Full-scale semantic segmentation of hyperspectral imaging based on spatial-spectral joint network

Hao Wu [1,2], Canhai Li [2], Yongchang Li [3]

[1] Liaoning Technical University,Faculty of Mapping Science and Technology ,Fu Xin,China-15140329681@163.com
[2] Land Satellite Remote Sensing Application Center,MNR,Beijing,China-licanhai@sohu.com
[3] DFH Satellite Co.,Ltd.,Beijing,China-liyongchang1231@163.com

Keywords: Hyperspectral image classification, Full-scale Skip Connections, Group convolution, Deep supervision Spectral attention mechanism

Abstract:

Hyperspectral images contain dozens or even hundreds of spectral bands, which contain rich spectral information and help distinguish different ground objects. Hyperspectral images have a wide range of applications in urban planning, environmental monitoring, and other fields. The semantic segmentation of hyperspectral images is one of the current research hotspots. The difficulty lies in the rich spectral information and strong correlation of hyperspectral images. Traditional semantic segmentation methods cannot fully extract information, which affects the accuracy of classification. This article utilizes an encoding decoding structure to simultaneously extract deep and shallow features of images. A REGCS convolution module was constructed using the idea of group convolution to extract spectral and spatial features of images. We compared the Salinas Valley dataset and MUUFL dataset with various classification algorithms. The experimental results show that compared with other classification models, the RESSU model has achieved stable and excellent results in hyperspectral image classification experiments. Among them, in the classification experiment of the Salinas Valley dataset, the accuracy of single class classification reached over 92%. In the effectiveness analysis experiment, we calculated different model parameter quantities to verify the performance of our method, and ultimately achieved good results.

1. Introduction

Hyperspectral imaging (HSI) consists of tens to hundreds of spectral bands (Du et al., 2016). The rich spectral information endows hyperspectral images with strong ground feature differentiation ability, and has a wide range of applications in environmental monitoring, urban planning, military reconnaissance, crop yield estimation, and other fields. In recent years, many methods have been applied to hyperspectral image classification tasks, including threshold based segmentation, support vector machine (SVM) (Bazi et al., 2006)、random forest (RF) (Yu et al., 2019), and polynomial logistic regression (Li et al., 2010). These methods have certain limitations, as they can only extract shallow features and lack research on extracting deep feature information. More and more scholars are using deep learning methods to study hyperspectral image classification, and convolutional neural networks play an important role in the field of image processing. The U-Net model (Ronneberger et al., 2015) has also achieved excellent results in image classification..

U-Net adopts an encoder decoder architecture. The encoding part is divided into multiple layers, and each layer performs convolution, normalization, activation, and pooling on the image before downsampling. In the decoding part, the FCN skip connection idea is improved, and the feature maps obtained from each downsampling layer in the encoder are merged with the corresponding feature maps obtained from the decoder layer for upsampling. To solve the problems that traditional semantic segmentation networks encounter when handling details and edges. Repeat the process of upsampling, fusion, and upsampling until a segmentation image with the same size as the input image is obtained.

The advantage of U-Net network's ability to quickly capture image features has made it widely used and improved in image segmentation processing.However, for the classification research of hyperspectral images, the U-Net network model still needs to improve its information processing for logarithmic tens or even hundreds of bands. Meanwhile, there is no clear and effective combination method for the deep and shallow semantics of hyperspectral images, resulting in unsatisfactory classification results in practical applications.

This article absorbs the ideas of the former's full scale skip connections and full scale deep supervision, and uses a U-shaped network structure to construct a model suitable for classifying hyperspectral images - RESSU. Its advantage lies in:

(1) effectively fusing semantic information of different layers of the image using full scale skip connections.
(2) In image downsampling, residual networks are used to first perform preliminary feature enhancement on the image and accurately extract feature information.
(3) Design a new convolutional module REGCS to simultaneously process hyperspectral images in both spectral and spatial domains.
(4) Closely combining spectral attention mechanism with newly designed convolutional module to improve the accuracy of feature information extraction.

2. RESSU Network Model

2.1 RESSU

RESSU adopts a U-shaped network structure, incorporating the ideas of full scale skip connections and deep supervision to reduce network depth. The overall network structure is shown in Figure 1.

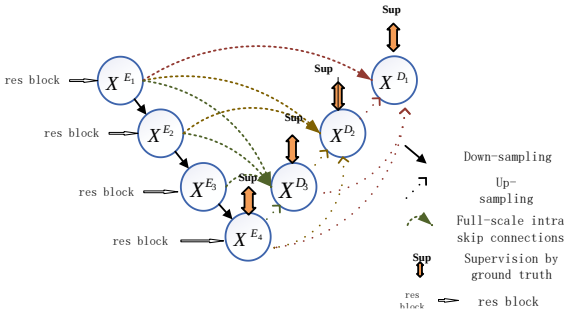


Figure 1. RESSU Structure

Traditional U-Net network models lack the ability to explore sufficient information from the full scale, making it difficult to clearly determine the position and boundaries of each classification. Each decoder layer in RESSU integrates feature maps encoded by same layer and shallow layer encoders and feature maps from deep decoders, known as full scale skip connections. They have good performance in finding the boundaries of classification objects and their positions in the image. Add res blocks during the network encoding process, use residual learning to alleviate feature degradation, and complete frequency band expansion. In the first convolution, the number of input frequency bands is m and the number of output frequency bands is n. After being activated by relu, it goes through 3x3 convolutional layers, and finally, the input image is dimensionalized using 1x1 convolution and added to the feature information of the residual module.

