Deep Learning Algorithm Based Remote Sensing Image Classification Research

Hengjie Zhu[#]
Chang'an University
Xi'an, China

Xuyao Wang[#]
Nankai University
Tianjin, China

Renxing Chen[#,*]
Jinan University
Zhuhai, China
crx672501109@gmail.com
[#]These authors contributed equally.

***Abstract*—In the field of remote sensing image recognition, compared with deep learning related image recognition algorithms, traditional image recognition algorithms have limitations in recognizing different types of land, so a large amount of deep learning is used for remote sensing image recognition. However, different models have different effects on different tasks, and selecting an excellent deep learning model is a matter of concern. This paper takes WHDLD dataset as the data source, DeepLabV3Plus, HRNet-OCR, ACGAN and other deep learning models as the experimental objects to study the performance ability of different deep learning models on WHDLD dataset. The experimental results show that ACGAN is the best among the experimental models. accuracy, precision, recall, F1-score are 0.83, 0.95, 0.92 and 0.90 respectively.This study can provide some reference values for remote sensing image land classification.**

Keywords—Deep learning, Image classification, Image segmentation, Remote sensing images

I. INTRODUCTION

With the development and maturity of remote sensing technology, remote sensing technology is more and more widely used in various fields, and the dimensionality of remote sensing image features is increasing, and the information contained in them is becoming richer and richer. Remote sensing image classification is an important direction in remote sensing image processing. Using image classification algorithms to classify remote sensing images according to their characteristics makes it easier for people to filter and analyze applications when facing a large number of remote sensing images. Unlike the classification tasks of natural images, remote sensing images are more difficult to distinguish their features due to their diverse scales and complex backgrounds and the presence of small target sizes [1-2]. With more deep learning models proposed, the models represented by convolutional neural networks are widely used in the field of remote sensing image classification with the advantages of automatic feature extraction and good robustness.

Traditional methods for semantic segmentation of remote sensing images mainly use image-based algorithms, which include image segmentation methods based on thresholding, region extraction, or edge detection. For example, the remote sensing image shadow detection method implemented by Yu et al [3] by setting segmentation thresholds on the grey-scale histogram of remote sensing images. However, these traditional segmentation methods can generally only achieve the segmentation of foreground and background, and their classification accuracy is also unsatisfactory, which cannot meet the needs of segmenting a variety of terrain and landforms in remote sensing images.

In recent years, the use of machine learning for the semantic segmentation of remote-sensing images has shown good results. For example, Liu et al [4] based on remote sensing impact features and maximum likelihood method, Tari G et al [5] used the ISO clustering method, and Thakur R et al [6] used support vector machine (SVM), all of which performed well in achieving semantic segmentation of remote sensing images. However, these methods are more dependent on spectral features, insufficient use of spatial features, and are not suitable for remote sensing images with low spectral analysis rates.

With the advancement of deep learning, an increasing number of researchers have used it to remote sensing picture classification. long et al first proposed the full convolutional neural network (FCN), which is a classical semantic segmentation network by discarding fully connected layers to improve segmentation efficiency while reducing computational complexity. For example, Wang [7] and Zhao[8] et al. used multi-scale images to construct multi-scale sample pyramids, which fully exploited the spatial information in remote sensing images. Based on the FCN, Cambridge modified the VGG-16 network to obtain the semantic segmentation network SegNet[9], which stores the position of the largest value during the maximum pooling operation in the Encoder section and performs non-linear upsampling by the appropriate pooling index at the decoder. Subsequently, Ronneberger O et al [10] proposed UNet, which features multi-scale feature fusion. Subsequently, PSPNet [11] and DeepLabV3+ [12] networks were proposed to perform the segmentation task better, while the discontinuity and overfitting problems of the segmentation results are also in need of a solution.

In this paper, DeepLabV3+[12], HRNet-OCR [13], ACGAN [14], and other models are used for training on the WHDLD dataset, and the segmentation effects of these models are evaluated.

II. DATASET

WHDLD is used for remote sensing image segmentation [15]. Its data imaging bands include R, G, and B bands, and the data coverage includes 6 types of landforms: buildings, roads, sidewalks, vegetation, bare ground, and water. The dataset contains 4940 remote sensing images and the corresponding feature classification tagging samples, the image size is 256x256 pixels, the images are stored in jpg, and the tagging data format is a single-channel png image, whose details are shown in Table I.

