International Journal of
Geo-Information

Article

Residual Multi-Attention Classification Network for A Forest Dominated Tropical Landscape Using High-Resolution Remote Sensing Imagery

Tong Yu [1,2], Wenjin Wu [1,3], Chen Gong [1,*] and Xinwu Li [1,3]

[1] Key Laboratory of Digital Earth Science, Aerospace Information Research Institute, Chinese Academy of Sciences, Beijing 100094, China; yutong@aircas.ac.cn (T.Y.); wuwj@radi.ac.cn (W.W.); lixw@aircas.ac.cn (X.L.)
[2] College of Resources and Environment, University of Chinese Academy of Sciences, Beijing 100049, China
[3] Sanya Institute of Remote Sensing, Sanya 572029, China
[*] Correspondence: gongchen@radi.ac.cn

Abstract: Tropical forests are of vital importance for maintaining biodiversity, regulating climate and material cycles while facing deforestation, agricultural reclamation, and managing various pressures. Remote sensing (RS) can support effective monitoring and mapping approaches for tropical forests, and to facilitate this we propose a deep neural network with an encoder–decoder architecture here to classify tropical forests and their environment. To deal with the complexity of tropical landscapes, this method utilizes a multi-scale convolution neural network (CNN) to expand the receptive field and extract multi-scale features. The model refines the features with several attention modules and fuses them through an upsampling module. A two-stage training strategy is proposed to alleviate misclassifications caused by sample imbalances. A joint loss function based on cross-entropy loss and the generalized Dice loss is applied in the first stage, and the second stage used the focal loss to fine-tune the weights. As a case study, we use Hainan tropical reserves to test the performance of this model. Compared with four state-of-the-art (SOTA) semantic segmentation networks, our network achieves the best performance with two Hainan datasets (mean intersection over union (MIoU) percentages of 85.78% and 82.85%). We also apply the new model to classify a public true color dataset which has 17 semantic classes and obtain results with an 83.75% MIoU. This further demonstrates the applicability and potential of this model in complex classification tasks.

Keywords: remote sensing; deep convolution network; image analysis; land use and land cover (LULC); tropical forest

check for updates

Citation: Yu, T.; Wu, W.; Gong, C.; Li, X. Residual Multi-Attention Classification Network for A Forest Dominated Tropical Landscape Using High-Resolution Remote Sensing Imagery. *ISPRS Int. J. Geo-Inf.* **2021**, *10*, 22. https://doi.org/10.3390/ijgi10010022

Received: 15 December 2020
Accepted: 7 January 2021
Published: 11 January 2021

1. Introduction

Tropical forests, which are the most abundant and complex forest ecosystem, are crucial in regulating the global climate and providing various ecosystem services. Despite their importance, these forests are also heavily threatened by deforestation, plantations, and other human activities [1,2]. Because of the inherent complexity and density while traversing tropical forests, field surveys are very inefficient and it is necessary to use an automatic classification method to dynamically monitor forest resources. As an efficient large-scale observation technology, remote sensing has obvious advantages in forest resource monitoring. By extracting and analyzing high-resolution spatial information and spectral information contained in images, such technology is able to classify land cover in forest areas with high precision. Many researchers have applied remote sensing-related technology to tropical forest research, such as mapping tropical forest classes [3,4], monitoring deforestation and degradation [5], and biomass estimation [6], among which the accurate classification of land cover is the basis of various studies.

Traditional remote sensing classification methods include random forest, k-nearest neighbor, support vector machine, and other machine learning algorithms. Since Hinton's article was published [7], deep learning methods have become a hot research area for image processing and have shown great advantages over traditional classification methods. Among deep learning methods, various convolutional neural network (CNN) models have gained high popularity for scene classification [8–12]. Semantic segmentation networks, also referred to classification networks in remote sensing contexts, have been developed from these networks. Fully convolutional networks (FCNs) [13], SegNet [14], U-Net [15], DeepLabv3+ [16], and various end-to-end networks have been proposed to improve the classification accuracy of deep neural networks. With the deepening of networks, feature extraction capabilities have also been strengthened. On this basis, some researchers have introduced attention mechanisms in image classification [17,18]. The attention modules make the network focus on the key areas and highlight the features and details. These gradually developed networks provide frameworks for remote sensing image classification [19]. Semantic segmentation networks, especially convolution neural networks, are widely used in land use and land cover (LULC) classifications [20–23], road extraction [24,25], building extraction [26,27], etc. At present, there are still many problems to be solved in the classification of tropical forest area and its surroundings based on deep neural networks. Large tropical forests are tall and dense, with multiple levels from canopy to understory trees. Beyond that, the texture and color of many small or artificially planted trees are quite different. Due to the complexity of tropical forest vegetation, the network needs a high representation ability for correctly extracting woodland. In addition, the land cover of the tropical forest is imbalanced. Woodland and water areas occupy the majority of the total area, while artificial land and farmland account for very small proportions. When using existing methods for training, imbalanced ground objects result in imbalanced classification results, thereby tending towards categories with large numbers and omitting small ones. It is necessary to optimize the existing methods to strengthen the feature extraction capabilities when using the high-resolution images and to improve algorithm robustness to both small and imbalanced datasets.

In this study, we propose a new residual multi-attention network to classify tropical forest regions and improve the efficiency and accuracy of large-scale tropical forest monitoring. A multi-scale CNN is introduced to extract multi-scale features from complex tropical landscapes and an attention mechanism is used to refine the feature map to highlight the spatial and spectral information of key features in high-resolution multi-spectral remote sensing images. To obtain smooth and continuous category boundaries, the model absorbs deep and shallow information with multi-stage upsampling. We also construct a joint loss function and use a two-stage strategy for network training to alleviate sample imbalances in tropical forests, which is also a common problem with small datasets. We test the model's performance with two self-built tropical forest datasets and a widely used public semantic segmentation dataset.

2. Materials and Methods

2.1. Study Area

We chose tropical nature reserves in Hainan Island, China as the study area. Hainan Island preserves China's largest tropical rainforest area and monsoon forest ecosystem [28] and has the only "continental island" rainforest in the tropical rainforest and monsoon evergreen broad-leaved forest transition zone. The island is located between 18°10′–20°10′ N and 108°37′–111°03′ E with a total area of approximately 33,900 km^2. The island's major axis runs from northeast to southwest and is about 290-km-long, while from northwest to southeast the length is 180 km. Hainan Island is low and flat around the periphery and features tall mountains in the central areas. Mountains and hills are the core landform of Hainan Island, accounting for more than 38% of the island's area. The Wuzhishan and Yinggeling Mountains represent the core of the uplifted areas. Around the mountains and hills, platforms and terraces of varying widths are widely distributed. The coastal areas are mostly coastal plains, mainly alluvial plains and marine plains. The topography forms a ring-shaped layered landform with a defined cascading structure that creates an altitudinal zonation of tropical forest vegetation. Hainan Island has a tropical monsoon climate with high annual average rainfall and perennial cloud cover and is characterized by dry and wet seasons. The wet season lasts from May to November and the dry season extends from November to April.

The tropical natural forests of Hainan Island are mainly distributed in mountainous areas above 500 m above sea level in the southeast, middle-south, southwest, and middle of the island. The types mainly include mountain rainforests, mountain evergreen forests, mangroves, and other forests. The 29 national and provincial nature reserves are the largest and most representative areas for tropical forests (Figure 1). These reserves are mainly distributed in the central mountainous areas and along the coast and offer protection to tropical rainforests, mangroves, rare wild animals, and plants. Among them, the Yinggeling National Nature Reserve is the most complete and concentrated tropical rainforest reserve with the largest area. The land cover of the reserve is dominated by woodland, including mangroves in Dongzhaigang and Qinglangang, Vatica hainanensis forests in the Qingpilin Nature Reserve, and many other characteristic forest species, followed by water, with agricultural land and artificial land occupying the least proportions.

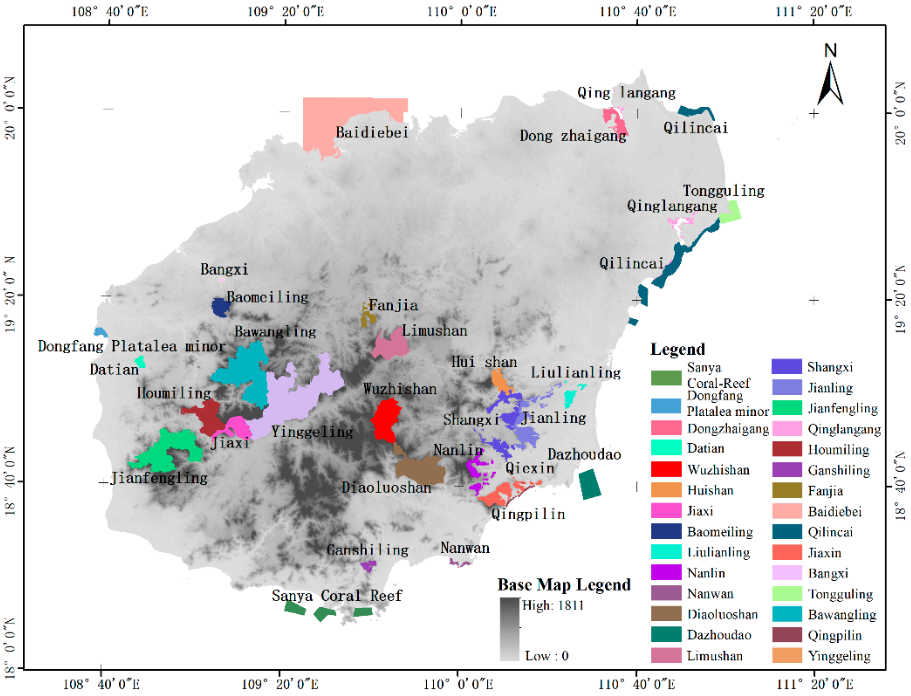


Figure 1. Distribution of 29 provincial and national nature reserves at Hainan Island.

We collected Hainan land cover data and unmanned aerial vehicle (UAV) images to assist with sample selection and result verification. By using these tropical reserves, we can have a better understanding of the tropical forest ecological environment.

2.2. Dataset

Three datasets were used, including two Hainan Nature Reserve datasets (datasets A and B) that were specifically built for this analysis and one public dataset. Dataset A is a small dataset that was used to verify the performance in the case of ultrahigh-resolution and various image qualities. Dataset B is a small dataset that was used to verify the performance in the case of multi-spectral data and multiple image sources. The influence of seasonal variation, such as the growth of crops and grass, is taken into account in Hainan datasets, and images in different seasons can improve the representation ability and test the performance of the networks. The public dataset has a higher resolution and a larger sample size than the other two datasets and can be used to verify the performance in the case of ultrahigh-resolution data and multiple semantic categories.