Reduce image resolution by maximizing pooling. In the process of network decoding, the feature fusion is achieved by using full scale skip connections to obtain the fused feature map of the encoding and decoding parts, fully utilizing multi-scale features to improve the extraction accuracy and efficiency of the network. Adopting a full scale deep supervised design, learning images from comprehensive aggregated feature maps.

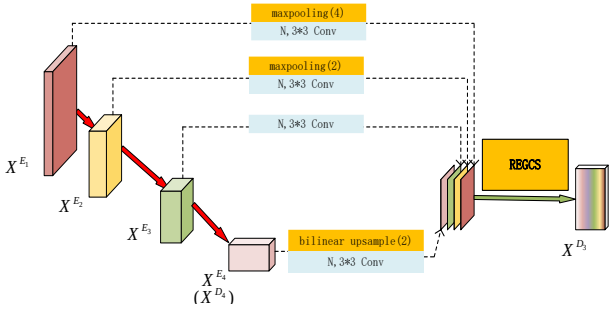


Figure 2. RESSU full scale jump connection structure

The RESSU full scale skip connection realizes the connection between different parts of the encoder and decoder, as well as the inherent connection between the decoder subnetworks. Taking the node X^{D_3} in Figure 2 as an example, its information comes from two aspects: one is the encoder that is shallower (including the same level), and the other is the decoder that is deeper than it. N is the number of bands input into the image. Each decoder layer in RESSU contains feature maps from different scales of the encoder and decoder, which capture fine-grained details and coarse-grained semantics at the full scale. The formula is as follows:

$$X_{De}^{i}=\begin{cases} X_{En}^{i} & i=\mathrm{N} \\ R\left(\left[\underbrace{C(D(X_{En}^{k}))_{k=1}^{i-1},C(X_{De}^{k})}_{scalaes:1^{th}\sim i^{th}},\underbrace{C(u(X_{De}^{k}))_{k=i+1}^{N}}_{scalaes:1^{th}\sim i^{th}}\right]\right) & ,i=1,\cdots,\ \mathrm{N}-1 \end{cases} \quad (1)$$

Function C represents convolution operation, function R represents REGCS module convolution operation, functions D and U-represent upsampling and downsampling operations, respectively, [ ] represents channel dimension concatenation fusion.

At the same time, RESSU adopts Deep Supervision as shown in Figure 3, and the feature maps of each decoding layer are processed by the REGCS convolution module for band fusion. We perform 3 x 3 convolution on the generated feature maps , and bilinear upsampling is used to restore the resolution of the feature map to the input image level. After sigmoid processing, it enters the loss function calculation and obtains the loss value for backpropagation, which is used to optimize model parameters.

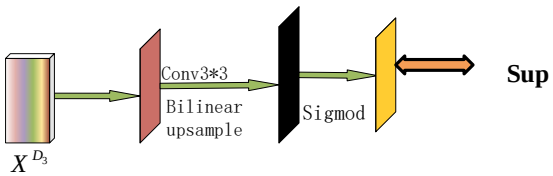


Figure 3. RESSU deep supervision structure

In terms of selecting the loss function, this article uses the softmax loss function based on the classification features of hyperspectral images to help the model generate predicted values that are close to the true value direction, thus achieving the purpose of learning. The formula is as follows:

$$l(y,z)=-\sum_{k=0}^{C}y_c\log(f(z_c)) \quad (2)$$

$$f(z_c)=e^{z_c}/(\sum_j e^{z_j}) \quad (3)$$

Where: y_c is the true value of the sample, which is the label value, $f(z_c)$ is the predicted value output by the softmax function, $l(y,z)$ is the loss value obtained.

2.2 REGCS convolutional module

The spectral band range of hyperspectral images is wide, extending from visible light to shortwave infrared, and even mid infrared, with hundreds of bands forming approximately continuous spectral curves. However, the bandwidth of a single band is relatively narrow, usually around 5-20nm. Therefore, when extracting features from images, it is necessary to consider all bands. This article designs a convolutional module REGCS,

which adopts a twice grouped convolution form to divide a large number of bands into different groups for analysis and feature extraction. Compared with 3D convolution (Yu et al., 2020), it reduces computational complexity while achieving ideal results. The specific process of the REGCS module is shown in Figure 4.

