This article has been accepted for publication in IEEE Access. This is the author's version which has not been fully edited and content may change prior to final publication. Citation information: DOI 10.1109/ACCESS.2023.3313390

IEEE Access
Multidisciplinary | Rapid Review | Open Access Journal

Date of publication xxxx 00, 0000, date of current version xxxx 00, 0000.

Digital Object Identifier 10.1109/ACCESS.2023.Doi Number

A Lightweight High-Resolution RS Image Road Extraction Method Combining Multi-scale and Attention Mechanism

Rui Wang1, Mingxiang Cai1,*, Zixuan Xia2

[1]China Transport Telecommunications & information Center, Beijing 100011, China; ruiwang_gd@whu.edu.cn

[2]College of Art and Architectural Engineering, Heilongjiang University of Technology, Jixi 158100, China; jshlu@shisu.edu.cn

Corresponding author: Mingxiang Cai (jamtsai@whu.edu.cn; Tel.: +86 155 7590 0185).

We thank Wei Yue and Xiangyuan Ding from Chinese Academy of Forestry for their help in data processing. The authors are also grateful to the editors and referees for their constructive criticism on this paper.

ABSTRACT Road information plays an indispensable role in human society's development. However, owing to the diversity and complexity of roads, it is difficult to obtain satisfactory road-extraction result. Some typical factors, such as discontinuity, loss of edge details, and long-time consumption, have negative impacts on obtaining accurate road information. These problems are particularly prominent during road extraction when high-resolution remote-sensing images are used. To obtain accurate road information, a novel lightweight deep learning neural network was pro-posed in this study by integrating a multiscale module and attention mechanisms. As an excellent multiscale segmentation module, the atrous spatial pyramid pooling was selected to enhance the road extraction ability of remote sensing images. In addition, an attention mechanism was employed to solve the problems of discontinuity and loss of edge details in road extraction, and MobileNet V2 was selected as the backbone of DeepLab V3+ because of its lightweight structure, which can help solve the problem of excessive training time consumption. The experimental verification was carried out on the Ottawa road dataset and the Massachusetts road dataset. Experimental results show that compared with U-Net, SegNet and MDeeplab v3+ networks, the proposed algorithm is the best in IoU, Recall, OA and Kappa. Among them, on the Ottawa road dataset, the OA and Kappa of the algorithm in this paper are 98.92 % and 95.02 %, respectively. On the Massachusetts road dataset, OA and Kappa 98.29% and 89.87%. In addition, the training time was significantly shorter than that of the other deep learning networks. The proposed method exhibited a good performance in road extraction.

INDEX TERMS road extraction, deep learning, CAM, SAM, ASPP, lightweight

I. INTRODUCTION

As important geo-information, roads are indispensable in modern society. Road information not only plays an important role in the transportation and insurance industry but also in domains such as vehicle navigation, unmanned driving, urban management planning, smart city construction, and geo-information database updating. Owing to their wide range of applications [1,2], road extraction has become a re-search hotspot worldwide. With the development of remote sensing (RS) technology, the resolution of RS images has increased, and road extraction based on high-resolution remote sensing images can be effectively and accurately achieved. However, obtaining road information from high-resolution RS images remains a challenging task owing to the presence of complex backgrounds, diverse topologies, and shadows.

In recent years, scholars have done a lot of research on road extraction by fully utilizing high-resolution RS images [4-15]. They proposed a variety of methods for different road extraction tasks. For example, road extraction can be divided into two categories according to different emphasis points: (1) paying attention to local feature information. Local feature information of the road itself is used, such as spectral, texture, shape, and structural features. Road skeletons are extracted using template matching [4-5] or knowledge-driven [6-9] algorithms. However, these methods rely on manual work and have a low degree of automation [10]. (2) Paying attention to the overall feature information. The road is regarded as the extraction object, and the road information in a high-resolution image is extracted through classical algorithms, such as image segmentation clustering [11], Support Vector Machine (SVM) [12], and conditional Random Field (RF) [13]. However, such methods generally suffer from poor stability and complicated operational steps. Extraction results from traditional methods often suffer from issues such as

This work is licensed under a Creative Commons Attribution 4.0 License. For more information, see https://creativecommons.org/licenses/by/4.0/

This article has been accepted for publication in IEEE Access. This is the author's version which has not been fully edited and
content may change prior to final publication. Citation information: DOI 10.1109/ACCESS.2023.3313390

discontinuous extraction and rough edge details, resulting in poor extraction results [14]. Consequently, these methods cannot meet the current requirements for road formation extraction from high-resolution RS images [15].

At present, the rapid development of neural networks has a promoting effect on RS information extraction technology. Long et al. [16] proposed a Fully Convolutional neural network (FCN) in 2015. Realizing from simple object classification to pixel-level classification greatly improves the classification accuracy and efficiency. Subsequently, there has been an increasing number of RS image classifications based on the FCN, especially in the task of road extraction. First, the U-Net and SegNet networks are widely used in road extraction after being proposed. Kong et al. [17] proposed an improved U-Net network to extract road information from high-resolution RS images. This method adds a stripe pooling module to the down-sampling part of the coding layer to focus on local information. A hybrid pooling module was added to the convolution of the coding layer to enhance its ability to obtain the network context. The algorithm was verified using the Gaofen-2 high-resolution RS image dataset. The results showed that this algorithm could effectively extract the road, and the extraction effect was better than that of the traditional network. Li et al. [18] proposed an improved U-Net network that combines core and global attention to extract roads from high-resolution RS images. Experiments were performed on the Massachusetts road dataset and the DeepGlobe-CVPR 2018 road dataset, and the results showed that the proposed method can effectively extract the road area blocked by the tree canopy and improve the connectivity of the road network. Ai [19] proposed an improved SegNet network to extract road information. An edge feature pyramid module was added to the original SegNet network. Through multi-scale convolution of the edge pyramid module, the purpose of enhancing the recognition of small targets was achieved. The experimental results on the Cityscapes dataset showed that the accuracy was im-proved by 7.5 % and 6.2 %, respectively, compared with the original SegNet and U-Net networks. The effectiveness of the pyramid module in this method was demonstrated in the ablation experiment. Subsequently, the Google team proposed the DeepLab series of network models [20-23], which scholars have widely used for road extraction. Han et al. [24] added dilated convolution based on the DeepLab V3 network to extract road information. The experiment verified that this method has a higher extraction accuracy than the other methods. Liu et al. [25] proposed the DeepLab V3+ road extraction method with an attention module. This method obtains more spatial context information through spatial attention to enhance the extraction of road information. The effectiveness of this method was verified by using the Cityscapes dataset. The results show that the ability of this method to extract road information was significantly better than that of the original network.

Although these methods have their own advantages in road extraction, they still have some shortcomings. First, the FCN has a large number of parameters, which not only depends on superior hardware equipment but also takes a long time for network training. Second, SegNet and U-Net networks have the problem of incomplete context information extraction during road extraction. Similarly, the DeepLab V3 network still has problems, such as missing details and discontinuous extraction. To solve these problems, the accuracy of road extraction was further improved. This study is based on the DeepLab V3+ network, and a lightweight high-resolution RS image road extraction method is proposed that integrates multi-scale and attention mechanisms. First, atrous spatial pyramid pooling (ASPP) is an excellent multi-scale feature extraction module. By using a multiscale convolution kernel, the ASPP can extract roads of different sizes. Second, the reference channel attention enhances the shallow features, and the reference space attention enhances the deep features to solve the problems of edge detail loss and extraction discontinuity. Finally, the lightweight MobileNet V2 network replaced the Xception network used in the DeepLab V3+ network as the backbone network. The problem of a long training time caused by a large number of model parameters was solved.

