RSUnet: A New Full-scale Unet for Semantic Segmentation of Remote Sensing Images

Songhua Chen (shchen97@stu.xjtu.edu.cn)
Xi'an Jiaotong University https://orcid.org/0000-0001-5490-445X
Bin Zhang
Xi'an Jiaotong University

Research

Keywords: remote sensing images, semantic segmentation, skip connection, feature fusion, adaptive feature selection module

Posted Date: February 14th, 2022

DOI: https://doi.org/10.21203/rs.3.rs-1211375/v1

RSUnet: A New Full-scale Unet for Semantic Segmentation of Remote Sensing Images

Songhua Chen and Bin Zhang
School of Software
Xian Jiaotong University, China
shchen97@stu.xjtu.edu.cn, bzhang82@mail.xjtu.edu.cn

Abstract

Robust semantic segmentation algorithm of remote sensing images is essential for flood detection, land use, land cover and mapping applications. However, high-resolution remote sensing images contain a large amount of ground object information, showing diversity and complexity with large intraclass variance, small inter-class variance and low-class discrimination, which makes semantic segmentation more difficult. Especially when segmenting the edge of objects in remote sensing images, it is easy to have irregular shapes and segmentation errors inside the objects. Recently, Unet series of semantic segmentation networks have become popular not only in the field of medical image segmentation, but also in the field of general semantic segmentation. Unet is based on the symmetrical encoder-decoder structure. As an improvement of Unet, Unet++ is composed of Unets of different depths. The skip connections in Unet++ are redesigned to achieve flexible feature fusion in the decoder. However, it still does not make full use of multiscale feature information and there is room for improvement. In this paper, we propose a new Unet called RSUnet for semantic segmentation of remote sensing images and Adaptive Feature Selection Module, which fuses deep feature and shallow feature to the extreme and selects useful feature during feature fusion to solve the problem of poor segmentation caused by the migration of Unet series of networks to the field of remote sensing image segmentation, especially the poor edge segmentation of objects. Experiments show RSUnet with adaptive feature selection module performs better than Unet series of networks and other mainstream semantic segmentation models in five public remote sensing image segmentation datasets.

Keywords: remote sensing images, semantic segmentation, skip connection, feature fusion, adaptive feature selection module

1 Introduction

In the last few decades, faced with increasing technical demands, various segmentation algorithms have continuously emerged. Common remote sensing im-

age segmentation algorithms are mainly divided into segmentation algorithms based on shallow feature, segmentation algorithms based on middle feature and segmentation algorithms based on deep feature. In view of the simplicity of early available remote sensing images, classical remote sensing image segmentation algorithms mostly rely on the shallow feature (edges, textures, colors, geometric shapes) of images. Common algorithms based on shallow feature are threshold segmentation algorithm based on image pixels, edge detection segmentation algorithm and area segmentation algorithm. Threshold segmentation algorithm's principle is simple and its calculation cost is low. But its anti-noise performance is poor and threshold is not easy to be determined when the background is complex. Edge detection segmentation algorithm has string anti-noise ability and objects with simple geometric shapes can be segmented better. But it is difficult to detect a complete edge. Region segmentation algorithm has good anti-noise performance, compact segmentation results and easy expansion of technology. But over segmentation occurs frequently and selection of seed regions and establishment of homogeneity criteria is not easy. Traditional segmentation algorithms based on shallow feature are suitable for remote sensing images with simple content. When external conditions such as light, sensors, clouds and fog change frequently, it will have adverse effects on pixel characteristics. The segmentation result is more susceptible to interference and robustness is not strong enough. As a result, researchers begin to use mathematical theory and graph theory to introduce prior information and contextual information or mix multiple shallow feature to obtain middle feature with better discrimination and robustness. Common segmentation algorithms based on middle feature include clustering segmentation, Markov random field segmentation and hybrid feature combination segmentation. Clustering segmentation algorithms have strong scalability. The common ones are based on K-means[1] and fuzzy C-means[2]. They are simple, unsupervised and fast. Their disadvantages are the center and number of clusters cannot be determined and they are sensitive to noise. Markov random field segmentation algorithm considers the neighborhood relationship of the image by obtaining the context limit of adjacent pixels, which is attractive for modeling of image texture and context. The introduction of context information, segmentation accuracy and noise immunity are significantly improved. But it still has high computational complexity. Hybrid feature combination segmentation is proposed by researchers to overcome the limitations of single shallow feature segmentation. Two or more segmentation algorithms complement each other to overcome the shortcomings of using a single feature for semantic segmentation of remote sensing images. With the increasement of resolution of remote sensing images, situation of foreign objects with the same spectrum and same objects with different spectrums in images appear more and more frequent. The improvement space of traditional shallow and middle feature representations is quite limited. With the development of artificial intelligence and deep learning, researchers have discovered deep neural networks can extract deep feature. Strong robustness is a typical advantage of deep neural network. In the deeper network layer, more abstract and global information will be extracted and stronger feature representation ability will