TABLE I. WHDLD DATASET

Label	Classification	Number
1	building	3722
2	road	3162
3	pavement	3881
4	vegetation	4631
5	bare soil	3539
6	water	3886

A partial example is shown in Fig.1.

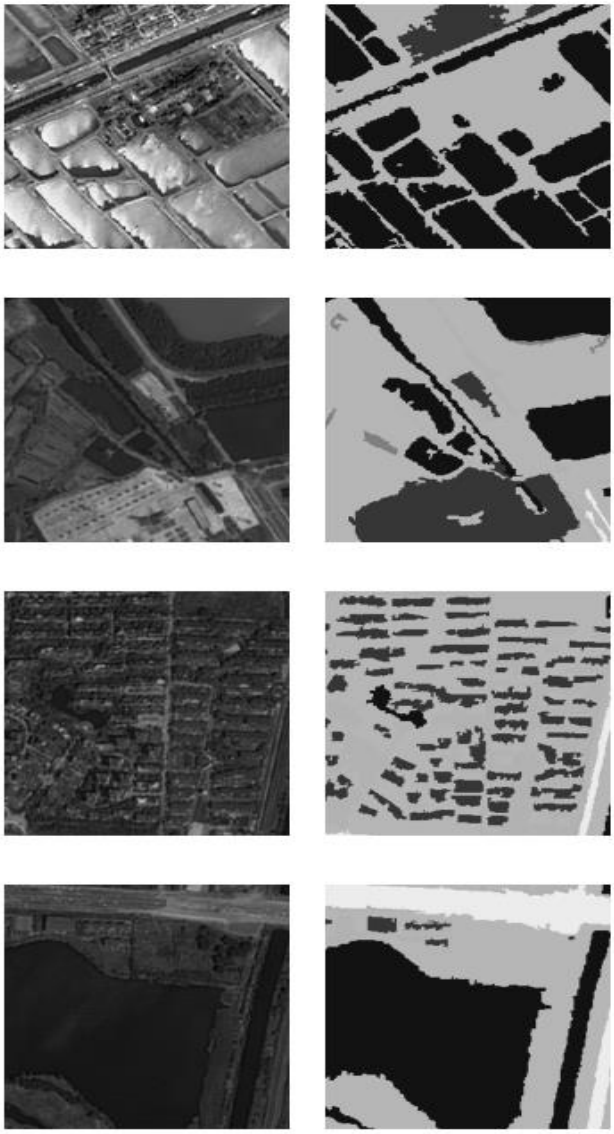

Fig. 1. Example of WHDLD Dataset

III. DEEPLABV3PLUS

DeepLabV3plus is a semantic segmentation model that presents a novel Encoder-Decoder architecture based on DeepLabv3 and a basic yet efficient decoder module. The model enables for resolution adjustment of the derived encoder features through Atrous convolution, allowing for a trade-off between accuracy and runtime. Furthermore, for segmentation, the network employs the Xception, ResNet, and MobileNet models, as well as Depthwise Separable Convolution on the ASPP and decoder modules, resulting in a quicker and more powerful Encoder-Decoder network. Fig. 2 depicts the total network structure.

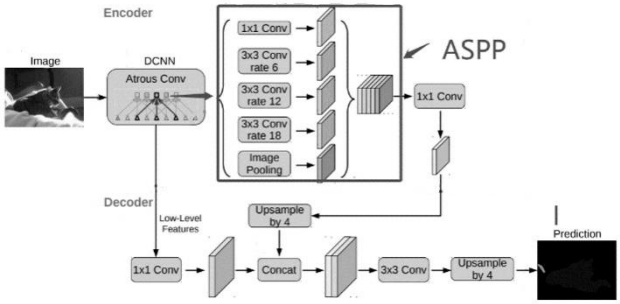

Fig. 2. The structure of DeepLabV3plus

The spatial pyramid pooling module gathers rich contextual information for semantic segmentation by combining feature maps of various resolutions, while the encoder-decoder structure yields clear object boundaries. Typically the encoder reduces the feature map size and captures higher-level semantic information, which the decoder is used to recover spatial information. The final feature map contains rich semantic information, but because to pooling or convolution processes inside the bone, specific information about object boundaries is lacking. To alleviate this problem, DeepLabV3plus introduces Atrous convolution, a relatively special convolution operation.

