remote sensing

MDPI

Article

Very High Resolution Images and Superpixel-Enhanced Deep Neural Forest Promote Urban Tree Canopy Detection

Yang Liu [1,2,3], Huaiqing Zhang [1,2,3,*], Zeyu Cui [1,2,3], Kexin Lei [1,2,3], Yuanqing Zuo [1,2,3], Jiansen Wang [1,2,3], Xingtao Hu [1,2,3,4] and Hanqing Qiu [1,2,3]

1 Institute of Forest Resource Information Techniques, Chinese Academy of Forestry, Beijing 100091, China
2 Key Laboratory of Forestry Remote Sensing and Information System, The National Forestry and Grassland Administration, Beijing 100091, China
3 Dongting Lake Remote Sensing Product Validation Station, Beijing 100091, China
4 School of Geography and Environmental Sciences, Guizhou Normal University, Guiyang 550025, China
* Correspondence: zhang@ifrit.ac.cn

Abstract: Urban tree canopy (UTC) area is an important index for evaluating the urban ecological environment; the very high resolution (VHR) images are essential for improving urban tree canopy survey efficiency. However, the traditional image classification methods often show low robustness when extracting complex objects from VHR images, with insufficient feature learning, object edge blur and noise. Our objective was to develop a repeatable method—superpixel-enhanced deep neural forests (SDNF)—to detect the UTC distribution from VHR images. Eight data expansion methods was used to construct the UTC training sample sets, four sample size gradients were set to test the optimal sample size selection of SDNF method, and the best training times with the shortest model convergence and time-consumption was selected. The accuracy performance of SDNF was tested by three indexes: F1 score (F1), intersection over union (IoU) and overall accuracy (OA). To compare the detection accuracy of SDNF, the random forest (RF) was used to conduct a control experiment with synchronization. Compared with the RF model, SDNF always performed better in OA under the same training sample size. SDNF had more epoch times than RF, converged at the 200 and 160 epoch, respectively. When SDNF and RF are kept in a convergence state, the training accuracy is 95.16% and 83.16%, and the verification accuracy is 94.87% and 87.73%, respectively. The OA of SDNF improved 10.00%, reaching 89.00% compared with the RF model. This study proves the effectiveness of SDNF in UTC detection based on VHR images. It can provide a more accurate solution for UTC detection in urban environmental monitoring, urban forest resource survey, and national forest city assessment.

Keywords: VHR; urban tree canopy; superpixel-enhanced deep neural forest; remote sensing

Citation: Liu, Y.; Zhang, H.; Cui, Z.; Lei, K.; Zuo, Y.; Wang, J.; Hu, X.; Qiu, H. Very High Resolution Images and Superpixel-Enhanced Deep Neural Forest Promote Urban Tree Canopy Detection. *Remote Sens.* **2023**, *15*, 519. https://doi.org/10.3390/rs15020519

Academic Editor: Austin Troy

Received: 13 December 2022
Revised: 7 January 2023
Accepted: 13 January 2023
Published: 15 January 2023

1. Introduction

Urban forest is an important material basis for urban economic development and can assure the physical and mental health of urban residents [1]. Canopy cover in urban regions is one of the most important indicators for monitoring planning and evaluating urban forests [2]. In 2007, China issued The National Forest City Assessment Indicators [3], it set tree canopy cover as an important indicator for evaluating the forest city. Urban tree canopy (UTC) is generally estimated as the area covered by tree leaf layers, branches and trunks when viewed vertically above the trees [4]. UTC is the most relevant stability optimization factor with the urban forest health assessment [5]. In this study, the term "UTC" refers to all trees in the urban region, including individual street trees and clusters of park trees, while peri-urban woods extend to the outskirts of the metropolitan area. City parks include urban forests larger than 0.5 ha, pocket parks and gardens with trees, trees on streets or in public squares, any other green areas with trees, and riparian corridors and roofs [1].

In recent years, the analysis of very high resolution (VHR) images has proven very useful for extracting UTC [2–7]. With the unique spatial location (streetside, roadside,

waterside, and building-side) [8], there have been great challenges for the segmentation of the accurate and consistent boundary based on VHR image classification [9]. The traditional remote sensing classification methods for medium- and low-resolution images are not suitable for VHR images, which contain abundant texture information [10]. Superpixel clusters image regions through grouping pixels; comparing the pixel, it provides image data in a more natural representation [11–13]. Learning-based superpixel segmentation method can effectively alleviate edge blur and classification noise [14–16].

Traditional machine learning methods heavily depend on the feature extraction process, such as Markov random fields (MRFs), random forest (RF) and conditional random fields (CRFs) [13]. However, deep convolutional neural networks (DCNNs) have better performance in image classification and learning feature representation simultaneously [17,18]. Although the powerful representation capability of DCNNs has promoted the development of image semantic segmentation, there are still many challenges: (1) The complexity of the surface feature spectrum, the features of high intraclass variance and low interclass variance are widely distributed, which easily confused in VHR images. DCNNs still exhibit a significant imbalance between the training of classifiers and representations for VHR images [19,20]. Simple categorization criteria are still insufficient to differentiate between ground objects. For instance, trees and submerged vegetation are frequently confused by spectral similarity [21–24]. (2) Ground objects in VHR images have strong requirements for consistency and boundary accuracy [25]. DCNNs have made remarkable achievements in semantic segmentation. Expanding the receptive field also increases edge blur and noise, since it employs a broad receptive field to obtain context information for classification [26]. In order to address these two challenges, we introduced a new VHR image classification framework—superpixel-enhanced deep neural forests (SDNF) [13].

The purpose of this study is to improve SDNF semantic segmentation based on Gaofen 2 (GF-2) VHR images, end-to-end integration of deep neural network and decision tree, and add a superpixel-enhanced region module (SRM) to conduct the framework increase the consistency and boundary precision for objects and to improve the sparse and small object detection ability for VHR images.

2. Materials and Methods

2.1. Study Area

The study area is Hui'an county, which is located on the southeast coast of Fujian province (Figure 1A,B), between Quanzhou Bay and Meizhou Bay. The total area of the study area is 793.04 km^2 (Figure 1C,D). The climate of Hui'an has four basic characteristics: high temperature, rich light and heat, abundant precipitation and monsoon climate. The precipitation is mainly concentrated in spring and summer, annual average precipitation approximately 1000.00 mm.

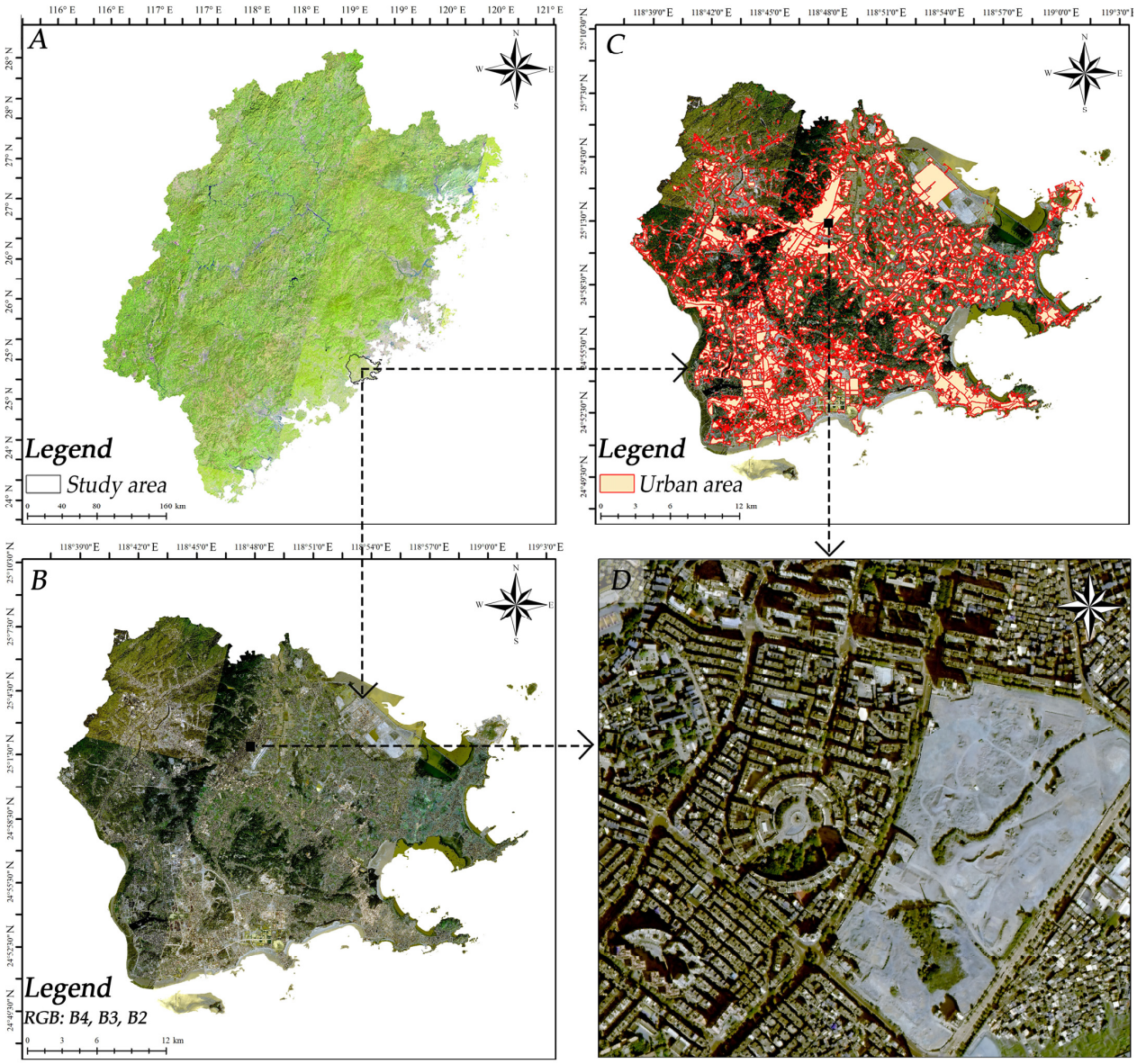


Figure 1. Study area: (**A**) the location of study area in Fujian province; (**B**) the GF-2 images of study area; (**C**) the distribution of urban areas of study area; (**D**) the images after superpixel-enhancement.