2.2.1. Hainan Nature Reserve Dataset with True Color Images (Dataset A)

True color samples (red, green, and blue) were selected from 0.5-m spatial resolution Google Earth images. These images were obtained for 2016 and 2017, and their wet to dry season ratio was 5:4. Two random samples are shown in Figure 2(a1,a2), and the geographic coordinates of their center points are about (18°43′30″ N, 109°52′5″ E) and (19°38′35″ N, 110°59′12″ E). Based on the main feature types of the nature reserves, we set up six semantic categories, namely farmland (SC1), woodland (SC2), grassland (SC3), water (SC4), artificial land (SC5), and bare land (SC6). The reserves with large areas and rich or unique surface features were chosen as typical areas to select samples. Most selected samples were able to highlight unique surface features and human activities, such as mangrove forests in Qinglangang, artificial buildings in Sanya, and roads in Tongguling, while ensuring that the samples were geographically dispersed. We manually labeled the images based on the results of field investigations, existing land cover maps, and UAV images, and divided dataset A into a training set and a validation set with a respective 4:1 ratio. Due to the limited ability of manual labeling, we selected 283 representative samples, such as forests and buildings, and carried out data augmentation via rotation and flipping to increase the number of samples and balance each class. This processing accelerated the convergence, alleviated over-fitting, and enhanced the generalization ability of the model. The final augmented dataset contained 1698 image blocks with a size of 256 × 256 pixels. Ultrahigh-resolution images were used to test the model's ability to finely classify surfaces. Using data with large quality differences increases the available data in practice and enhances the learning ability and application value.

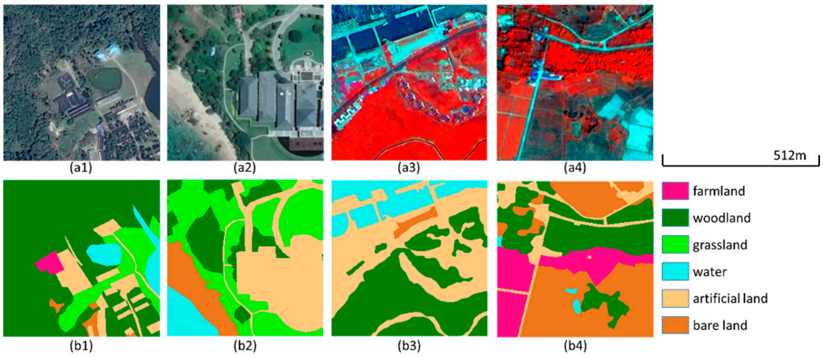


Figure 2. Examples for datasets A and B. (**a1**,**a2**) True color images. (**a3**,**a4**) False color composites of the near-infrared, green, and blue bands (NIR, G, B). (**b1**–**b4**) Corresponding labels.

2.2.2. Hainan Nature Reserve Dataset with Multi-Spectral Images (Dataset B)

We built the multi-spectral dataset to highlight the vegetation, whose reflectance increases rapidly in the near-infrared band. We also aimed to increase the distinction between woodland, artificial land, and water. Data were obtained from multi-spectral and panchromatic images from the Gaofen-1 (GF-1), Gaofen-2 (GF-2), and Ziyuan-3 (ZY-3) satellites. The relevant parameters for the three raw datasets are shown in Table 1. The three satellites have the same wavelength range in the multi-spectral bands but different spatial resolutions. This ensures the consistency of the spectral information while providing multi-scale features, which is helpful for extracting primary forest and water areas with few spatial features. After resampling panchromatic band of GF-2 to match the other bands (with resolution of 2 m), the panchromatic band has slightly different wavelengths but equal spatial resolutions. After radiometric calibration and QUick Atmospheric Correction (QUAC), the panchromatic and multi-spectral images were fused by Gram–Schmidt panchromatic sharpening to obtain multi-spectral images with spatial resolutions of 2 m × 2 m.

Table 1. Basic parameters of the satellite imagery.

Satellite		B1 (μm)	B2 (μm)	B3 (μm)	B4 (μm)	Panchromatic (μm)
GF-1	Wavelength	0.45–0.52	0.52–0.59	0.63–0.69	0.77–0.89	0.45–0.90
	Resolution	8 m				2 m
GF-2	Wavelength	0.45–0.52	0.52–0.59	0.63–0.69	0.77–0.89	0.45–0.90
	Resolution	4 m				1 m
ZY-3	Wavelength	0.45–0.52	0.52–0.59	0.63–0.69	0.77–0.89	0.45–0.80
	Resolution	6 m				2 m

The samples were selected according to the same principle as dataset A. The images were obtained in 2018, and the ratio of images in the dry and wet seasons was about 1:3. We manually labeled the images based on the results of field investigations and existing land cover maps and divided dataset B into a training set and a validation set according to a 4:1 ratio. The samples were scattered, and the labor and time costs for processing the multi-spectral images preprocessing were relatively high, resulting in fewer samples than dataset A. We used rotation and flip to augment the selected 125 samples. The final augmented dataset contained 750 image blocks with a size of 256 × 256 pixels and two random data samples are shown in Figure 2(a3,a4), and the geographic coordinates of their center points are about (18°13′54″ N, 109°29′39″ E) and (18°21′18″ N, 109°41′41″ E). The multi-spectral dataset was used for verifying the network's ability to extract spectral information and the applicability with multi-source data, which also increased the amount of available data for actual tasks.

2.2.3. Public Dataset

To verify the applicability of the improved model, we chose public datasets from the UC Merced Land Use Dataset [29] and the dense labeling remote sensing dataset (DLRSD) [30,31] for training. The UC Merced Land Use Dataset (a public dataset) is a scene classification dataset with land use images for 21 classes with 100 images for each class. Each image measures 256 × 256 pixels and the spatial resolution is 1 foot. Corresponding to the UC Merced Land Use Dataset, the DLRSD segments each image with 17 semantic classes. The following 17 class labels were considered in this dataset: Airplanes, bare soil, buildings, cars, chaparral, courts, docks, fields, grass, mobile homes, pavement, sand, sea, ships, tanks, trees, and water. We augmented the data with rotation, flipping, gamma transformation, and other methods for images. The final dataset contained 16,448 images. Figure 3 shows four images with corresponding pixel-wise labeling results that do not represent all 17 classes. We randomly assigned 80% of the images to the training set and 20% to the validation set. Very high-resolution (VHR) images were used to verify the model's ability to finely classify ground areas, and 17 semantic categories were used to verify the model's classification performance for complex tasks with multiple semantic categories.

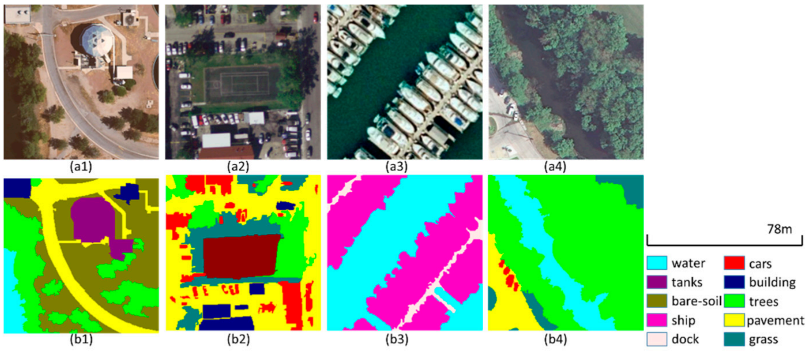


Figure 3. Examples in the public dataset. (**a1**–**a4**) True color images. (**b1**–**b4**) Corresponding labels.

2.3. Methods

2.3.1. Structure of ResMANet

The residual multi-attention network (ResMANet) proposed here adopts an encoder–decoder structure. The encoder extracts image features layer-by-layer. The decoder upsamples the feature map and concatenates it with shallow features to gradually recover the original size. The pixel level classification result is obtained through a softmax function. The overall architecture of ResMANet is shown in Figure 4.

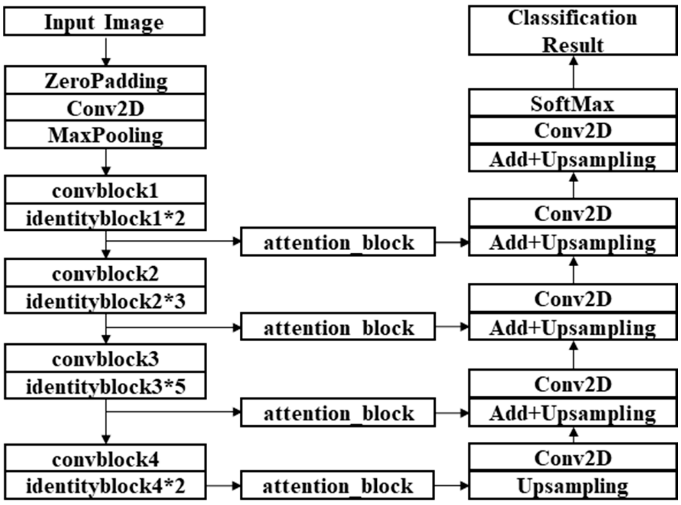


Figure 4. Structure of ResMANet (residual multi-attention network).

The encoder module refers to ResNet-50 and contains 5 convolution stages. The feature extraction part includes two residual structures, namely, "conv_block" and "identity_block" (Figure 5). We replaced the original 3 × 3 convolution kernel of conv_block to 1 × 1, 3 × 3, 5 × 5, and 7 × 7 convolution kernels and concatenated the feature map obtained by the four convolution kernels, which expanded the receptive field and extracted the multi-scale spatial features of ground objects in order to improve the discrimination of different landscapes.

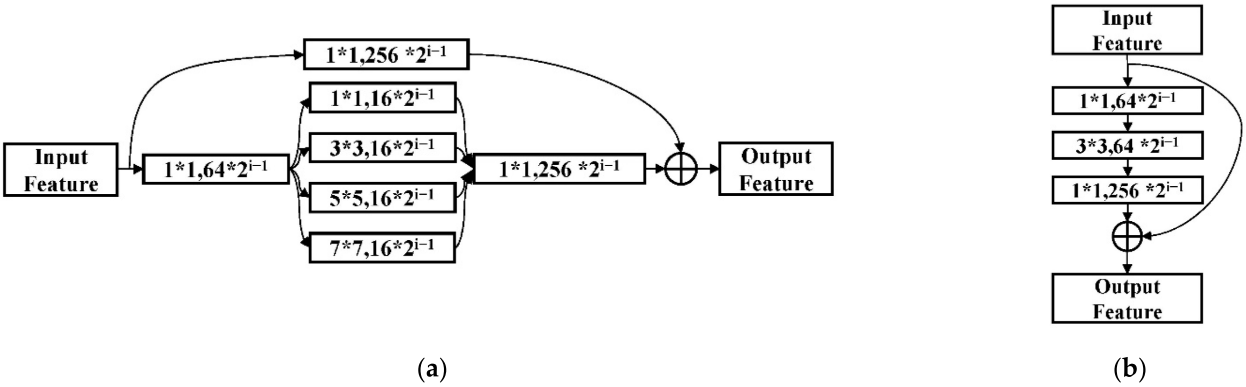
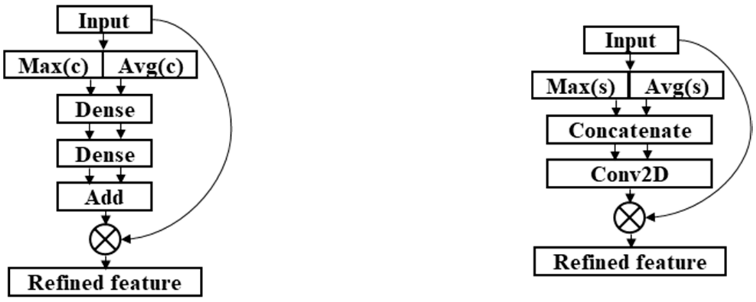


Figure 5. Structures of the modules in the encoder. (**a**) Conv_block. (**b**) Identity_block. i: The convolution stage.