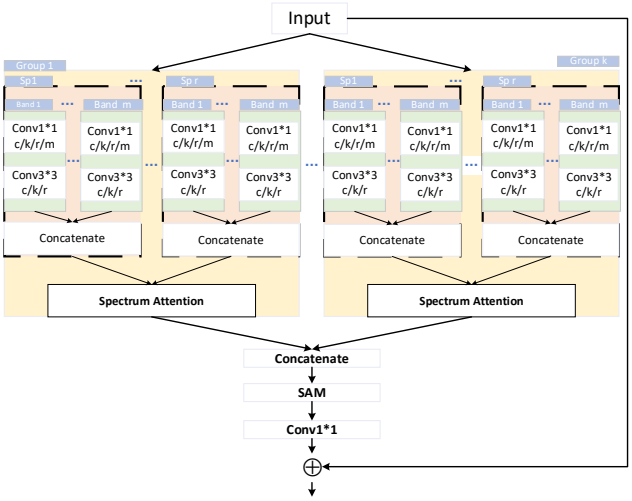


Figure 4. REGCS module

As shown in the figure, the REGCS module located in the decoding part of the decoder includes the fusion of deep and shallow image feature information, and adds spectral and spatial attention mechanisms to increase the weight of important feature information, thereby obtaining new feature maps. Firstly, input the encoded hyperspectral image with a size of $h \times w \times c$ (c is the number of bands) into the REGCS convolution module. Firstly, according to the number of bands, they are divided into k groups. Different datasets have different k values, and in the experiment, the specific value of k is specifically tested. Each group contains four different feature information layers, therefore, r=4. In each Sp group, it contains m bands, representing the number of bands in the same feature level, where m x k represents the total number of bands in a single level c. In each group, the deep and shallow feature information of the same band is convolved separately, and cascaded to obtain the complete feature description of the image in that band group. Perform spectral attention calculation on the completed bands within the group, extract spectral weight information of similar images, and perform spatial attention calculation on all grouped bands after stitching to obtain image spatial weight information. Finally, multiply the weight parameters containing all information with the input image to obtain the final feature map. The spectral attention mechanism and spatial attention mechanism will be introduced in the next section.

2.3 Spectral and spatial attention mechanisms

The attention mechanism can enable the model to focus on information that is more critical to the current task among numerous input information, reduce attention to other information, solve the problem of information overload, improve the efficiency and accuracy of task processing (Woo et al., 2018), and improve classification accuracy. This article proposes a new spectral attention mechanism (SA) suitable for hyperspectral image classification, and utilizes the existing spatial attention mechanism (SPA) (Fang et al., 2021) to closely link it with the REGCS convolutio module. It can allocate computing resources to more important tasks in feature extraction of hyperspectral image datasets.

First, perform spectral attention model calculations on the feature maps, and then connect the new feature maps obtained and input them into the spatial attention model for calculation. The calculation formula is:

$$F^{''} = M_s(F_{con}^{'}) \otimes F_{con}^{'} \quad (4)$$

$$F^{'} = M_c(F) \otimes F \quad (5)$$

Where $F^{''}$ is the spatial spectral joint feature refined by the convolutional block attention mechanism to reassign weights, $F_{con}^{'}$ is the sum of spatial spectral joint features of different groups after being processed by the spectral attention module, $F^{'}$ is the feature calculated by the spectral attention module for hyperspectral images, and F is the feature extracted from hyperspectral images; $M_c(\cdot)$ and $M_s(\cdot)$ are spectral attention submodules and spatial attention submodules, respectively; $\otimes$ represents multiplication operation. Below are the spectral attention module and spatial attention module, respectively. The introduction of spectral attention mechanism is as follows, and the structure is shown in Figure 5:

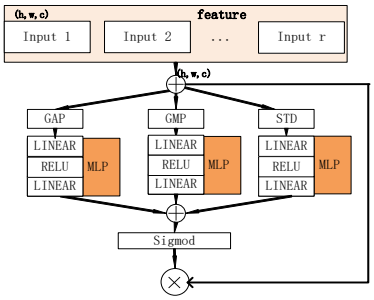


Figure 5. Spectral attention mechanism structure

In order to improve the computing efficiency of spectral attention, the feature maps were respectively subjected to global maximum pooling operations based on width and height, global standard

differential pooling operations and global average pooling operations to obtain the maximum pooling feature map F_{GMP}^{C}, global standard differential pooling feature map F_{STD}^{C} and average pooling feature map F_{GAP}^{C} (C is the number of channels). The three parties obtain 3 spectral weights through a multi-layer perceptron with shared parameters of two layers, and sum the 3 output spectral weights to get M_C, and normalize them by sigmod activation function to get $M_C \in F$. Spectral attention is calculated as follows:

$$\begin{aligned} M_c(F) &= \sigma(MLP(GAP(F)) + \\ &MLP(GMP(F)) + MLP(STD(F))) \\ &= \sigma(W_1(W_0(F_{GAP}^{C}))) + \sigma(W_1(W_0(F_{GMP}^{C}))) \\ &+\sigma(W_1(W_0(F_{STD}^{C}))) \end{aligned} \tag{6}$$

In the formula: σ Represents the sigmod activation function; $W_0 \in \mathbf{R}^{C/r \times C}$ and $W_1 \in \mathbf{R}^{C \times C/r}$ represent the weights r of the first and second hidden layers in a multi-layer perceptron, respectively, representing the feature compression rate. After obtaining spectral weights and their results through spectral attention mechanism in different bands, they are added in the concatenate module and then subjected to spatial attention calculation.