2.2. Materials

2.2.1. Data Source

In order to accurately detect the distribution of tree crowns in the urban area, seven VHR Gaofen-2 (GF-2) images with no cloud coverage were collected in 2021. GF-2 includes two PMS sensors that can obtain multispectral image with four bands, the spectral range is 0.45–0.89 μm and the spatial resolution is 4 m. In addition, PMS sensor can also obtain panchromatic images with higher spatial resolution, for which spectral range is 0.45~0.89 μm with one band, and the spatial resolution is 1 m. See Table 1 for details. The auxiliary data include digital elevation model (DEM) data of 30 m, the forest resource planning and design survey data at the study area in 2020 (FRPDSD2020) and a Google Earth visual selection of 40,000 images of sample data.

Table 1. GF-2 image band information.

ID	Acquisition Date	Center Longitude and Latitude	Scene Identifier	Senor	Spatial Resolution (Multispectral and Panchromatic)
1	2021/11/25	E118.7_N25.1	9600277	PMS 1	4 m and 1 m
2	2021/11/25	E118.6_N24.9	9600278	PMS 1	4 m and 1 m
3	2021/11/30	E119.1_N25.1	9617729	PMS 1	4 m and 1 m
4	2021/11/30	E119.1_N24.9	9617730	PMS 1	4 m and 1 m
5	2021/11/30	E119.0_N24.8	9617731	PMS 1	4 m and 1 m
6	2021/11/25	E118.9_N25.1	9600565	PMS 2	4 m and 1 m
7	2021/11/25	E118.8_N24.9	9600566	PMS 2	4 m and 1 m

2.2.2. Data Preprocessing

(1) The GF-2 absolute radiometric calibration coefficient provided by China Centre for Resources Satellite Data and Application (http://www.cresda.com/CN/ (accessed on 17 November 2021) to perform radiometric calibration calibrated panchromatic data as surface reflectance (SR). (2) The FLAASH module was used to calibrate the atmospheric correction on SR images. (3) The DEM and FRPDSD2020 were used to process orthorectification and geo-spatial registration for GF-2 images, registration error < 1 pixel. (4) Fused the corrected multispectral and panchromatic images. (5) Clipped the fused images and used the study area boundary to obtain the region of interest (ROI). The above operations were implemented in ENVI 5.3 and ArcGIS 10.2. Finally, the ROI images with 1m spatial resolution and four multispectral bands, and the root mean square error is 0.039 m. The image size of the study area was 34,620 rows and 41,948 columns, with a total area of 793.04 km^2, as shown in Figure 1C.

2.3. Method

In this study, based on the GF-2 VHR images, the SDNF framework was used [13] to solve the problems in UTC extraction where detection difficulty and boundary blur always occur.

2.3.1. SDNF

SDNF combined the advantages of DCNNs and tree-based model in encoder-decoder framework, to enhance decision tree (DT) and deep neural forest (DNF) image classification ability. There are two main parts: DNF and SRM. In order to improve the classification ability of DCNNs, the encoder was based on atrous spatial pyramid pooling (ASPP) and the residual neural network 101 (ResNet101), and closely combined with deep neural forest (DNF). The image features encoded in ResNet101's first four modules were delivered to the ASPP. As the following part, we focused on the interpretation of ASPP to represent different proportions of local and global data at various scales. Then the differentiable forest decision nodes were predicted through random paths using the fully differentiable forest as the decoder. In the SRM module, a learning-based superpixel segmentation task was introduced to emphasize the homogeneity inner the superpixel and the heterogeneity between superpixels to reduce classification noise and edge blur. Superpixel constraints were added in an end-to-end manner with designing superpixel enhanced segmentation loss. The SDNF framework is shown in Figure 2.

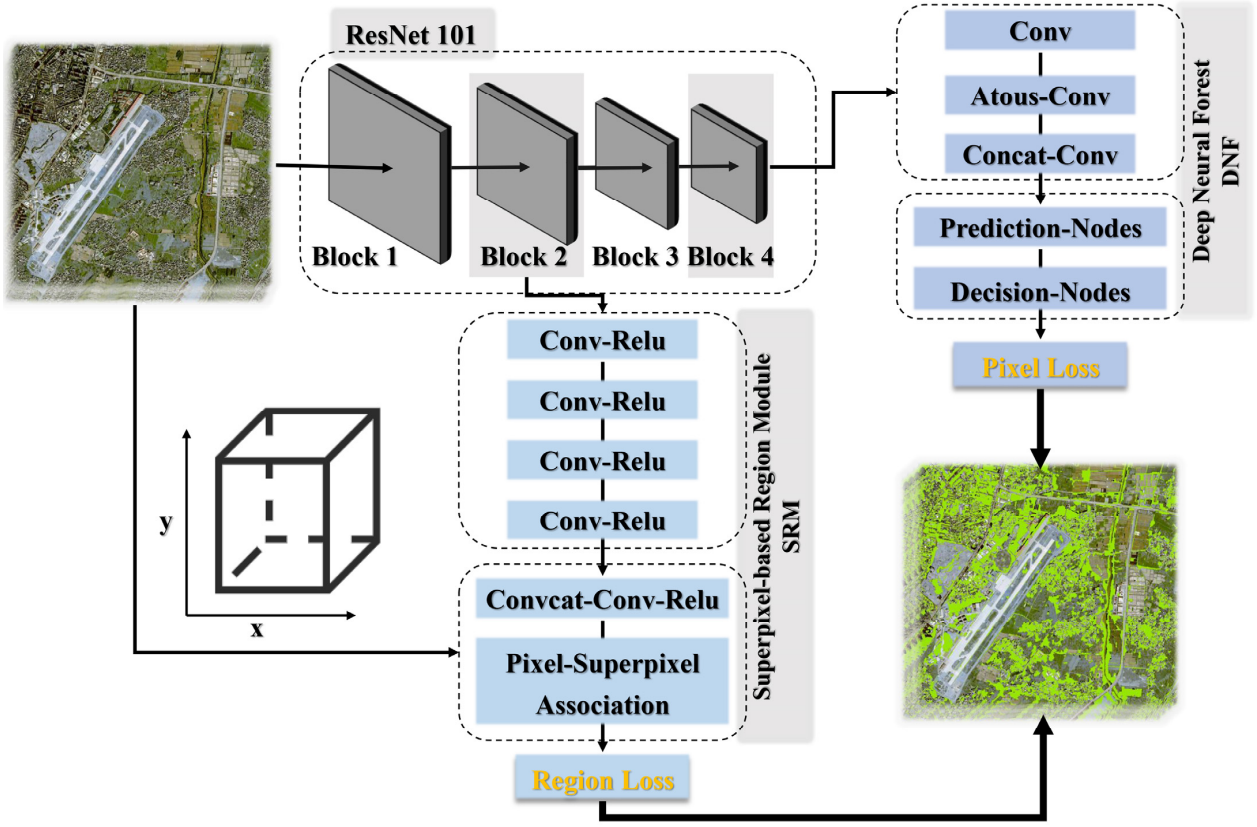


Figure 2. The structure of SDNF model.

(1) Deep neural forest (DNF). Based on ASPP, the feature representation of image classification considered more context information from the image scenario. However, if the number of convolution kernels was increased, the repeated combination of maximum pool and step size would lead to the number of parameters increased while the size of feature mapping would decrease. The former made network training more difficult, while the latter made difficult application of the decoder for upsampling. Atrous-Conv can avoid these problems (Figure 3). For each location (i, j) on mapping of input feature x_{conv} and the filter "f", the formulation of atrous convolution be expressed as following:

$$y_{conv}[i\ ,\ j] \equiv \sum_{k=1}^{|K|} x_{conv}[i + rk\ ,\ j + rk]\ f[k\ ,\ k], \tag{1}$$

where K was the kernel size of filter f, y_{conv} was the output convolution result and r was the corresponding atrous rate.