The attention module adopts a sequential combination of channel attention and spatial attention [17], and the specific structures are shown in Figure 6. The channel attention module first aggregates the spatial information of a feature map by using both global average pooling (avg(c)) and global maximum pooling (max(c)) operations, generating two different features with 1 × 1 channels. Both features are then forwarded to two dense shared layers. After applying the shared network, we added the two feature vectors in for each element and the result was multiplied by the original input to obtain the final channel attention map. This module highlights the key spectral bands, and for multi-spectral data it can significantly improve the distinction between water, vegetation, and land in nature reserves. The spatial attention module first aggregates the information of a feature map by using both average pooling (avg(s)) and maximum pooling (max(s)) along the channel axis, generating two different features whose number of channels convert to 1. After concatenating the two feature maps, the result is forwarded to a convolutional layer with activation function. The result is multiplied by the original input to obtain the final spatial attention map. This module emphasizes key objects and channels, suppresses repetitive information, and optimizes the class boundary by focusing on location information.



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

Figure 6. Structures of the attention modules. (**a**) Channel attention module. (**b**) Spatial attention module.

The decoder module includes upsampling, summation, and convolutional layers (Figure 7). After concatenating the input feature map and the refined feature map obtained with the attention module, the result is forwarded to the upsampling layer with bilinear interpolation and 1 × 1 convolution, and then to the 3 × 3 convolutional layer. The input feature of the module is the output of previous stage, and the input of the first upsampling module is only the refined feature. The above module is repeated to gradually recover the feature map to the original size and finally input the softmax layer for probability prediction. The decoder integrates multi-layer shallow features to make up for the missing position information in the deep features and improves the boundary accuracy for small objects.

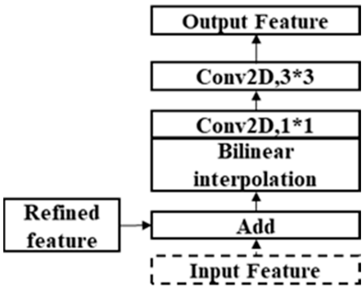


Figure 7. Upsampling module.

2.3.2. Four SOTA Semantic Segmentation Networks

We chose four semantic segmentation network structures as comparison networks, U-Net, PSPNet (ResNet-50) [32], DeepLabv3+ (ResNet-50), and U-Net (ResNet-50). They are all encoder–decoder structures, but the last three have a ResNet-50 backbone. Original U-Net contains four symmetrical maxpooling layers and upsampling layers to form a U-shaped structure. The encoder extracts features and sends them to the decoder. Through concatenate, convolution and upsampling layers, features of different scales are merged and recovered to the input size. We replaced the U-Net's decoder with ResNet-50 to get U-Net (ResNet-50). PSPNet (ResNet-50) has a pyramid pooling module to extract the global context on the top of the ResNet-50 and the kernel sizes of the module are 1 × 1, 2 × 2, 3 × 3, 6 × 6. DeepLabv3+ (ResNet-50) includes feature extraction network, atrous spatial pyramid pooling (ASPP) module and decoder. The feature extraction network outputs a shallow feature and a deep feature. After refined by ASPP, the deep feature is concatenate with the shallow feature and then recovered to the input size in the decoder.

2.3.3. Joint Loss Function

Sample imbalance is a common problem that is encountered when using deep learning methods in practice. Due to large differences in the proportions of various ground objects, this problem also appeared for the manually labeled Hainan datasets. To solve this problem, a joint loss function was constructed and a two-stage training strategy was applied to alleviate the imbalance, reduce overfitting, and improve the extraction accuracy for categories with small numbers. The first stage is the joint loss function, created by the cross-entropy loss (CEL) and generalized Dice loss (GDL) [33] (Equation (4)); the second stage uses the focal loss [34] to fine-tune the weights and makes the model tend towards objects that are difficult to classify. The GDL is developed from Dice loss, and Dice is a measurement function used to calculate the similarity of two samples [35] (Equation (1)). If the sample is not balanced, the Dice loss is unstable in training and results in rapid gradient changes. The GDL can solve this problem to an extent. When performing multi-class classification, each class has a Dice loss value and these values are weighted and integrated to form a generalized Dice loss (Equation (2)).

$$DICE\ LOSS = 1 - 2\frac{|A \cap B|}{|A + B|} \tag{1}$$

$$GDL = 1 - 2\frac{\sum_{l=1}^{m} w_l \sum_n r_{ln} p_{ln}}{\sum_{l=1}^{m} w_l \sum_n (r_{ln} + p_{\mathrm{ln})}} \tag{2}$$

$$w_l = \frac{1}{\sum_{i=1}^{N} {r_{ln}}^2} \tag{3}$$

$$Joint\ Loss = GDL + CEL \tag{4}$$

$|A|$ and $|B|$ represent the prediction and label, respectively. As common elements between A and B are calculated twice in the denominator, the numerator has a coefficient of 2. m is the number of classes; N is the number of pixels; r_{ln} represents the ground truth of category l at the nth pixel, while p_{ln} represents the corresponding predicted probability; and w_l represents the weight of each class.

The focal loss is an improvement of CEL for dense object detection tasks. It is mainly used to measure the difference in information between two probability distributions. In multi-classification problems, unbalanced datasets cause two problems, namely, many samples provide useless information to the model and the dominance of a certain kind of sample will degrade the performance of the model. The focal loss aims to solve the problem by reducing the weight of simple and dominant samples so that the model focuses on training with sparse and difficult samples. When the value of focal loss is small it is suitable for fine-tuning the model in the second stage. The calculation formula is as follows:

$$Focal\ loss = -\sum (1 - p_t)^{\gamma} * T * \log p_t \tag{5}$$

where γ is set to 2, T is the ground truth, and p_t represents the predicted probability.

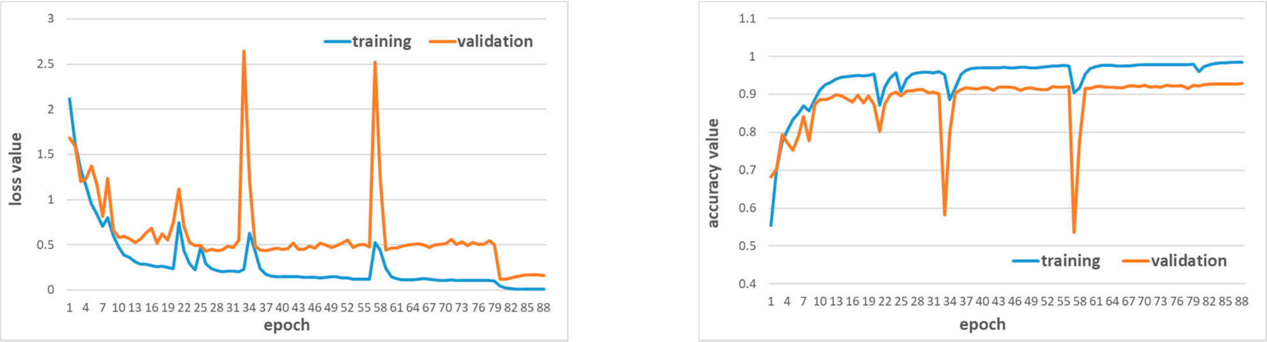
3. Results and Discussion

In this section, we present the detailed experimental setup and experimental results along with reasonable analysis and comparison. Our experiments were performed on Python 3.5 and Keras 2.2 with a NVIDIA GeForce GTX 1080 graphics card with 8 GB of memory and an E5-2637 v4 CPU with 16 GB RAM. We used four accuracy indicators to evaluate the classification accuracy of the trained network, namely, the producer's accuracy (PA), user's accuracy (UA), overall accuracy (OA), and mean intersection over union (MIoU). The PA and UA are the average of each category. The MIoU is the average IoU score of each category, while the IoU score is a standard performance measure for the object category segmentation problem [36].

3.1. Experiments on Dataset A

The convolutional layers adopted the normal distribution initialization method [37], which works well with rectified linear unit (ReLU) functions and improves the convergence speed. The learning rate is a hyperparameter that controls the convergence process. After repeated tests, the learning rate of the first stage was set to 2×10^{-4}, the learning decay rate was set to 2×10^{-5}, and the batch size was set to 8. To ensure that it was as close as possible to the best point during training, the number of epochs was set to 100. Training could be manually interrupted or the epochs could be increased according to the accuracy changes while training. The optimal weight was selected as the initial weight for the second stage, the learning rate was set to 2×10^{-4}, and the learning decay rate was set to 2×10^{-5}. We terminated the training when the accuracy no longer improved after approximately 5 to 10 epochs.

The training process for Hainan dataset A is shown in Figure 8. The loss of training and validation decreased in terms of fluctuations, which were caused by the update of weights of shallow layers, and the validation accuracy increased from 68.21% to 92.39%. The optimal weight of the first stage (the 79th percentile result) was used as the initial weight of the second stage. The input sequence was shuffled and then the second learning stage was entered. The training stopped when the accuracy of the validation set no longer increased, finding an accuracy of up to 92.81%.