Atrous convolution explicitly controls the resolution of the feature maps computed in deep convolutional neural networks and adjusts the field-of-view of the convolutional kernel to capture multi-scale information.

Depthwise separable convolution, which is the same as Depthwise convolution and Pointwise convolution, is an operation that minimizes the computation and amount of parameters while keeping equal (or slightly higher) performance than normal convolution.

In the design of Backbone, DeepLabV3plus chose three feature extraction networks, ResNet [16], MobileNet [17], and Xception [18], respectively.

The improved Xception with DeepLabV3plus has more layers [19], in addition, all max pooling operations have been changed to Deepwise separable convolution and batch normalization and ReLU activation functions have been added after the 3x3 DW convolution operation.

The ResNet101 and ResNet51 used by DeepLabV3plus are the same as the original ResNet101 and ResNet51 in terms of the number of network layers, except that the later convolution operation uses dilation to specify the expansion rate [20].

The backbone network of MobileNet uses an inverted residual structure, which switches the order of downsampling and upsconcerninghe residual structure, and replaces the 3x3 convolution with a 3x3 DW convolution [21]. In addition, the activation function in the network uses ReLU.

IV. HRNET-OCR

HRNet-OCR uses HRNet [22] as the Backbone and then the final result is obtained by the OCR method, which has the advantage that it transforms the pixel classification problem into the object region classification problem and solves the problems of low resolution missing contextual information/border errors. The OCR method is more efficient and accurate than other semantic segmentation methods. Because the OCR method solves the object region classification problem instead of the pixel classification problem, i.e., the OCR method can effectively and explicitly enhance the object information. In terms of performance and complexity, OCRNet is also superior, and the "HRNet + OCR + SegFwon the first place in 2020ECCV Cityscapes.

A. *HRNet*

A characteristic of HRNet is that it maintains a high-resolution representation throughout the learning process by connecting multiple resolutions in parallel and repeatedly exchanging information in parallel multi-resolution sub-networks.HRNet achieves multi-scale fusion by interacting information across resolutions to achieve high resolution for semantically rich feature output. The network structure of HRNet is shown in Fig.3.Bello used it for high resolution remote sensing imagery [23].

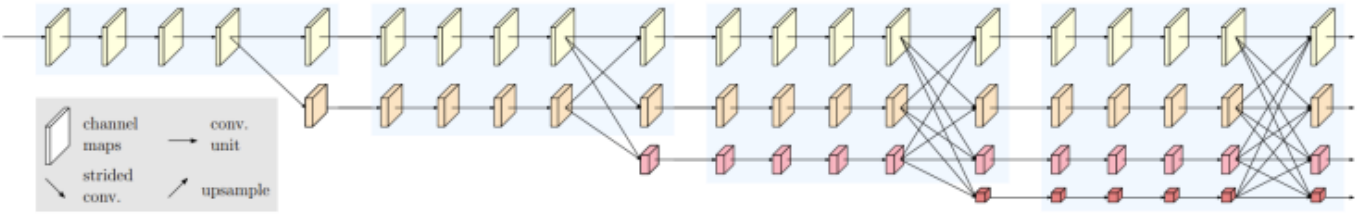

Fig. 3. The structure of HRNet

It can be seen that like the traditional networks such as UNet that downsampling and then upsampling, HRNet has multiple resolution branches, high-resolution is consistently maintained from the beginning to the end of the network, ensuring that there is a high resolution from beginning to end, greatly reducing information loss. In addition, there is a constant exchange of information between different resolutions, so that the low-resolution information can also receive the high-resolution information and improve the amount of information within the low-resolution branches.

B. *Object-Contextual Representations*

The main idea of OCRNet is to use the target region representation to enhance its pixel representation. The network before OCRNet is usually pixel-level for context, i.e., it considers the relationship between context pixels. Context is limited to the pixel level, which can result in not fully utilizing the features of the target region. The network structure of OCRNet is shown in Fig.4. Li R used for semantic segmentation [24].

