RESEARCH ARTICLE OPEN ACCESS

Semantic Segmentation Of Remote Sensing Images Based On Deep Learning

M. Rega [1], Dr. S. Sivakumar [2]

[1] Research Scholar,
Department of Computer and Information Science,Annamalai University, Annamalainagar
[2] Assistant Professor,
PG Department of Computer Science,Government Arts College, Chidambaram

ABSTRACT

Deep learning is a method in artificial intelligence (AI) that teaches computers to process data in a way that is inspired by the human brain. Deep learning models can recognize complex patterns in pictures, text, sounds, and other data to produce accurate insights and predictions. Image segmentation is a computer vision technique that partitions a digital image into discrete groups of pixels—image segments—to inform object detection and related tasks. By parsing an image's complex visual data into specifically shaped segments, image segmentation enables faster, more advanced image processing. Remote sensing is the process of detecting and monitoring the physical characteristics of an area by measuring its reflected and emitted radiation at a distance (typically from satellite or aircraft). Special cameras collect remotely sensed images, which help researchers "sense" things about the Earth. Semantic segmentation is a deep learning algorithm that associates a label or category with every pixel in an image. It is used to recognize a collection of pixels that form distinct categories.

Keywords: - Deep learning, Remote Sensing images, semantic segmentation, Image segmentation

I. INTRODUCTION

Semantic segmentation is a fundamental but challenging problem of pixel-level remote sensing (RS) data analysis. Semantic segmentation tasks based on aerial and satellite images play an important role in a wide range of applications. Recently, with the successful applications of deep learning (DL) in the computer vision (CV) field, more and more researchers have introduced and improved DL methods to the task of RS data semantic segmentation and achieved excellent results. Although there are a large number of DL methods, there remains a deficiency in the evaluation and advancement of semantic segmentation techniques for RS data. Semantic segmentation is a very important direction in the CV. Unlike target detection and recognition, semantic segmentation achieves image pixel-level classification. It can divide a picture into multiple blocks according to the similarities and differences of categories. Semantically related pixels are annotated with the same label. The semantic segmentation algorithm can comprehensively complete the recognition, detection, and segmentation of visual elements in the scene, and improve the efficiency and accuracy of image understanding. Compared with image classification and target detection, the semantic segmentation results can provide richer information about image parts and details. Semantic segmentation algorithms have extensive applications and long-term development prospects.

A large number of segmentation methods have been proposed before the deep learning era, such as the partial differential equation-based methods. With sufficient training data, the supervised learning strategy is able to greatly extend the capacity of a segmentation model, such as the random forest and visual grammar applied in natural scene understanding. The emergence of the deep learning technique has greatly promoted semantic segmentation research. To date, numerous novels deep-learning-based methods have been proposed, which are based on different technical roadmaps and target different applications.

II. RELATED WORKS

Some surveys and review papers have addressed advancements and innovations on the subject of deep learning and semantic segmentation. A Survey by Zhu et al. [15] covering a wide range of the papers and areas of semantic segmentation topics including, interactive methods, recent development in the super-pixel, object proposals, semantic image parsing, image co-segmentation, semi & weakly supervised, and fully supervised image segmentation. Thoma [16] presented a taxonomy of segmentation algorithms and overview of completely automatic, passive, semantic segmentation algorithms.

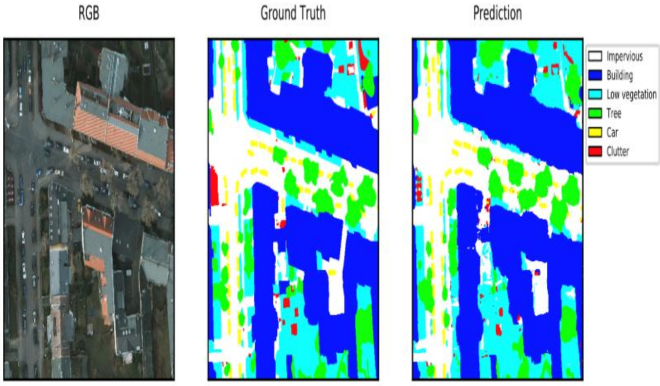
Niemeijer et al. [17] presented a review of neural network based semantic segmentation for scene understanding in the context of the autonomous driving. Guo et al. [18] provided a review of semantic segmentation approaches, i.e., region-based, FCN-based and weakly supervised approaches. They have summarized the strengths, weaknesses and major challenges in image semantic segmentation.