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

Figure 8. Training process for Hainan dataset A. (**a**) Loss curve. (**b**) Accuracy curve

In order to verify the performance of the algorithms, we compared the training and prediction efficiency of the five networks (Table 2). Although our network has the most layers, the training efficiency is better than that of PSPNet (ResNet-50) and U-Net, which is close to that of DeepLabv3 + (ResNet-50), and the prediction efficiency is better than that of DeepLabv3 + (ResNet-50) and close to that of PSPNet (ResNet-50). Beyond that, five models were also used to predict the two images in the validation set and the predictions were compared with the ground truths. The purpose of classifying the surface is to monitor forest resources and detect the human activities on forest areas. It is necessary to accurately detect human activities, such as buildings, transportation facilities, and farmland. Therefore, we selected these two images based on two principles, i.e., containing as many categories as possible and representing human activities. The comparison results and difference maps are shown in Figure 9. The predictions obtained by different networks were quite dissimilar. U-Net integrated shallow information and the classification performances for ground objects and boundary restoration were good, but the shallow network depth affected the feature extraction ability, resulting in classification errors between categories. PSPNet

(ResNet-50) focused on extracting global context information and did not extract shallow feature information, which led to small features being ignored and blurred boundaries being formed. DeepLabv3+ (ResNet-50) extracted deep-level information through the atrous spatial pyramid pooling (ASPP) module and merged the second-stage output of the encoder during the restoration process. There was little internal noise in each class and the boundaries were slightly improved. U-Net (ResNet-50) did not perform many operations with deep features and recovered the features with a layer-by-layer method to achieve a good classification map. Similar to DeepLabv3+, there were few errors, although a jagged boundary appeared. The ResMANet method uses multi-scale convolution to extract richer feature information. The two attention modules retain the channel and spatial information of prominent objects so that the model distinguishes between different land cover types well. Among these models, our model had the fewest classification errors and the best boundary recovery.

Table 2. Comparison of training and prediction efficiency on true color images.

Nets	Training (Seconds/Epoch)	Prediction (Seconds/1,000,000 Pixel)
U-Net	80	0.69
PSPNet (ResNet-50)	85	0.83
DeepLabv3+ (ResNet-50)	73	0.89
U-Net (ResNet-50)	50	0.63
ResMANet	74	0.85

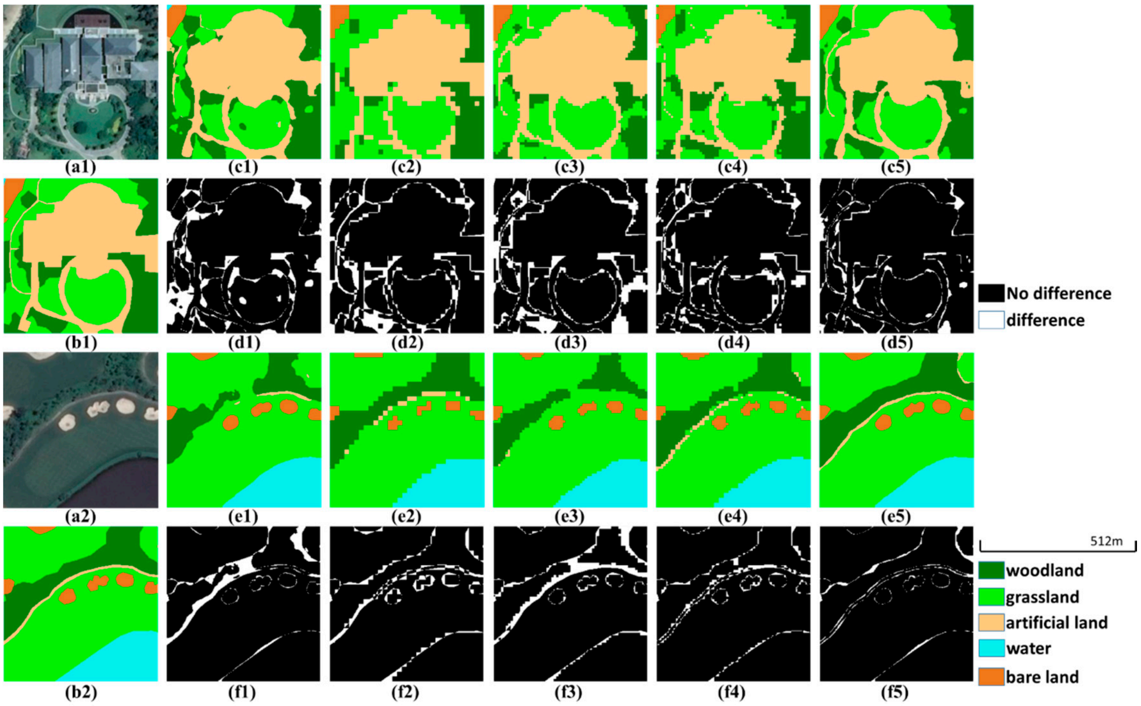
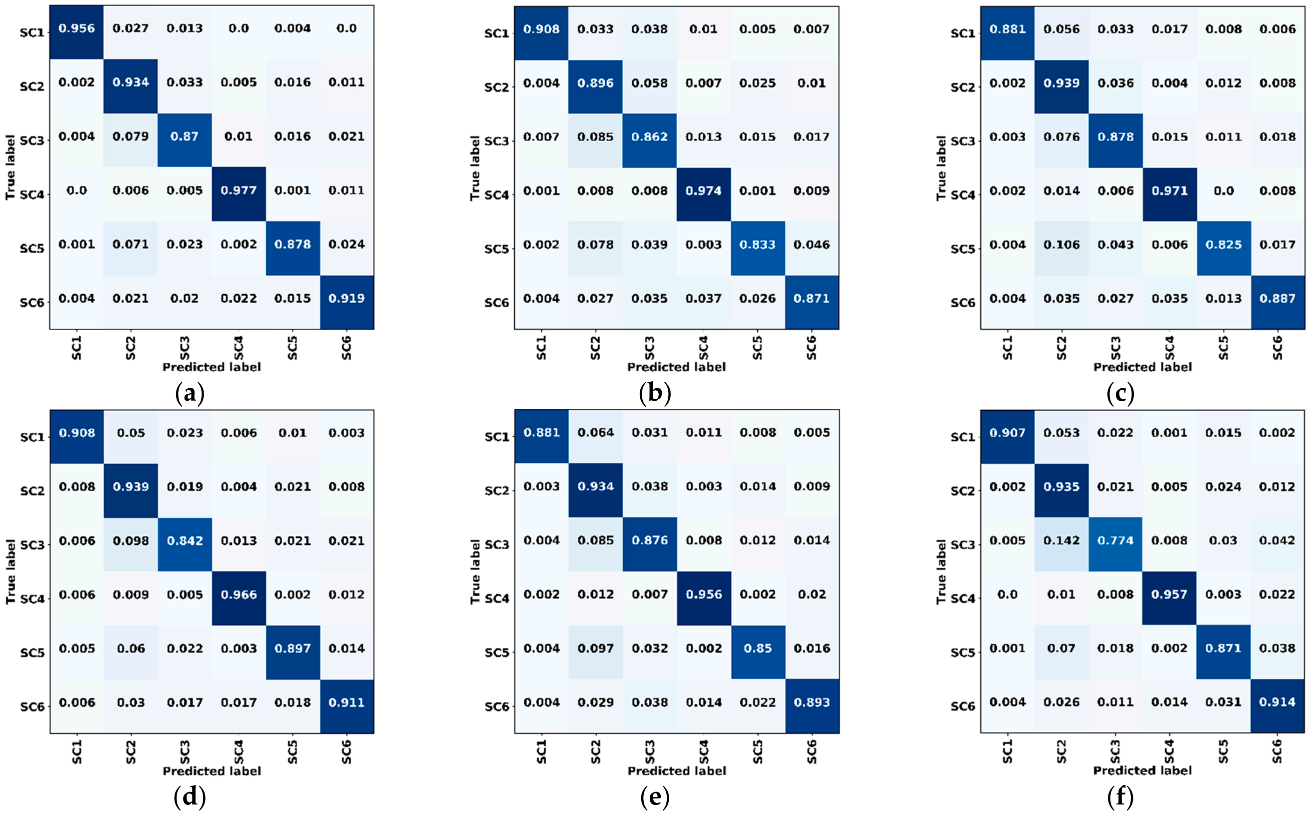


Figure 9. Prediction examples in the validation set of dataset A. (**a1**,**a2**) True color images. (**b1**,**b2**) Ground truth images. (**c1**–**c5**,**e1**–**e5**) Predictions of U-Net, PSPNet (ResNet-50), DeepLabv3+ (ResNet-50), U-Net (ResNet-50) and ResMANet. (**d1**–**d5**,**f1**–**f5**) Corresponding difference maps.

Based on the validation set, confusion matrices of the five networks were calculated and are shown in Figure 10, where (a) to (e) are the results of using the joint loss function and (f) is the result of using the cross-entropy loss. Because of the advantage in sample size, the classification results generally tilted toward the woodland (SC2) and water (SC4) categories, and many pixels were misclassified into these two categories. Our network had the highest accuracy for farmland, water, and bare land, and was close to the best in the other categories. Comparison between the results obtained by the joint loss function and the results obtained by the cross-entropy loss proved that the joint loss function we constructed improved the classification accuracies for categories with small numbers, such as farmland, and reduced the sample imbalance impact.



(**a**)

	SC1	SC2	SC3	SC4	SC5	SC6
SC1	0.956	0.027	0.013	0.0	0.004	0.0
SC2	0.002	0.934	0.033	0.005	0.016	0.011
SC3	0.004	0.079	0.87	0.01	0.016	0.021
SC4	0.0	0.006	0.005	0.977	0.001	0.011
SC5	0.001	0.071	0.023	0.002	0.878	0.024
SC6	0.004	0.021	0.02	0.022	0.015	0.919

(**b**)

	SC1	SC2	SC3	SC4	SC5	SC6
SC1	0.908	0.033	0.038	0.01	0.005	0.007
SC2	0.004	0.896	0.058	0.007	0.025	0.01
SC3	0.007	0.085	0.862	0.013	0.015	0.017
SC4	0.001	0.008	0.008	0.974	0.001	0.009
SC5	0.002	0.078	0.039	0.003	0.833	0.046
SC6	0.004	0.027	0.035	0.037	0.026	0.871

(**c**)

	SC1	SC2	SC3	SC4	SC5	SC6
SC1	0.881	0.056	0.033	0.017	0.008	0.006
SC2	0.002	0.939	0.036	0.004	0.012	0.008
SC3	0.003	0.076	0.878	0.015	0.011	0.018
SC4	0.002	0.014	0.006	0.971	0.0	0.008
SC5	0.004	0.106	0.043	0.006	0.825	0.017
SC6	0.004	0.035	0.027	0.035	0.013	0.887

(**d**)

	SC1	SC2	SC3	SC4	SC5	SC6
SC1	0.908	0.05	0.023	0.006	0.01	0.003
SC2	0.008	0.939	0.019	0.004	0.021	0.008
SC3	0.006	0.098	0.842	0.013	0.021	0.021
SC4	0.006	0.009	0.005	0.966	0.002	0.012
SC5	0.005	0.06	0.022	0.003	0.897	0.014
SC6	0.006	0.03	0.017	0.017	0.018	0.911

(**e**)

	SC1	SC2	SC3	SC4	SC5	SC6
SC1	0.881	0.064	0.031	0.011	0.008	0.005
SC2	0.003	0.934	0.038	0.003	0.014	0.009
SC3	0.004	0.085	0.876	0.008	0.012	0.014
SC4	0.002	0.012	0.007	0.956	0.002	0.02
SC5	0.004	0.097	0.032	0.002	0.85	0.016
SC6	0.004	0.029	0.038	0.014	0.022	0.893

(**f**)

	SC1	SC2	SC3	SC4	SC5	SC6
SC1	0.907	0.053	0.022	0.001	0.015	0.002
SC2	0.002	0.935	0.021	0.005	0.024	0.012
SC3	0.005	0.142	0.774	0.008	0.03	0.042
SC4	0.0	0.01	0.008	0.957	0.003	0.022
SC5	0.001	0.07	0.018	0.002	0.871	0.038
SC6	0.004	0.026	0.011	0.014	0.031	0.914

Figure 10. Confusion matrices of validation data for dataset A. (**a**) ResMANet. (**b**) U-Net. (**c**) PSPNet (ResNet-50). (**d**) DeepLabv3+ (ResNet-50). (**e**) U-Net (ResNet-50). (**f**) ResMANet (CEL). SC1: Farmland, SC2: Woodland, SC3: Grassland, SC4: Water, SC5: Artificial land, SC6: Bare land, CEL: Cross-entropy loss.

