z. Building extraction in very high resolution remote sensing imagery using deep learning and guided filters. *Remote Sens.* **2018**, *10*, 144. [CrossRef]
60. Ma, G.; He, Q.; Shi, X.; Fan, X. Automatic Vectorization Extraction of Flat-Roofed Houses Using High-Resolution Remote Sensing Images. In Proceedings of the IGARSS 2019–2019 IEEE International Geoscience and Remote Sensing Symposium, Yokohama, Japan, 28 July–2 August 2019; pp. 44–47.
61. Teng, L.; Xue, F.; Bai, Q. Remote sensing image enhancement via edge-preserving multiscale retinex. *IEEE Photonics J.* **2019**, *11*, 1–10. [CrossRef]
62. Chen, L.-C.; Papandreou, G.; Schroff, F.; Adam, H. Rethinking atrous convolution for semantic image segmentation. *arXiv* **2017**, arXiv:1706.05587.
63. Xu, L.; Chen, Q. Remote-sensing image usability assessment based on ResNet by combining edge and texture maps. *IEEE J. Sel. Top. Appl. Earth Obs. Remote Sens.* **2019**, *12*, 1825–1834. [CrossRef]
64. 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; pp. 770–778.
65. Li, D.; Deng, L.; Lee, M.; Wang, H. IoT data feature extraction and intrusion detection system for smart cities based on deep migration learning. *Int. J. Inf. Manag.* **2019**, *49*, 533–545. [CrossRef]
66. Girshick, R.; Donahue, J.; Darrell, T.; Malik, J. Rich feature hierarchies for accurate object detection and semantic segmentation. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Columbus, OH, USA, 23–28 June 2014; pp. 580–587.
67. Girshick, R. Fast r-cnn. In Proceedings of the IEEE International Conference on Computer Vision, Santiago, Chile, 7–13 December 2015; pp. 1440–1448.
68. Ren, S.; He, K.; Girshick, R.; Sun, J. Faster r-cnn: Towards real-time object detection with region proposal networks. *Adv. Neural Inf. Processing Syst.* **2015**, *28*, 91–99. [CrossRef] [PubMed]
69. Tuia, D.; Muñoz-Marí, J.; Camps-Valls, G. Remote sensing image segmentation by active queries. *Pattern Recognit.* **2012**, *45*, 2180–2192. [CrossRef]
70. Li, M.; Wu, P.; Wang, B.; Park, H.; Yang, H.; Wu, Y. A Deep Learning Method of Water Body Extraction From High Resolution Remote Sensing Images With Multisensors. *IEEE J. Sel. Top. Appl. Earth Obs. Remote Sens.* **2021**, *14*, 3120–3132. [CrossRef]
71. Guo, H.; Shi, Q.; Du, B.; Zhang, L.; Wang, D.; Ding, H. Scene-driven multitask parallel attention network for building extraction in high-resolution remote sensing images. *IEEE Trans. Geosci. Remote Sens.* **2020**, *59*, 4287–4306. [CrossRef]
72. Chen, L.-C.; Zhu, Y.; Papandreou, G.; Schroff, F.; Adam, H. Encoder-decoder with atrous separable convolution for semantic image segmentation. In Proceedings of the European Conference on Computer Vision (ECCV), Munich, Germany, 8–14 September 2018; pp. 801–818.
73. Zhao, H.; Shi, J.; Qi, X.; Wang, X.; Jia, J. Pyramid scene parsing network. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Honolulu, HI, USA, 21–26 July 2017; pp. 2881–2890.
74. Zhu, Y.; Liang, Z.; Yan, J.; Chen, G.; Wang, X. ED-Net: Automatic Building Extraction From High-Resolution Aerial Images With Boundary Information. *IEEE J. Sel. Top. Appl. Earth Obs. Remote Sens.* **2021**, *14*, 4595–4606. [CrossRef]
75. Diamond, N.B.; Armson, M.J.; Levine, B. The truth is out there: Accuracy in recall of verifiable real-world events. *Psychol. Sci.* **2020**, *31*, 1544–1556. [CrossRef]
76. Yang, X.; Sun, H.; Fu, K.; Yang, J.; Sun, X.; Yan, M.; Guo, Z. Automatic ship detection in remote sensing images from google earth of complex scenes based on multiscale rotation dense feature pyramid networks. *Remote Sens.* **2018**, *10*, 132. [CrossRef]