In Figure 4, when rate = 6, the receptive field of the Atrous Conv convolution nucleus was 23 × 23; when rate = 12, the receptive field of Atrous Conv was 47 × 47; when rate = 18, the receptive field of Atrous Conv was 71 × 71; when rate = 24, the receptive field of Atrous Conv was 99 × 99, respectively.

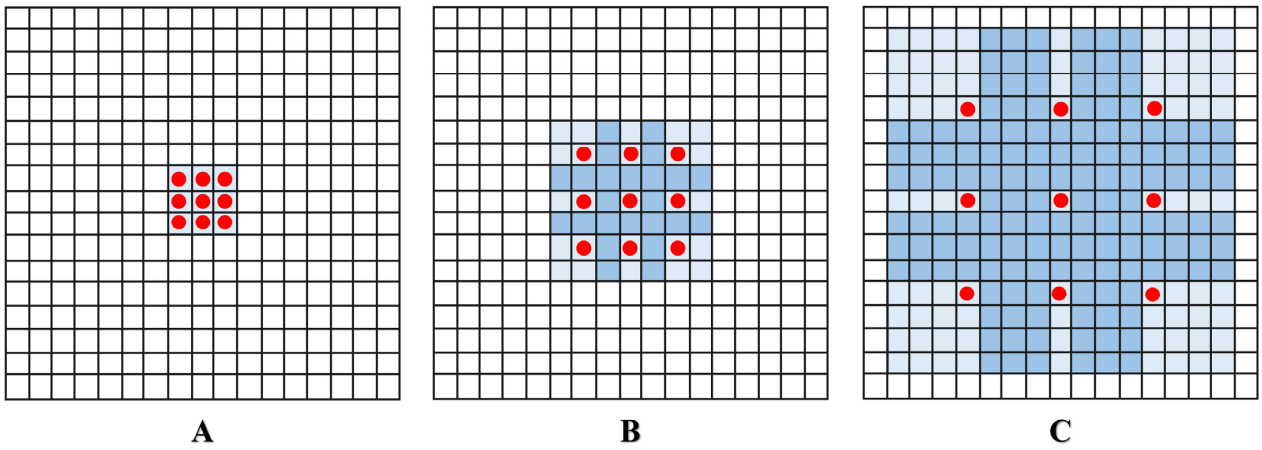


Figure 3. Conceptual graph of atrous convolution. (**A**) The 1-dilated conv of 3 × 3; (**B**) the 2-dilated conv of 3 × 3; (**C**) the 4-dilated conv of 3 × 3.

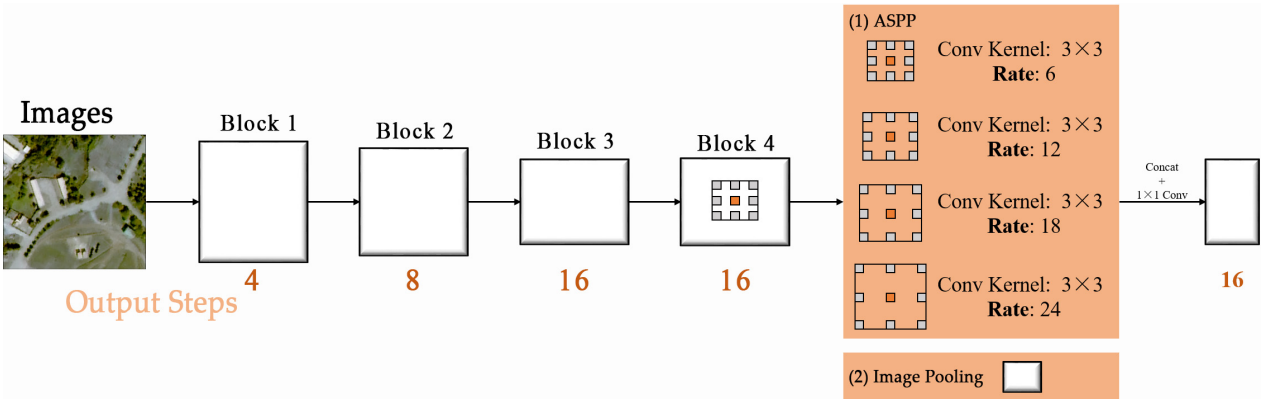


Figure 4. The structure of ASPP.

A differentiable forest was introduced as a decoder to conduct the representation learning of the DCNN. The decision tree was a tree-structured classifier consisting of split nodes and leaf nodes. The differentiable forest can be implemented to the overall framework of DCNNs through stochastic routing, which was different from traditional decision forest with deterministic routing rules. Decision nodes indexed by $\mathcal{N}$ were internal nodes of the tree, while prediction nodes indexed by $\mathcal{L}$ were the final nodes of the tree. Each prediction node $l \in \mathcal{L}$ held a probability distribution π_l over y. Each decision node $n \in \mathcal{N}$ was instead assigned a decision function $d_n(\,;\Theta) : x \to [0,1]$ parametrized by Θ, which was responsible for routing samples along the tree. When a sample $x \in \mathcal{X}$ reached a decision node n it was sent to the left or right subtree based on the output of $d_n(\,;\Theta)$. In standard decision forests, d_n was binary and the routing is deterministic. Rather, in this paper, we consider a probabilistic route. Once a sample ends in a leaf node l, the related tree prediction is provided by the class label distribution π_l. In the case of stochastic routings, the leaf predictions are averaged by the probability of reaching the leaf [27]. The final prediction for sample x from tree $\mathcal{T}$ with decision nodes parametrized by Θ as the following:

$$P_T(y|x,\Theta,\pi) = \sum_{l \in \mathcal{L}} \pi_{ly}\mu_l(x|\Theta), \tag{2}$$

where $\mu_l(x|\Theta)$ is regarded as the routing function providing the probability that sample x will reach leaf l. $\sum_l \mu_l(x|\Theta) = 1$ for all $x \in \mathcal{X}$, $\pi = (\pi_l)_{l\in\mathcal{L}}$ and π_{ly} denotes the probability of a sample reaching leaf l to take on y.

(2) Superpixel-based region module (SRM). A superpixel is a small area composed of adjacent pixels with similar color, brightness, texture and other features. Most of these regions have effective information for image segmentation and have generally not destroyed the boundary information of objects in the image. A superpixel is the abstraction of basic information elements, dividing a pixel level image into a district level image. To enhance the target boundary and ensure the regional continuity, SDNF introduced superpixel segmentation as an auxiliary module. The block 2 mapping features of residual network (ResNet) was sent to a various of convolution layers to activate was rectified linear unit (ReLU) and batch normalization (BN). The maximum pool's window size was set to 2, which was used to expand the receptive field after secondary convolution. The convolution filter output channel was set to 64, and the convolutional neural network (CNN) output channel was set to m. Concatenation was conducted on the original image, the input of ResNet block 2, and two pooled features in XYLab. The outcome after combining XYLab channels was m+5 dimensional pixel features, m+5 dimensional feature was transferred to two modules of simple linear iterative cluster (SLIC) [28], which have updated the correlation between the v iteration pixel and the super pixel center. The complexity of SLIC was linear in the number of pixels in the image O (N), while the conventional k-means algorithm was O (kNI), where I was the number of iterations. This provided a search space for each cluster center in the allocation step. Figure 5A shows that in the conventional k-means algorithm, the distance from each cluster center to each pixel in the image was calculated. Figure 5B shows that SLIC only calculates data from each cluster center to 2S × Distance of pixels in 2S area. The expected superpixel size here was only S × S, represented by a smaller square. This method not only reduces the distance calculation, but also makes the complexity of SLIC independent of the number of superpixels.

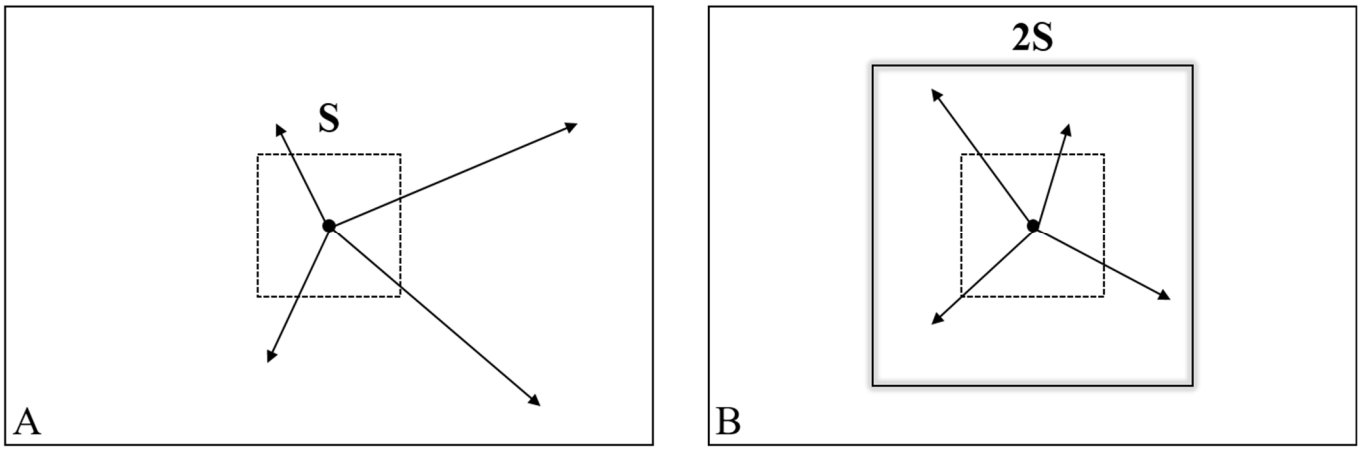


Figure 5. Superpixel reduction search region. (**A**) The standard k-means search region; (**B**) SLIC model search region.