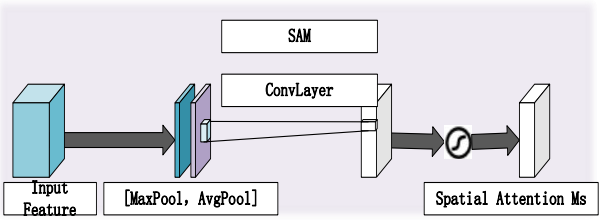


Figure 6. Structure of spatial attention mechanis

The spatial attention submodule focuses on the spatial position relationship of adjacent pixels in hyperspectral images, generating different weights related to spatial features. As shown in Figure 6. Perform a channel based global max pooling operation and a global average pooling operation on the feature map, and connect these two results based on the channel. Then, a convolution operation is performed to compress the dimensions and reduce them to one channel. Afterwards, the weight $M_s \in \mathbf{R}^{W \times H \times 1}$, which is $M_s(F^{'})$ (where s represents the result of attention processing on the spatial dimension of the hyperspectral image), is redistributed through convolution operations and activation functions. The formula for calculating spatial attention is:

$$\begin{aligned} \mathrm{M_s(F)} &= \sigma(f^{7\times7}([AvgPool(F); \\ &MaxPool(F)])) = \sigma(f^{7\times7}([F_{avg}^{C}, F_{\max}^{C}]) \end{aligned} \tag{7}$$

In the formula: σ Indicates the activation function; $f^{7\times7}$ indicates that using convolution kernels with a size of 7×7 .during the convolution process. F_{avg}^{C} represents the feature map after average pooling, and $F_{\max}^{C}$ represents the feature map after maximum pooling.This method adds a spatial attention model to the end of the RESSU model, which can increase the weight of spatial key feature information on the image while reducing the overall computational complexity of the model

2.4 The overall architecture of this article

The overall architecture of this article is shown in Figure 7:

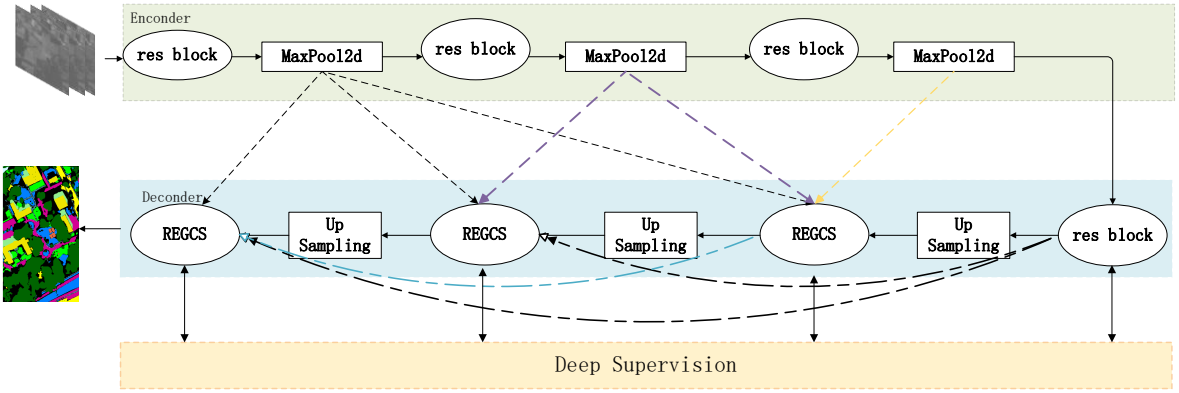


Figure 7. RESSU Overall Network Structure

This article uses a four layer U-shaped Encoder Decoder encoding and decoding structure. The image to be classified is input into the network. During the Encoder encoding stage, the image features are roughly extracted through res blocks and downsampled. In the Decoder decoding stage, different levels of feature information are concatenated and stacked, and the deep and shallow semantic feature information is fully communicated through the REGCS convolution module proposed in this article. After a completely symmetrical upsampling operation with the encoding stage, the image is restored in size. After each decoding layer, a full scale deep supervision mechanism is used to upsample the current feature image and compare it with the label.

Backpropagation is used to optimize model parameters and improve the accuracy of hyperspectral image classification.

3. Experimental Process and Analysis

3.1 Dataset introduction

The experiment used two hyperspectral datasets to test the classification performance of our model, and the specific information is shown in Table 1:

	Hyperspectral Datasets	
	Salinas Valley	MUUFL
Collection time	1992	2010
Locality	California,USA	Mississippi,USA
collecting device	AVIRIS	CASI-1500
Data size/pixel	512×217	325×220
Spatial resolution/m	3.7×3.7	0.54×1.0
Spectral resolution/nm	5	10
Band number	204	64
Number of categories	16	11
Sample size	54129	53687

Table 1. Sample types and quantity of hyperspectral datasets

3.2 Experimental environment

In order to improve the generalization ability of classification, the dataset is randomly divided proportionally and the partitioned data is randomly selected as the input for the image. Epoch is set to 100, the model optimizer uses Adam optimizer, and the initial learning rate of the network is set to 0.001. To ensure experimental fairness, all methods do not preprocess the training set and are implemented on the same graphics workstation. The detailed configuration parameters are shown in Table 2.

experimental environment	configuration parameter
operating system	Windows
GPU	NVIDIA GeForce RTX3090
Memory	32G
Application Framework	Pytorch 1.8.0
Programming Language	Python

Table 2. Experimental environment configuration parameters

3.3 The impact of the number of packets on the network

Compared to conventional convolution, group convolution can improve the learning efficiency and performance of the model, affecting the classification performance of RESSU networks. This article conducted experiments on the impact of the number of groups on the classification accuracy of RESSU on those datasets. The optimal input size of the image in section 2.3 was selected as the input size. In addition, in the setting of network parameters, the number of channels in group convolution must be divided by the number of groups. Therefore, in Tables 3-4, Observe the impact of grouping number on model performance in three datasets using g=[2,3,4,6,12]/g=[2,4,8,16,32].