be gained. Segmentation algorithms based on deep feature are gradually being migrated to remote sensing images. The most prominent representative is convolutional neural network. It has gradually become the mainstream algorithm in the field of semantic segmentation due to its strong ability to extract image feature. In the field of semantic segmentation, many CNN-based algorithms have appeared, such as FCN[3], SegNet[4], MANet[5], PSPNet[6], DeepLab series, LinkNet[7] and Unet series. In 2015, the full convolutional network FCN as a representative was of epoch-making significance for image segmentation. It achieves pixel-level image semantic segmentation and uses CNN structure to fully connect classification and mapping. The obtained feature heat map is upsampled to the original input image size through deconvolution operation. At the same time, the image prediction segmentation map is generated by combining the information of intermediate pooling layers. Based on the structure of FCN's downsampling and upsampling, a deep convolutional encoder-decoder structure SegNet was proposed to identify and segment objects such as streets and vehicles in cities. Its first five modules correspond to encoder, which is a process of gradual downsampling for feature extraction. The latter corresponds to decoder, which uses the idea of upsampling to restore the image size to the original input size. It saves memory by retaining only the pooling index value of the encoder structure. The target edge can be restored by using max pooling index value in the decoding stage and detail information of the image is preserved, which effectively improves the segmentation accuracy. MANet is different from the Unet series in that it introduces self-attention mechanism to adaptively integrate local feature with global feature. MANet can capture rich contextual feature based on the attention mechanism. PSPNet mainly proposes Pyramid Scene Parsing network, which merges feature at different scales to achieve the fusion of semantic information and edge details and combines feature of four different pyramid scales. There are four versions of DeepLab series, namely v1, v2, v3 and v3+. DeepLabV1[8] mainly solves two main problems of standard deep convolutional neural network. One of them is striding operation reduces the output size and the other is the invariance of pooling to small changes in the input. Atrous convolution and conditional random field solve these problems. To solve the multiscale problem of objects, DeepLabV2[9] proposes the atrous spatial pyramid pooling module, which performs resampling on a given feature layer at multiple sampling rates before convolution and uses multiple parallel convolutional layers with different sampling rates. The multiscale segmentation objects obtained by sampling at multiple sampling rates make the model obtain better segmentation results. DeepLabV3[10] proposes an enhanced ASPP module, which includes a 1*1 convolution and three 3*3 atrous convolutions to capture feature under different receptive fields and extracts denser feature. However, due to the existence of pooling and stride convolution, the boundary information of segmentation targets is seriously lost. DeepLabV3+[11] proposes an encoder-decoder structure, the decoder can repair the sharp object boundary. It also tries to improve Xception[12], which eventually performs better than ResNet[13]. LinkNet is also an encoder-decoder structure and solves a major pain point of semantic segmentation task that is not real-time enough. By

bypassing the spatial information, encoder and decoder are directly connected to improve the accuracy, which reduces the processing time. In this way, the lost information in different layers during encoding will be retained. At the same time, no additional parameters and operations are added when relearning the lost information. Unet[14] is originally used in medical images. Its idea borrows from FCN. Its network structure includes two symmetrical parts. The previous part of the network is the same as the normal convolutional network, using 3*3 convolution and downsampling as max pooling, which can capture the pixel relationship. The latter part is basically symmetrical with the previous one, using 3*3 con-volution and upsampling to achieve the purpose of output image segmentation. In addition, feature fusion is also used in the network, which can lead model to obtain more accurate context information and achieve better segmentation result. Unet++[15] improves feature fusion on the basis of Unet and proposes a pruning scheme to accelerate the model inference. To meet the demand for more accurate semantic segmentation and enhance the effect of model on edge segmentation of remote sensing images, we further study how to reduce false positive and false negative. In summary, our main contributions are two-fold: (i) we design a novel Unet structure called RSUnet that makes full use of multiscale feature by introducing full-scale skip connections and fusing low-level detail feature and high-level semantic feature of full-scale feature maps. (ii) We propose an adaptive feature selection module, which can adaptively select semantic feature that is beneficial to the segmentation result and suppress other feature that is unfavorable to the segmentation result. Figure 1 shows the similarities and differences between the proposed model RSUnet, Unet and Unet++. Unet has four downsampling layers, four upsampling layers and four skip connections. Compared with Unet, Unet++ has a richer feature fusion architecture. RSUnet takes feature fusion to the extreme and fully integrates low-level edge feature and high-level semantic feature.

2 Methods

2.1 Motivation behind this new architecture

In the field of medical image, Unet and Unet++ can segment target organs and lesions. But when they are migrated to the field of remote sensing image segmentation, their performances seem to be bad. We have done experiments on multiple remote sensing image segmentation datasets. Figure 2 shows the segmentation effect diagrams of segmentation truth, DeepLabV3, MANet, PSPNet, Unet and Unet++ on the Berlin street dataset. All of them do not perform well on the edge segmentation of remote sensing image targets. The worst model is PSPNet. There are segmentation errors inside the segmented streets. DeepLabV3 is also not good at edge segmentation and has irregular segmentation result compared to segmentation truth. The segmentation effects of MANet and Unet are better, but the segmentation edge appear jagged. Compared to other models, Unet++ segments best, but it is also lacking in the regularity of