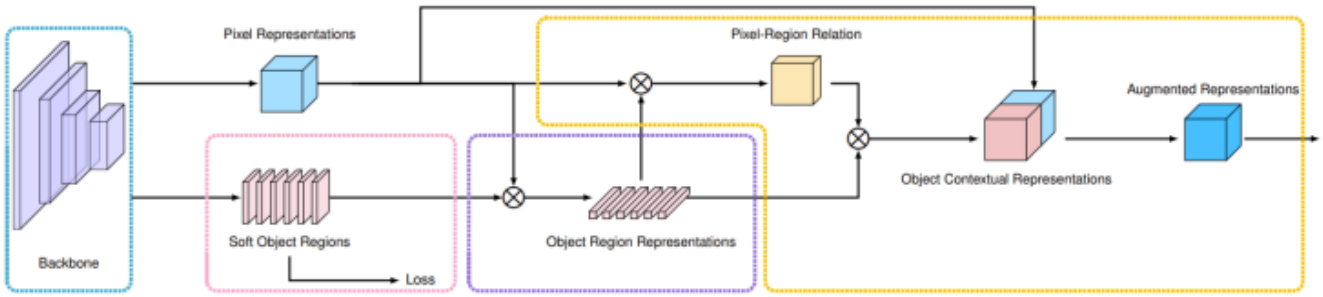

Fig. 4. The structure of OCR

The Soft Object Regions are created in the pink dashed box, the Object Region Representations in the purple dashed box, and the orange dashed box contains the Object Context Representations and Augmented Representations.

OCR initially creates a collection of soft object areas out of the contextual pixels. This segmentation is learned under the supervision of ground-truth segmentation. A coarse semantic segmentation result based on the feature estimation in the middle layer of the network is used as the input to OCR. After that, Object Region Representations are calculated based on the coarse semantic segmentation results and the pixel representations of the deepest layer of the network output. Finally, the relationship matrix between the Pixel Representations and the Object Region Representation is calculated, and then the Object Region Representation is weighted and summed according to the value of each pixel and the Object Region Representation in the relationship matrix to obtain when the network's deepest input is stitched together with the OCR, the Augmented Representation is used as the contextual information. Based on the augmented feature representation, it is possible to anticipate the semantic category of each pixel. OCR methods are more efficient and accurate than other semantic segmentation methods. Because the OCR method solves the object region classification problem rather than the pixel classification problem, i.e., the OCR method can effectively and explicitly enhance the object information.

V. ACGAN

The original GAN model can only generate sample data without labels, and the discriminator can only perform binary classification, which is not applicable to multi-classification of images. ACGAN adds classifiers to the original GAN, introduces labels in the generator training process, and constrains the generator to generate data according to the labels by the cross-entropy loss function of the discriminator. Renato Cardoso used it for satellite image generation[25]. Its loss function consists of two components, L_s and L_c.

$$L_s = E[logP(S = real|X_{real})] + E\left[logP\left(S = fake\middle|X_{fake}\right)\right] \quad (1)$$

$$L_c = E[logP(C = c|X_{real})] + E\left[logP\left(C = c\middle|X_{fake}\right)\right] \quad (2)$$

Where S denotes the sample true or false and C denotes the sample category. The objective of discriminator training is to maximize $L_s + L_c$, and the objective of generator training is to maximize $L_c - L_s$。

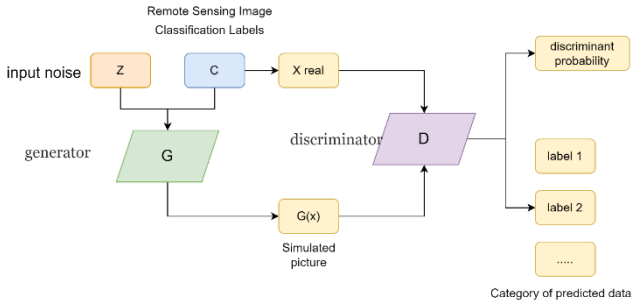

Fig. 5. Classification structure of ACGAN in remote sensing images

A. Model structure

Observing Fig.5, it can be seen that the input to the generator G consists of noise vectors and image categories, while discriminator D takes the real image with labels and the simulated image data G(z,c) generated by the generator. In the model used, the discriminator returns only the probability of a specific category or outputting a true or false label. The generator G is trained to generate blocks of images that match the desired category labels. Based on this, the discriminator D is trained to maximize the log-likelihood:

$$L_D = E[\log P(C = c|X_{real})] + E[logP(C = fake|X_{real})] \quad (3)$$

The first of these expects discriminator D to be able to correctly identify the true sample and match a true label, while the other expects a false label to be returned for the generated sample. The generator, in turn, is trained to maximize the maximum likelihood of：

$$L_G = E\left[logP\left(C = c\middle|X_{fake}\right)\right] \quad (4)$$

Generator G, on the other hand, wants to generate image data that is closest to the real sample. Both learn by adversarial learning, and the generator can eventually capture the real data distribution of the desired class.

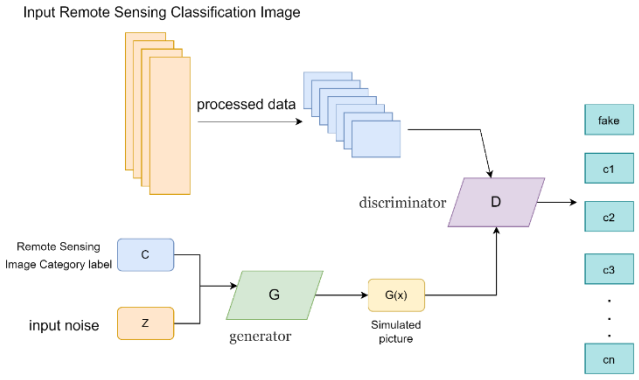

Fig. 6. The model structure used in this paper

In this model, the discriminator has only one output, which is either the label of the real data or a false label. Discriminator D has been trained to link genuine data samples with the category to which they belong, as shown in Fig. 6. In addition, discriminator D also tries to associate the samples generated by generator G with false labels. Based on the idea of adversarial, generator G is then trained to avoid false labels and match the generated samples with the class they belong to. In this way, samples from a few classes are avoided to be judged as false images in the discriminator, and the problem that the two components cannot be optimized due to sampling balance is solved. Moreover, since the discriminator in this model has a single output rather than a combination of two targets, it does not appear to be self-contradictory.

B. AdaptDrop

Nowadays, the training of deep neural networks often encounters the problem of overfitting and model collapse in GAN. Therefore, AdaptDrop [26] is introduced in this paper to alleviate the above problems. AdaptDrop is a regularization method with an attention mechanism. dropout operation is shown schematically in Fig.7 (a), and the AdaptDrop method is shown schematically in Fig.7 (b)(c). In which, the drop action is shown by a black circle, while the valid information region is denoted by a blue square. In all locations, the dropout procedure randomly discards pixels with a predetermined probability, as shown in Fig.7, apparently without using spatial information. Before performing the adaptive dropout operation, the input feature map A(t) is first obtained by normalizing the current feature map D(t), and then a set of pixels from each feature map are sampled using a Bernoulli distribution (the yellow circles in Fig.7 mark the sampled elements). For each element at position $M_{i,j}$, a block of space of size $block\_size \times block\_size$ is created centered on $M_{i,j}$. Then the kth percentile element is discarded, the number of discarded features is controlled by γ. and the rest of the elements are kept and set to 1. This results in an irregularly shaped adaptive mask. The parameter can be calculated as:

$$\gamma = \frac{1 - keep_{prob}}{block_{size}^2} \frac{size_{feature_{map}}^2}{\left(size_{feature_{map}} - block_{size} + 1\right)^2} \quad (5)$$

Where $keep\_prob$ is set 0.75 and 0.95 as in the dropout operation, and $size_{feature\_map}$ indicates the size of the $feature\_map$ where the AdaptDrop operation is performed. Finally, the generated adaptive mask is applied and the output is

$$A^{(t+1)} = A^t \times \frac{count(M)}{count_{ones(M)}} \quad (6)$$

where count(M) denotes the number of elements in the mask and $count\_ones(M)$ denotes the number of elements in the mask is 1.