For pixel (P) and super pixel (SP) in iteration t, the correlation between pixel and superpixel can be expressed as following:

$$Q^{t}_{(p,sp)} = e^{-\|I_p - C^{t-1}_{sp}\|}, \quad (3)$$

where I represented feature mapping from convolution layer, super pixel center C^{t}_{i} was modified to the weighted sum of pixel characteristics: $C^{t}_{i} = \frac{1}{M^{t}_{i}} \sum^{n}_{p=1} Q^{t}_{pi} I_p$ and $M^{t}_{i} = \sum_{p} Q^{t}_{pi}$ were a general term.

2.3.2. Expand Training Samples

Remote sensing images provide a wealth of spectral information that is simple to operate, such as color transformation, band combination and band computation. Due to

the limitation of time and space in the study area, in order to obtain enough sample data to train the depth network model, this study expanded the sample on the basis of the original image data. For model training, the image of the whole experimental area was cut into 256 × 256 pixel blocks, the number of pixel blocks at the entire study area was 14,077 blocks, and eight-image transformation processing was designed on each block. The transformation results are shown in Figure 6, where Figure 6A is the original image of RGB band combination B1, B2, B3; Figure 6B is the original image of B1, B2 and B3 after high contrast retention processing; Figure 6C is the maximum processed image of the original image based on RGB band combination B1, B2, B3; Figure 6D is the minimum processed image of the original image based on RGB band combination B1, B2, B3; Figure 6E is the uniform color processed image of the original image based on RGB band combination B1, B2, B3; Figure 6F is the processed image with Gaussian noise added to the original image based on RGB band combination B1, B2, B3; Figure 6G is RGB band combination B3, B2, B1—false color image; Figure 6H is the RGB band combination B3, B4, B2—false color image; Figure 6I is the RGB band combination B4, B3, B2—false color image. To evaluate the effect of sample size on the model, using random sampling, 5000, 10,000, 20,000 and 40,000 different gradient sample size gradients were set, respectively, in which 2/3 of the samples in each gradient were used for training the model and 1/3 of the samples were used for model accuracy verification.

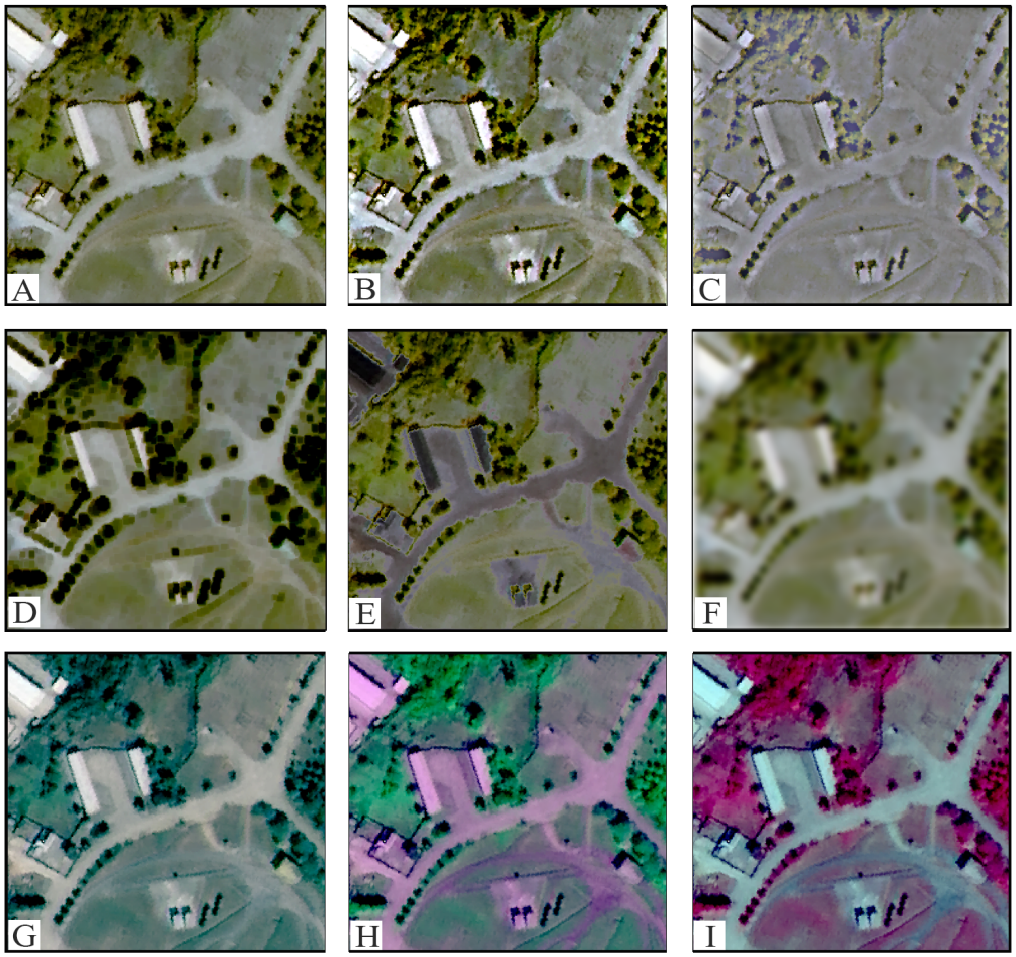


Figure 6. Sample expanded images: (**A**) original picture (band 1, band 2, band 3 combination); (**B**) high contrast retention; (**C**) maximum value; (**D**) minimum value; (**E**) overexposure; (**F**) Gaussian noise; (**G**) band 3, band 2, band 1 combination; (**H**) band 3, band 4, band 2 combination; (**I**) band 4, band 3, band 2 combination.

2.3.3. Model Parameter Setting and Training

After expanding samples, L2 regularization was used, the single input sample size of the training model was 64. The initial learning rate of Adam Optimizer was set a 0.000001. In the BN layer, ε was set as 0.000001, and attenuation was set as 0.9997. In ASPP, the multi-grids were set as [2,4,8], the atrous rate was [1,6,12,18], and the output step size was 16. For DNF module, the number of trees was set as 3 and the depth of trees was 9. For the SRM module, the superpixel value was set as 64 and the number of iterations was 10. The memory of graphics processing unit (GPU) was 24 G, which was used to speed up the training process. The details of parameters were shown in Figure 7.

Module	Name	Value
Global	Batch Size	64
	Starting learning rate	0.00001
ASPP	Multi-grids	[2，4，8]
	Atrous rate	[1，6，12，18]
	Output stride	16
DNF	Number of trees	3
	Depth of tree	9
SRM	Number of superpixels	64
	Iteration	10

Figure 7. Model parameters.

2.3.4. Comparison Method

In this study, VHR image semantic segmentation methods have three significant characteristics: (1) The segmentation process in one stage: SDNF combined differentiable forest and deep neural network in an end-to-end manner to improve model performance. (2) Without post-processing. (3) Without other information. Therefore, RF classifier was selected as the comparison method, which used the training samples in Section 2.2.2 to extract UTC synchronously. RF classifier is the mainstream method of remote sensing image supervised classification at present. Based on decision tree learner, which was constructed with Bagging integration, introduced random attribute selection.

2.3.5. Accuracy Assessment

In order to evaluate the performance of SDNF in UTC extraction, three overall benchmark indicators were used, such as F1 score (F1), intersection over union (IoU) and overall accuracy (OA). The formula of indicators are as follows:

$$F_1 = 2 \times \frac{P \times R}{P + R}, \; P = \frac{TP}{TP + FP}, \; R = \frac{TP}{TP + FN}, \tag{4}$$

$$IoU = \frac{TP}{TP + FP + FN}, \tag{5}$$

$$OA = \frac{TP + FN}{N}, \tag{6}$$

where TP, FP and FN are the number of true positives, false positives and false negatives, respectively. P represents precision, R represents recall rate and N represents the total number of pixels.