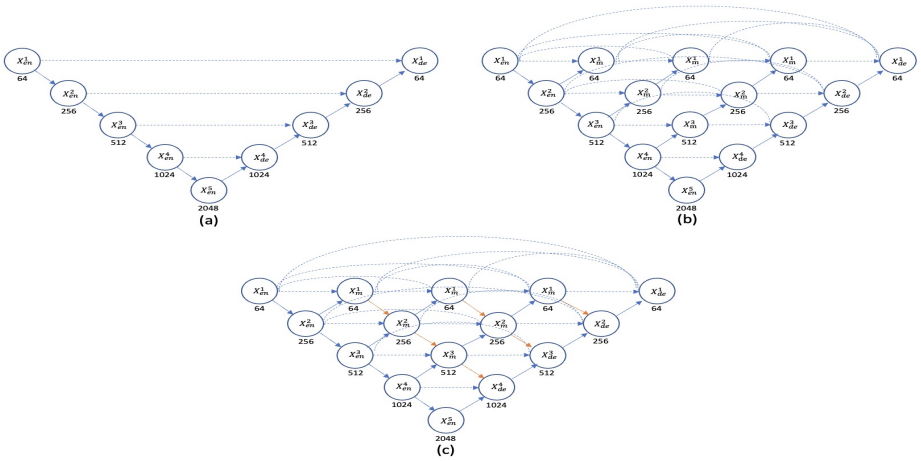

Figure 1: Model structure:(a)Unet; (b)Unet++; (c)RSUnet. Each node in the graph represents a convolution block. Downward arrows indicate downsampling. Upward arrows indicate upsampling. Dot arrows indicate skip connection. The number under node represents the number of channels. Xen represents the encoder layer. Xm represents the intermediate layer between encoder and decoder. Xde represents the decoder layer. Orange arrows indicates the difference between RSUnet and Unet++.

street shape and the effect of edge segmentation. We wonder if we can propose a new semantic segmentation model that can better segment remote sensing image targets and obtain edge information segmentation effect is not bad. Therefore, we have made some improvements in feature fusion and feature selection.

2.2 Richer feature fusion structure

As shown in above Figure 1, the overall structure of Unet is to encode, then decode. It can also be said that it first downsamples, upsamples and then returns to the classification of pixels of the same size as the original image. It contains four simple skip connections and fuses feature maps during downsampling. Unet++ is composed of Unets of different depths. Its decoder is densely connected with the same resolution through redesigned skip connections. The redesigned skip connections introduced in Unet++ provide feature maps of different proportions at the decoder nodes. So that the aggregation layer can decide how to merge the various feature maps carried in the skip connections with the decoder's feature maps. Both Unet with plain skip connections and Unet++ with nested and dense skip connections are short of exploring sufficient feature from all kinds of scales, failing to learn the edge target feature in the remote sensing images. There are many feature fusion methods in deep learning. The most commonly used ones are feature fusion used in FPN[16] and DenseNet[17]. The main problem FPN solves is the insufficiency of target

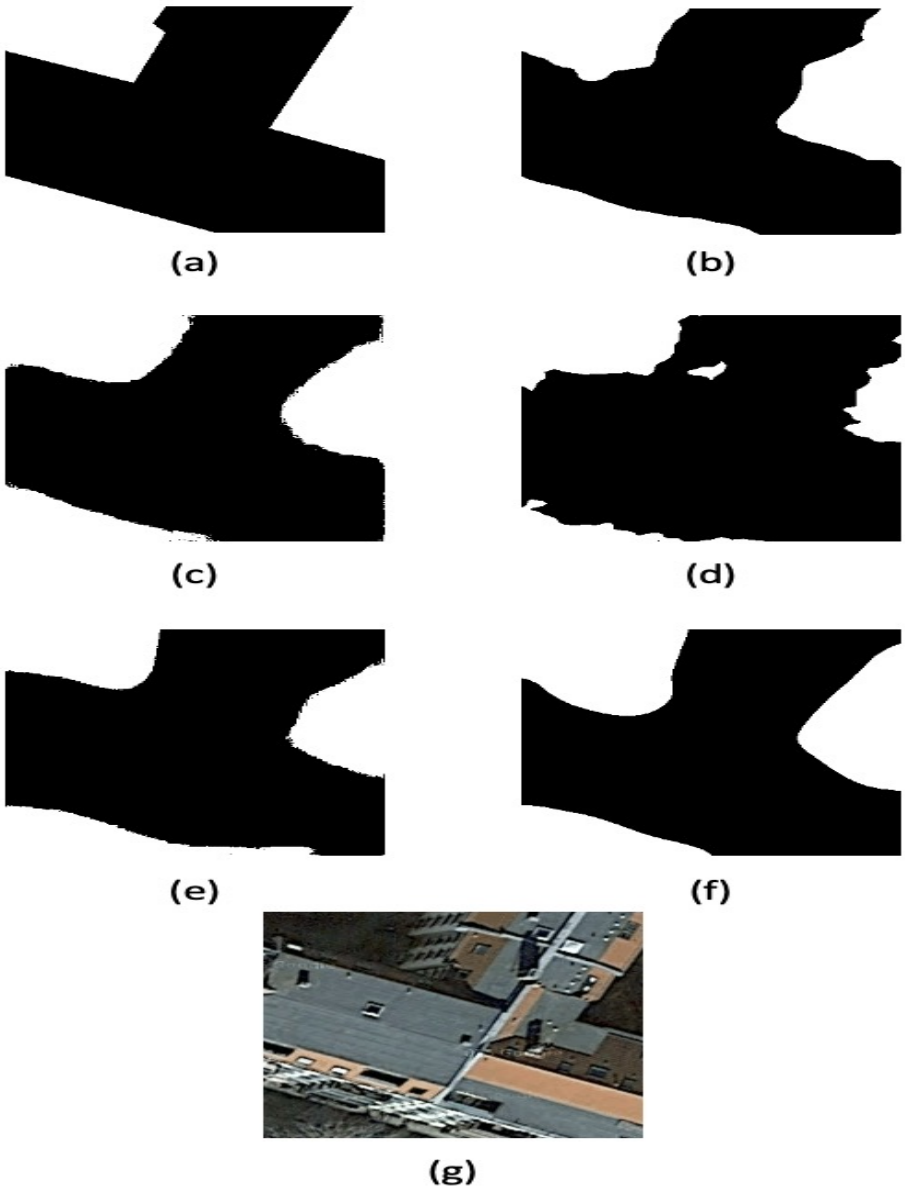

Figure 2: Visual comparison of the segmentation effects of (a) Truth; (b) DeepLabV3; (c) MANet; (d) PSPNet; (e) Unet; (f) Unet++; (g) Original Image on berlin street dataset.