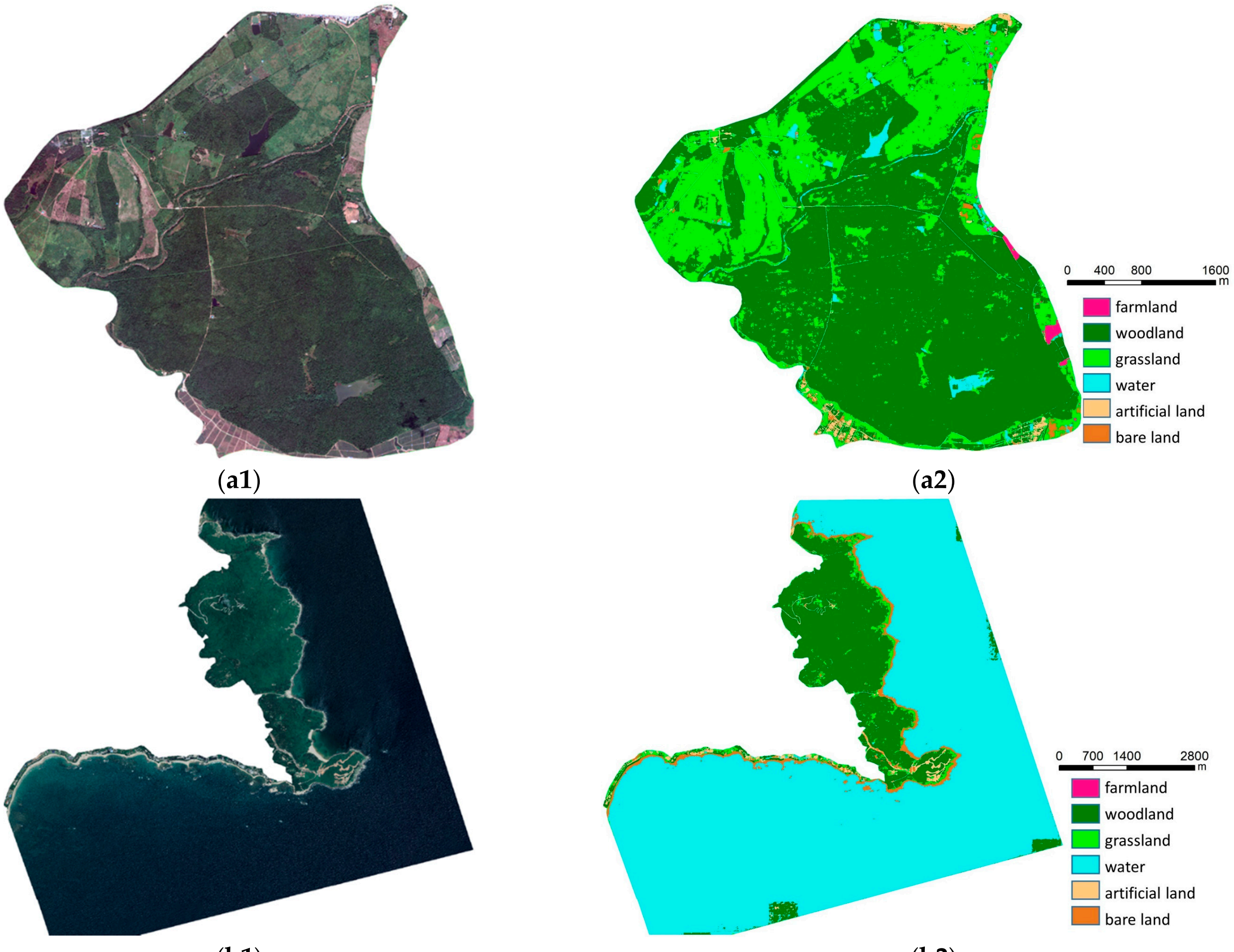
The confusion matrices were used to calculate the overall accuracy indicators shown in Table 3. The producer's accuracy represents the probability that a certain land cover of an area on the ground is classified as such and the user's accuracy is referred to as the reliability. The two indicators are average values. The overall accuracy shows the correct proportion of result. We achieved 0.50–3.22%, 1.18–3.18%, and 0.70–2.78% improvements for these three indicators. As an important indicator in computer vision, the MIoU represents the ratio of the intersection and union of the ground truth and predictions. Due to the small dataset and the small differences with true color images, the accuracies of all the networks were generally high; however, an improvement of 2.39–5.27% for the MIoU was achieved by the proposed network, which proves the excellent performance of the network.

We used the well-trained network to classify the Datian and Tongguling nature reserves, which differ in their geographical location and feature types. The Datian Nature Reserve is located in the northwest of Hainan Island, far away from the coast and its surface is mostly covered with woodland and grassland. The Tongguling Nature Reserve is located in the east of Hainan Island and is covered with woodland and water, and some artificial lands along the coast. We cropped the image into 256 × 256 pixel blocks for input into network and to create a mosaic of predictions within a 30-pixel-wide boundary. The final classification maps of the two nature reserves are shown in Figure 11.

Table 3. Accuracy assessment of Hainan dataset A.

Nets	PA	UA	OA	MIoU
U-Net	88.98%	89.05%	90.02%	80.51%
PSPNet (ResNet-50)	91.70%	89.67%	91.75%	83.06%
DeepLabv3+ (ResNet-50)	90.70%	91.05%	92.10%	83.39%
U-Net (ResNet-50)	91.03%	89.84%	91.51%	82.71%
ResMANet (CEL)	89.73%	89.31%	90.51%	81.16%
ResMANet	92.20%	92.23%	92.80%	85.78%

PA: Producer's accuracy; UA: User's accuracy; OA: Overall accuracy; MIoU: Mean intersection over union.



(**a1**) (**a2**)

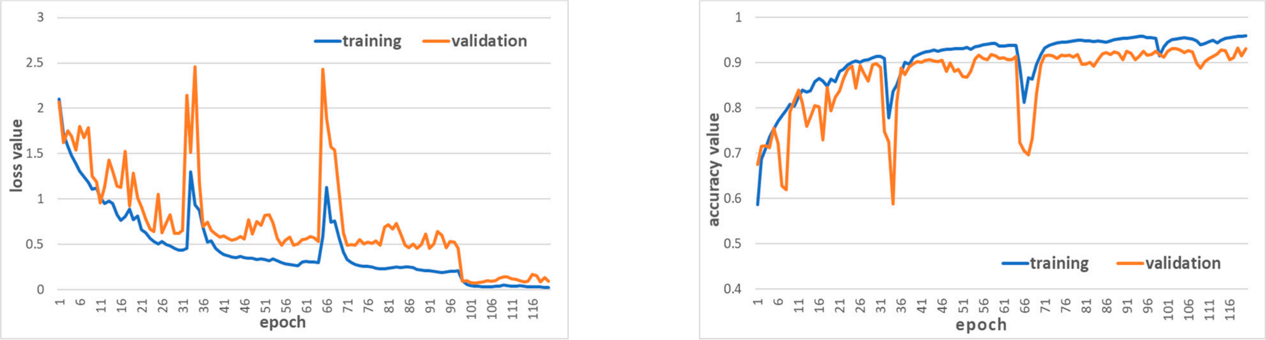
(**b1**) (**b2**)

Figure 11. Classification results for nature reserves based on true color images. (**a1**) Image of Datian Nature Reserve. (**a2**) Label of Datian Nature Reserve. (**b1**) Image of Tongguling Nature Reserve. (**b2**) Label of Tongguling Nature Reserve.

There were no large-scale misclassifications for these two typical reserves. The detection of artificial land, large woodland, and grassland areas in the Datian Nature Reserve was accurate. Few small features such as rivers and unhardened roads were confused with adjacent objects, resulting in discontinuous features. The classification performance for Tongguling Nature Reserve was better and the classification of bare land along the coast, woodland, and most paved roads was accurate, while some unhardened roads were confused with bare land. Due to the high resolution of the images and mixed ground features, some objects were fragmented and sometimes it was impossible to accurately determine their class. With only a small proportion of the typical area selected, most of the reserves were not involved in the training. In this case, the great overall results also prove the excellent performance of the model proposed here. Although the problems pertaining to unobvious spatial characteristics and the large difference between woodland and water have been alleviated, there is still some noise in the woodland areas and at the edges of water. By adding spectral information (e.g., near-infrared band information), the related errors may be further reduced.

3.2. Experiments on Dataset B

Hyperparameter settings such as the learning rate were the same as in the previous experiment. The training could be manually interrupted or the number of epochs could be increased according to the desired accuracy. The training process for Hainan dataset B is shown in Figure 12. The loss during training and validation decreased with fluctuation and the validation accuracy increased from 67.55% to 92.57%. The optimal weight in the first stage (the 97th result) was used as the initial weight in the second stage. The input sequence was shuffled and then entered into the second learning stage. The training stopped when the accuracy of the validation set no longer increased, reaching up to 93.20%.



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

Figure 12. Training process for Hainan dataset B. (**a**) Loss curve. (**b**) Accuracy curve.