II. GUIDELINES FOR MANUSCRIPT PREPARATION

A. Experimental Data

The experimental data of this paper are Ottawa road dataset and Massachusetts road dataset. Among them, the Ottawa road dataset has a resolution of 0.2 meters. The Massachusetts road dataset has a resolution of 1 meter. The image data with dense roads in the two data sets is selected as the training samples of roads. The training sample is cropped to a size of 512×512 pixels, and the data set is expanded by rotating 90°, 180° horizontal mirroring and vertical mirroring. The Ottawa-Datase and Massachusetts-Datase images and labels are shown in Figure 1.



FIGURE 1. Dataset images and labels (a) Image of Ottawa-Dataset; (b) Corresponding label of Ottawa-Dataset; (c) Image of Massachusetts -Dataset; (d) Massachusetts -Dataset corresponding label.

B. Parameter Settings

The experimental environment in this study was a Windows 10 operating system with 128G memory. The processor is an

This work is licensed under a Creative Commons Attribution 4.0 License. For more information, see https://creativecommons.org/licenses/by/4.0/

This article has been accepted for publication in IEEE Access. This is the author's version which has not been fully edited and
content may change prior to final publication. Citation information: DOI 10.1109/ACCESS.2023.3313390

Intel(R) CPU E5-2640 v3@2.60GHz. The graphics card was a Nvida GTX 1600 super 6 GB graphics card. The deep learning framework is a high-level neural network application interface Kearas 2.1.6 of Tensorflow 1.14.0. SegNet and U-Net were selected as comparison algorithms. According to the experimental environment of this study, the batch size was set to 2, and the epoch was set to 100 for the training network.

III. METHODS

The method in this study is based on the DeepLab V3+ network [26-28] to make two improvements. The first is that the lightweight MobileNet V2 network replaces the Xception of the DeepLab V3+ network as the backbone network, and the second is to enhance the deep features by referring to channel attention and spatial attention at the jump connection structure.

A. Lightweight Road Extraction Network

The method in this study adopts an encoder-decoder structure. First, in the encoder layer, the shallow road feature information in high-resolution RS images is extracted through the different channel-depth separable convolution layers of MobileNet V2. Shallow feature information is obtained using atrous spatial pyramid pooling (ASPP) to obtain deep semantic road feature information. In the ASPP module, shallow features are processed through a 1×1 convolution, three dilated convolutions with dilation rates of 6, 12, and 18, and a global-average pooling layer. After processing, the deep feature information with multiscale features was obtained. Second, the deep feature information is input to the decoder layer together with the shallow feature information under the enhancement of the channel attention and spatial attention. Finally, in the decoder layer, the deep features are up-sampled by 4 times bilinear interpolation four times and combined with the shallow feature information of the remote sensing image road obtained by the encoder layer. The combined feature information passes through two 3×3 convolutional layers to restore detailed feature information and obtain fine target boundary information through four bilinear interpolation upsamplings. Through the above, the extraction results of the high-resolution remote sensing image road are finally obtained.

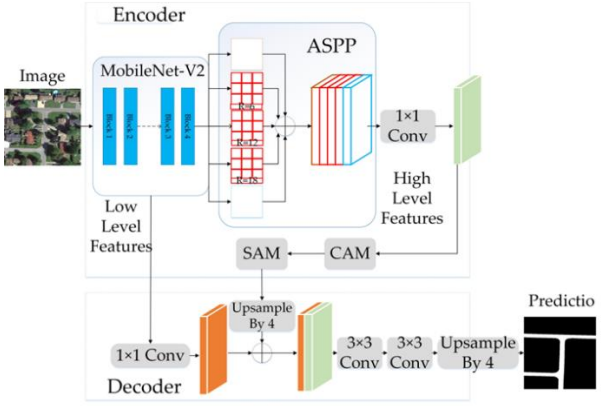


FIGURE 2. Lightweight DeepLab V3+ network.

B. MibleNet V2

As a lightweight deep neural network, the MobileNet network [29] has the advantages of smaller volume, less computation, higher accuracy, and faster speed, and is suitable for a variety of application scenarios. Its core is a depth-wise separable convolution, and its structure is shown in Figure 3. First, a 3×3 convolution kernel separated the input feature channels by traversing the individual data in each channel. Second, a 1×1 convolution kernel was used to traverse each feature graph to integrate the feature information to obtain the output feature. Compared with standard convolution, depth-separable convolution can effectively reduce the number of model parameters.

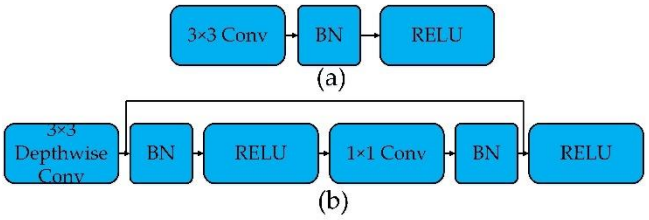


FIGURE 3. Deeply separable convolutional structures. (a) Standard convolution;(b) Depthwise sep-arable convolution.

MobileNet V2 [30-33] is an improved version of the MobileNet model, as shown in Figure.4(a) in Figure.4(b). It has two distinct features. The first is to invert the residual structure. This structure first uses a 1×1 convolution before the 3×3 network structure to increase the dimensions. This structure then uses a 1×1 convolution after the 3×3 network structure to achieve dimensionality reduction and compression. Compared with the direct use of a 3×3 convolution, the effect is better, and the number of parameters is effectively reduced. The second is a linear bottleneck structure. The bottleneck structure of MobileNet V2 is shown in Figure 4. Because the linear ReLU function operation causes feature loss, to reduce the information loss, the ReLU operation is no longer performed after the 1×1 convolution dimension reduction. However, the addition of the residual network was performed directly. The expansion coefficient was designed in MobileNet V2, the purpose of which is to achieve better control of the network size.

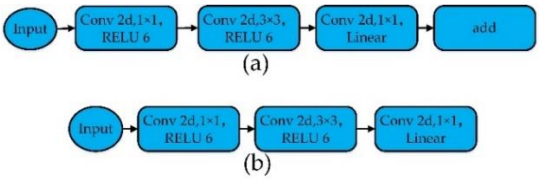


FIGURE 4. The bottleneck structure of MobileNet V2. (a) Step size is 1;(b) Step size is 2.

Compared with MobileNet, MobileNet V2 has a deeper network structure, a complete convolutional layer with 32 channels, and 17 bottleneck structures, as shown in Table I. where *t* is a multiple of the internal dimensions of the bot-tleneck structure, *c* is the number of channels, *n* is the number of repetitions of the bot-tleneck structure, and *s* is the step size.