According to experimental data tables 3 and 4, classification experiments were conducted on the MUUFL dataset. The number of groups in the network convolutional layer was 4, while in the Salinas Valley dataset classification experiment, the number of groups in the network convolutional layer was 3.

Number of groups	2	3	4	6	12
OA (%)	94.22	**95.44**	94.31	93.82	93.46
AA (%)	92.68	**93.39**	92.91	92.73	92.30
Kappa*100	91.15	**93.01**	91.74	91.39	91.03

Table 3. Salinas Valley grouping accuracy

Number of groups	2	4	8	16	32
OA (%)	95.14	**95.80**	95.45	95.05	94.97
AA (%)	91.40	**92.65**	92.17	91.94	91.83
Kappa*100	92.55	**92.83**	92.56	92.44	92.13

Table 4. MUUFL grouping accuracy

3.4 Comparative experimental analysis

To verify the classification performance of the proposed method, the RESSU method was compared with various hyperspectral image classification methods, including 3D-Unet(É i ç ek et al., 2016),pResNet(Paoletti et al., 2019),VIT(Dong et al., 2022), SSGCN(Qin et al., 2017),WFCG(Dong et al., 2022), and CBAM-Res-HybridSN(Yang et al., 2023). Compare the accuracy based on commonly used accuracy indicators in hyperspectral analysis, including overall accuracy (OA), average accuracy (AA), Kappa coefficient, and accuracy for each category, i.e. the number of correctly judged samples in that category/the number of samples in that category.

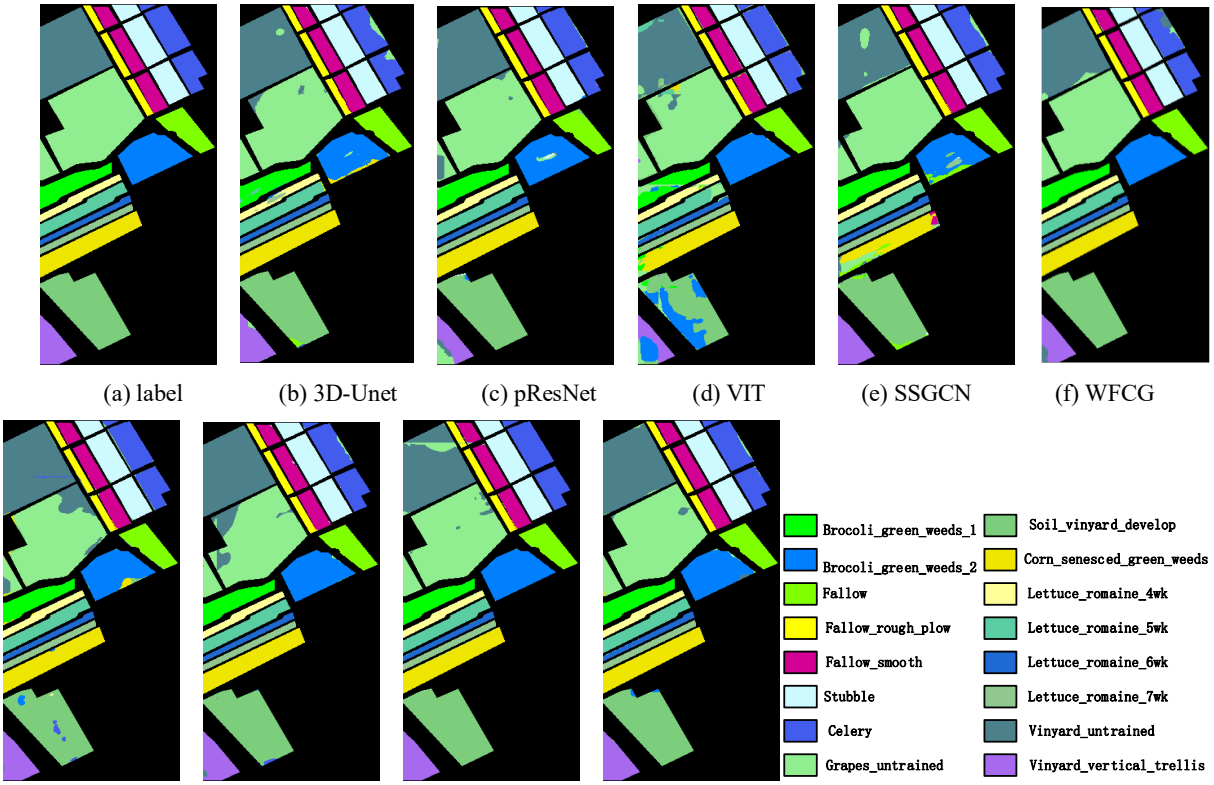
The test results of the Salinas Valley dataset are shown in Figure 8 and Table 5. The Salinas dataset itself has a relatively balanced sample distribution, and there are significant differences in spectral and spatial dimension information between different categories, which is beneficial for classification. Most methods, such as WFCG, pResNet, SSGCN, SSRN, CBAM-Res-HybridSN, etc., have achieved good results in this dataset experiment. Compared with other algorithms, our method achieved the highest accuracy results on OA/AA/Kappa during the testing process. Compared with 3D-Unet, pResNet, VIT, SSGCN, WFCG, SSRN, MS3D-CNN-A, and CBAM-Res-HybridSN, the OA values of RESSU on the Salinas Valley dataset were 2.66%, 3.93%, 8.91%, 4.17%, 2.51%, 1.18%, 6.17%, and 1.13%, respectively. The VIT method has achieved poor results in classification experiments with limited sample data. Although the MS3D-CNN-A method can extract depth and multi-scale features, it has not achieved ideal results due to the same reasons. The overall classification accuracy of pResNet, SSRN, and CBAM-Res-HybridSN methods with residual structures exceeds 91%, achieving good results in small sample datasets. However, their accuracy in specific classifications is not stable.