detection in dealing with multiscale changes. It proposes a network structure that uses the inherent muitscale pyramid structure of deep convolutional neural networks to construct feature pyramids with a very small amount of calculation. This is a top-down network structure with lateral connections, which uses to construct feature maps of different sizes with high-level semantic information. The feature fusion adopted in FPN performs feature extraction on images of each scale, which can produce multiscale feature representations and feature maps of all levels. But at the same time, the inference time has increased significantly. DenseNet breaks away from the fixed thinking of deepening the number of network layers (ResNet) and widening the network structure (Inception[18]) to improve network performance. From the perspective of feature reuse and bypass design, it greatly reduces the network's parameters and alleviates the emergence of gradient vanishing. Compared with ResNet, DenseNet proposes a more radical dense connection mechanism. All layers are connected to each other. Specifically, each layer will accept all the previous layers as its additional input. These feature fusion mechanisms have achieved good results in the field of general image segmentation, but the performance of edge segmentation in remote sensing image segmentation is still not good. To eliminate this shortcoming, we design a richer feature fusion structure, which can fuse both feature of the same scale and feature of different scales and capture fine-grained details and coarse-grained semantic in full scales. To better merge the shallow edge feature with deep semantic feature, we further perform a feature fusion mechanism on the concatenated feature map from five scales, which consists of 320 filters of size 3*3, a batch normalization and a ReLU activation function. Formally, we formulate the skip connections as follows: let i indexes the downsampling layer along the encoder, N refers to the total number of the encoder. Function C denotes a convolution operation, H denotes the feature fusion with a convolution followed by a batch normalization and a ReLU activation function. D and U denote upsampling and downsampling operation respectively. [] denotes the concatenation of feature maps. The stack of feature maps represented by Xde is computed by the following equation:

$$X_{De}^{i} = H(C(D(X_{En}^{k})_{k=1}^{i-1}), C(X_{En}^{i}), C(U(X_{De}^{k})_{k=i+1}^{N})), i = 1, ..., N \quad (1)$$

2.3 Adaptive feature selection module

We have studied many semantic segmentation algorithms such as PSPNet, MANet, Unet series and DeepLab series and find that they simply concatenate or add low-level edge feature and high-level semantic feature when fusing feature, but this is unreasonable. Just as we humans observe the world around us, we do not directly pay attention to all the objects around us. We always first observe the objects we want to notice. This is also called Attention Mechanism in the field of image classification. The classic algorithms include SENet[19], SKNet[20], CBAM[21] etc. For convolution network, its core calculation is the convolution operator, which learns a new feature map from the input feature

map through convoution kernel. Essentially, convolution is a feature fusion of the local area, which includes spatial (h and w dimensions) and inter-channel (c dimension) feature fusion. For covolution operations, a large part of work is to improve the receptive field to fuse more feature fusion spatially or to extract multiscale spatial information, just like the multi-branch structure of Inception. For feature fusion of channel dimensions, convolution operations are fusion of all channels of the input feature map. The group convolution and depthwise separable covolution in the MobileNet[22] mainly to make model more lightweight and reduce the amount of calculation. The innovation of SENet is to focus on the relationship between channels and hope that the model can automatically learn the importance of different channel features. SENet proposes the Squeeze-and-Excitation module. The SE module first performs convolution on the feature map obtained. Then, it performs a squeeze operation to get the channel-level global feature and performs an excitation operation on the global features. SENet implements the attention mechanism on the channels. SKNet implements the attention mechanism on the convolution kernel, which lets the network choose appropriate covolution kernels adaptively. Specifically, SKNet is divided into three steps, namely Split, Fuse, Select. Split is the stage uses different convolution kernels to convolve the original image. Fuse means combine and aggregate information from multiple paths to obtain a global selection weight and comprehensive representation. Select is to aggregate feature maps of kernels of different sizes according to selection weights. CBAM includes two independent submodules, Channel Attention Module (CAM) and Spatial Attention Module (SAM), which perform channel and spatial attention respectively. This not only saves parameters and computing power, but also ensures that it can be integrated into the existing network architecture as a plug module. In our research, we find that in the field of semantic segmentation, we can also perform adaptive feature selection when fusing feature to enhance the feature that are beneficial to the semantic segmentation effect and suppress those that are unfavorable to the segmentation result. As shown in Figure 3, we call this adaptive feature selection module. It shows how feature fusion layer of the last layer of the decoder implements adaptive feature fusion. The input are four feature maps with the same scale and different layers containing 64 channels and a feature map with different scales and different layers containing 256 channels and then each feature map is subjected to global average pooling to obtain global feature and then through 1*1 convolution for feature scoring and then normalized by softmax and then multiply the normalized value with the original feature map to get the feature map after adaptive feature selection.