Geng et al. [19] presented a survey of recent progress in semantic segmentation with CNN's, and newly developed strategies that have achieved promising results on the Pascal VOC 2012 semantic segmentation challenge. Detail review provided by Garcia-Garcia et al. [20] on deep learning methods for semantic segmentation with their contributions and significance in the field. An extensive review presented by Hongshan et al. [21], categorized different methods based on hand engineered features, learned features, and weakly supervised learning.

The earlier approaches used for semantic segmentation were textonforest [12], random-forest based classifiers [13], whereas deep learning techniques allowed precise and much faster segmentation [14]. Variation of intensity/gray level defines their shape and size.

III. THE PROPOSED MODEL

In this paper, an RSI semantic segmentation model called the Adaptive Feature Fusion UNet (AFF-UNet) is proposed (shown in Fig).



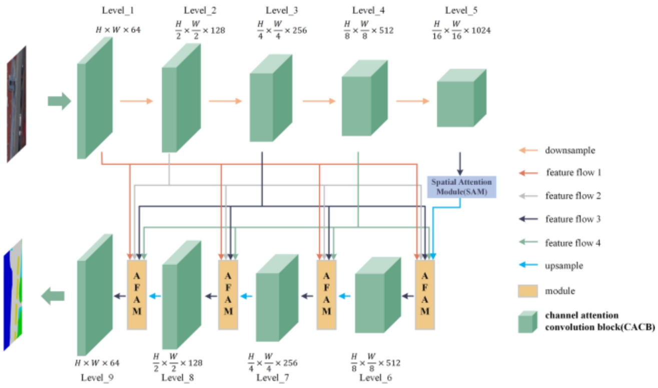
The proposed model is developed from UNet, and includes three parts: 1. DSC and AFAM were designed to make the model have a stronger context aggregation ability and could adaptively emphasize or suppress different levels of feature blocks instead of feature channels during fusion; 2. CACB was applied to obtain the relationship between the different channels; and 3. SAM was applied to obtain the relationship between the different positions. Details are as follows:

1. The feature maps of the first, second, third, and fourth levels in the encoder are directly connected to the adaptive fusion attention module (AFAM) in the decoder through a dense connection structure (consisting of feature flow 1, feature flow 2, feature flow 3, and feature flow 4). The dense skip connections (DSC) improves the feature fusion ability and feature utilization of the model, which in turn allows the model to obtain more accurate feature representations and reduce the misidentification of confused classes.

2. The AFAM can differentiate between features from different levels and assign weights to these features to improve the segmentation of different-sized objects. The input of the AFAM comes from both the encoder and the decoder feature maps. The module primarily performs feature map fusion and assigns weights to the input's five feature map blocks to improve the segmentation of different-sized categories.

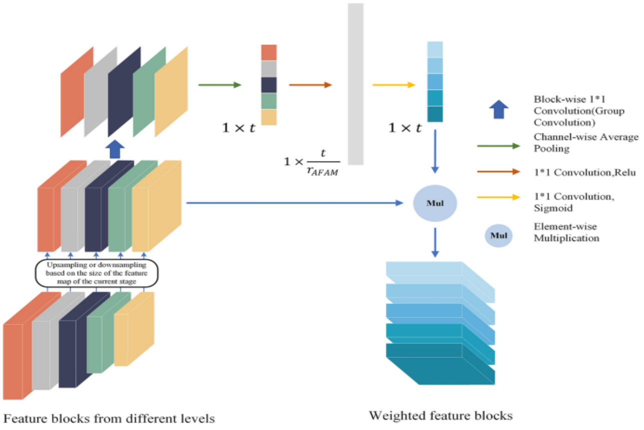
3. The Channel Attention Convolution Block (CACB) consists of convolution, batch normalization, activation function, and compress-activation operations. It does not change the size of the feature map but alters the number of channels in the feature map. In this paper, the model's input is 3 or 4

channels, and the encoder’s channel changes as follows: 3 or 4 channels, 64 channels, 128 channels, 256 channels, 512 channels, 1024 channels, accompanied by a decrease in the size of the feature map. The decoder's channel changes are 512 channels, 256 channels, 128 channels, and 64 channels, accompanied by an increase in the size of the feature map.