For example, in Vinyard_vertical_trellis, pResNet only has the 69.39% about Accuracy. In Grapes_untrained, SSRN is only 83.57% about Accuracy. we adopts spectral band grouping and adopts a residual structure of res blocks in the encoding part, effectively alleviating the phenomenon of gradient vanishing. At the same time, using the full scale skip method for feature extraction significantly improves classification accuracy. Stable and effective results were achieved in various categories, with a classification accuracy of over 92%.

Finally, experiments were conducted on MUFFL datasets with higher spatial resolution and lower spectral dimensions. The results are shown in Figure 9 and Table 6. Except for the 3D-Unet model, the OA values of all other models are above 90, achieving ideal results. However, in the classification of water and architectural shadows, their accuracy is often poor. In terms of

This contribution has been peer-reviewed. The double-blind peer-review was conducted on the basis of the full paper.
https://doi.org/10.5194/isprs-annals-X-1-2024-267-2024 | © Author(s) 2024. CC BY 4.0 License.

OA value, RESSU leads CBAM-Res-HybridSN with the second highest classification performance by 1.13% in Overall Accuracy, which is 7.96% higher than 3D-Unet, which is also a U-Net network model framework.



(a) label (b) 3D-Unet (c) pResNet (d) VIT (e) SSGCN (f) WFCG

(g) SSRN (h) MS3D-CNN-A (i) CBAM-Res-HybridSN (j) RESSU （k）Sample category and color

Figure 8 Classification results of 10% training samples on Salinas Valley images using labels and different methods

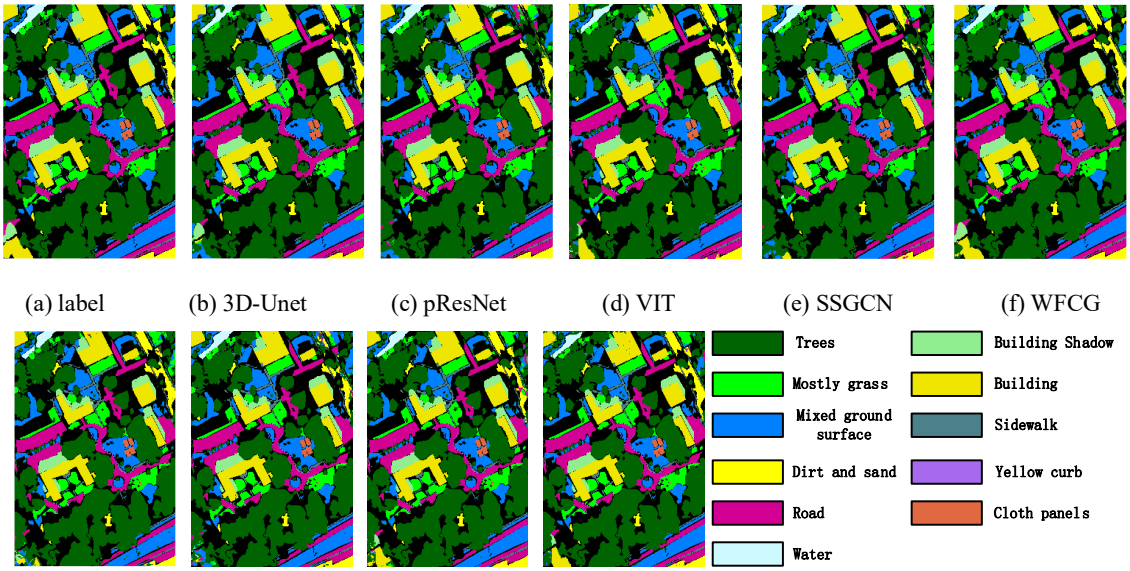
((a)label；(b)3D-Unet；(c) pResNet；(d)VIT；(e) SSGCN；(f) WFCG；

(g) SSRN；(h) MS3D-CNN-A；(i) CBAM-Res-HybridSN；(j)RESSU； （k）Sample category and color)

