2021 IEEE 4th Advanced Information Management, Communicates, Electronic and Automation Control Conference (IMCEC) | 978-1-7281-8535-4/21/$31.00 ©2021 IEEE | DOI: 10.1109/IMCEC51613.2021.9482266

E-Unet++: A Semantic Segmentation Method for Remote Sensing Images

Yintu Bao, Wei Liu, Ouyang Gao, Zhikang Lin, Qing Hu
Information Engineering University, Zhengzhou, China
bao258456@163.com, greatliuliu@163.com, gaoouyang997@126.com, 2780813701@qq.com, 948107836@qq.com
Corresponding Author: Wei Liu Email: greatliuliu@163.com

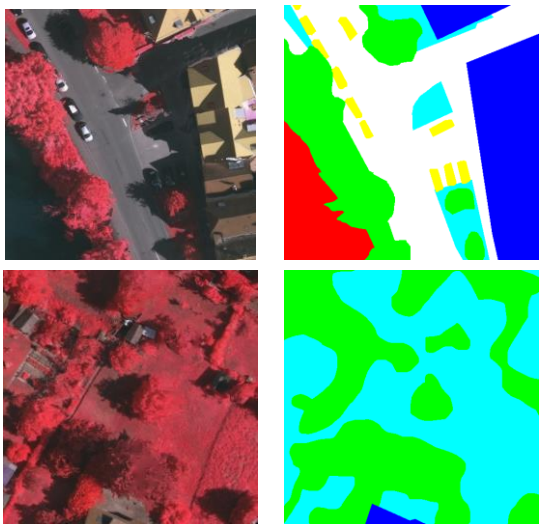
***Abstract*—Semantic segmentation can distinguish objects in remote sensing images at the pixel level. However, traditional semantic segmentation algorithms are more and more difficult to meet people's needs. With the rapid development of deep learning, especially its application in remote sensing images has greatly improved the parsing ability and efficiency. But, the complexity and diversity of remote sensing image content make the accuracy of semantic segmentation still need to be improved. Thus, a semantic segmentation method that combines the characteristics of EfficientNet and UNet++ is proposed in this paper. The method can make the segmentation boundary clearer and improve the segmentation effect of densely distributed objects. The results show that the proposed method achieves good performance in the Vaihingen dataset.**

Keywords—deep learning; semantic segmentation; remote sensing image; unet; efficientnet

I. INTRODUCTION

Semantic segmentation is a basic task in computer vision. This task assigns a label to each pixel, which is also called pixel-level classification. It is an important part of computer vision-based applications[1]. Image semantic segmentation technology can segment and label specific targets in remote sensing images, to extract specific information in remote sensing images research, and realize the fine analysis of the whole scene of remote sensing images. For example, semantic segmentation technology can segment and extract buildings or vegetation in remote sensing images, providing basic support for other research. Semantic segmentation of remote sensing images is an important research work, which can promote the development of important research fields such as military, agriculture, and environmental protection[2].

In recent years, the Convolutional Neural Network (CNN) shows its ability to capture features, especially in computer vision. Researchers have successively constructed CNN such as FCN, Unet, and DeepLab for semantic segmentation[3], these encoder-decoder frameworks can capture richer semantic information, the performance of semantic segmentation has been rapidly improved. However, with the development of remote sensing technology, image clarity has improved, and feature information has become more abundant, which has led to the problems of similarity between classes and differences within classes[4]. As shown in Fig. 1, the example is taken from the ISPRS Vaihingen dataset. There are great differences in the size of objects, such as Car and Building, and even the shapes of objects in the same category, such as Tree. In addition, the boundary of the target is usually very irregular, and the distribution of various types of targets is unbalanced. For these reasons, the result of segmentation of smaller target and target boundary is usually not very good[5]. Thus, there are still challenges in how to improve the quality of semantic segmentation of small-sized objects and object boundaries in remote sensing images.



(a) Original image (b) Ground truth

Fig. 1. An example is taken from the ISPRS Vaihingen dataset. The label includes six categories: Impervious Surfaces (white), Building (blue), Low Vegetation (cyan), Tree (green), Car (yellow), and Clutter/background(red).

To address the need for more accurate segmentation in remote sensing images, we propose E-Unet++, a new segmentation architecture that can capture semantic features more accurately, so that the relationship between the pixels is clearer, so as to get more accurate object boundaries. Because the information of each pixel is more accurate, the segmentation result of small-sized objects is also improved. The contributions of this work primarily include the following three points:

Authorized licensed use limited to: Indraprastha Institute of Information Technology. Downloaded on July 24,2024 at 06:16:39 UTC from IEEE Xplore. Restrictions apply.