Level_1
$H \times W \times 64$
Level_2
$\frac{H}{2} \times \frac{W}{2} \times 128$
Level_3
$\frac{H}{4} \times \frac{W}{4} \times 256$
Level_4
$\frac{H}{8} \times \frac{W}{8} \times 512$
Level_5
$\frac{H}{16} \times \frac{W}{16} \times 1024$
Spatial Attention Module(SAM)
AFAM
AFAM
AFAM
AFAM
$H \times W \times 64$
Level_9
$\frac{H}{2} \times \frac{W}{2} \times 128$
Level_8
$\frac{H}{4} \times \frac{W}{4} \times 256$
Level_7
$\frac{H}{8} \times \frac{W}{8} \times 512$
Level_6
downsample
feature flow 1
feature flow 2
feature flow 3
feature flow 4
upsample
module
channel attention convolution block(CACB)

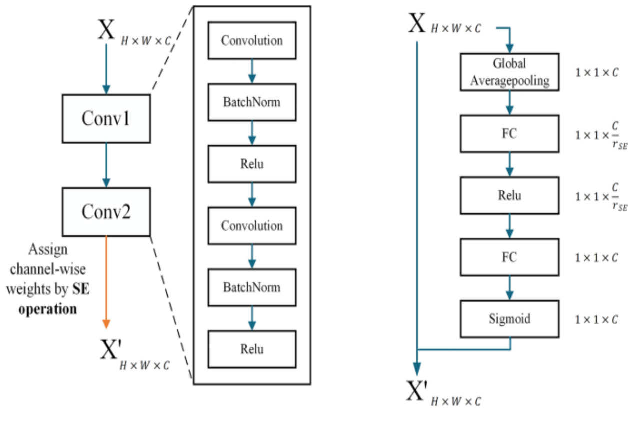
DSC architecture and the AFAM

To enhance the feature fusion ability, the proposed model applies the DSC structure that makes that the information of each level in the encoder be connected to each level in the decoder. The motivation is because it is difficult to predict, which context information to aggregate, and this is more conducive for improving model accuracy. Hence, we used DSC to make the model have the potential to fuse contextual information at any stage. We then let the model automatically select which levels of information should be “paid more attention” to by AFAM


$1 \times t$
$1 \times \frac{t}{r_{AFAM}}$
$1 \times t$
Mul
Upsampling or downsampling based on the size of the feature map of the current stage
Block-wise 1*1 Convolution(Group Convolution)
Channel-wise Average Pooling
1*1 Convolution,Relu
1*1 Convolution, Sigmoid
Mul
Element-wise Multiplication
Feature blocks from different levels
Weighted feature blocks

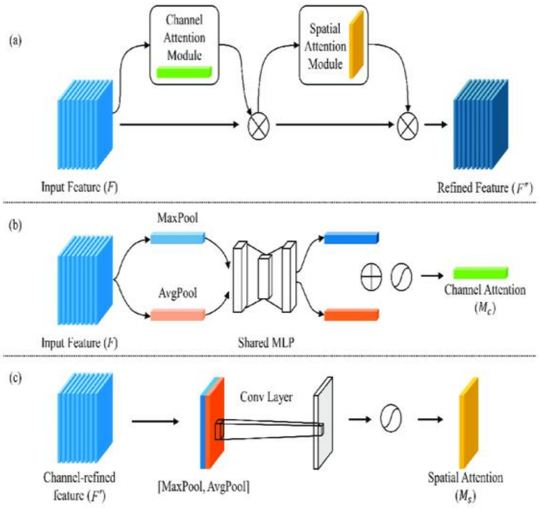
Channel attention convolution block (CACB)

The CACB contains two sets of combination (convolution, batch normalization, Relu activation) operations and a SE operation. The SE operation includes two steps: Squeeze and Excitation. Squeeze compresses the features of the H × W × C into 1 × 1 × C through the squeeze operator, that is, compresses each channel into a number. The Excitation operation obtains the weight of each channel using two convolution-activation operations (the second activation function is called the excitation operator) and finally applies the obtained weight to the corresponding channel, that is, reweight. The difference between AFAM and the SE Operation is that the SE assigns weights to each feature channel, while AFAM assigns weights to different levels of feature blocks. CACB can automatically obtain the importance of each feature channel for segmentation and then choose to enhance or suppress each feature channel. rSE is a hyperparameter that allows us to vary the capacity of the SE operation. In the proposed model, rSE was set to two, which is the optimal configuration obtained by experiments.