2.4 Dataset and implementations

Our model has been metrically verified on five public datasets. These datasets are Massachusetts Build, Massachusetts Road, Google online map in berlin street, Urban Drone Dataset and BH dataset. Massachusetts Build and Massachusetts Road are an aerial image segmentation dataset released by the University of Toronto in 2013. It contains two types of remote sensing feature type,

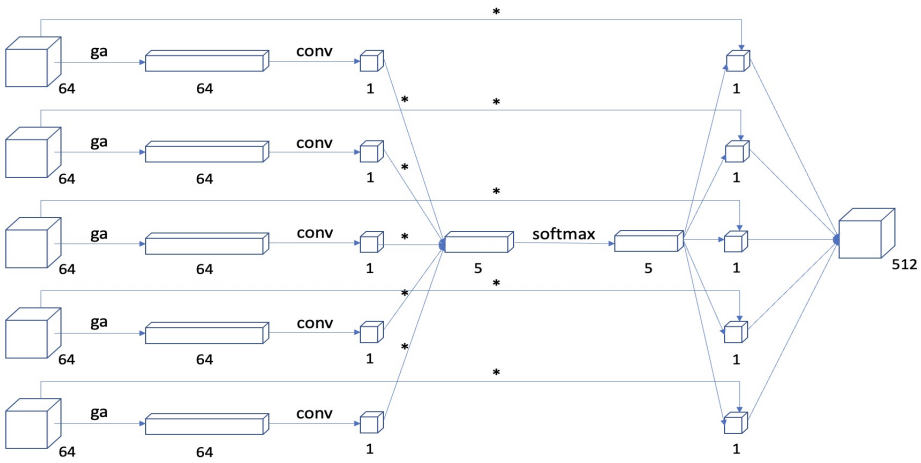

Figure 3: Adaptive feature selection module. Ga is global average pooling operation. Conv denotes convolution operation. The number under box represents the number of feature maps' channels.

and the image size is 1500*1500*3. Google online map in berlin street is extracted from Google Maps and OpenStreetMap organizations and released by ETH Zürich in 2017. It contains two types of remote sensing feature. The image size is 3000*3000*3. Urban Drone Dataset includes two data subsets UDD5 and UDD6, which is extracted from drone data (DJI Phantom 4), released by Peking University in 2018. They contain five and six types of remote sensing feature respectively and the image size is 4196*2160*3 with 120 images as the training dataset and 40 images as the validation dataset. BH-DATASET includes 2 data subsets, BH-POOLS & BH-WATERTANKS, extracted from Google Earth and released by Federal University of Minas Gerais in 2020. Each includes two type of remote sensing feature type. The image size is 3840*2160*3. It contains 200 pictures and 150 pictures respectively. In order to speed up training, the input image has three channels including the slice to be segmented and the upper and lower slices, which is cropped to 300*300. There are many optimizers that can be used to train neural networks such as SGD, Adam, AdamW. We utilize the AdamW to optimize our neural network because it can make the neural network converge quickly and it is not easy to make the network fall into the local minima. Its learning rate is set to 1e-4 and other hyperparameters are set to default values. The learning rate decline strategy during training is ReduceLROnPlateau. For train loss, as long as three epochs are not reduced, the learning rate will be reduced to half of the previous. The model is trained for a total of 200 epochs. Because miou is the most commonly used evaluation index for remote sensing image segmentation, we also choose it as an index to evaluate the effects of our model and other sota models. Miou is mean intersection over union. TP is true positive, that is, the prediction class is correct. FP is false positive, that is, the model classifies it as a positive example but the truth

is not a positive example. FN is false negative, that is, the model classifies it as a negative example, but the truth is not a negative example. Miou can be computed by the following equation:

$$miou = \frac{1}{k+1} \sum_{i=0}^{k} \frac{TP}{TP + FP + FN} \tag{2}$$

3 Results

3.1 Comparison with Unet and Unet++

In this section, we first compare RSUnet with Unet and Unet++ in three public datasets. The loss functions used in our method are CrossEntropy Loss, Focal loss and Dice loss. The encoders used in our model are resnet18 and resnet50. The performance of RSUnet is better than Unet and Unet++. Then we visualize the segmentation results of Unet, Unet++ and RSUnet in Figure 4. 1-6 are the results of experiments using CrossEntropy Loss, Focal Loss and Dice Loss on the Berlin street dataset, Massachusetts Build dataset, Massachusetts Road dataset, BH-pool dataset and UDD dataset with Unet, Unet++ and RSUnet. On the Berlin street dataset, RSUnet based on Resnet18 encoder with AFS module has the best segmentation result. The segmentation effect based on the Resnet50 encoder is the same, but when using Dice Loss as the loss function, the RSUnet with AFS module has poor performance compared to the original RSUnet. We analyze it may be that Dice Loss has made model have better generalization performance and the feature selection will interface with the training of the model. On the Massachusetts Build dataset, RSUnet with AFS module also performs best and the miou index drops slightly under Dice Loss. On the Massachusetts Road dataset, the RSUnet with AFS module also performs best and the index drops a little under Dice Loss and Focal Loss. We think that this dataset's class is already relatively balanced. If focal Loss allows the model to learn hard samples, it will reduce the generalization performance of the model. Unet and Unet++ are not good at segmentation of street edge on the Berlin Street dataset and Massachusetts Road dataset. Straight streets are divided into curved ones. RSUnet solves this problem. On the Massachusetts Build Dataset, Unet and Unet are not good at building segmentation. The boundary of different buildings is not obvious. The segmentation effect of RSUnet is improved compared with Unet and Unet++.