3. Results

3.1. Sample Size

In order to evaluate the impact of the sample size on the SDNF model, four gradients: 5000, 10,000, 20,000 and 40,000 were input. The model fitting accuracy, classification accuracy and running time of SNDF models were tested, respectively. The final classification accuracy OA of the SNDF model and RF classifier are shown in Figure 8A. The performance of SNDF and RF in training time Figure 8B. The performance of the SNDF model in F1, IoU and OA are shown in Figure 8C. The overall results shown that when the sample size was 5000, the overall classification of the two models was lower, and the stitching line of the extracted results was obvious. With the increase in sample size, the accuracy of both models improved, but the RF model trend of improvement was small. After the sample size was increased eightfold, the accuracy had only increased by 0.13, compared with SNDF model, which increased by 0.21. In terms of training time, with the increase of sample size, the training time increased significantly. But in all four sample gradients, the RF model took more time than the SNDF model. For the SNDF model, when 40,000 samples were selected, the OA had reached a higher level, up to 0.89.

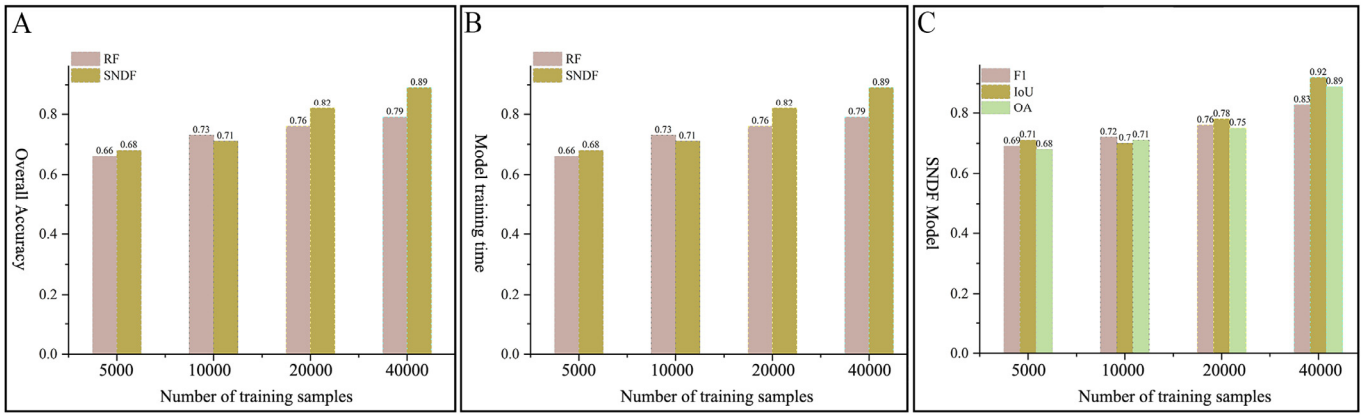


Figure 8. The influence of sample size on model accuracy and efficiency. (**A**) The final classification accuracy OA of the SNDF model and RF classifier; (**B**) the performance of SNDF and RF in training time; (**C**) the performance of SNDF model in F1, IoU and OA.

3.2. Training Loss and Elapsed Time

In the SNDF, the loss function was used to assess the model training procedure, the model convergence, and if the model was overfitted, it may also have reflected the error between the target detection model's final prediction result and the actual true value [29]. Figure 9 shows the change of the loss function in the model training process with different epoch times. With the model training, the loss function started to converge gradually, and after 160 training times, it was at the stable point of 0.14. Figure 10 shows the pictures of pixel reconstruction with different training data. Line 1 to line 4 shows the low-resolution image, after model-reconstructed images and the original images. When epoch = 30, 80, 160 and 220, with the increase in iteration times, the better the image restoration effect. The image restored after the 80th epoch had reduced the impact of the noise block greatly, and the images were more smooth and clear.

Regarding the operation environment of this study, the training time of SDNF was longer than that of RF. On a single GPU, it took 2 h and 45 min to the test study area based on RF, while the process of SDNF only took 1 h and 37 min. There was no significant difference between the SDNF and the RF in the test time of each sample. RF took an average of 0.012 s to predict the semantic segmentation results of samples, while SDNF took 0.014 s at the same time.

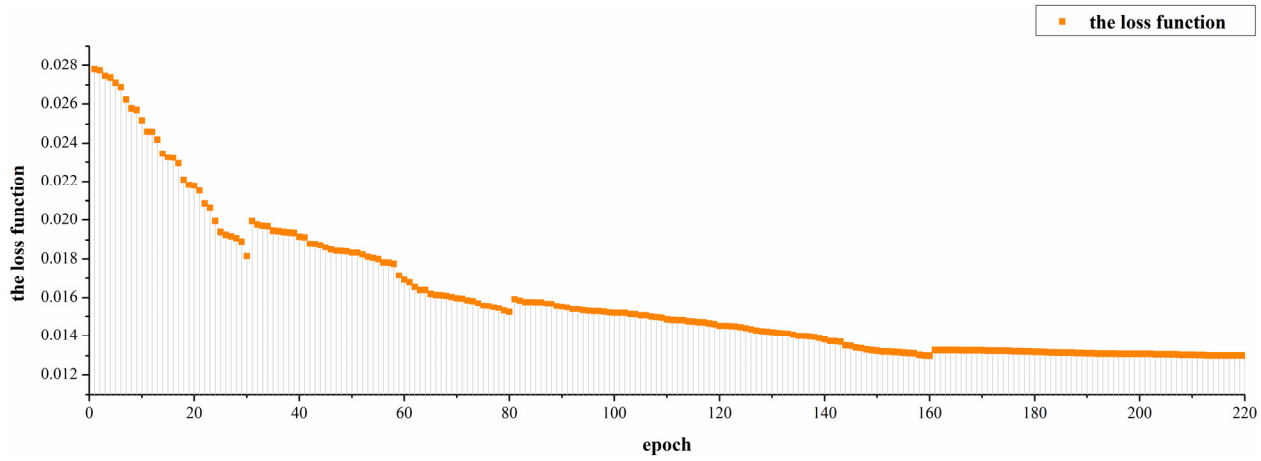


Figure 9. The variation curve of loss function value with the number of training rounds.

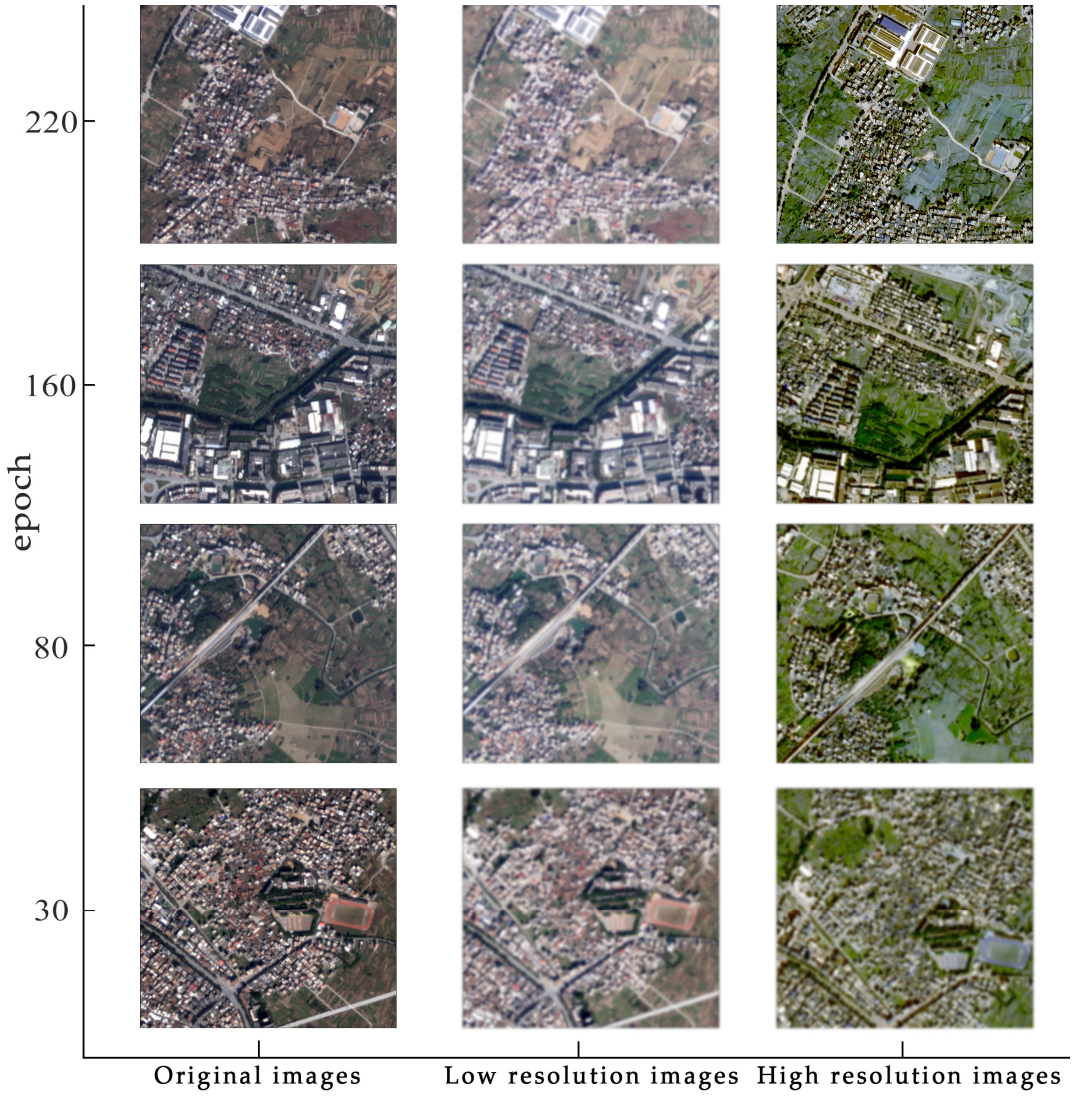


Figure 10. Superpixel reconstruction of training sample change.