	3D-Unet	pResNet	VIT	SSGCN	WFCG	SSRN	MS3D-CNN-A	CBAM-Res-HybridSN	RESSU
Brocoli_green_weeds_1	94.41	97.21	92.65	**97.80**	97.63	97.25	97.75	96.92	97.21
Brocoli_green_weeds_2	90.05	90.37	95.62	82.69	96.12	92.39	96.31	**96.80**	92.80
Fallow	**98.03**	97.63	96.51	95.42	97.23	95.06	97.38	96.94	97.12
Fallow_rough_plow	94.57	97.03	95.42	**97.61**	92.76	94.34	96.97	92.76	96.92
Fallow_smooth	97.24	98.03	96.2	97.51	94.71	97.16	95.15	96.51	**98.17**
Stubble	97.67	96.38	96.41	97.09	**97.92**	97.01	96.92	97.11	97.21
Celery	89.39	92.24	91.57	91.97	**97.41**	94.12	93.33	95.33	92.30
Grapes_untrained	94.69	91.23	91.87	**95.03**	78.88	83.57	81.76	93.13	93.33
Soil_vinyard_develop	95.58	95.32	61.24	92.56	94.71	93.52	94.37	**96.99**	95.39
Corn_senesced_green_weeds	97.77	97.78	91.14	80.69	97.57	97.51	98.16	97.01	**98.63**
Lettuce_romaine_4wk	81.05	**98.24**	92.04	97.21	98.02	94.87	96.97	97.06	97.12
Lettuce_romaine_5wk	96.18	97.24	83.05	96.51	**97.58**	95.82	96.92	95.85	96.72
Lettuce_romaine_6wk	96.95	98.15	92.21	97.27	97.45	96.47	**98.37**	96.58	95.17
Lettuce_romaine_7wk	97.14	97.36	90.52	92.05	97.07	96.57	**98.25**	96.10	97.31
Vinyard_untrained	95.89	95.57	90.98	86.78	95.91	95.78	97.62	89.05	**96.05**
Vinyard_vertical_trellis	94.07	69.39	50.13	**97.37**	92.46	96.72	86.85	97.16	97.03
Overall Accuracy	92.78	91.51	86.53	91.27	92.93	94.26	89.27	94.31	**95.44**
Average Accuracy	90.04	89.16	83.36	85.94	89.75	93.17	86.36	91.58	**93.39**
Kappa*100	89.07	88.88	82.96	89.46	89.98	92.56	87.71	92.32	**93.01**

Table 5 Classification accuracy evaluation of 5% training samples on Salinas Valley images using different methods (%)

This contribution has been peer-reviewed. The double-blind peer-review was conducted on the basis of the full paper.
https://doi.org/10.5194/isprs-annals-X-1-2024-267-2024 | © Author(s) 2024. CC BY 4.0 License.



(a) label (b) 3D-Unet (c) pResNet (d) VIT (e) SSGCN (f) WFCG

(g) SSRN (h) MS3D-CNN-A (i) CBAM-Res-HybridSN (j) RESSU （k） Sample category and color

Figure 9. Classification results of 10% training samples on MUUFL images using labels and different methods

((a)label；(b)3D-Unet；(c) pResNet；(d)VIT；(e) SSGCN；(f) WFCG；

(g) SSRN；(h) MS3D-CNN-A；(i) CBAM-Res-HybridSN；(j)RESSU； （k） Sample category and color)

	3D-Unet	pResNet	VIT	SSGCN	WFCG	SSRN	MS3D-CNN-A	CBAM-Res-HybridSN	RESSU
Trees	83.20	95.92	96.68	95.99	95.28	96.19	96.12	96.61	**97.03**
Mostly grass	90.69	89.59	91.06	**92.37**	89.80	90.12	87.62	89.77	91.51
Mixed ground surface	84.83	88.90	90.05	90.72	89.71	89.38	**94.35**	90.28	92.21
Dirt and sand	90.40	64.73	86.08	66.72	**93.30**	82.68	70.99	82.95	90.68
Road	94.41	94.02	94.98	93.35	94.78	93.66	95.08	94.24	**96.23**
Water	81.62	94.57	75.82	89.85	**97.53**	91.73	95.57	91.27	87.33
Building Shadow	95.37	83.72	87.98	84.39	**95.49**	85.65	85.79	88.74	88.79
Building	86.56	85.71	88.24	88.17	93.37	88.68	87.35	**93.46**	93.05
Sidewalk	89.97	89.42	93.87	86.83	92.29	92.65	88.24	92.07	**93.95**
Yellow curb	71.68	94.63	96.27	90.80	88.07	91.35	90.81	**97.36**	95.72
Cloth panels	92.68	96.88	96.43	97.26	95.56	95.65	97.88	98.26	**98.63**
Overall Accuracy	87.84	92.03	94.33	92.17	94.14	93.78	94.12	94.67	**95.80**
Average Accuracy	85.40	88.05	90.59	87.26	93.34	90.75	89.91	92.32	**92.65**
Kappa*100	85.81	89.69	91.63	88.75	93.37	91.02	91.86	92.14	**92.83**

Table 6. Classification accuracy evaluation of 5% training samples on MUUFL images using different methods (%)

4. Validity Analysis

In order to evaluate the performance of the network model in this article, the number of parameters and FLOP calculations were performed on the MUUFL dataset, mainly compared with 3D-Unet.

4.1 Parameter quantity and FLOPs analysis

The number of parameters (Params) and floating-point operations per second (FLOPs) are two crucial indicators for assessing the complexity of deep learning networks. In this study, we selected the MUUFL dataset for classification experiments and presented the parameter results in Table 7, which include non-attention model (RESTU), non-spectral attention model (UnSA-RESS), non-space attention model (UnSPA-RESS), our proposed RESSU model, and 3D-Unet. Among them, the first four models achieved the best results in terms of group numbers. The settings of 3D-Unet were consistent with those used in our experiment. Our proposed model and its different combinations have significantly fewer parameters and FLOPs than 3D-Unet, indicating a clear advantage in overall classification accuracy..

Model	Parameter quantity /M	FLOPs/G	OA/%
RESTU	4.89	272.93	93.51
UnSA-RESS	5.02	281.62	94.37
UnSPA-RESS	5.07	283.02	94.75
RESSU	5.13	287.38	**95.80**

3D-Unet	14.68	822.36	87.84

Table 7. The parameter quantities, FLOPs, and OA of the 3D Unet model with different combinations in this mode.