3.2 Comparison with other State of the Art models

In order to further illustrate the effect of our proposed model RSUnet, we compare the effect of other sota models on the public dataset. As shown in below, RSUnet performs better than other sota models on all public datasets. On the Berlin Street dataset, Masschusetts Road dataset and Masschusetts Build dataset, RSUnet has a point miou improvement compared to other models. After adding AFS module, RSUnet also has a point miou improvement. On

Table 1: Miou of Unet, Unet++, RSUnet and RSUnet with AFS on Berlin street dataset.

Model	Resnet18			Resnet50		
	CrossEntropy	Focal	Dice	CrossEntropy	Focal	Dice
Unet	0.8426	0.8560	0.8565	0.8630	0.8629	0.8642
Unet++	0.8555	0.8640	0.8596	0.8666	0.8672	0.8690
RSUnet	0.8622	0.8682	0.8602	0.8746	0.8698	**0.8708**
AFS	**0.8704**	**0.8696**	**0.8622**	**0.8822**	**0.8782**	0.8702

Table 2: Miou of Unet, Unet++, RSUnet and RSUnet with AFS on Massachusetts Build dataset.

Model	Resnet18			Resnet50		
	CrossEntropy	Focal	Dice	CrossEntropy	Focal	Dice
Unet	0.5479	0.5441	0.5747	0.5823	0.5223	0.6094
Unet++	0.5577	0.5828	0.6259	0.6326	0.6356	0.6442
RSUnet	0.5722	0.5926	**0.6402**	0.6404	0.6404	0.6553
AFS	**0.5784**	**0.6014**	0.6400	**0.6406**	**0.6501**	**0.6598**

Table 3: Miou of Unet, Unet++, RSUnet and RSUnet with AFS on BH-pool dataset.

Model	Resnet18			Resnet50		
	CrossEntropy	Focal	Dice	CrossEntropy	Focal	Dice
Unet	0.6505	0.6556	0.6652	0.7027	0.7124	0.7222
Unet++	0.6678	0.6604	0.6802	0.7422	0.7333	0.7506
RSUnet	0.6772	0.6723	**0.6894**	0.7401	**0.7456**	0.7566
AFS	**0.6802**	**0.6866**	0.6804	**0.7455**	0.7425	**0.7602**

Table 4: Miou of Unet, Unet++, RSUnet and RSUnet with AFS on UDD5 dataset.

Model	Resnet18			Resnet50		
	CrossEntropy	Focal	Dice	CrossEntropy	Focal	Dice
Unet	0.7934	0.7804	0.8012	0.8431	0.8202	0.8545
Unet++	0.8012	0.7992	0.8156	0.8681	0.8302	0.8669
RSUnet	0.8001	0.8045	**0.8215**	0.8691	0.8493	0.8652
AFS	**0.8192**	**0.8225**	0.8207	**0.8802**	**0.8546**	**0.8842**

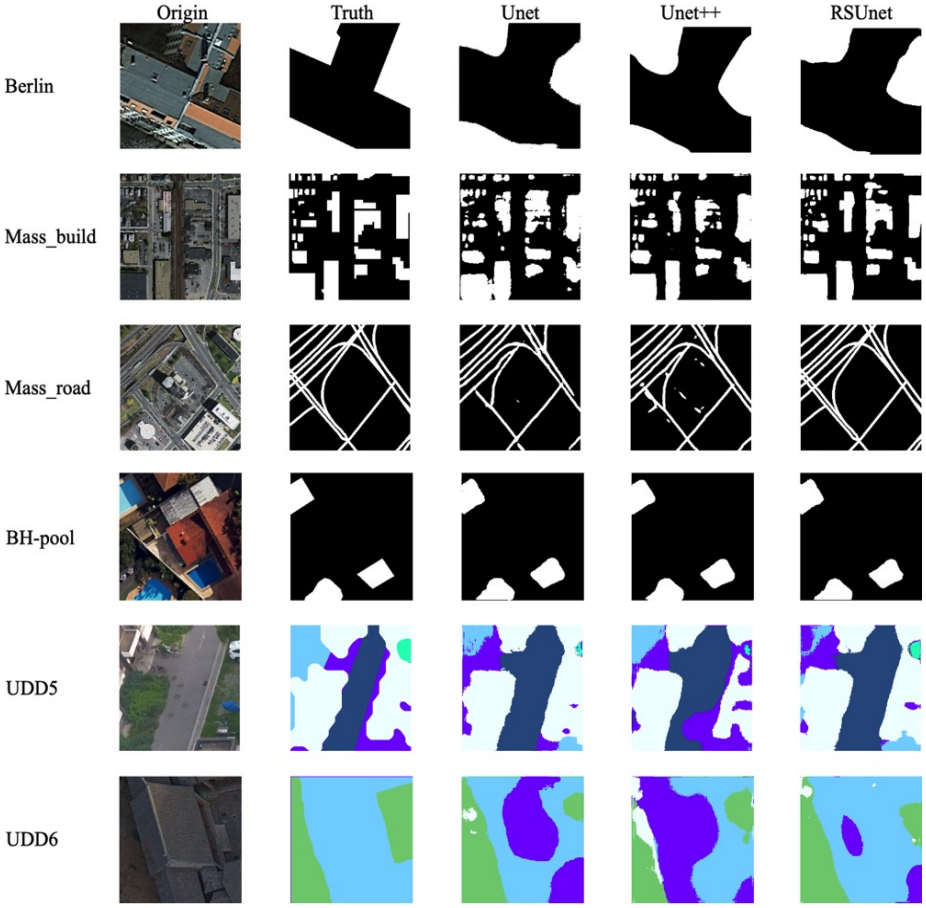

Figure 4: Qualitative comparison of Resnet50-based Unet, Unet++ and RSUnet on Berlin street dataset, Massachusetts Road dataset, Massachusetts Build dataset, BH-Pool dataset, UDD5 dataset and UDD6 dataset. On the Berlin street dataset, the black area is the street segmented by models. On the Massachusetts Road dataset, Massachusetts Build dataset, BH-Pool dataset, the white area is building, road, pool segmented by models. On the UDD dataset, the dark blue area is the street segmented by models. The light blue area is roof segmented by models. The purple area is other class segmented by models. The white area is vegetation segmented by models. The light green area is vehicle segmented by models.