This work is licensed under a Creative Commons Attribution 4.0 License. For more information, see https://creativecommons.org/licenses/by/4.0/

This article has been accepted for publication in IEEE Access. This is the author's version which has not been fully edited and
content may change prior to final publication. Citation information: DOI 10.1109/ACCESS.2023.3313390

TABLE I
MOBILENET V2 NETWORK ARCHITECTURE

Input	Operation	t	c	n	s
2242×3	Covn2d	-	32	1	2
1122×32	bottleneck	1	16	1	1
562×16	bottleneck	6	24	2	2
282×24	bottleneck	6	32	3	2
142×32	bottleneck	6	64	4	2
142×64	bottleneck	6	96	3	1
72×96	bottleneck	6	160	3	2
72×160	bottleneck	6	320	1	1
72×320	Covn2d 1×1	-	1 280	1	1
72×1280	avgpool 7×7	-	-	1	-
1×1×1280	Covn2d 1×1	-	k	-	

C. ASPP Module

Figure 5 shows the ASPP module [34-36]. The ASPP module in the DeepLab V3+ network contained a 1×1 convolution layer, hole convolution with hole rates of 6, 12, and 18, and a global average pooling layer. In addition, a 1×1 convolutional layer was added after the global average pooling layer and each dilated convolutional layer to adjust the dimensionality of the feature maps. The purpose of this method was to achieve the same dimensionality as the output feature map. Void convolution with different void-rate sizes can capture road feature information of various sizes and enhance the ability of the network to extract roads of different sizes. The 1×1 convolution in the ASPP is used to capture smaller targets, whereas global average pooling can integrate the information of the entire feature map. Finally, the concat operation is performed on the feature maps obtained by the 1×1 convolution, hole convolution with hole ratios of 6, 12, and 18, and global average pooling layer. The dimension of the output feature map is adjusted using a 1×1 convolution so that it is equivalent to the input.

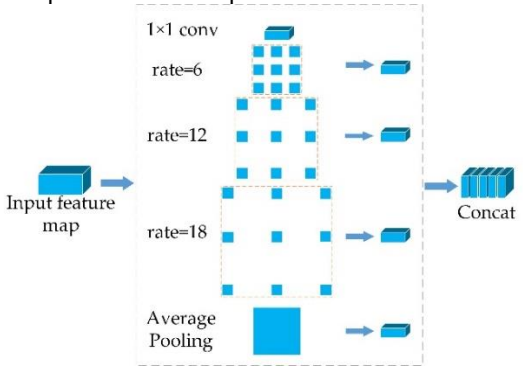


FIGURE 5. ASPP module.

D. Attention Mechanism

The core idea of the channel attention module (CAM) [37-39] is to learn the features of different feature channels to obtain weights. Then, CAM enhances the useful feature channels according to the weight, suppresses the useless feature channels, and realizes attention to the channel. The formula is as follows:

$$M_c(F) = \sigma\left(MLP\left(AvgPool(F)\right) + MLP\left(MaxPool(F)\right)\right) = \sigma\left(W_1(W_0)\left(FC_{Avg}\right)\right) + \left(W_1\left(W_0\left(FC_{max}\right)\right)\right) \quad (1)$$

The specific implementation and structure of the CAM used in this study are shown in Figure 6. First, the feature graph F is processed by global average pooling and global maximum pooling. Global average pooling integrates the global spatial information of the input feature map F. Global maximum pooling reduces unnecessary information by extracting the maximum value of pixel points in the neighborhood. Second, the dimension is reduced through the first full connection layer to reduce the complexity. Subsequently, the number of channels is restored through the second full connection layer to build the correlation between channels. Finally, the weight of each channel was activated by the sigmoid function. It is applied to the feature channel corresponding to feature map F. Feature map MC is obtained after weighting processing by the channel attention module.

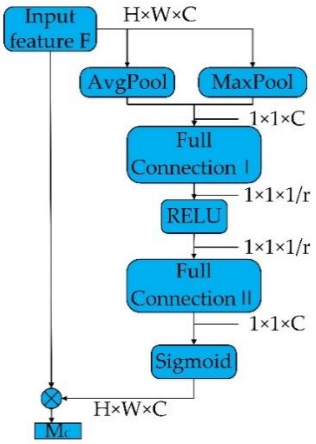


FIGURE 6. Channel attention module.

E. Spatial Attention Module

The spatial attention module (SAM) [40-42] is to further improves the screening ability of salient features by focusing on spatial features and according to their importance. Its expression is shown in (2).

$$M_S(F) = \sigma\left(f^{7*7}\left([AvgPool(F)), MaxPool(F)]\right)\right) = \sigma\left(f^{7*7}\left([F^C_{Avg}; F^S_{\max}]\right)\right) \quad (2)$$

The implementation and structure of the SAM adopted in this study are shown in Figure 7. First, SAM compresses the channel-domain features of the input feature map F through global maximum pooling and global average pooling. Second, multiple channels are compressed into a single channel by convolution to reduce the influence of inter-channel information on spatial attention. Finally, after the Sigmoid function was activated, the weight containing the spatial information feature was obtained. It is multiplied by the input feature map F pixel-by-pixel to obtain the feature map Ms.

This work is licensed under a Creative Commons Attribution 4.0 License. For more information, see https://creativecommons.org/licenses/by/4.0/

This article has been accepted for publication in IEEE Access. This is the author's version which has not been fully edited and
content may change prior to final publication. Citation information: DOI 10.1109/ACCESS.2023.3313390

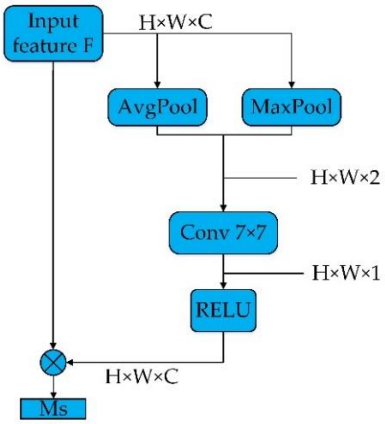


FIGURE 7. **Spatial attention module.**

F. Accuracy evaluation

To effectively verify the effectiveness of the algorithm in this study, the average intersection-over-union ratio (IoU), Kappa coefficient, overall precision (OA), and Recall rate as the evaluation indicators of the experimental re-sults. As shown in (3)–(6). The evaluation indices of network model complexity are model Parameters, Floating-Point operations (FLOPS), and training time. The number of floating-point operations represents the number of FLOPS operations during the forward inference process of the model.

$$IoU=TP/(TP+FN+FP) \quad (3)$$

$$Recall=TP/(TP+FN) \quad (4)$$

$$OA=(TP+TN)/(TP+FN+FP+TN) \quad (5)$$

$$Kappa=(P0-Pe)/(1-Pe\ ) \quad (6)$$

where TP represents the number of positive samples correctly extracted from the road. FN represents the number of positive samples directly extracted from the road. FP represents the number of negative samples incorrectly extracted from the road. TN represents the number of negative samples for which the road was extracted correctly. P0 is the sum of the correctly extracted road pixels divided by the total number of pixels. Pe is the sum of the product of the real number of road samples and number of predicted samples divided by the square of the total number of samples.

IV. Experiment and Analysis

A. Comparison of Traditional Methods

To verify the superiority of the proposed method in road extraction, it was compared with two traditional RS image segmentation methods, namely support vector machine and random forest. SVM [42-43]and RF [44], as shallow machine learning algorithms, exhibit good results in RS images [45]. In this study, through multi-ple verifications and referring to the relevant experimental results. In the SVM classification method in this study, the radial basis function was selected as the kernel function and the gamma value was set to 0.333. The penalty coefficient was maintained at a default value of 100. Finally, the number of decision trees in the RF classification was set to 100. The number of feature variables participating in the construction of decision trees was set as the square root of the number of all features of the classification. The image classification results are shown in Figure 8, where (a) is the image, (b) is the label map, (c) is the SVM extraction result, (d) is the RF extraction result, and (e) is the extraction result of the proposed method.