We also compared the training and prediction efficiency of the five networks (Table 4), and ResMANet achieved high efficiency on processing multi-spectral data. The training efficiency is better than that of PSPNet (ResNet-50), DeepLabv3+ (ResNet-50) and U-Net, and the prediction efficiency is better than that of DeepLabv3+ (ResNet-50) and PSPNet (ResNet-50), which is close to that of U-Net. The accuracies of the four classic networks were also evaluated with the validation set and the results were compared in terms of their performance against ResMANet (Figure 13). According to purpose of this study, we selected two validation images that contained human activities and as many categories as possible to present the classification performances of different networks. We also built difference maps to show the classification errors, which were mainly concentrated at the boundaries between categories. The predictions and difference maps obtained with the various networks showed large differences. The results for U-Net had the greatest classification errors and discontinuous ground object classifications. There was also confusion between artificial land and bare land. Without any shallow information, PSPNet (ResNet-50) exhibited the most serious loss of boundary information and jagged boundaries appeared. DeepLabv3+ (ResNet-50) and U-Net (ResNet-50) had more convolutional layers for extracting and recovering features than the previous two models, which reduced the internal noise for categories and enhanced the information for small features. Our network fully extracted the near-infrared band information so that the speckles of vegetation and artificial land were incorrectly classified significantly less than with the other networks. The proposed network not only had good consistency within classes, but also the smoothest boundaries, which was helpful for distinguishing between the internal features of large objects and detecting small and narrow objects.

Table 4. Comparison of training and prediction efficiency on multi-spectral data.

Nets	Training (Seconds/Epoch)	Prediction (Seconds/1,000,000 Pixel)
U-Net	35	0.68
PSPNet (ResNet-50)	37	0.79
DeepLabv3+ (ResNet-50)	32	0.88
U-Net (ResNet-50)	22	0.60
ResMANet	30	0.71

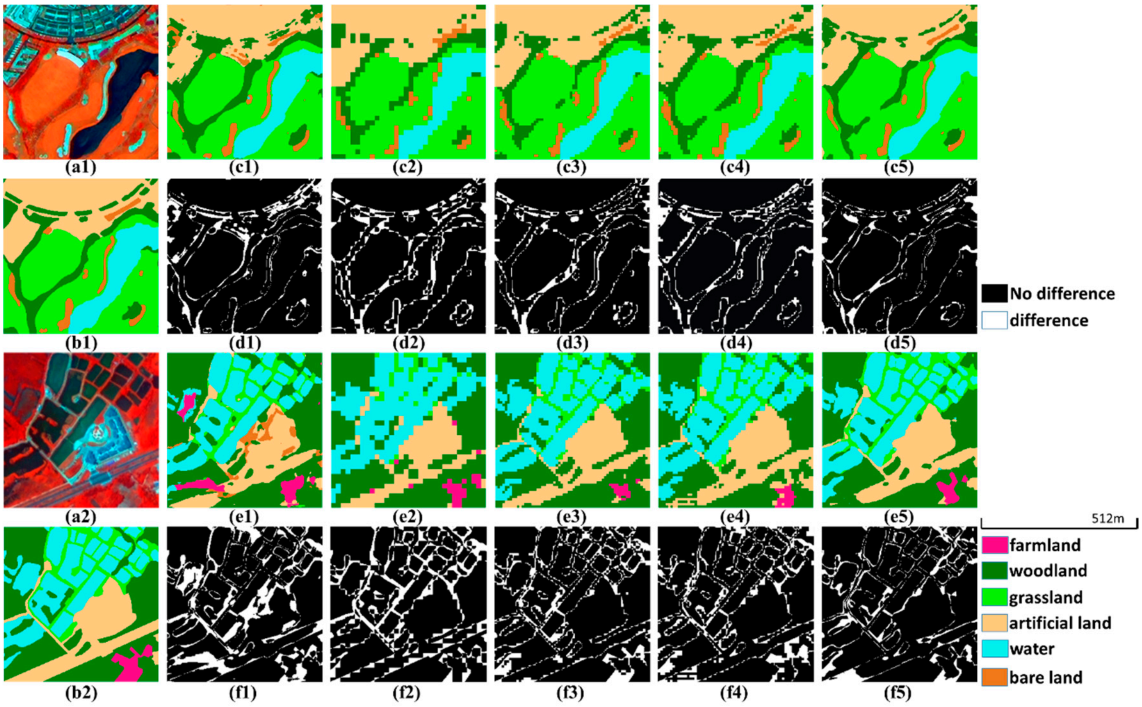


Figure 13. Prediction examples in the validation set of dataset B. (**a1**,**a2**) False color images composited by the NIR, G, and B bands. (**b1**,**b2**) Ground truth images. (**c1**–**c5**,**e1**–**e5**) Predictions of U-Net, PSPNet (ResNet-50), DeepLabv3+ (ResNet-50), U-Net (ResNet-50) and ResMANet. (**d1**–**d5**,**f1**–**f5**) Corresponding difference maps.

Based on the validation set, confusion matrices of the five networks were calculated and are shown in Figure 14, where (a) to (e) are the results of using the joint loss function and (f) is the result of using the cross-entropy loss. With the advantage in sample size, the classification results also tilted toward woodland (SC2) and water (SC4) areas, and many pixels in grassland (SC3), artificial land (SC5), and bare land (SC6) areas were misclassified into these two categories. ResMANet had the smallest percentage of misclassifications and achieved the best classification accuracy between the five categories, especially for grassland, artificial land, and bare land. The use of a joint loss function significantly improved the classification accuracy for categories with small numbers, such as grassland and bare land, and made the results more balanced. The confusion matrices were used to calculate the overall accuracy indices shown in Table 5. We achieved 1.6–2.82%, 1.86–6.8%, and 1.34–3.17% improvements for the PA, UA, and OA, respectively. The significant improvement of the MIoU (3.25–7.38% improvement) proves the high consistency between the proposed method's predictions and the ground truth.