3.3. Superpixel Segmentation

In this study, the superpixel segmentation results did not involve the training and testing process of the SDNF model. To present the effectiveness of the SRM module more visually, and better analyze the contribution of SLIC in this study, Figure 11 shows the comparison of the superpixel segmentation results before and after simple noniterative clustering (SNIC), SLIC and modified SLIC (MSLIC) were put into the SRM module. Figure 11A1,A2 show the segmentation results before and after SNIC, respectively; Figure 11B1,B2 show the segmentation results before and after SLIC, respectively; Figure 11C1,C2 show the segmentation results before and after MSLIC. The experimental results show that the segmentation results of SLIC-based superpixel segmentation in the SRM module were more accurate and suitable for VHR images, especially for the edge confusion areas of buildings and their shadows.

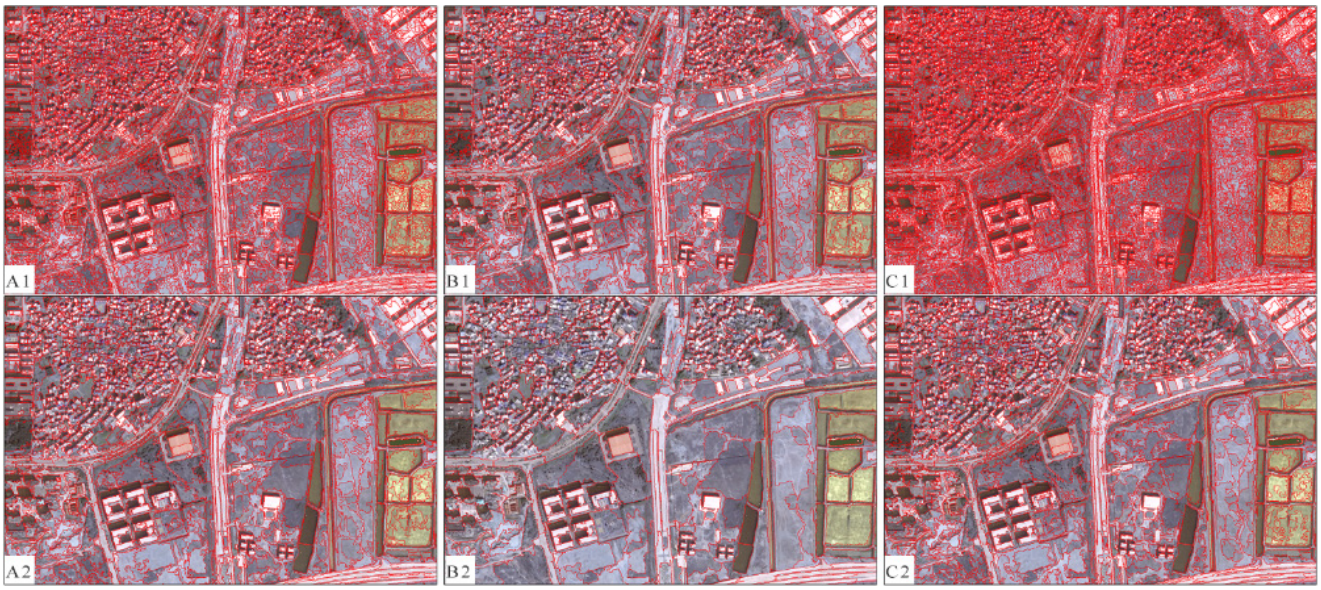


Figure 11. Comparison of SRM module segmentation effect. (**A1**,**A2**) The segmentation results before and after SNIC, respectively; (**B1**,**B2**) The segmentation results before and after SLIC, respectively, and (**C1**,**C2**) are the segmentation results before and after MSLIC.

3.4. Model Accuracy

The SDNF model was trained with the training sample, which was designed in Section 2.2, and the accuracy of the trained model was verified with the verification sample. After iterative training, the SDNF model converged when the epoch was 200, with training accuracy of 95.16% and verification accuracy of 94.87%; the RF model converged in the 160th epoch, with the training accuracy being 83.16% and verification accuracy being 87.73%.

Compared with the training accuracy, (1) the SDNF model has three deeper network layers, and the training accuracy improved significantly; (2) combining the semantic segmentation with semantic information edge detection can improve the prediction accuracy of object boundaries. The SDNF model enhanced object boundaries using superpixel segmentation; (3) the SDNF model end-to-end self-cascade network, and the sequential context aggregation from global to local improved the consistency of sample labels, the strategy from coarse to thinning improved the accuracy of the labels. Compared with the verification accuracy, the results show that the convergence effect of the SDNF model was better when it was at deeper training depths.

3.5. Classification Accuracy

As shown in Table 2, compared with RF model, the OA of SDNF improved significantly from 0.79 to 0.89, achieving higher detection accuracy, particularly for low and sparse trees (Figure 11). However, SDNF still made some error pixel in impervious surfaces and UTCs. Figure 12 shows the classification results, where column A is the superpixel-enhanced images; column B is the training sample images; column C is the classification result based on SDNF; D is the classification result of RF. The results show that compared with RF, SDNF improved its performance through end-to-end training, and combined the region loss with superpixels enhanced into the whole framework, emphasized the effective of object boundary zonation through edge detection. The performance of SDNF model varies with specific objects, and the confusion of low trees, sparse trees and building shadows improved significantly. For SNDF, the F1 improved 5%, IoU improved 12% and OA improved 10%. The results (Figure 13) show that, through the representation learning of DCNN under the guidance of differentiable forests (DF), the classification accuracy of easily confused objects was improved. This indicates that the SDNF made significant progress in UTC extraction.

Table 2. Comparison of classification accuracy between SDNF and RF Model.

Model	Accuracy Assessment		
	F1	**IoU**	**OA**
RF	0.78	0.80	0.79
SDNF	0.83	0.92	0.89

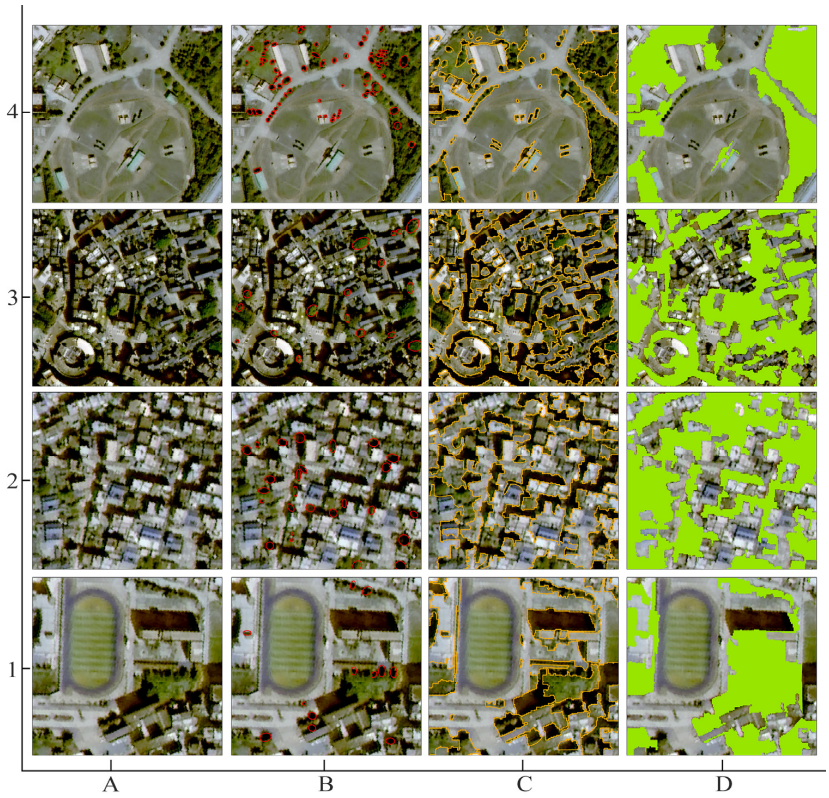


Figure 12. Detailed display of classification results of SDNF and RF methods. (**A**) The superpixel-enhanced images; (**B**) training sample images; (**C**) classification result based on SDNF; (**D**) classification result of RF.