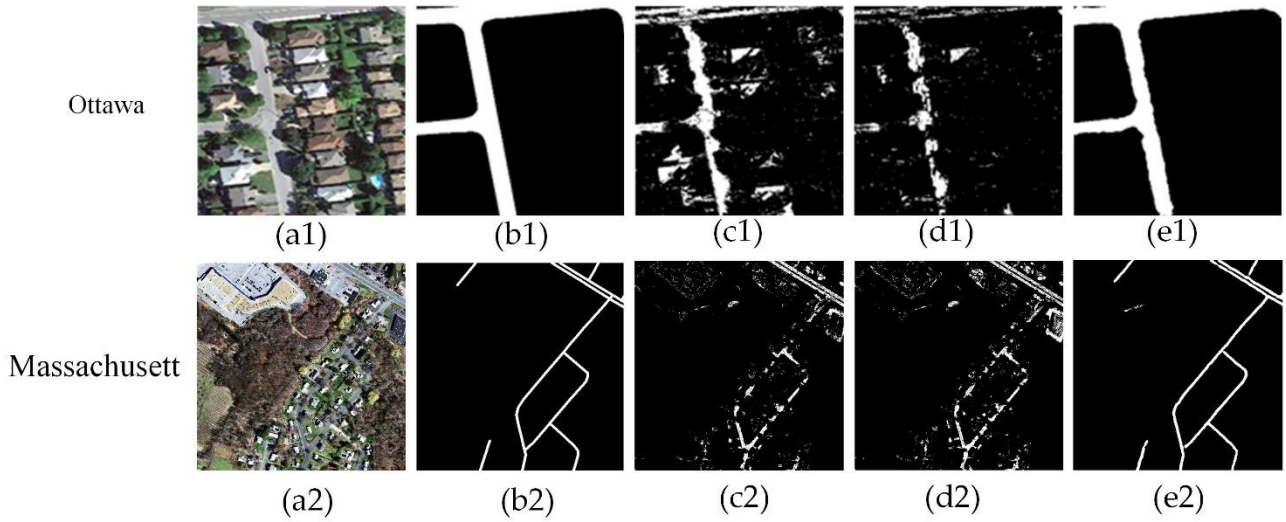


FIGURE 8. **Traditional method results. (a) Image map; (b) Label map; (c) SVM extraction results; (d) RF extraction results; (e) Extraction results of this paper.**

From Figure 8(c1) and (c2), the SVM road result extraction is discontinuous, and the road edge details are significantly lost. Buildings have a significant influence on the extraction of SVM. When the color of the roof is similar to that of the road, the extraction error becomes more serious. In places that are partially or completely blocked by obstacles, such as trees and buildings, the extraction effect is poor, and there are many misextraction parts. In Figure 8(d1) and (d2), the road

This work is licensed under a Creative Commons Attribution 4.0 License. For more information, see https://creativecommons.org/licenses/by/4.0/

This article has been accepted for publication in IEEE Access. This is the author's version which has not been fully edited and
content may change prior to final publication. Citation information: DOI 10.1109/ACCESS.2023.3313390

extraction re-sults of RF are more serious than those of SVM, and the mis-extracted part of SVM is slightly improved, but the problem of misallocated buildings still exists. In Figure 8(e1) and (e2), the algorithm in this study can effectively solve the problems of discontinuous road extraction and loss of road-edge details. There were no problems, such as false mentions. Buildings and roads can be effectively distinguished, and the occlusion problem does not affect the extraction results, which are basically consistent with the annotated data.

The precision calculation of pixel-by-pixel matching was performed on the experimental results and RS image label map. the accuracy evaluation results are shown in Table II. It can be seen from Table 1 that the algorithm in this study is the best in terms of the accuracy indicators of IoU, Recall, OA, and Kappa. Among them, in the comparison experiment of Ottawa road dataset, compared with SVM, IoU and Recall increased by 44.65 and 49.04% respectively. OA increased by 12.97% and Kappa increased by 0.4904. Compared with RF, IoU increased by 48.34%, recall rate increased by 52.54%, OA increased by 12.97%, kappa increased by 0.5269. In the comparative experiment of the Massachusett road dataset, IoU increased by 26.37 %, Recall increased by 41.01 %, OA increased by 2.23 %, and Kappa increased by 0.4278. Compared with RF, IoU increased by 28.13 %, recall rate increased by 41.01 %, OA increased by 2.98 %, kappa increased by 0.4635. It can be seen from the above indicators that the method in this study has more advantages than the traditional method for high-resolution road extraction. The method used in this study does not require post-processing and can achieve automatic and accurate extraction.

TABLE II
ACCURACY VERIFICATION RESULTS OF TRADITIONAL METHODS

DATA	Method	IoU/%	Recall/%	OA/%	Kappa
Ottawa	SVM	46.98	46.03	85.95	0.4601
	RF	43.29	42.53	77.87	0.4233
	Ours	91.63	95.07	98.92	0.9502
Massachusett	SVM	64.28	49.10	96.06	0.4709
	RF	62.52	51.32	95.31	0.4352
	Ours	90.65	90.11	98.29	0.8987

B. Comparison of Semantic Segmentation Methods

In order to verify the advantages of the algorithm in this paper in road extraction, the Ottawa road data set is used for experiments. Through U-Net, SegNet, the DeepLab V3+ network (hereinafter referred to as MDeepLab V3+) with MobileNet V2 as the backbone network and the algorithm in this paper are compared and tested. To ensure the comparability and validity of the results, the training and test datasets are the same. The experimental results of the U-Net, SegNet, MDeepLab V3+ network, and algorithm used in this study are shown in Figure 9. Where (a) is the image, (b) is the label map, (c) is the U-Net extraction result, (d) is the SegNet extraction result, (e) is the MDeepLab V3+ extraction re-sult, and (f) is the extraction result of the proposed method.

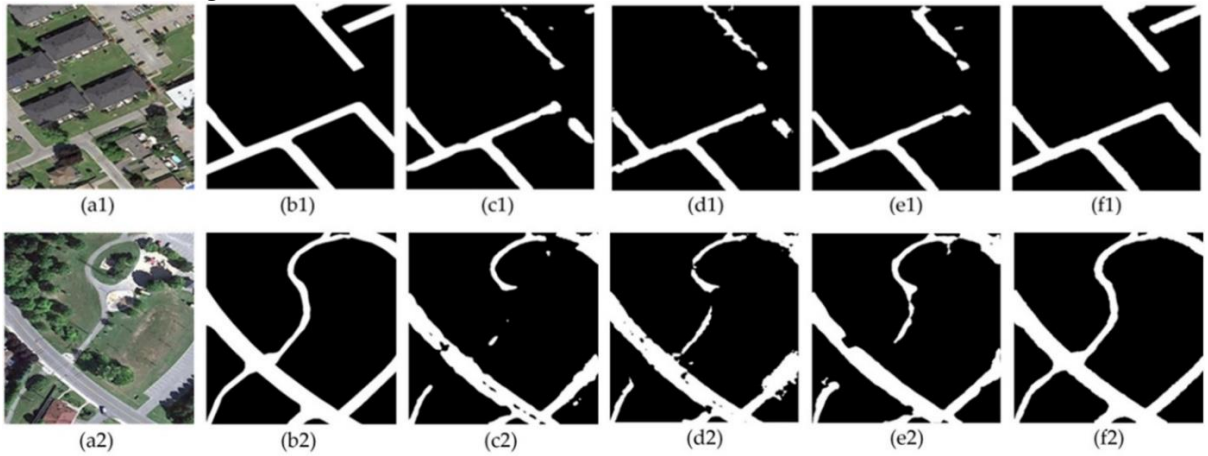


FIGURE 9. Compare the experimental results. (a) Image map; (b) Label map; (c) U-Net extraction results; (d) SegNet extraction results; (e) MDeepLab V3+ extraction results; (f) Extraction results of this paper.