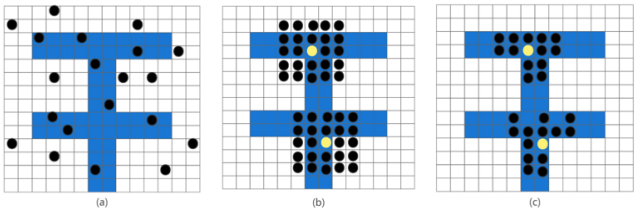

Fig. 7. The model structure used in this paper

C. Network structure

Table II shows the information on the generator and discriminator network layers in the ADGAN of the paper. Five convolutional layers make up the CNNs used by the discriminator D and generator G. The size of the input noise is 100 × 1 × 1. The input is changed by the generator G to have the size 64 × 64 × 3. In generator G, AdaptDrop is used for the second convolutional layer, while AdaptDrop is used for the fourth transposed convolutional layer in the discriminator.

D. block_size parameter analysis

An essential parameter that influences the categorization accuracy in AdaptDrop is $block\_size$. Classification uses contextual information that is susceptible to environmental noise. The classification performance of the dataset for various $block\_size$ is shown in Fig. 8. In implementation, the $block\_size$ varies from 3 to 11. The experiment result shows that when $block\_size$ increases from 3 to 11, as more contextual information is taken into account, classification accuracy increases. The best accuracy is reached when $block\_size = 11$. Therefore, $block\_size$ is set to 11 in the next experiments.

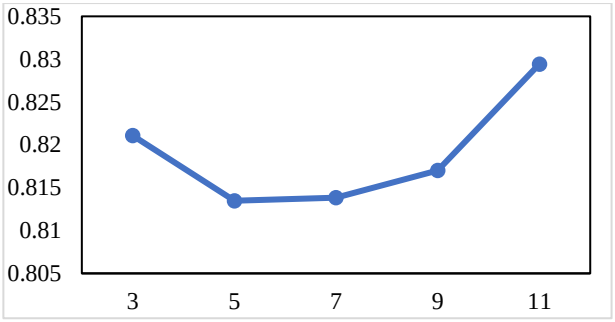

Fig. 8. Classification accuracy for different block_size

E. Comparison with different regularization methods

To explore the advantages of AdaptDrop compared to Dropout, the accuracy of classification is chosen to compare the two in this paper. As shown in Fig.9, the AdaptDrop cited in this paper has superior performance compared to Dropout, which is less effective because it randomly discards individual pixels in the feature map and can easily retrieve the discarded information by neighboring pixels. The network can efficiently learn robust features of ground objects in remote sensing image classification thanks to the proposed AdaptDrop, which eliminates informative areas from the feature map.

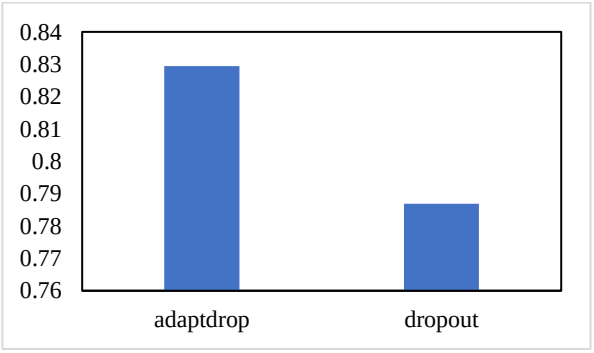

Fig. 9. Accuracy of different regularization methods for classification

TABLE II. NETWORK ARCHITECTURE USED IN THE EXPERIMENT

Network	Number of layers	Type	Convolution kernel	AdaptDrop	Stride	Padding	Activation Function
G	1	Conv	4*4*512	×	1	×	ReLU
G	2	Conv	4*4*256	√	2	√	ReLU
G	3	Conv	4*4*128	×	2	√	ReLU
G	4	Conv	4*4*64	×	2	√	ReLU
G	5	Conv	4*4*3	×	2	√	Tanh
D	1	TConv	4*4*64	×	2	√	LeakyReLU
D	2	TConv	4*4*128	×	2	√	LeakyReLU
D	3	TConv	4*4*256	×	2	√	LeakyReLU
D	4	TConv	4*4*512	√	2	√	LeakyReLU
D	5	TConv	4*4*128	×	1	×	LeakyReLU
D	6	FC		×			Softmax

VI. EXPERIMENT

For DeepLabV3plus [12], this paper replaces various backbones for training, such as resenet50, resnet101, mobilenet, and so on. In addition, segnet and mobilenet themselves were chosen as the baseline, and accuracy, precision, recall, and F1-score were chosen as the metrics for the experiments. Evaluation metrics are the key to measuring

a good or bad model, and the evaluation metrics used in this paper are Precision, Recall, F1-score, and Accuracy.