1) The distribution of various objects in remote sensing images is unbalanced. By using the weighted cross-entropy loss function, the feedback contribution of various objects to the neural network can be balanced.

2) Replace the convolution block of Unet++ with a more efficient network, thereby enhancing the ability of the model to capture information unction.

3) Adopting the learning rate of cosine annealing decay makes it more likely to get the optimal solution during model training.

II. METHODOLOGY

A. The Weighted Cross-entropy Loss Function

The loss function can measure the gap between the predicted label and the true label. For multi-classification problems such as semantic segmentation, cross-entropy loss is usually used. However, this cross-entropy loss does not work well in semantic segmentation of remote sensing images, because the distribution of various objects in remote sensing image data is unbalanced, and the size and shape of the objects are changeable, which leads to the imbalance of data becomes more prominent.

To solve the problem of class imbalance, we propose a weighted cross-entropy loss, which is defined as:

$$L(p,q) = -\sum_{j=1}^{K} W_j \times q_j \times \log\left(p_j\right) \qquad (1)$$

where K is the number of classes, $W_j \left(0 < W_j < 1 \text{ and } \sum_j^K W_j = 1\right)$ is a specific weight assigned to the class j, q_j is the j-th element of the normalized ground truth vector, p_j is the j-th element of the estimated vector for the class j, which is described as:

$$p_j = \frac{\exp^{z_j}}{\sum_{k=1}^{K} \exp^{z_k}} \qquad (2)$$

where z_j represents the classification results of the model, p_j converted result is divided by the sum of all the converted results, which can be understood as the percentage of the converted result in the total and that gives the probability of approximation.

Because of the weight factor, different weights can be given to each class as needed, thereby affecting the contribution of each class. For example, a class with a small proportion can be given larger weight to increase its contribution, which can reduce the impact of unbalanced data distribution. Therefore, the advantage of the weighted cross-entropy loss function is that it can calculate the error of each category more accurately.

B. Backbone for E-Unet++

UNet++ is a segmentation architecture based on nested and dense skip connections[6]. Fig. 2 is the architecture of UNet++, as seen, UNet++ consists of an encoder and decoder that are connected through a series of nested dense convolutional blocks, skip connections to associate feature maps of different depths, and fully integrate multi-scale feature maps, which is beneficial to improve the accuracy of semantic segmentation. However, as a result, the number of parameters increases, and the required computing resources increase. To solve the problem of parameters, the deep supervision method is adopted, and the model is pruned when a better result is obtained.

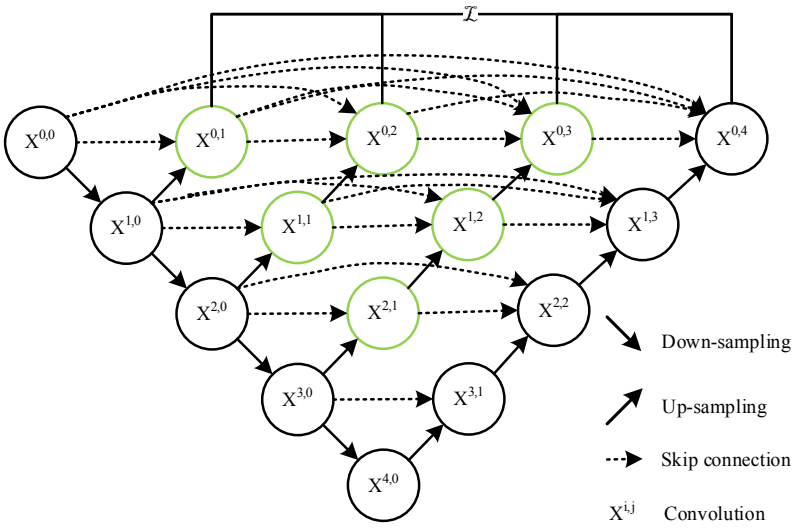


Fig. 2. The architecture of UNet++.

Authorized licensed use limited to: Indraprastha Institute of Information Technology. Downloaded on July 24,2024 at 06:16:39 UTC from IEEE Xplore. Restrictions apply.