As shown in Figure 9(a1), the road was clear and regular. There are trees blocking it, and there are buildings with colors similar to the road in the figure. As shown in Figure 9(c1), U-Net can avoid the extraction discontinuity problem caused by obstacles such as trees and buildings in road results. However, the pro-cessing of edge details is not ideal, and holes remain in the middle. As shown in Figure 9(d1), the extraction result of SegNet is slightly worse than that of U-Net. There are discontinuities in the extraction results, and the detailed processing was slightly worse. However, the hole phenomenon is weaker. As can be seen in Figure 9(e1), the edge details of the extraction results of MDeepLab V3+ are better processed. The discontinuities and holes in the extraction results of U-Net and SegNet are weakened. This is owing to the multi-scale extraction of road details in the ASPP module in the MDeepLab V3+ network. However, some problems remain, such as incomplete extraction. As shown in Figure 9(f1), the method proposed in this study can effectively extract roads. The edge details were handled well, and the extraction was complete and continuous. There were no problems such as false mentions or holes. As shown in Figure 9(a2), the road is irregular and there are many trees. There

This work is licensed under a Creative Commons Attribution 4.0 License. For more information, see https://creativecommons.org/licenses/by/4.0/

This article has been accepted for publication in IEEE Access. This is the author's version which has not been fully edited and
content may change prior to final publication. Citation information: DOI 10.1109/ACCESS.2023.3313390

were no occlusions, such as buildings or trees. From Figure. 9(c2), it can be observed that the road extraction is discontinuous. There were large fractures associated with this problem. There are many holes in the middle of the road. There is a noticeable loss in the edge detail. In Figure 9(d2), SegNet shows obvious improvement compared with the U-Net extraction results, but the above problems still exist. It can be observed from Figure 9(e2) that the extraction result of MDeepLab V3+ is better. It can be accurately extracted, and there is no extraction discontinuity. However, the processing of edge details is still imperfect. After the method of this study integrates multiscale and attention mechanisms, the extraction effect is better than the extraction results of SegNet, U-Net, and MDeepLab V3+. As shown in Figure 9(f2), the method proposed in this study handles the edge details perfectly, and the extraction results are continuous. The problems with the extraction of the other three methods were solved. The extraction results are basically consistent with the labeled samples.

TABLE III
COMPARISON OF EXPERIMENTAL ACCURACY VERIFICATION RESULTS

Method	IoU/%	Recall/%	OA/%	Kappa	Params/MB	FLOPs/G	Training time/t
U-Net	61.67	64.57	95.34	0.7381	7.77	30.51	13.55
SegNet	77.60	80.76	97.40	0.8595	30.5	75.20	15.99
MDeepLabv 3+	80.21	81.80	97.23	0.8776	2.76	10.517	10.06
Ours	91.63	95.07	98.92	0.9502	2.85	12.76	10.08

As can be seen from Table III, the method presented in this paper is optimal for all indices. Among them, Compared with U-Net, IoU increased by 29.96%, Recall in-creased by 30.5%, OA increased by 3.58%, and kappa increased by 0.1821. Compared with SegNet, IoU increased by 14.03%, Recall increased by 14.31%, OA increased by 1.52%, and Kappa increased by 0.0907. Compared with MDeepLab V3+, all indicators have also been greatly improved, among which IoU and Recall have increased by more than 10%. The method in this study is much smaller than the U-Net and SegNet methods in terms of Params and FLOPs, which effectively reduces the training time. Compared with the MDeepLab V3+ network, the method in this study increases the attention mechanism. Although the performance is slightly improved in Params and FLOPs, the training time is not greatly increased. In addition to the shorter training time, it is advantageous to obtain a higher accuracy. Based on the above accuracy indices, the proposed method can effectively extract high-resolution RS images. Compared with other methods, not only can it achieve the highest accuracy, such as Kappa, but it can also effectively solve the problem of training time, which shows that the method in this study performs well.

In order to verify the superiority and generalization ability of the algorithm in this paper in road extraction, the public dataset Massachusetts Roads road dataset is used for experimental verification. The model of U-Net, SegNet, MDeep Lab V3 + and this paper are respectively trained. To ensure comparability and validity of results, the training and test data sets are the same. The experimental results of U-Net, SegNet and MDeepLab V3 + networks and algorithms are shown in Figure 10. Where (a) is the image, (b) is the label map, (c) is the U-Net extraction result, (d) is the SegNet extraction result, (e) is the MDeepLab V3+ extraction re-sult, and (f) is the extraction result of the proposed method.

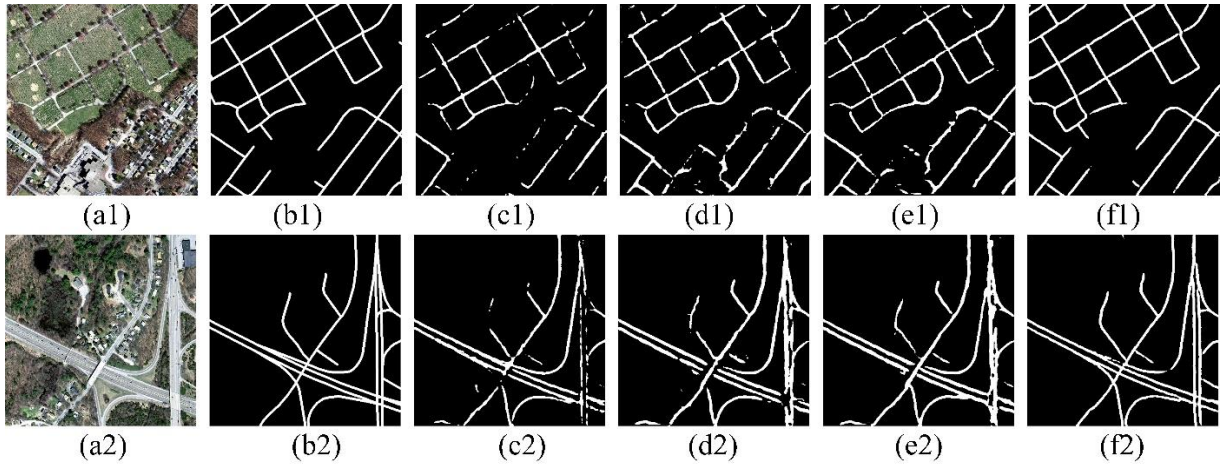


FIGURE 10. Compare the experimental results. (a) Image map; (b) Label map; (c) U-Net extraction results; (d) SegNet extraction results; (e) MDeepLab V3+ extraction results; (f) Extraction results of this paper.