$X_{H \times W \times C}$
Conv1
Conv2
Assign channel-wise weights by SE operation
$X'_{H \times W \times C}$
Convolution
BatchNorm
Relu
Convolution
BatchNorm
Relu
$X_{H \times W \times C}$
Global Averagepooling
$1 \times 1 \times C$
FC
$1 \times 1 \times \frac{C}{r_{SE}}$
Relu
$1 \times 1 \times \frac{C}{r_{SE}}$
FC
$1 \times 1 \times C$
Sigmoid
$1 \times 1 \times C$
$X'_{H \times W \times C}$

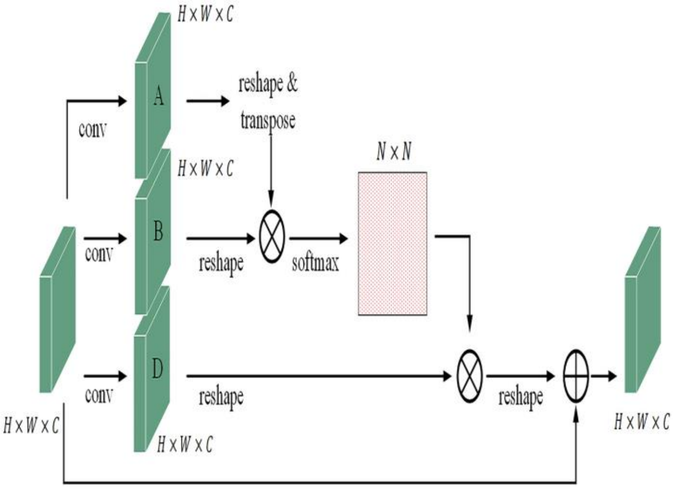
(a) Channel Attention Convolution Block(CACB)

(b) SE Operation



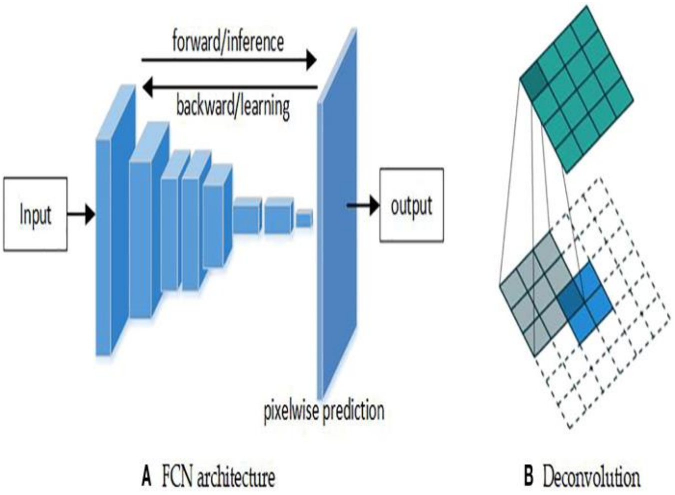
Spatial attention module (SAM)

The SAM consists of four steps. First, convolution operations on the input are implemented to obtain three feature maps: A, B, and C. The sizes of A, B, and D are H × W × C (where H represents height, W represents the width, and C represents the number of channels). Then the shape of A is changed into N × C (N = H × W) through reshaping and transposing, and the size of B is changed into C × N through reshaping. Matrix multiplication and SoftMax activation on A and B are implemented to obtain the attention weight map of size N × N. The third step is to change the shape of D to C × N by reshaping and performing matrix multiplication with the attention weight map obtained in the second step to obtain the optimized feature map. Finally, the original feature map and the optimized feature map are subjected to matrix addition to obtain the final output.