The confusion matrix is used to calculate the evaluation metrics.

TP predicts positive classes as positive class numbers.

TN predicts the negative class as the number of negative classes.

FP predicts negative classes as positive class numbers.

FN predicts positive classes as negative classes.

Precision is defined as $P = \frac{TP}{TP+FP}$ denotes the proportion of the examples classified as positive that are actually positive.

Recall is defined as $R = \frac{TP}{TP+FN}$ as a measure of coverage and a measure of how many positive cases are classified as positive cases.

Accuracy is defined as: $Acc = \frac{TP+TN}{TP+TN+FP+FN}$ denotes the proportion of the total number of correctly predicted outcomes.

F1-score is defined as: $F_1 = \frac{2PR}{P+R}$ where P is Precision and R is Recall. it is a weighted average of the model precision and recall.

TABLE III. RESULT

NETWORK	accuracy	precision	recall	F1-score
Segnet	0.78	0.39	0.37	0.38
DeeplabV3plus mobilenet	0.82	0.57	0.58	0.57
DeeplabV3plus resnet50	0.83	0.53	0.60	0.56
DeeplabV3plus resnet101	0.82	0.56	0.58	0.57
DeeplabV3plus xception	0.82	0.50	0.57	0.52
Mobilenet	0.82	0.69	0.73	0.46
HRNet-OCR	0.90	0.82	0.69	0.75
ACGAN	0.83	0.95	0.92	0.90

VII. CONCLUSION

This paper investigates the effectiveness of different models in resolving remote sensing mapping images. From Table III, we can see that the ACGAN model has the best results. The precision of ACGAN is at least 13 percent better than other models. Except for the accuracy, which is lower than that of HRNet, the remaining three indices are the best among all the models. It is hoped that the experimental results in this paper can help researchers in the field of remote sensing and mapping to have a reference basis when selecting a model. The models used in this paper are some of the more representative models in recent years. They are more prominent in the traditional image segmentation field. It is hoped that these models can be applied to actual production life and better help workers in the field of remote sensing to map the terrain and landscape. The work in this paper has some limitations, such as fewer and simpler indicators for testing. In the future, additional metrics can be considered for experimentation to better test the effects of different models. In addition, more models can be used for experiments to test the effect between these models and help select better models.

REFERENCES

[1] Mehmood M, Shahzad A, Zafar B, et al. Remote sensing image classification: A comprehensive review and applications[J]. Mathematical Problems in Engineering, 2022, 2022: 1-24..

[2] Huang H, Shi G, He H, et al. Dimensionality reduction of hyperspectral imagery based on spatial–spectral manifold learning[J]. IEEE transactions on cybernetics, 2019, 50(6): 2604-2616.

[3] Qian H, Li Y, Yang J, et al. Segmentation and analysis of cement particles in cement paste with deep learning[J]. Cement and Concrete Composites, 2023, 136: 104819.

[4] Liu Huanjun, Yang Haoxuan, Xu Mengyuan, et al. Soil classification based on maximum likelihood method and features of multi-temporal remote sensing images in bare soil period[J]. Transactions of the Chinese Society of Agricultural Engineering (Transactions of the CSAE), 2018, 34(14): 132 －139. (in Chinese with English abstract)

[5] Tari G, Jessen L, Kennelly P, et al. Surface mapping of the Milh Kharwah salt diapir to better understand the subsurface petroleum system in the Sab'atayn Basin, onshore Yemen[J]. Arabian Journal of Geosciences, 2018, 11(15): 428－438.

[6] Thakur R, Panse P. Classification Performance of Land Use from Multispectral Remote Sensing Images using Decision Tree, K-Nearest Neighbor, Random Forest and Support Vector Machine Using EuroSAT Data[J]. International Journal of Intelligent Systems and Applications in Engineering, 2022, 10(1s): 67-77.