As shown in Figure 10 (a1), the road is clear and regular, but there are also trees blocking it. As shown in Figure 9 (c1), U-Net extracts roads incompletely, and occluded roads cannot be fully extracted. As shown in Figure 9 (d1), the extraction

This work is licensed under a Creative Commons Attribution 4.0 License. For more information, see https://creativecommons.org/licenses/by/4.0/

This article has been accepted for publication in IEEE Access. This is the author's version which has not been fully edited and
content may change prior to final publication. Citation information: DOI 10.1109/ACCESS.2023.3313390

results of SegNet are slightly better than U-Net. However, there are still discontinuities in the extraction results, and the detail processing is slightly poor. Figure 9 (e1), The extraction results of MDeepLab V3 + are better in continuity than U-Net and SegNet. However, it is also affected by other ground features, resulting in inaccurate extraction. As shown in Figure 9 (f1), the method proposed in this paper can effectively extract roads. The edge details are well processed, and the extraction is complete and continuous. As shown in Figure 9 (a2), the road is irregular and the road structure is relatively complex. But the road is not blocked by buildings or trees etc. From Figure 9 (c2), the discontinuity of the extraction results of the U-Net network is serious, and the obvious road details are not accurately extracted. In Figure 9 (d2), SegNet showed obvious improvement compared with U-Net extraction results, but the above problems still existed. As can be seen from Figure 9 (e2), the extraction results of MDeepLab V3 + are improved compared with the previous two methods, but the extraction results are still inaccurate. After the method of this study integrates multi-scale and attention mechanism, the extraction effect is better than the extraction results of SegNet, U-Net and MDeepLab V3+. As shown in Fig. 9 (f2), the proposed method handles the edge details well and the extraction results are continuous. The extraction problems that existed with the other three methods are solved. The extraction results are basically consistent with the labeled samples.

TABLE IV
COMPARISON OF EXPERIMENTAL ACCURACY VERIFICATION RESULTS

Method	IoU/%	Recall/%	OA/%	Kappa	Params/MB	FLOPs/G	Training time/t
U-Net	74.95	64.27	95.09	0.6842	7.77	30.51	16.20
SegNet	76.88	64.88	96.13	0.7125	30.5	75.20	18.47
MDeepLabv 3+	77.25	83.12	96.83	0.7200	2.76	10.517	12.24
Ours	90.65	90.11	98.29	0.8987	2.85	12.76	12.30

It can be seen from Table IV that the method proposed in this paper is optimal in all indicators. Among them, compared with U-Net, IoU increased by 15.7%, Recall increased by 25.84%, OA increased by 3.2% and kappa increased by 0.2145. Compared with SegNet, IoU increased by 13.77%, Recall increased by 25.23%, OA increased by 2.16%, and Kappa increased by 0.1862. Compared with MDeepLab V3 +, all indicators have also been greatly improved. The performance of the proposed method in Params and FLOPs is much lower than that of the U-Net and SegNet methods, which effectively reduces the training time. Compared with the MDeep Lab V3+ network, our method adds an attention mechanism. Although there is a slight improvement in the Params and FLOPs indicators, the training time does not increase to a large extent. Based on the above accuracy indicators, the method proposed in this paper can effectively extract high-resolution remote sensing images. Compared with other methods, it can not only achieve the highest accuracy, but also effectively solve the problem of training time. And the method in this paper has certain generalization ability.

C. Ablation Experiment

To verify the effectiveness of the module in this study, an algorithm and its variants were used for experimental verification. MDeepLab V3+ is a DeepLab V3+ network with MobileNet V2 as the backbone network, and CDeepLab V3+ is an MDeepLab V3+ network that fuses channel attention. The experimental results are shown in Figure 11, where (a) is the image, (b) is the label, (c) is the extraction result of MDeepab v3+, (d) is the extraction result of CDeepLab V3+, and (e) is the algorithm used in this study.

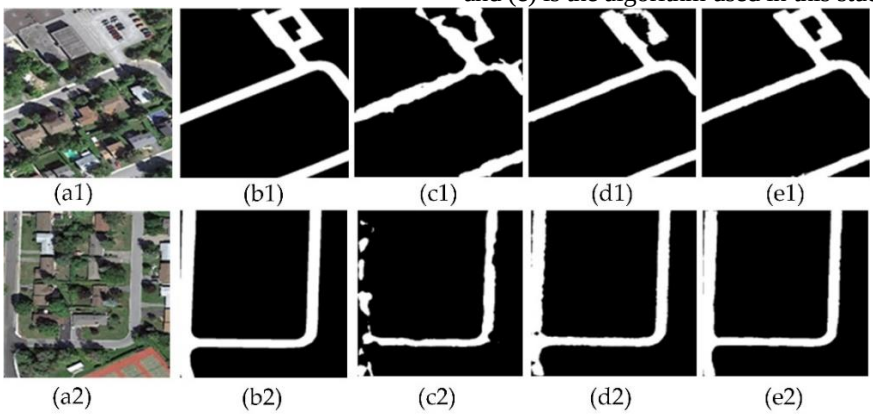


FIGURE 11. Ablation experiment results. (a) Image map; (b) Label map; (c) MDeepLab V3+ extraction results; (d) CDeepLab V3+ extraction results; (e) Extraction results of this paper.

As shown in Figure 11 (c1), MDeepLab V3+ is not accurate for extraction around the parking lot. At the bend of the road, there is an inaccurate extraction problem caused by tree occlusion. The edge detail processing of the overall road is not

This work is licensed under a Creative Commons Attribution 4.0 License. For more information, see https://creativecommons.org/licenses/by/4.0/

This article has been accepted for publication in IEEE Access. This is the author's version which has not been fully edited and
content may change prior to final publication. Citation information: DOI 10.1109/ACCESS.2023.3313390

accurate, and there is a loss of detail. Compared to the MDeepLab V3+ network, the CDeepLab V3+ network adds a channel attention mechanism. The problem in Figure 11(d1) is improved compared to that in Figure 11 (c1). The edge details are better and more accurate, but there are still inaccuracies in the extraction around the parking lot. In Figure 11 (e1), the proposed algorithm integrates the spatial attention and channel attention mechanisms, and the extraction results of the proposed algorithm are more accurate than those of the MDeepLab V3+ and CDeepLab V3+ networks. As shown in Figure 11 (c2), the extraction results of the MDeepLab V3+ network exhibit serious discontinuity problems, and the processing of edge details is poor. It can be seen from Figure 11 (d2) that the extraction results of CDeepLab V3+ are compared to those in Figure 11 (c2), and the discontinuity problem has been effectively improved, but there are holes. As can be seen in Figure 11 (e2), when the MDeepLab V3+ network simultaneously increases channel attention and spatial attention at the same time, the problem of extracting discontinuity and edge loss is optimized. The method presented in this paper obtained a good extraction effect. Based on the above, the modules of the algorithm in this study are effective, and the method in this study is more applicable.

TABLE V
ACCURACY VERIFICATION RESULTS OF ABLATION EXPERIMENTS

Method	IoU/%	Recall/%	OA/%	Kappa	Params/M	FLOPs/G	Training time /t
MDeepLab V3+	80.21	81.80	97.23	0.8776	2.76	10.517	10.06
CDeepLab V3+	88.20	90.31	98.50	0.9288	2.83	11.74	10.08
Ours	91.63	95.07	98.92	0.9502	2.85	12.76	10.08

As can be seen in Table V, after MDeepLab V3+ increased the channel attention, IoU, Recall, OA and Kappa increased significantly. Compared with MDeepLab V3+, the CDeepLab V3+ network has an 11.42% increase in IoU and a 13.27% increase in recall. OA increased by 1.69%, and Kappa reached 0.9288. When spatial attention and channel attention are increased at the same time, the IoU value of the proposed method reaches 91.63%. OA reached 98.92%, and Kappa reached 0.9502. It was shown that channel attention and spatial attention can effectively solve the problems of road extraction discontinuity and edge detail loss. However, there was not much improvement in the two individual cators of Params and FLOPs, resulting in no change in training time. The above indicators show that the improved part of the method in this study can effectively solve the problems of discontinuous extraction and loss of edge details. The algorithm in this study can effectively reduce training time and achieve accuracy and efficiency.