Table 5: Miou of Unet, Unet++, RSUnet and RSUnet with AFS on UDD6 dataset.

Model	Resnet18			Resnet50		
	CrossEntropy	Focal	Dice	CrossEntropy	Focal	Dice
Unet	0.5479	0.5441	0.5747	0.6240	0.6340	0.6094
Unet++	0.5577	0.5828	0.6259	0.6326	0.6498	0.6442
RSUnet	0.5722	0.5926	**0.6402**	0.6345	0.6404	0.6553
AFS	**0.5784**	**0.6014**	0.6400	**0.6404**	**0.6501**	**0.6598**

Table 6: Miou of Unet, Unet++, RSUnet and RSUnet with AFS on Massachusetts Road dataset.

Model	Resnet18			Resnet50		
	CrossEntropy	Focal	Dice	CrossEntropy	Focal	Dice
Unet	0.5340	0.5317	0.5444	0.5776	0.5812	0.5979
Unet++	0.5679	0.5664	0.5880	0.5968	0.5844	0.6097
RSUnet	0.5706	0.5728	**0.5921**	0.6018	0.5866	0.6092
AFS	**0.5774**	**0.5798**	0.5901	**0.6111**	**0.5848**	**0.6111**

BH-pool Dataset, UDD5 Dataset and UDD6 Dataset, RSUnet has an improvement of half a point miou. After adding AFS module, miou is slightly reduced. We analyze it may be on segmentation tasks with more than two classes, feature fusion and feature selection are affected by classes leading to insignificant improvement in segmentation effect.

Table 7: Miou of DeepLabV3, MANet, PSPNet, Unet, Unet++, RSUnet and RSUnet with AFS on five public datasets.

Model	Berlin	BHPOOL	Mass_build	Mass_road	UDD5	UDD6
DeepLabV3	0.8684	0.6868	0.6048	0.5776	0.8527	0.6048
MANet	0.8637	0.6413	0.6034	0.5653	0.8453	0.6034
PSPNet	0.8461	0.6245	0.5167	0.5511	0.8263	0.5167
Unet	0.8630	0.7027	0.5823	0.5776	0.8431	0.6240
Unet++	0.8666	0.7422	0.6326	0.5968	0.8681	0.6326
RSUnet	0.8746	**0.7476**	0.6404	0.6018	0.8742	0.6404
AFS	**0.8822**	0.7471	**0.6406**	**0.6111**	**0.8756**	**0.6424**

4 Discussion

Experiments show on the Berlin street dataset, RSUnet uses resnet18 as the encoder and loss uses CrossEntropy Loss, Focal Loss and Dice Loss, the miou index has been improved compared with Unet and Unet++. With AFS module, RSUnet's segmentation effect can be further improved. It can be seen that the useful feature selection of AFS module we proposed has a gain in the segmentation result. For Resnet50 and Dice Loss, the segmentation miou of AFS module is reduced. We analyze because the performance limit of the model has reached with Dice Loss, the feature selection will not have much effect. 2-6's results are similar to 1. Compared with Unet and Unet++, RSUnet equipped with AFS module has an effect gain. Figure 4 shows RSUnet has a better segmentation effect than Unet and Unet++. On the Berlin street dataset, the street segmented by Unet and Unet++ are inferior to RSUnet in terms of shape regularity. On the Massachusetts Build dataset, it can be clearly seen that Unet and Unet++ split some buildings into fragments. In contrast, the result segmented by RSUnet is easier to be recognized as buildings. On the Massachusetts Road dataset, Unet and Unet++ will incorrectly segment some results that should not be road in the middle area. At the same time, the road lines segmented by Unet and Unet++ are not straight. The result segmented by RSUnet is almost as same as the truth. On the multiclass UDD dataset, all models perform unsatisfactorily. This is also what we need to explore in model structure improvement in the future. From above experimental results and visualization results, RSUnet and RSUnet with AFS module are improved compared to Unet and Unet++.

5 Conclusions

In this paper, we import Unet++ for medical image segmentation into remote sensing image segmentation and make two improvements based on it. we propose a richer feature fusion structure and adaptive feature selection module to make full use of feature maps in all scales for accurate segmentation and select them adaptively. We do ablation experiments on five public image segmentation datasets. Experimental results show that the proposed model RSUnet can outperform other mainstream sota algorithms such as PSPNet, Unet, Unet++ in the five public datasets of remote sensing image segmentation. We believe that these improvements can play a main role in flood detection, land change detection and other fields.

Abbreviations

FCN: Fully convolutional networks; SegNet: Segmentation network; MA-Net: Multi-Scale attention network; PSPNet: Pyramid scene parsing network; LinkNet: Link network; CRF: Contional random field; Unet: 'U' network; SENet: Squeeze-and-Excitation networks; SKNet: Selective kernel networks; CBAM: Convolutional block attention module; AFS: Adaptive feature selection.