[7] Wang L, Li R, Duan C, et al. A novel transformer based semantic segmentation scheme for fine-resolution remote sensing images[J]. IEEE Geoscience and Remote Sensing Letters, 2022, 19: 1-5..

[8] Zhao W, Du S. Learning multiscale and deep representations for classifying remotely sensed imagery[J]. ISPRS Journal of Photogrammetry and Remote Sensing, 2016, 113: 155－165.

[9] Alex Kendall, Vijay Badrinarayanan and Roberto Cipolla "Bayesian SegNet: Model Uncertainty in Deep Convolutional Encoder-Decoder Architectures for Scene Understanding." arXiv preprint arXiv:1511.02680, 2015.

[10] Khabarlak K. Post-Train Adaptive U-Net for Image Segmentation[J]. arXiv preprint arXiv:2301.06358, 2023.

[11] Zhao H, Shi J, Qi X, et al. Pyramid Scene Parsing Network[C]// 2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR). IEEE, 2017:6230- 6239.

[12] Zhao Weiiyu, Zhang Honghai, Zhong Bo. A Deep Learning Based Method for Remote Sensing Image Parcel Segmentation[J]. Frontiers of Data&Computing,2021, 3(2): 133-141. (in Chinese with English abstract)

[13] Yuhui Yuan et al. "Segmentation Transformer: Object-Contextual Representations for Semantic Segmentation" arXiv: Computer Vision and Pattern Recognition(2019): n. pag.

[14] Odena, Augustus, Christopher Olah, and Jonathon Shlens. "Conditional image synthesis with auxiliary classifier gans." International conference on machine learning. PMLR, 2017.

[15] Gu X, Li S, Ren S, et al. Adaptive enhanced swin transformer with U-net for remote sensing image segmentation[J]. Computers and Electrical Engineering, 2022, 102: 108223.

[16] K. He, X. Zhang, S. Ren and J. Sun, "Deep Residual Learning for Image Recognition," 2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Las Vegas, NV, USA, 2016, pp. 770-778, doi: 10.1109/CVPR.2016.90.

[17] Yuan J, Ma X, Han G, et al. Research on lightweight disaster classification based on high-resolution remote sensing images[J]. Remote Sensing, 2022, 14(11): 2577..

[18] François Chollet, "Xception: Deep Learning with Depthwise Separable Convolutions" arXiv:1610.02357, 2017

[19] Lu J, Zhou B, Wang B, et al. Land cover classification of remote sensing images based on improved DeeplabV3+ network[C]//Journal of Physics: Conference Series. IOP Publishing, 2022, 2400(1): 012035.

[20] Tilak T, Braun A, Chandler D, et al. Very High Resolution Land Cover Mapping of Urban Areas at Global Scale with Convolutional Neural Networks[J]. arXiv preprint arXiv:2005.05652, 2020.

[21] Mousavi S. Bio-Inspired Fossil Image Segmentation for Paleontology[J]. International Journal of Computational Engineering Science, 2022, 12: 5243-5249.

[22] J. Wang et al., "Deep High-Resolution Representation Learning for Visual Recognition," in IEEE Transactions on Pattern Analysis and Machine Intelligence, vol. 43, no. 10, pp. 3349-3364, 1 Oct. 2021, doi: 10.1109/TPAMI.2020.2983686.

[23] Bello I M, Zhang K, Su Y, et al. Densely multiscale framework for segmentation of high resolution remote sensing imagery[J]. Computers & Geosciences, 2022, 167: 105196.

[24] Li R, Zheng S, Zhang C, et al. Multiattention network for semantic segmentation of fine-resolution remote sensing images[J]. IEEE Transactions on Geoscience and Remote Sensing, 2021, 60: 1-13.

[25] Renato Cardoso, Sofia Vallecorsa, Edoardo Nemni, “Conditional Progressive Generative Adversarial Network for satellite image generation” arXiv:2211.15303, 2022

[26] Lee S, Kim D, Kim N, et al. Drop to adapt: Learning discriminative features for unsupervised domain adaptation[C]//Proceedings of the IEEE/CVF International Conference on Computer Vision. 2019: 91-100.