V. Conclusions

As important geo-information for a country, roads play an important role in the transportation and insurance industries. However, in the process of extracting road information from high-resolution RS images, problems such as incompleteness and inaccuracy of road information have become problems that need to be solved. This paper aims at the low precision of the traditional high-resolution RS image road extraction method, the post-processing requires a lot of manual participation. The FCN network method has problems, such as many parameters, time consumption, discontinuous extraction, and loss of edge details. A lightweight road extraction method that integrates multiscale and attention mechanisms is proposed. Solve the problems of discontinuous extraction and loss of edge details through the ASPP and attention mechanism. By using the MobileNet V2 network instead of Xception as the backbone network of DeepLab V3+, the problem of a long training time is solved. The Ottawa Road dataset was verified. In the comparison test of traditional methods, SVM and RF were not suitable for road extraction, and they were easily affected by occluding objects. Compared with the two traditional methods, the method proposed in this paper is superior in terms of extraction. In the comparative experiment on the Ottawa road dataset, although the U-Net and SegNet networks are not affected by occlusions in the process of road extraction, there are discontinuities, edge detail losses, holes, and other phenomena in the road extraction results. MDeepLab V3+ integrates an ASPP module. This solves the phenomenon of holes in the extraction process, but there are still problems such as discontinuous extraction and loss of edge details. The method presented in this paper can effectively solve the above problems, and the road extraction problem can be effectively solved. Compared with the other three comparison methods, this study not only has a greater advantage in accuracy but also can save a lot of training time, effectively solving the problem of lengthy road extraction. In the comparison experiment on the Massachusetts road dataset, the extraction results of U-Net and SegNet both have discontinuous extraction. Although the extraction results of MDeeplab v3+ have been improved compared to the previous two methods, this phenomenon still exists. The method in this paper can effectively extract roads. It shows that the method in this paper has certain superiority and generalization ability. In addition, ablation experiments verified that the improved part in this study is effective. It can be observed from the above that the proposed method can effectively extract roads from high-resolution RS images and has advantages over the comparison method. The time-consuming problem is solved, which shows that the method in

This work is licensed under a Creative Commons Attribution 4.0 License. For more information, see https://creativecommons.org/licenses/by/4.0/

This article has been accepted for publication in IEEE Access. This is the author's version which has not been fully edited and
content may change prior to final publication. Citation information: DOI 10.1109/ACCESS.2023.3313390

this study has a certain applicability. However, it can be seen from the verification of different data sets in this paper that the accuracy of the experimental results on the Ottawa road data set is better than that on the Massachusetts road data set. It shows that the method in this paper is more suitable for the research of road extraction from high-resolution remote sensing images. Although the method proposed in this paper has improved the number of parameters in the model, the long training time is still a problem faced by FCN networks. In future studies, lightweight networks can be used for road extraction. High-resolution RS images require a large amount of storage space, which is required in large-scale experiments. In future research, a suitable resolution RS image will be selected for large-scale road extraction research. The method in this study was only researched for the extraction of roads. However, it is not known whether it can effectively extract other ground features. In future research, this method will be applied to more target-extraction tasks.

Acknowledgments

We thank Wei Yue and Xiangyuan Ding from the Chinese Academy of Forestry for their help with the data processing.

The authors are also grateful to the editors and referees for their constructive criticism of this study.

REFERENCES

[1] Cheng, G, L. Zhu, F, Y. Xiang, S, M. and Pan, C. Road centerline extraction via semisupervised segmentation and multidi rection nonmaximum suppression. *IEEE Geoscience and remote sensing Letters*. 2016, *13(4)*: 545-549.

[2] Zhang, Y, H. He, J. Kan, X. and Li, J, Y. Summary of road extraction methods for RS images. Computer Engineering and Ap plications, 2018, 54(13): 1-10,51.

[3] Liu, Y. Kang, J. Guan, H, Y. and Wang, H, Y. High-resolution RS image road extraction model based on dual-attention residual network. Journal of Geo-Information Science, 2023, 25(02): 396-408.

[4] Lin, X, G. Zhang, J, X. Liu, Z, J. et al. Semi-automatic road tracking by template matching and distance transform, Joint Urban RS Event. IEEE, 2009: 1- 7.

[5] Udomhunsakul, S. Kozaitis, S, P. Sritheeravirojana, U. Semi automatic road extraction from aerial images, RS. ProcSPIE 5239, RS for Environ mental M onitoring, GIS Applications, and Geology III, Barcelona, Spain. 2004, 5239: 26-32.

[6] Mayer, H. Laptev, I. Baumgartner, A. Steger, C. Automatic road extraction based on multi-scale modeling, context, and snakes. International Archives of Photogrammetry and remote sensing. 1997, XXXII(3-2W3): 106-113.

[7] Baumgartner, A. Steger, C. Mayer, H. Ebner H. Automatic road extraction based on multi-scale, grouping, and context. Photogrammetric Engineering and Remote Sensing. 1999, 65(7): 777-785.

[8] Treash, K. Amaratunga, K. Automatic road detection in grayscale aerial images. Journal of Computing in Civil Engineering. 2000, 14(1): 60-69.

[9] Gaetano, R. Zerubia, J. Scarpa, G. Poggi, G. Morphological road segmentation in urban areas from high resolution sat ellite images, 17th International Conference on Digital Signa l Processing (DSP). IEEE, 2011: 1-8.

[10] Wang, L, F. Yan, C, M. A review of road scene semantic segmentation. Progress in Laser and Optoelectronics.2021, 58(12): 1200002.

[11] Alshehhi, R. Marpu, P, R. Hierarchical graph-based segmentation for extracting road networks from high-resolu tion satellite images. ISPRS Journal of Photogramme try and Remote Sensing. 2017, 126(APR.): 245-260.

[12] Song, M, J. Civco, D. Road extraction using SVM and im age segmentation. Photogrammetric Engineering and remote sensing. 2004, 70(12): 1365-1371.

[13] Xiao, L. Dai, B. Liu, D, X. and Wu T. CRF based road detection with multi-sensor fusion, IEEE Intelligent Vehicles Symposium. IEEE, 2015: 192-198.

[14] Dai, J, G. Wang, Y. Du, Y. Du, T, T. Xie, S, Z. Li, C, C. and Fang, X, X. Development and prospect of road extraction method for optical RS image. Journal of remote sensing. 2020, 24(7): 804-823.

[15] Guo, M, Q. Liu, H. Xu, Y, Y. and Huang Y. Building extraction based on U-Net with an attention block and multiple losses. Remote Sensing. 2020, 12(9): 1400.

[16] Long, J. Shelhamer, E. and Darrell T. Fully convolutional networks for semantic segmentation.2015 IEEE Confer ence on Computer Vision and Pattern Recognition (CVPR). 2015: 3431-3440.

[17] Kong, X, W. Wang, C, Y. Zhang, S, C. Li, J, H. and Sui, Y. Application of Improved U-Net Network in Road Extraction from RS Image Remote Sensing Information. 2022, 37(02): 97-104.

[18] Li, J. Liu, Y. and Zhang, Y. Cascaded Attention DenseUNet (CADUNet) for Road Extraction from Very-High-Resolution Images. International Journal of Geo-Information. 2021, 10(5):329.

[19] Luis Alfredo Rodriguez UmanaJavier Eduardo Martinez Baquero Robinson Jimenez Moreno. Semantic Segmentation for Applications in Autonomous Vehicles. International Journal of Applied Engineering Research. 2018, 13(22aPta5).