5. Conclusion

We propose a new semantic segmentation model, RESSU, which can make full use of the spatial and spectral features of hyperspectral images to solve the problem of poor performance of classification models with limited training samples. We conducted experiments using two hyperspectral datasets with different characteristics and compared them with other methods, and the classification results were superior to the other methods. At the same time, in the validity analysis, different combinations of parameters, FLOP and OA are compared with the 3D-Unet model, and the parameters and flops of this model and its different combinations are much smaller than the 3D Unet model. Compared with other classification methods, the RESSU model proposed in this paper has significant advantages in learning small hyperspectral data sets, improving classifier performance and classification accuracy. However, there are also shortcomings, and future studies will continue to improve the network structure so that the model has more outstanding classification performance in larger hyperspectral data sets.

Acknowledgements

Fund Project: The Pre researvh of "14th Five-Year Plan" technology in civil aerospace.

Project Number:D010206.

References

Bazi Y , Melgani F.2006.Toward an optimal SVM classification system for hyperspectral remote sensing images. IEEE Transactions on Geoscience and Remote Sensing, 44(11).

Çiçek, Ö.; Abdulkadir, A.; Lienkamp, S.S.; Brox, T, Ronneberger, O.2016.3D U-Net: Learning Dense Volumetric Segmentation from Sparse Annotation. Medical Image Computing and Computer-Assisted Intervention–MICCAI 2016: 19th International Conference, Proceedings,Part II 19.: 424-432. doi.org/10.1007/978-3-319-46723-8_49.

Dong X Y, Bao J, Chen D D, Zhang W M, Yu N H ,Yuan L, Chen D and Guo B N.2022.CSWin Transformer: A General Vision Transformer Backbone with Cross-Shaped Windows. 2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition ,12114-12124. doi: 10.1109/CVPR52688.2022.01181.

DONG Y N, LIU Q W, DU B and Zhang L P. 2022.Weighted feature fusion of convolutional neural network and graph attention network for hyperspectral image classification. IEEE Transactions on Image Processing, 31: 1559-1572.

Du P J, Xia J S, Xue Z H, Tan K, Su H J and Bao R. 2016. Review of hyperspectral remote sensing image classification. Journal of Remote Sensing, 20(2):236–256.

Fang Y, Huang H, Yang W, Xu X, Jiang W and Lai X. 2022. Nonlocal convolutional block attention module VNet for gliomas automatic segmentation. International Journal of Imaging Systems and Technology, 32(2): 528-543.

Li J, Bioucas-dias J M amd Plaza A.2010. Semisupervised Hyperspectral Image Segmentation Using Multinomial Logistic Regression With Active Learning. IEEE Transactions on Geoscience and Remote Sensing,48 (11).

Paoletti M E, Haut J M, Fernandez-Beltran R, Plaza J, Plaza A J and Pla F. 2018. Deep pyramidal residual networks for spectral-spatial hyperspectral image classification. IEEE Transactions on Geoscience and Remote Sensing, 57(2): 740-754.

Qin A, Shang Z, Tian J, Wang Y L, Zhang T P and Tang Y Y.2018.Spectral–Spatial Graph Convolutional Networks for Semisupervised Hyperspectral Image Classification. IEEE Geoscience and Remote Sensing Letters,16(2):241-245.

Ronneberger O, Fischer P and Brox T.2015. U-Net: Convolutional Networks for Biomedical Image Segmentation. Lecture Notes in Computer Science. Medical Image Computing and Computer Assisted Intervention MICCAI 2015: 18th International Conference. Proceedings, Part III.234-241.doi:10.1007/978-3-319-24574-4_28.

Roy S K, Krishna G, Dubey S R and Chauhuri BB.2020. HybridSN: Exploring 3-D–2-D CNN Feature Hierarchy for Hyperspectral Image Classification. IEEE Geoscience and Remote Sensing Letters,17(2):277-281.

Woo S, Park J, Lee J Y and Kweon I S.2018.CBAM: Convolutional Block Attention Module. Proceedings of the European conference on computer vision.3-19.doi:10.1007/978-3-030-01234-2_1.

Wu Qinggang, Liu Zhongchi and He Mengkun.2023. Fusion of MS3D-CNN and attention mechanism for hyperspectral image classification. Journal of Chongqing University of Technology(Natural Science),37(2):173 - 182.

Yang Z W, Zhang H B, Du W B and Pan Y S.2023. A Study of Small Sample Hyperspectral Image Classification Based on CBAM-Res-HybridSN. Spacecraft Recovery & Remote Sensing, 44(3): 85-96.

Yu C, Han R, Song M, Liu C Y and Chang C L.2020. A Simplified 2D-3D CNN Architecture for Hyperspectral Image Classification Based on Spatial–Spectral Fusion. IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing,13: 2485-2501.

Yu X Y,Zhao G X,Chang C Y,Yuan X J,Wang Z R.2019.Random Forest Classifierin Remote Sensing Information Extraction: A Review of Applicationsand Future Developmen. Remote Sensing Information,34(2):8-14

Zhong Z , Li J , Luo Z and Chapman M.2017.Spectral-Spatial Residual Network for Hyperspectral Image Classification: A 3-D Deep Learning Framework.IEEE Transactions on Geoscience and Remote Sensing, 56(2):847-858.