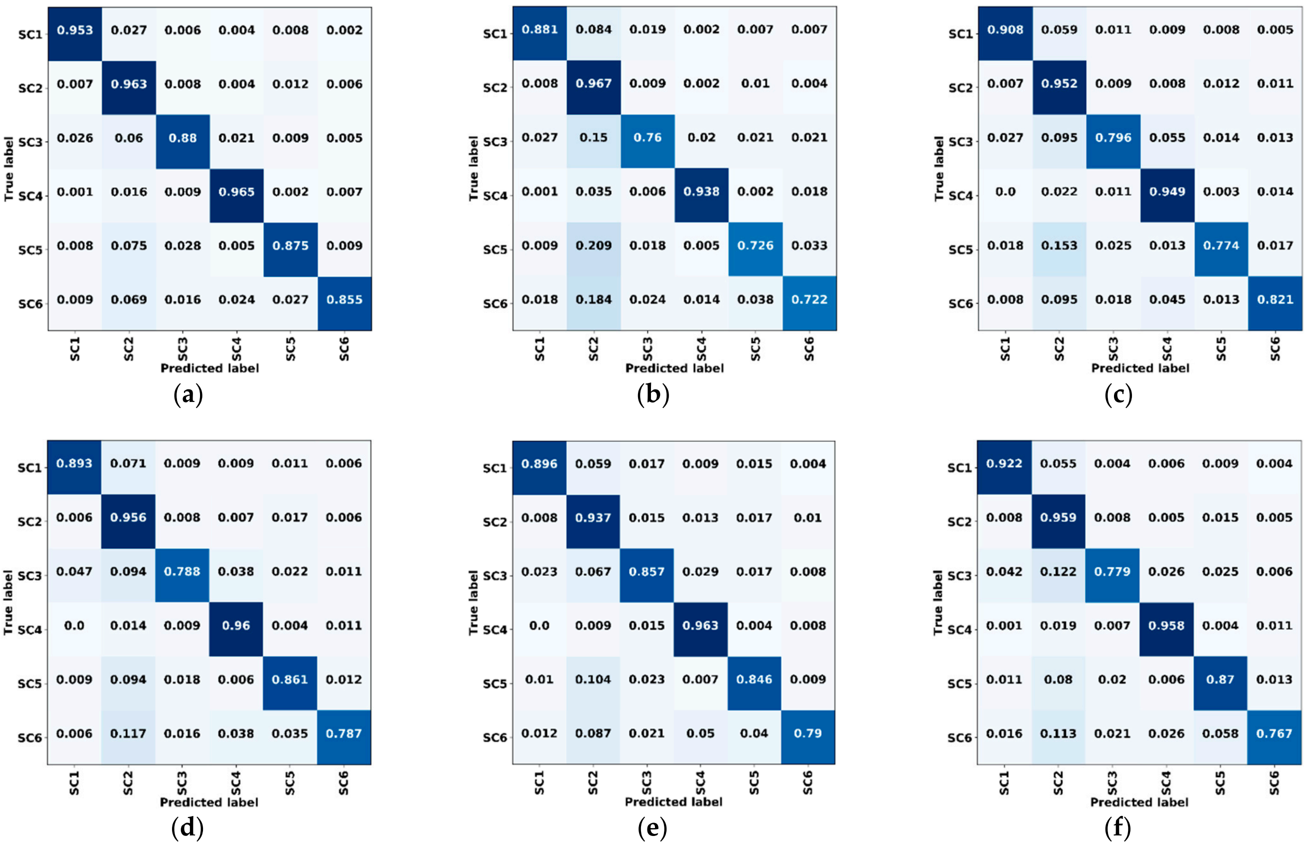


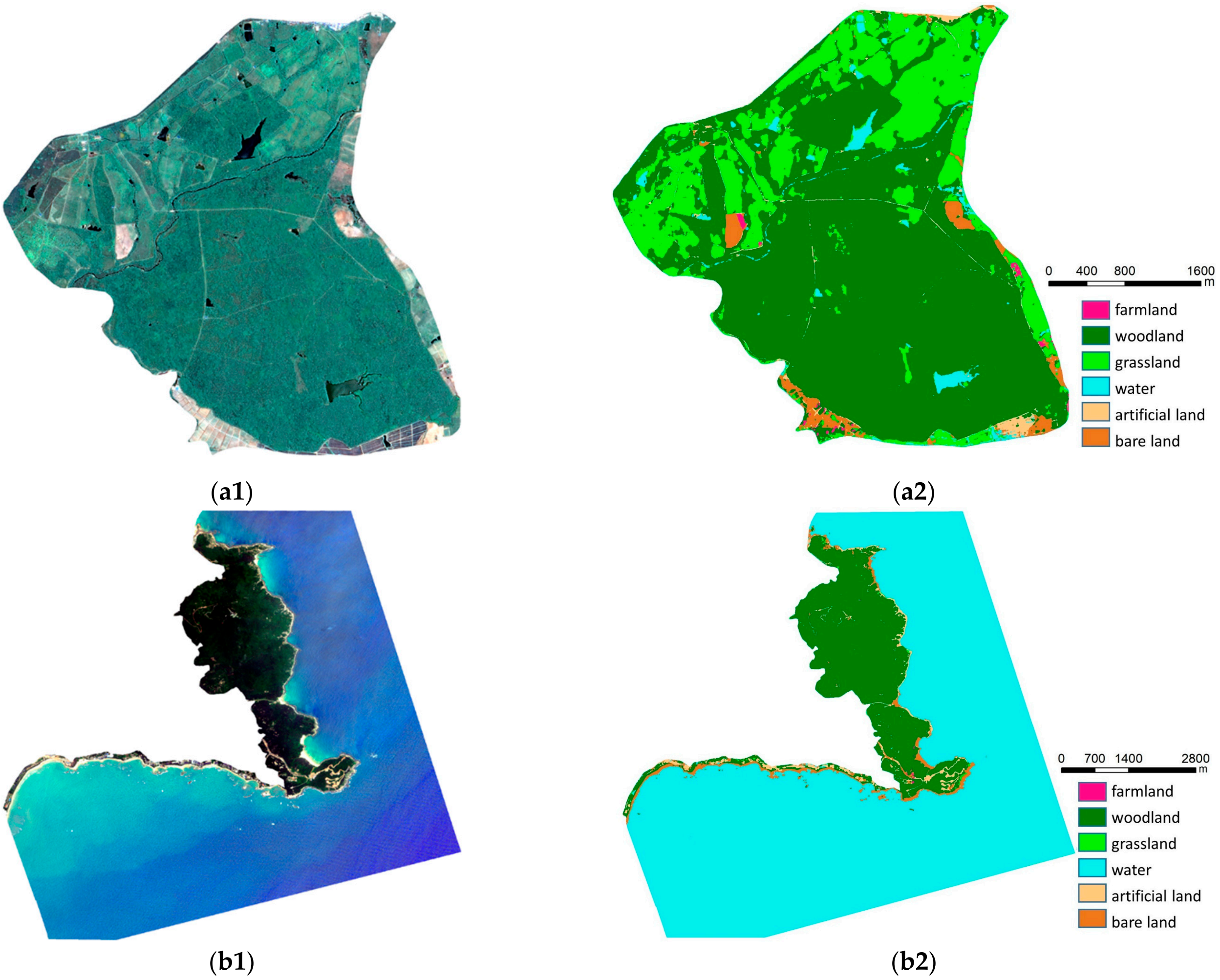
Figure 14. Confusion matrices of validation data in Dataset B. (**a**) ResMANet. (**b**) U-Net. (**c**) PSPNet (ResNet-50). (**d**) DeepLabv3+ (ResNet-50). (**e**) U-Net (ResNet-50). (**f**) ResMANet (CEL). SC1: Farmland, SC2: Woodland, SC3: Grassland, SC4: Water, SC5: Artificial land, SC6: Bare land, CEL: Cross-entropy loss.

Table 5. Accuracy assessment of Hainan dataset B.

Nets	PA	UA	OA	MIoU
U-Net	88.28%	83.22%	90.00%	75.47%
PSPNet (ResNet-50)	88.47%	86.67%	90.95%	78.15%
DeepLabv3+ (ResNet-50)	89.50%	87.43%	91.83%	79.60%
U-Net (ResNet-50)	88.43%	88.16%	91.45%	79.38%
ResMANet (CEL)	89.83%	87.58%	92.08%	80.00%
ResMANet	91.10%	90.02%	93.17%	82.85%

PA: Producer's accuracy; UA: User's accuracy; OA: Overall accuracy; MIoU: Mean intersection over union.

With the network trained on dataset B, we applied the same prediction and mosaic methods to present the final classification maps for the two reserves (Figure 15). The multi-spectral data with a 2-m resolution were more blurred than the true color data and featured less patch fragmentation. ResMANet's channel attention module fully extracted the near-infrared band information and reduced the internal speckles of vegetation and water areas. The woodland and grassland areas in the Datian Nature Reserve were well distinguished, and the water areas were accurately extracted. There was some confusion between bare land and farmland with small areas of vegetation. There were almost no classification errors for water areas in the Tongguling Nature Reserve. The uniformity of woodland areas was very good, and the classification map for narrow bare land at the coast was acceptable, while the unhardened roads and bare soil were partially confused. Dataset B had a smaller size and our model still obtained accurate classification results, proving that the proposed network is able to effectively extract spectral information and maintain excellent performance with small datasets.



(**a1**) (**a2**) (**b1**) (**b2**)

Figure 15. Classification results for nature reserves based on multi-spectral images. (**a1**) Image of Datian Nature Reserve. (**a2**) Label of Datian Nature Reserve. (**b1**) Image of Tongguling Nature Reserve. (**b2**) Label of Tongguling Nature Reserve.

3.3. Experiments on the Public Datasets

In addition to the tropical reserves, we also verified the applicability of our model with the UC Merced Land Use Dataset and DLRSD public datasets. The same training strategy and loss function were applied. The public datasets include 17 semantic classes with ultrahigh-resolution, which has higher requirements for the model. While the Hainan datasets only contained nature reserves and nearby areas, these datasets have a wider acquisition area, and therefore they can be used to verify the applicability of the model for complex classification tasks. We also selected two validation images that contained as many categories as possible to present the classification performances of different networks. Using the trained network to predict the two images in the validation dataset, the results and difference maps are shown in Figure 16.

It can be seen from the figure that the representation ability of shallow networks was weak. There were some mixed categories with U-Net, although the boundary information was well preserved due to the addition of multi-stage shallow features. The deep networks had stronger representation abilities and higher classification performances, but some errors still occurred due to the ground feature complexity. Similar to the previous experiments, the results of networks based on ResNet-50 had fewer internal errors within categories; however, the differences in extracting deep information and utilizing shallow information led to big differences in the results. The boundaries of PSPNet (ResNet-50) were jagged, especially for bare soil and buildings. The boundaries of DeepLabv3+ (ResNet-50) and U-Net (ResNet-50) were clearer, although with severe categorical errors, such as the misclassification of pavement into grass (c3) and the misclassification of bare soil into pavement (e4). Due to the advantages of multi-scale convolutional and attention modules, our model had the best multi-scale feature extraction capabilities and the most accurate results. The boundaries of bare soil, buildings, and trees were highly consistent with the ground truths; however, the misclassification of bare soil into pavement also occurred (e5). The overall accuracy is shown in Table 6. U-Net's accuracy indicators were lower than other deep networks, which was determined by its fewer convolutional layers. The performance of U-Net (ResNet-50) was better than U-Net, but worse than the other two networks with ResNet-50, which is because of the insufficient ability to extract global information with U-Net. The accuracy difference between DeepLabv3+ (ResNet-50) and PSPNet (ResNet-50) was not significant. Our network achieved the best performance and best accuracy, especially in terms of the improvement of the MIoU, which indicates a great discrimination ability for the 17 semantic categories.

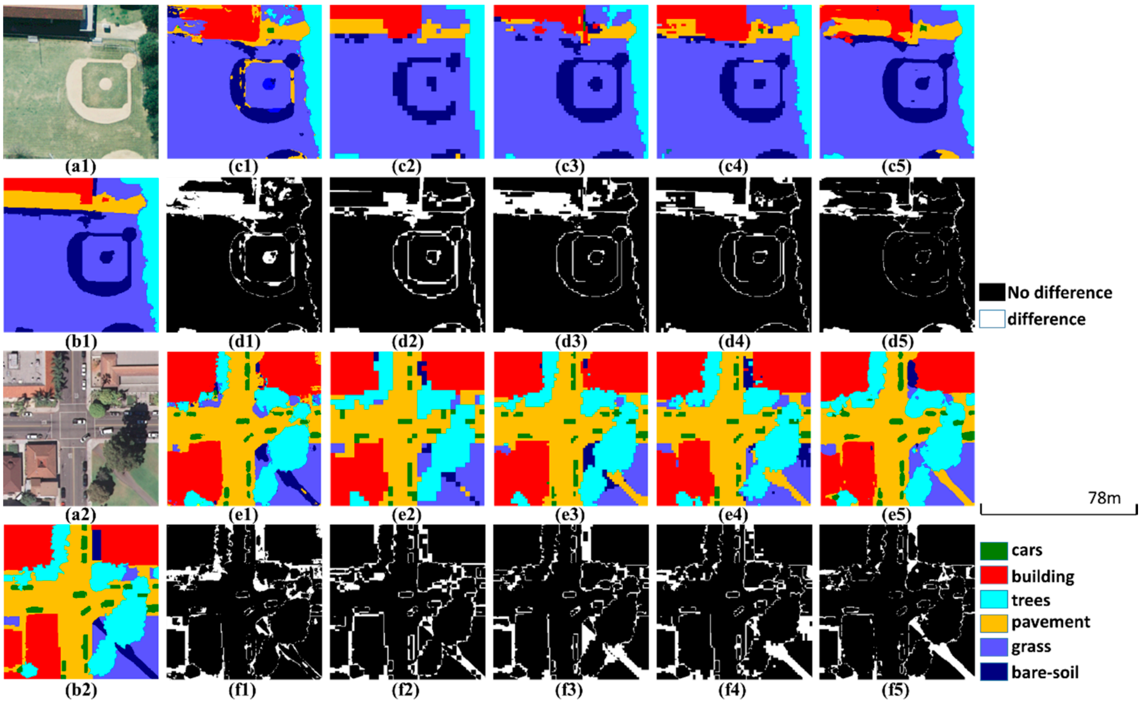


Figure 16. Prediction examples in the validation set of the public dataset. (**a1**,**a2**) True color images. (**b1**,**b2**) Ground truth images. (**c1**–**c5**,**e1**–**e5**) Predictions of U-Net, PSPNet (ResNet-50), DeepLabv3+ (ResNet-50), U-Net (ResNet-50) and ResMANet. (**d1**–**d5**,**f1**–**f5**) Corresponding difference maps.

Table 6. Accuracy assessment of the public dataset results between methods.

Nets	PA	UA	OA	MIoU
U-Net	85.63%	83.04%	83.89%	73.37%
PSPNet (ResNet-50)	87.79%	87.58%	90.65%	79.03%
DeepLabv3+ (ResNet-50)	89.47%	89.71%	90.82%	80.04%
U-Net (ResNet-50)	87.86%	85.26%	88.41%	76.79%
ResMANet	90.97%	90.90%	91.52%	83.75%

PA: Producer's accuracy; UA: User's accuracy; OA: Overall accuracy; MIoU: Mean intersection over union.

These image results and accuracy indicators prove that our network is sufficient for handling complex semantic segmentation tasks and has wider applicability.

4. Conclusions

The tropical forests and environments in Hainan Island feature unusual biogeographical positions in a Chinese context and certain regional characteristics. The in-depth study of tropical forests and environments requires accurate land cover classification. To automatically obtain the land use and land cover information for tropical forests and surrounding areas, we have proposed an improved classification model based on a deep convolutional network, constructing an end-to-end model, and expanding the application scope of related technologies. This classification model can reduce manual work such as visual interpretation and manual delineation. By applying multi-scale convolution, an attention module, and layer-by-layer upsampling to improve the original network, our model expands the receptive field, enhances feature extraction capabilities, and fully integrates both deep and shallow information. The multi-stage training strategy enables the network's weights to converge quickly without overfitting. With these loss functions, the model tends to extract the surface feature information of bare land and artificial land, balancing the influence of dominant samples such as woodland.