[20] Chen, L, C. Papandreou, G. Kokkinos, I. Murphy, K. and Yuille A. Semantic image segmentation with deep convolutional nets and fully connected CRFs. Computer Science. 2014(4): 357-361.

[21] Chen, L, C. Papandreou, G. Schroff, F. and Adam H. Rethinking atrous convolution for semantic image segmentation, Conference on Computer Vision and Pattern Recognition (CVPR). IEEE/CVF. 2017.

[22] Chen, L, C. Papandreou, G. Kokkinos, I. Murphy, K. and Yuille, A DeepLab: Se mantic image segmentation with deep convolutional nets, atrous convolution, and fully connected CRFs. IEEE Transactions on Pattern Analysis and Machine Intelli gence. 2018, 40(4): 834-848.

[23] Chen, L, C. Zhu, Y. Papandreou, G. Schroff, F. Adam, H. Encoder-decoder with atrous separable convolution for semantic image seg mentation, Proceedings of the European conference on computer vision (ECCV). 2018: 801-818.

[24] Han, L. Yang, Z, H. LI, L, Z. Liu, Z, X. and Huang, B, W. Road extraction of high resolution RS imagery based on DeepLab V3. remote sensing Information. 2021, 36(1): 22-28.

[25] Liu, R. He, D. Semantic Segmentation Based on DeepLabv3+ and Attention Mechanism, 2021 IEEE 4th Advanced Information Management, Communicates, Electronic and Automation Control Conference (IMCEC). IEEE, 2021.

[26] Ding, Y, L. Hong X, J. Ni, N, R. Cheng, S, L. and Tao, G. Computer Aided Annotation of Early Esophageal Cancer in Gastroscopic Images based on Deeplabv3+ Network, the 2019 4th International Conference. 2019.

[27] Su, A. Using improved DeepLabv3+network integrated with normalized difference water index to extract water bodies in Sentinel-2A urban RS images. Journal of Applied remote sensing. 2021, 15(1).

[28] Yang, W. Zhang, J, L. Xu, Z, Y. and Hu, K. Real-time DeepLabv3+ for pedestrian segmentation. Journal of Optical Technology c/c of Opticheskii Zhurnal. 2019, 86(9):570.

[29] Howard, A, G. Zhu, M. Chen, B. Wang, W. and Weyand, T. Andreetto, M.; Adam, H. MobileNets: efficient convolutional neural networks for mobile vision applications.

This work is licensed under a Creative Commons Attribution 4.0 License. For more information, see https://creativecommons.org/licenses/by/4.0/

This article has been accepted for publication in IEEE Access. This is the author's version which has not been fully edited and
content may change prior to final publication. Citation information: DOI 10.1109/ACCESS.2023.3313390

[30] Huang, S. He, Y. Chen, X, A. M-YOLO: A Nighttime Vehicle Detection Method Combining MobileNet- V2 and YOLO v3. Journal of Physics Conference Series. 2021, 1883(1):012094.
[31] Li, J, X. Zhao, H. Zhu, S, P. Jiang, Z, Y. An improved lightweight network architecture for identifying tobacco leaf maturity based on Deep learning. Journal of Intelligent & Fuzzy Systems: Applications in Engineering and Technology. 2021(2):41.
[32] Huu, P, N. Thu, H. Minh, Q, T. Proposing a Recognition System of Gestures Using MobilenetV2 Combining Single Shot Detector Network for Smart-Home Applications. Journal of Electrical and Computer Engineering. 2021, 2021(3):1-18.
[33] Lu, S. Wang, S. Zhang, Y, D. A Classification method for Brain MRI via MobileNet and Feedforward Network with Random Weights. Pattern Recognition Letters. 2020, 140.
[34] Chen, L, C. Papandreou, G. Kokkions, I. Murphy, K. Yuille, A. DeepLab: semantic image segmentation with deep convolutional nets, atrous convolution, and fully connected CRFs. IEEE Transactions on Pattern Analysis and Machine Intelligence. 2018, 40(4): 834-848.
[35] Zhu, W. ASPP-DF-PVNet: Atrous Spatial Pyramid Pooling and Distance-Filtered PVNet for occlusion resistant 6D estimation. Signal Processing. Image Communication: A Publication of the the European Association for Signal Processing. 2021, 95(1).
[36] Zhu, Y. Wan, L. Xu, W. ASPP-DF-PVNet: Atrous Spatial Pyramid Pooling and Distance-Filtered PVNet for occlusion resistant 6D object pose estimation. Signal Processing: Image Communication. 95.
[37] Hu, J. Shen, L. Albanie, S. Squeeze-and-excitation networks. IEEE Transactions on Pattern Analysis and Machine Intelligence. 2020, 42(8): 2011-2023.
[38] Li, J. Tian, Y. Lee, T. Convolution-Based Channel-Frequency Attention for Text-Independent Speaker Verification. 2022. https://arxiv.org/abs/2210.17310
[39] Park, J, H. Seo, S, M. Ji, H, Y. Channel Attention Module in Convolutional Neural Network and Its Application to SAR Target Recognition Under Limited Angular Diversity Condition. The Korea Institute of Military Science and Technology. 2021(2).
[40] Zhu, X, Z. Cheng, D, Z. Zhang, Z. Lin, S. Dai, J. An empirical study of spatial attention mechanisms in deep networks, 2019 IEEE/CVF International Conference on Computer Vision (ICCV). Seoul, Korea (South): IEEE, 2020: 6687-6696.
[41] LI, Z, Q. Huang, Y, F. Cai, M, J. Manipulation-skill assessment from videos with spatial attention network, 2019 IEEE/CVF International Conference on Computer Vision Workshop (ICCVW). Seoul, Korea (South): IEEE, 2020: 4385-4395.
[42] Cui, Z, Y. Li, Q. Cao, Z, J. Dense attention pyramid networks for multi-scale ship detection in SAR images, IEEE Transactions on Geoscience and remote sensing. [S. l.]: IEEE, 2019: 8983-8997.
[43] Hao, Z. Berg, A, C. Maire, M. Malik J. SVM-KNN: Discriminative Nearest Neighbor Classification for Visual Category Recognition, 2006 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR'06). IEEE, 2006.
[44] Liaw, A. Wiener, M. Classification and Regression by Random Forest. R News, 2002, 23(23).
[45] Heremans, S. Orshoven, J. Machine learning methods for sub-pixel land-cover classification in the spatially heterogeneous region of flanders (belgium): a multi-criteria comparison. International Journal of Remote Sensing. 2015, 36(11-12), 2934-2962.



RUI WANG. China Transport Telecommunications & information Center, Beijing 100011, China. Dr. Wang Rui graduated from Wuhan University. His research interests include the application of traffic data in the insurance industry, spatial big data management and the study of image recognition algorithms based on machine learning. He currently works at the China Transport Telecommunications and information Center.



MINGXIANG CAI. China Transport Telecommunications & information Center, Beijing 100011, China. Dr. Mingxiang, Cai, graduated from Wuhan University. His research interests include spatio-temporal information mining and application of fundamental models in traffic big data.



ZIXUAN XIA entered Heilongjiang University of Technology in 2020 and will receive a degree in surveying and mapping engineering in 2024. His current research interests include flight calibration and validation of forestry applications of high-definition aerial systems and the study of vegetation cover differentiation based on the image dichotomous model.

This work is licensed under a Creative Commons Attribution 4.0 License. For more information, see https://creativecommons.org/licenses/by/4.0/