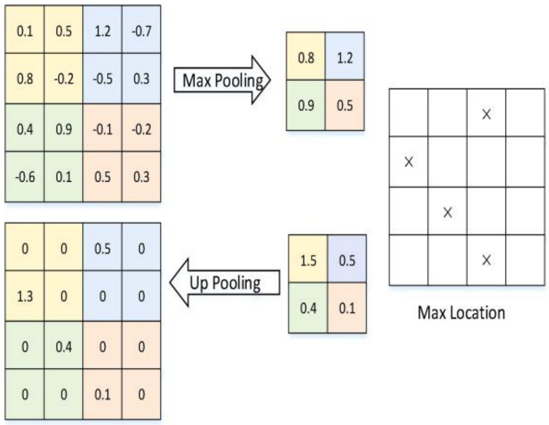
FCN architecture

In the FCN, 1×1 convolution replaces the full connection in the CNN. Then, the probability category value of each pixel is obtained through the SoftMax layer. FCN introduces the deconvolution shown in Figure . The true category of each pixel is the category with the largest corresponding probability value. Finally, a segmented image is obtained, whose size is the same resolution as the input image. The deconvolution uses known convolution kernels and convolutional output to restore images, thereby obtaining refined features. The reason that FCN is more efficient than CNNs is that computing convolutions are avoided one by one for each pixel block, in which adjacent pixel blocks are repeated.



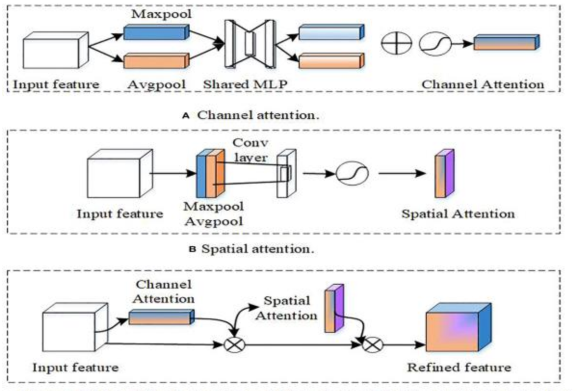
SegNet architecture

The SegNet network includes an encoder network, a symmetric decoder network, and a classification layer pixel-wise. It has 13 convolutional layers that are the same as the VGG16. Up-pooling, which applies the index of Max Pooling, is used in the encoder to the decoder. It improves the recognition effect of the segmentation task on the segmentation boundary. The positions of the maximum values of the four colours are recorded. In the up-pooling block, these positions are marked, and the other positions are filled with zeros. In this way, the recognition effect of the segmentation task can be improved on the boundary.



Channel attention

Channel attention generates an attention mask in the channel domain to select important channels. Channel attention focuses on the channel dimension, which is shown in figure. A feature detector detected feature maps of each channel. For a feature map, the importance of each channel is calculated, and the weighted feature map is obtained by multiplying weights with the feature maps.



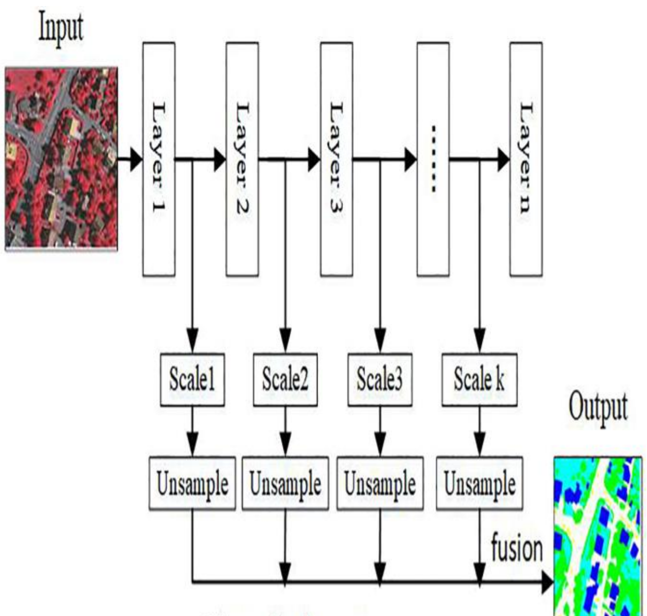
A Channel attention.