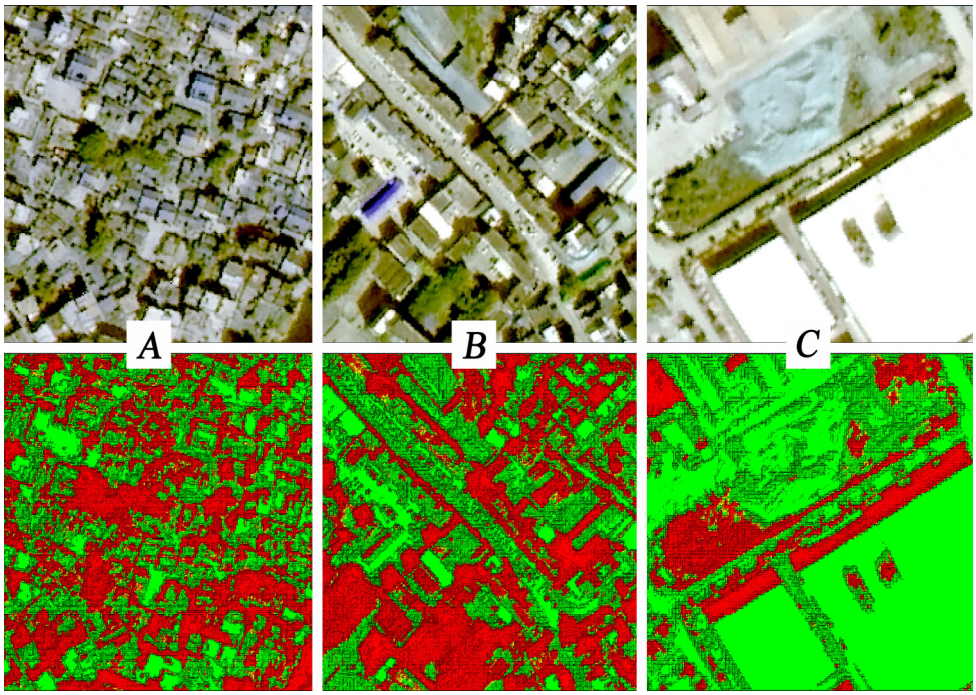


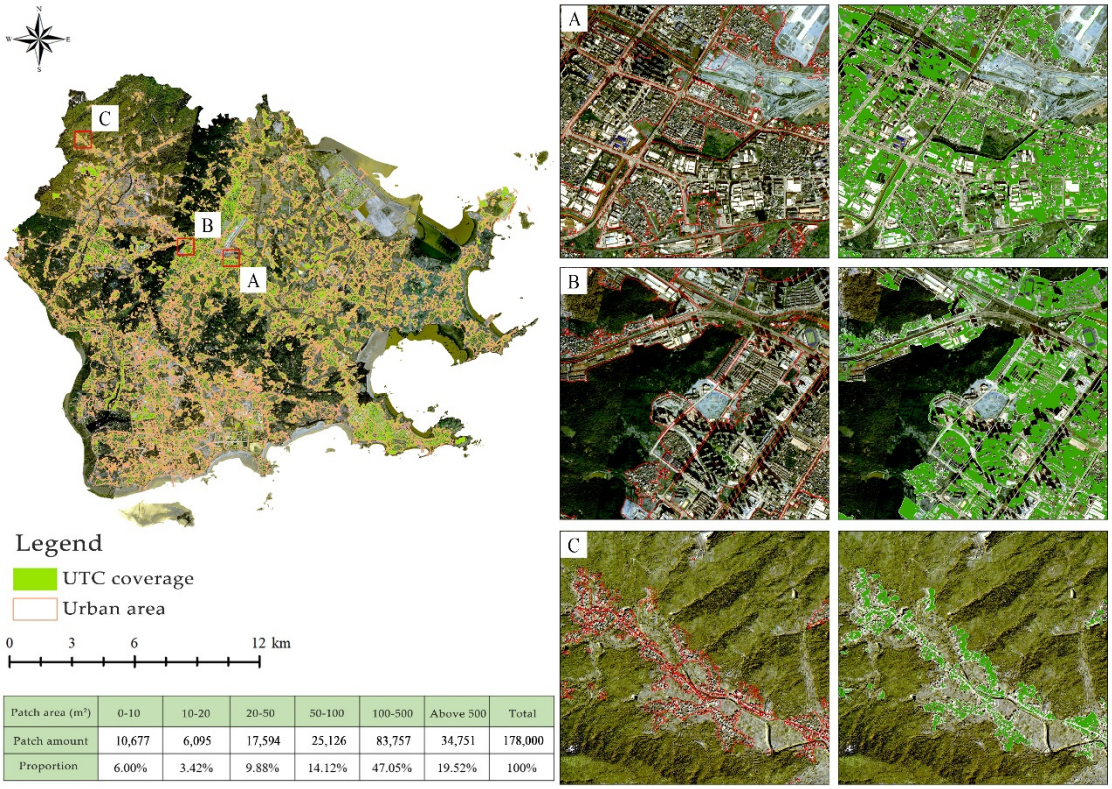
Figure 13. Detailed display of classification results of SDNF methods. Red represents the canopy coverage area; green represents the other land coverage area. (**A**) Urban residential region; (**B**) urban road region; (**C**) urban large building region.

4. Discussion

According to data published by pertinent authorities, nearly 70% of the world's population will reside in urban settings by 2050 [30]. The rapid growth of the global urban population drives the dynamic development of cities [31]; urban areas provide a home to 10 billion trees [1], which will play an increasingly essential role in preserving livable cities in the face of population and climatic pressures. Assessment methods are crucial for preserving and enhancing the ecosystem services offered by urban forests as well as determining the efficacy of policies intended to improve this important resource. Here, a new VHR image semantic segmentation framework—SDNF—was introduced to improve the sparse and small objects detection ability for VHR images.

In this study, satellite submicron VHR images were used to detect UTC. The traditional methods for VHR image classification, representation learning and classifier training were unbalanced, and the edge blur and noise problems of sparse and low objects were present (Figure 14). The SDNF framework can be used in any urban area for UTC detection. The framework is designed in two modules: DNF and SRM, which are responsible for pixel loss and region loss, respectively. SDNF is emphasized for improving DCNNs classification ability, combining DNF in the coding and decoding architecture, and reducing classification edge blur and noise. The introduced SRM module is used for learning-based superpixel segmentation tasks to dissolve the homogeneity within the superpixel and the heterogeneity between superpixels. The SDNF method goes beyond the capability of traditional remote sensing monitoring technology in UTC detection.

The availability and quality of training data are likely to be the most important impediments to applying this method more generally. Here, the sample data were expanded in the limited study region to produce a dense training dataset, which contains derived metrics in a limited spatial range (i.e., not directly measured), provides direct but limited observations of canopy structure throughout the study area. Other training data sources include current government tree datasets and Google Street View [32]. When substituting training data sets, the understanding of expected changes in model uncertainty and variance should be attained [33].



Patch area (m²)	0-10	10-20	20-50	50-100	100-500	Above 500	Total
Patch amount	10,677	6,095	17,594	25,126	83,757	34,751	178,000
Proportion	6.00%	3.42%	9.88%	14.12%	47.05%	19.52%	100%

Figure 14. Detailed results of UTC coverage in the study area. (**A**) The center of city; (**B**) The suburban region of the city; (**C**) Outermost region of the city.

Owing to the special location of urban forest distribution, unclear boundaries, and ease of obfuscation by the shadows of buildings, the regions with high canopy density of single tree segmentation in the process of target detection using the SDNF method cannot be addressed. How to detect single tree target of urban forest through optimization model or combining with other auxiliary data will be the content of the next step.

5. Conclusions

In this study, a method based on semantic segmentation object extraction—superpixel-enhanced deep neural forest (SDNF)—was used to solve the representation learning and classifier training unbalanced problem and the edge blur and noise of sparse and low objects in VHR images, such as UTC. The SDNF model balanced the generalization ability of the DCNNs model in classification and deep learning, introducing a completely differentiable forest to control the learning ability of deep convolution layer. At the same time, a superpixel-enhanced region module (SRM) was used to reduce the classification noise and enhance the edge details of objects in VHR images. The experimental results show that the training accuracy of the superpixel-enhanced deep neural forest (SDNF) was 95.16%, the verification accuracy was 94.87%, and the OA of UTC extraction was 89.17%, evidencing that SDNF reached a new level in UTC extraction. It has obvious advantages compared with the existing unified detection methods using the RF model: it can quickly, accurately and automatically detect the canopy cover in urban regions. Therefore, the SDNF method will provide technical support for improving the location extraction, spatial distribution planning and protection of canopy cover distribution in urban areas.

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

Funding: This research was funded by the Foundation Research Funds of Institute of Forest Resource Information Techniques (IFRIT), grant number CAFYBB2019SZ004 and the National Natural Science Foundation of China, grant number 32071681.