Declarations

Ethics approval and consent to participate

Not applicable

Consent for publication

Not applicable

Availability of data and materials

Please contact author for data requests.

Competing interests

The authors declare that they have no competing interests.

Funding

This paper is not supported by any funding.

Authors' contributions

All the experiments in the paper and the first draft of the paper were completed by Songhua Chen. Bin Zhang gave valuable opinions on the improvement of the experiment and the revision of the paper.

Acknowledgements

In addition to the authors, the following individuals contributed to guiding suggestions in the model training stage: Wujiang Xu, Kui He.

Authors' information

Songhua Chen is a third-year graduate student of the School of Software, Xi'an Jiaotong University and Bin Zhang is an associate professor of the School of Software, Xi'an Jiaotong University.

References

[1] Macqueen. J.B, "Some Methods for Classification and Analysis of Multi-Variate Observations. Mathematical Statistics and Probability" 1967, 281-297.

[2] Dunn. J.C, "A Fuzzy Relative of the ISODATA Process and Its Use in Detecting Compact Well-Separated Clusters", Journal of Cybernetics 3, 1973, 32-57.

[3] Jonathan. L, Evan. S, Trevor. D, "Fully Convolutional Networks for Semantic Segmentation" International Conference on Computer Vision and Pattern Recognition, 2015.

[4] Badrinarayanan. V, Kendall. A, Cipolla. R, "SegNet: A Deep Convolutional Encoder-Decoder Architecture for Image Segmentation", International Conference on Computer Vision and Pattern Recognition, 2015.

[5] Tongle. F, Guanglei. W, Yan. L, Hongrui. W, "MA-Net: A Multi-Scale Attention Network for Liver and Tumor Segmentation", IEEE Access, 2020.

[6] Hengshuang. Z, Jianping. S, Xiaojuan. Q, Xiaogang. W, Jiaya. J, "Pyramid Scene Parsing Network", International Conference on Computer Vision and Pattern Recognition, 2017.

[7] Abhishek. C, Eugenio. C, "LinkNet: Exploiting Encoder Representations for Efficient Semantic Segmentation", International Conference on Computer Vision and Pattern Recognition, 2017.

[8] Liang-Chieh. C, George. P, Iasonas. K, Kevin. M, Alan. L.Y, "Semantic image segmentation with deep convolutional nets and fully connected CRFs", International Conference on Computer Vision and Pattern Recognition, 2015.

[9] Liang-Chieh. C, George. P, Iasonas. K, Kevin. M, Alan. L.Y, "DeepLab: Semantic Image Segmentation with Deep Convolutional Nets, Atrous Convolution, and Fully Connected CRFs", IEEE Transactions on Pattern Analysis and Machine Intelligence, 2016.

[10] Liang-Chieh. C, George. P, Florian. S, Hartwig. A, "Rethinking Atrous Convolution for Semantic Image Segmentatio", International Conference on Computer Vision and Pattern Recognition, 2017.

[11] Liang-Chieh. C, Yukun. Z, George. P, Florian. S, Hartwig. A, "Encoder-Decoder with Atrous Separable Convolution for Semantic Image Segmentation", Europe Conference on Computer Vision, 2018.

[12] François. C, "Deep Learning with Depthwise Separable Convolutions", International Conference on Computer Vision and Pattern Recognition, 2017.

[13] Kaiming. H, Xiangyu. Z, Shaoqing. R, Jian. S, "Deep Residual Learning for Image Recognition", International Conference on Computer Vision and Pattern Recognition, 2015.

[14] Olaf. R, Philipp. F, Thomas. B, "U-Net: Convolutional Networks for Biomedical Image Segmentation", International Conference on Medical Image Computing and Computer Assisted Intervention, 2015.

[15] Zongwei. Z, Md. Mahfuzur, Rahman. S, Nima. T, Jianming. L, "UNet++: A Nested U-Net Architecture for Medical Image Segmentation", International Conference on Medical Image Computing and Computer Assisted Intervention, 2018.

[16] Tsung-Yi. L, Piotr. D, Girshick. R, Kaiming. H, Hariharan. R, Belongie. S, "Feature Pyramid Networks for Object Detection", International Conference on Computer Vision and Pattern Recognition, 2017.

[17] Gao. H, Zhuang. L, Van-Der-Maaten. L, Q. Weinberger, K, "Densely Connected Convolutional Networks", International Conference on Computer Vision and Pattern Recognition, 2017.

[18] Christian. S, Wei. L, Yangqing. J, Pierre. S, Scott. R, Dragomir. A, Dumitru. E, Vincent. V, Andrew. R, "Going Deeper with Convolutions", International Conference on Computer Vision and Pattern Recognition, 2015.

[19] Jie. H, Li. S, Samuel. A, Gang. S, Enhua. W, "Squeeze-and-Excitation Networks", International Conference on Computer Vision and Pattern Recognition, 2018.

[20] Xiang. L, Wenhai. W, Xiaolin. H, Jian. Y, "Selective Kernel Networks", International Conference on Computer Vision and Pattern Recognition, 2019.

[21] Sanghyun. W, Jongchan. P, Joon-Young. L, In. So. K, "CBAM: Convolutional Block Attention Module", European Conference on Computer Vision, 2018.

[22] Andrew. G.H, Menglong. Z, Bo. C, Dmitry. K, Weijun. W, Tobias. W, Marco. A, Hartwig. A, "MobileNets: Efficient Convolutional Neural Networks for Mobile Vision Applications", International Conference on Computer Vision and Pattern Recognition, 2017.