B Spatial attention.

C The convolutional block attention combines channel attention and spatial attention.

Hierarchical structure

The algorithm based on the hierarchical structure obtains multi-scale information through different stage features of the CNN, which is shown in Figure . During the forward propagation process of the CNN, the receptive field increases continuously with the convolution and pooling operations. Multi-scale features from channel and spatial can be captured by fusing the features from CNN’s different stages



A Hierarchical structure

High-resolution RS data have larger dimensions than typical natural images. Nodes in each path originate from leaf nodes. They designed a new loss function to optimize boundaries. It learned scale-invariant and small objects context information.

CNN Architectures for semantic segmentation

There have been a number of deep network architectures devised for image classification and segmentation. This section first gives a brief introduction of the fundamental ideas of CNNs and then we focus on variants of CNN designed toward semantic segmentation and present their network structures and key ideas. This section also includes deep learning architectures that are not directly applicable to semantic segmentation problems, but their ideas have been adopted in a few methods in the field.

IV. CONCLUSION

In this study, we proposed the AFF-UNet, which exhibits better performance in handling confusion between object classes and segmentation of different sizes of object classes, particularly buildings and vehicles, in RSI. To achieve this, we utilized a tailored CACB to obtain the channel relationship. We performed context aggregation and obtained the relationship between different levels of feature blocks using DSC structures and AFAM. Moreover, we utilized SAM to further improve segmentation accuracy by obtaining the relationship between different positions. The AFF-UNet was evaluated on two datasets, and compared with other models, it demonstrated improvements in (1) the confusion of the object classes; (2) achieving better segmentation results for different sizes of object classes; and (3) improving object class integrity. Our proposed model has potential to optimize binary classification tasks, such as extracting vehicles or buildings. However, the class imbalance in RSI segmentation datasets typically negatively impacts performance, and resolving this will be the focus of future work.

REFERENCES

[1] Study and Comparison of Different Edge Detectors for Image Segmentation By Pinaki Pratim Acharjya, Ritaban Das & Dibyendu Ghoshal.

[2] Methods of Image Edge Detection: A Review Dharampal and Vikram Mutneja

[3] Raman Maini & Dr. Himanshu Aggarwal International Journal of Image Processing (IJIP), Volume (3) : Issue (1) 1 Study and Comparison of Various Image Edge Detection Techniques.

[4] https://ieeexplore.ieee.org/abstract/document/6885761#:~:text=An%20improved%20Canny%20edge%20detection%20algorithm

[5] Pustokhina IV, Pustokhin DA, Kumar Pareek P, Gupta D, Khanna A, Shankar K (2021) Energy-efficient cluster-based unmanned aerial vehicle networks with deep learning-based scene classification model. Int J Commun Syst 34(8):e4786

[6] Saeed S, Latif MA, Rajput MA (2021) Fuzzy-based multi-crop classification using high resolution UAV imagery. Quaid-EAwam Univ Res J Eng Sci Technol Nawabshah 19(1):1–8

[7] Lin D, Lin J, Zhao L, Wang ZJ, Chen Z (2021) Multilabel aerial image classification with a concept attention graph neural network. IEEE Trans Geosci Remote Sens 23:1–12

[8] Alshehri A, Bazi Y, Ammour N, Almubarak H, Alajlan N (2019) Deep attention neural network for multi-label classification in unmanned aerial vehicle imagery. IEEE Access 7:119873–119880

[9] Bashmal L, Bazi Y, Al Rahhal MM, Alhichri H, Al Ajlan N (2021) UAV image multi-labeling with data-efficient transformers. Appl Sci 11(9):3974

[10] Ragab, M., 2023. Leveraging mayfly optimization with deep learning for secure remote sensing scene image classification. *Computers and Electrical Engineering*, *108*, p.108672.

[11] Laban, N., Abdellatif, B., Ebied, H.M., Shedeed, H.A. and Tolba, M.F., 2020. Multiscale satellite image classification using deep learning approach. *Machine Learning and Data Mining in Aerospace Technology*, pp.165-186.

[12] https://www.mdpi.com/2072-4292/15/5/1229