We manually delineated the true color and multi-spectral Hainan datasets to verify the performance of ResMANet with ultrahigh-resolution conditions and various image qualities, as well as multi-spectral information and multiple image sources. Compared with four state-of-the-art models, the proposed model was more accurate and obtained 92.80% and 93.17% overall accuracy for the two validation sets. It not only had good consistency within classes, but also the smoothest boundaries. The near-infrared band was fully extracted to improve the uniformity and further reduce errors in the vegetation and water areas and some artificial land classes. The overall refined classification maps for the reserves also intuitively show the excellent performance of the model for Hainan tropical reserves and the prospects for application in other tropical forest areas. Besides the Hainan dataset, ResMANet achieved a 91.52% overall accuracy with the public dataset, which proved its good capability for complex multi-class problems and its broad application prospects. Through the application of a joint loss function and a two-stage training strategy, the effect of sample imbalance in the three datasets was significantly reduced.

Adding near-infrared information to the training data improved the network performance for land use and land cover classification. In the next steps, we will continue to optimize the algorithm in terms of two aspects. The first aspect is to increase the information richness for inputs, such as adding digital surface model information and hyperspectral information for tropical flora; the second is to further improve the network feature extraction and recovery capabilities to make full use of the training samples. At the same time, we hope to expand the application scope of the algorithm based on small datasets and low hardware requirements.

Author Contributions: Conceptualization, Wenjin Wu; Methodology, Tong Yu; Investigation, Wenjin Wu and Chen Gong; Validation, Tong Yu and Chen Gong; Writing—Original Draft Preparation, Tong Yu; Writing—Review AND Editing, Tong Yu and Wenjin Wu; Project Administration, Xinwu Li; Funding Acquisition, Xinwu Li. All authors have read and agreed to the published version of the manuscript.

Funding: This research was funded by the Key Research and Development Program of Hainan Province, grant number ZDYF2019005, and the Aerospace Information Research Institute, Chinese Academy of Sciences, grant number Y951150Z2F. And the APC was funded by the Key Research and Development Program of Hainan Province, grant number ZDYF2019005.

Data Availability Statement: The dataset A and dataset B presented in this study are available on request from the corresponding author. The public dataset presented in this study are available in [29–31].

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

References

1. The Food and Agriculture Organization (FAO). Global Forest Resources Assessment. 2015. Available online: http://www.fao.org/forest-resources-assessment/past-assessments/fra-2015 (accessed on 9 December 2020).
2. Cabrera-Barona, P.F.; Bayón, M.; Durán, G.; Bonilla, A.; Mejía, V. Generating and Mapping Amazonian Urban Regions Using a Geospatial Approach. *ISPRS Int. J. Geo-Inf.* **2020**, *9*, 453. [CrossRef]
3. Häme, T.; Kilpi, J.; Ahola, H.; Rauste, Y.; Antropov, O.; Rautiainen, M.; Sirro, L.; Bounepone, S. Improved Mapping of Tropical Forests with Optical and SAR Imagery, Part I: Forest Cover and Accuracy Assessment Using Multi-Resolution Data. *IEEE J. Sel. Top. Appl. Earth Obs. Remote Sens.* **2013**, *6*, 74–91. [CrossRef]
4. Häme, T.; Stenberg, P.; Andersson, K.; Rauste, Y.; Kennedy, P.; Folving, S.; Sarkeala, J. AVHRR-based forest proportion map of the Pan-European area. *Remote Sens. Environ.* **2001**, *77*, 76–91. [CrossRef]
5. Bullock, E.L.; Woodcock, C.E.; Olofsson, P. Monitoring tropical forest degradation using spectral unmixing and Landsat time series analysis. *Remote Sens. Environ.* **2020**, *238*, 110968. [CrossRef]
6. Ghosh, S.M.; Behera, M.D. Aboveground biomass estimation using multi-sensor data synergy and machine learning algorithms in a dense tropical forest. *Appl. Geogr.* **2018**, *96*, 29–40. [CrossRef]
7. Hinton, G.E.; Salakhutdinov, R.R. Reducing the Dimensionality of Data with Neural Networks. *Science* **2006**, *313*, 504–507. [CrossRef]
8. Krizhevsky, A.; Sutskever, I.; Hinton, G.E. Imagenet Classification with Deep Convolutional Neural Networks. *Commun. ACM* **2017**, *60*, 84–90. [CrossRef]
9. Simonyan, K.; Zisserman, A. Very Deep Convolutional Networks for Large-Scale Image Recognition. *arXiv* **2014**, arXiv:1409.1556.
10. Szegedy, C.; Liu, W.; Jia, Y.; Sermanet, P.; Reed, S.; Anguelov, D.; Erhan, D.; Vanhoucke, V.; Rabinovich, A. Going Deeper With Convolutions. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Boston, MA, USA, 7–12 June 2015; pp. 1–9.
11. 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, 26 June–1 July 2016; pp. 770–778.
12. Huang, G.; Liu, Z.; Van Der Maaten, L.; Weinberger, K.Q. Densely Connected Convolutional Networks. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Honolulu, HI, USA, 21–26 July 2017; pp. 4700–4708.
13. Long, J.; Shelhamer, E.; Darrell, T. Fully Convolutional Networks for Semantic Segmentation. In Proceedings of Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Boston, MA, USA, 7–12 June 2015; pp. 3431–3440.
14. 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]
15. 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.
16. 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.
17. Woo, S.; Park, J.; Lee, J.-Y.; So Kweon, I. CBAM: Convolutional Block Attention Module. In Proceedings of the European Conference on Computer Vision (ECCV), Munich, Germany, 8–14 September 2018; pp. 3–19.
18. Park, J.; Woo, S.; Lee, J.-Y.; Kweon, I.S. BAM: Bottleneck Attention Module. *arXiv* **2018**, arXiv:1807.06514.
19. Hoeser, T.; Kuenzer, C. Object Detection and Image Segmentation with Deep Learning on Earth Observation Data: A Review-Part I: Evolution and Recent Trends. *Remote Sens.* **2020**, *12*, 1667. [CrossRef]
20. Kussul, N.; Lavreniuk, M.; Skakun, S.; Shelestov, A. Deep Learning Classification of Land Cover and Crop Types Using Remote Sensing Data. *IEEE Geosci. Remote. Sens. Lett.* **2017**, *14*, 778–782. [CrossRef]
21. Zhu, L.; Chen, Y.; Ghamisi, P.; Benediktsson, J.A. Generative Adversarial Networks for Hyperspectral Image Classification. *IEEE Trans. Geosci. Remote Sens.* **2018**, *56*, 5046–5063. [CrossRef]
22. Liu, Y.; Gao, L.; Xiao, C.; Qu, Y.; Zheng, K.; Marinoni, A. Hyperspectral Image Classification Based on a Shuffled Group Convolutional Neural Network with Transfer Learning. *Remote Sens.* **2020**, *12*, 1780. [CrossRef]
23. Zou, Q.; Ni, L.; Zhang, T.; Wang, Q. Deep Learning Based Feature Selection for Remote Sensing Scene Classification. *IEEE Geosci. Remote Sens. Lett.* **2015**, *12*, 2321–2325. [CrossRef]
24. Zhang, Z.; Liu, Q.; Wang, Y. Road Extraction by Deep Residual U-Net. *IEEE Geosci. Remote Sens. Lett.* **2018**, *15*, 749–753. [CrossRef]
25. Xu, Y.; Xie, Z.; Feng, Y.; Chen, Z. Road Extraction from High-Resolution Remote Sensing Imagery Using Deep Learning. *Remote Sens.* **2018**, *10*, 1461. [CrossRef]
26. 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]
27. Liu, P.; Liu, X.; Liu, M.; Shi, Q.; Yang, J.; Xu, X.; Zhang, Y. Building Footprint Extraction from High-Resolution Images via Spatial Residual Inception Convolutional Neural Network. *Remote Sens.* **2019**, *11*, 830. [CrossRef]
28. Han, X.; Zhiyun, O.; Xiaoke, W.; Jingzhu, Z. Spatial Distribution Characteristics of Soil Erosion in Hainan Island by GIS. *J. Soil Water Conserv.* **1999**, *5*, 75–80.

29. Yang, Y.; Newsam, S. Bag-Of-Visual-Words and Spatial Extensions for Land-Use Classification. In Proceedings of the 18th SIGSPATIAL International Conference on Advances in Geographic Information Systems, San Jose, CA, USA, 2–5 November 2010; pp. 270–279.
30. Shao, Z.; Yang, K.; Zhou, W. Performance Evaluation of Single-Label and Multi-Label Remote Sensing Image Retrieval Using a Dense Labeling Dataset. *Remote Sens.* **2018**, *10*, 964. [CrossRef]
31. Shao, Z.; Zhou, W.; Deng, X.; Zhang, M.; Cheng, Q. Multilabel Remote Sensing Image Retrieval Based on Fully Convolutional Network. *IEEE J. Sel. Top. Appl. Earth Obs. Remote Sens.* **2020**, *13*, 318–328. [CrossRef]
32. 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.
33. Sudre, C.H.; Li, W.; Vercauteren, T.; Ourselin, S.; Cardoso, M.J. Generalised Dice Overlap as a Deep Learning Loss Function for Highly Unbalanced Segmentations. In *Deep Learning in Medical Image Analysis and Multimodal Learning for Clinical Decision Support*; Springer: Cham, Switzerland, 2017; pp. 240–248.
34. Lin, T.-Y.; Goyal, P.; Girshick, R.; He, K.; Dollár, P. Focal Loss for Dense Object Detection. In Proceedings of the IEEE International Conference on Computer Vision, Venice, Italy, 22–29 October 2017; pp. 2980–2988.
35. Milletari, F.; Navab, N.; Ahmadi, S.-A. V-Net: Fully Convolutional Neural Networks for Volumetric Medical Image Segmentation. In Proceedings of the 2016 Fourth International Conference on 3D Vision (3DV), Stanford, CA, USA, 25–28 October 2016; pp. 565–571.
36. Rahman, M.A.; Wang, Y. Optimizing Intersection-Over-Union in Deep Neural Networks for Image Segmentation. In Proceedings of the International Symposium on Visual Computing, Las Vegas, NV, USA, 12–14 December 2016; pp. 234–244.
37. He, K.; Zhang, X.; Ren, S.; Sun, J. Delving Deep into Rectifiers: Surpassing Human-Level Performance on ImageNet Classification. In Proceedings of the IEEE International Conference on Computer Vision, Santiago, Chile, 13–16 December 2015; pp. 1026–1034.