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

References

1. Endreny, T.A. Strategically Growing the Urban Forest Will Improve Our World. *Nat. Commun.* **2018**, *9*, 10–12. [CrossRef] [PubMed]
2. Ucar, Z.; Bettinger, P.; Merry, K.; Siry, J.; Bowker, J.M.; Akbulut, R. A Comparison of Two Sampling Approaches for Assessing the Urban Forest Canopy Cover from Aerial Photography. *Urban For. Urban Green.* **2016**, *16*, 221–230. [CrossRef]
3. The State Forestry Administration of the People's Republic of China. *National Forest City Evaluation Indicators*; Urban Forestry in China: Beijing, China, 2007; Volume 5, p. 2.
4. Jia, B.Q.; Liu, X.P. *Canopy Coverage Characteristics and Landscape Ecological Changes in the First Green Isolation Area in Beijing*; Scientia Silvae Sinicae: Beijing, China, 2017; Volume 53, p. 10.
5. Krajter Ostoić, S.; Salbitano, F.; Borelli, S.; Verlič, A. Urban Forest Research in the Mediterranean: A Systematic Review. *Urban For. Urban Green.* **2018**, *31*, 185–196. [CrossRef]
6. Erker, T.; Wang, L.; Lorentz, L.; Stoltman, A.; Townsend, P.A. A Statewide Urban Tree Canopy Mapping Method. *Remote Sens. Environ.* **2019**, *229*, 148–158. [CrossRef]
7. Alonzo, M.; Bookhagen, B.; Roberts, D.A. Urban Tree Species Mapping Using Hyperspectral and Lidar Data Fusion. *Remote Sens. Environ.* **2014**, *148*, 70–83. [CrossRef]
8. Mi, L.; Chen, Z. Superpixel-Enhanced Deep Neural Forest for Remote Sensing Image Semantic Segmentation. *ISPRS J. Photogramm. Remote Sens.* **2020**, *159*, 140–152. [CrossRef]
9. Kuffer, M.; Barros, J.; Sliuzas, R.V. The Development of a Morphological Unplanned Settlement Index Using Very-High-Resolution (VHR) Imagery. *Comput. Environ. Urban Syst.* **2014**, *48*, 138–152. [CrossRef]
10. Achanta, R.; Shaji, A.; Smith, K.; Lucchi, A.; Fua, P.; Süsstrunk, S. SLIC superpixels compared to state-of-the-art superpixel methods. *IEEE Trans. Pattern Anal. Mach. Intell.* **2012**, *34*, 2274–2282. [CrossRef]
11. Niu, S.F.; Qiu, E.F.; Zhang, Z.Y.; Xi, L. Gradient change of population diversity of woody plants in the urban riverbank forest of Beijing. *Sciatica Silvia Sin.* **2020**, *56*, 198–206.
12. Liu, B.; Hu, H.; Wang, H.; Wang, K.; Liu, X.; Yu, W. Superpixel-Based Classification with an Adaptive Number of Classes for Polarimetric SAR Images. *IEEE Trans. Geosci. Remote Sens.* **2013**, *51*, 907–924. [CrossRef]
13. Ren, C.Y.; Reid, I. *gSLIC: A Real-Time Implementation of SLIC Superpixel Segmentation*; University of Oxford, Department of Engineering: Oxford, UK, 2011; pp. 1–6.
14. Wei, J.; Huang, W.; Li, Z.; Sun, L.; Zhu, X.; Yuan, Q.; Liu, L.; Cribb, M. Cloud Detection for Landsat Imagery by Combining the Random Forest and Superpixels Extracted via Energy-Driven Sampling Segmentation Approaches. *Remote Sens. Environ.* **2020**, *248*, 112005. [CrossRef]
15. Huang, X.; Cao, Y.; Li, J. An Automatic Change Detection Method for Monitoring Newly Constructed Building Areas Using Time-Series Multi-View High-Resolution Optical Satellite Images. *Remote Sens. Environ.* **2020**, *244*, 111802. [CrossRef]
16. Wang, S.; Guan, K.; Zhang, C.; Zhou, Q.; Wang, S. Remote Sensing of Environment Cross-Scale Sensing of Field-Level Crop Residue Cover: Integrating Field Photos, Airborne Hyperspectral Imaging, and Satellite Data. *Remote Sens. Environ.* **2023**, *285*, 113366. [CrossRef]
17. Aloysius, N.; Geetha, M. A Review on Deep Convolutional Neural Networks. In Proceedings of the 2017 International Conference on Communication and Signal Processing (ICCSP 2017), Chennai, India, 6–8 April 2017; pp. 588–592. [CrossRef]
18. Liu, M.; Shi, J.; Li, Z.; Li, C.; Zhu, J.; Liu, S. Towards Better Analysis of Deep Convolutional Neural Networks. *IEEE Trans. Vis. Comput. Graph.* **2017**, *23*, 91–100. [CrossRef]
19. Waldner, F.; Diakogiannis, F.I. Deep Learning on Edge: Extracting Field Boundaries from Satellite Images with a Convolutional Neural Network. *Remote Sens. Environ.* **2020**, *245*, 111741. [CrossRef]
20. Gallwey, J.; Robiati, C.; Coggan, J.; Vogt, D.; Eyre, M. A Sentinel-2 Based Multispectral Convolutional Neural Network for Detecting Artisanal Small-Scale Mining in Ghana: Applying Deep Learning to Shallow Mining. *Remote Sens. Environ.* **2020**, *248*, 111970. [CrossRef]
21. Huang, B.; Zhao, B.; Song, Y. Urban Land-Use Mapping Using a Deep Convolutional Neural Network with High Spatial Resolution Multispectral Remote Sensing Imagery. *Remote Sens. Environ.* **2018**, *214*, 73–86. [CrossRef]
22. Zhang, D.; Pan, Y.; Zhang, J.; Hu, T.; Zhao, J.; Li, N.; Chen, Q. A Generalized Approach Based on Convolutional Neural Networks for Large Area Cropland Mapping at Very High Resolution. *Remote Sens. Environ.* **2020**, *247*, 111912. [CrossRef]

23. Zhang, C.; Sargent, I.; Pan, X.; Li, H.; Gardiner, A.; Hare, J.; Atkinson, P.M. An Object-Based Convolutional Neural Network (OCNN) for Urban Land Use Classification. *Remote Sens. Environ.* **2018**, *216*, 57–70. [CrossRef]
24. Persello, C.; Tolpekin, V.A.; Bergado, J.R.; de By, R.A. Delineation of Agricultural Fields in Smallholder Farms from Satellite Images Using Fully Convolutional Networks and Combinatorial Grouping. *Remote Sens. Environ.* **2019**, *231*, 111253. [CrossRef]
25. Wang, J.; Ma, A.; Zhong, Y.; Zheng, Z.; Zhang, L. Cross-Sensor Domain Adaptation for High Spatial Resolution Urban Land-Cover Mapping: From Airborne to Spaceborne Imagery. *Remote Sens. Environ.* **2022**, *277*, 113058. [CrossRef]
26. Cao, Y.; Huang, X. A Deep Learning Method for Building Height Estimation Using High-Resolution Multi-View Imagery over Urban Areas: A Case Study of 42 Chinese Cities. *Remote Sens. Environ.* **2021**, *264*, 112590. [CrossRef]
27. Kontschieder, P.; Fiterau, M.; Criminisi, A.; Bulo, S.R. Deep neural decision forests. In Proceedings of the IEEE International Conference on Computer Vision, Santiago, Chile, 7–13 December 2015; pp. 1467–1475. [CrossRef]
28. Chen, L.C.; Papandreou, G.; Kokkinos, I.; Murphy, K.; Yuille, A.L. DeepLab: Semantic Image Segmentation with Deep Convolutional Nets, Atrous Convolution, and Fully Connected CRFs. *IEEE Trans. Pattern Anal. Mach. Intell.* **2018**, *40*, 834–848. [CrossRef] [PubMed]
29. Ma, Y.K.; Liu, H.; Lin, C.X.; Zhao, F.; Jiang, X.; Zhang, Y.T. Study on target detection of mangrove single tree based on improved YOLOv5. *Adv. Lasers Optoelectron.* **2022**, *29*, 1828003.
30. Al-Surf, M.; Balabel, A.; Alwetaishi, M.; Abdelhafiz, A.; Issa, U.; Sharaky, I.; Shamseldin, A.; Al-Harthi, M. Stakeholder's Perspective on Green Building Rating Systems in Saudi Arabia: The Case of LEED, Mostadam, and the SDGS. *Sustainability* **2021**, *13*, 8463. [CrossRef]
31. Parnell, S.; Walawege, R. Sub-Saharan African Urbanisation and Global Environmental Change. *Glob. Environ. Chang.* **2011**, *21*, S12–S20. [CrossRef]
32. Li, X.; Zhang, C.; Li, W.; Ricard, R.; Meng, Q.; Zhang, W. Assessing Street-Level Urban Greenery Using Google Street View and a Modified Green View Index. *Urban For. Urban Green.* **2015**, *14*, 675–685. [CrossRef]
33. Baines, O.; Wilkes, P.; Disney, M. Quantifying Urban Forest Structure with Open-Access Remote Sensing Data Sets. *Urban For. Urban Green.* **2020**, *50*, 126653. [CrossRef]