The ability of the convolution block to extract features also affects the segmentation results of Unet++. According to this principle, we use EfficientNet[7] as the convolution block to capture features. EfficientNet converts width, depth, and resolution of the network into optimization problems, which is described as:

$$\begin{aligned} &\max_{d,w,r} \ Accuracy(N(d,w,r)) \\ &s.t. \quad N(d,w,r) = \bigodot_{i=1\cdots s} \hat{F}_i^{d\cdot\hat{L}_i}\left(X_{\langle r\cdot\hat{H}_i, r\cdot\hat{W}_i, w\cdot\hat{C}_i\rangle}\right) \\ &\qquad \text{Memory}(N) \le \text{target\_memory} \\ &\qquad \text{FLOPS}(N) \le \text{target\_flops} \end{aligned} \tag{3}$$

where d, w, r are coefficients for width, depth, and resolution of the network, N is the network, $\hat{F}_i, \hat{L}_i, \hat{H}_i, \hat{W}_i, \hat{C}_i$ are predefined parameters.

By constantly adjusting the width, depth, and resolution of the network, EfficientNet can effectively use resources and has a strong ability to capture features.

C. Cosine Annealing Decay

Cosine annealing decay uses the cosine sparking schedule to set the learning rate of each parameter group, gradually use the cosine function to reduce the learning rate in the cycle, and restore the learning rate to the maximum after the training times of the cycle step length is over[8]. The way that the cosine annealing scheduler adjusts the learning rate during the training process is described as:

$$\eta_t = \eta_{\min}^i + \frac{1}{2}\left(\eta_{\max}^i - \eta_{\min}^i\right)\left(1+\cos\left(\frac{T_{cur}}{T_i}\pi\right)\right) \tag{4}$$

where i is the index value, represents the *i-th* learning process; $\eta_{\max}^i$ and $\eta_{\min}^i$ defines the range of learning rate, $\eta_{\max}^i$ represents the peak of the *i-th* learning rate, $\eta_{\min}^i$ represents the trough value of the *i-th* learning rate. T_i means the current epochs that have been iterated, it will update its value after each batch is run. T_{cur} represents the total number of epochs. When it comes to restarting, in order to improve performance, a relatively small T_{cur} will be initialized at the beginning, after each restart, T_i will be multiplied by T_{cur} to increase its value, if no restart is involved, that is to fix T_{cur} as the number of epochs of our training model.

In the traditional training process, the learning rate gradually decreases, so the model gradually finds the local optimum. In this process, because the initial learning rate is high, the model will not step into the steep local optimum, but quickly move to the flat local optimum. As the learning rate gradually decreases, the model finally converges to a better best point. Since the learning rate of cosine annealing drops rapidly, the model will quickly step into the local optimal point (regardless of the magnitude of the slope of the decline), and save the local optimal point model. After saving the model, the learning rate returns to a larger value, escapes the current local optimum, and finds a new optimum. Because the models with different local best points are stored in greater diversity, the effect will be better after the collection. Comparing the two methods, the starting point and ending point of model training are basically the same. But compared with the traditional training process, the difference is that the cosine annealing learning rate makes the training process of the model more tortuous. During the training process, we use the cosine annealing scheduler to adjust the learning rate.

III. Experimental Results

This section introduces the datasets and experimental settings to verify the effectiveness of E-Net++, and then compares the performance between different frameworks.

A. Datasets

We validate our proposed method on the ISPRS Vaihingen dataset. This dataset was captured over Vaihingen in Germany, which has a resolution of 9 cm/pixel with tiles of approximately 2400 × 2000 pixels. The data set contains 33 remote sensing images, 16 of them are annotated with ground truth labels. There are 6 classes of labels, five foreground classes (Impervious Surfaces, Building, Low Vegetation, Tree, Car) and one background class (Clutter, which includes water bodies and other objects). We divide the annotated dataset into training set (1, 3, 5, 7, 13, 17, 21, 23, 26, 32, and 37) and test set (11, 15, 28, 30, and 34). Because the image size is too large, in order to facilitate the network model training, the picture and label will be divided into many small patches of 512 * 512 pixels. However, only about 300 images are obtained, which is difficult to train. The data is expanded by vertical flipping, horizontal flipping, distortion, etc., and finally, a dataset of 2000 images is used for training.

B. Experimental Setting

Our implementations are based on the publicly available PyTorch deep learning framework and tested on a workstation with Windows10, 128 GB RAM, an Intel Xeon Gold 5218 processor, and one NVIDIA Quadro RTX 5000 GPU card. During the training process, the optimizer is set as Adam with weight decay 0.0001, and 4 batch size. We train all models for 200 epochs, the base learning rate set 0.0001, and the learning rate strategy uses the cosine annealing decay. The weighted cross-entropy loss function is used as an evaluation to measure the disparity between the segmentation maps and ground truth. For quantitative evaluation, we report the Overall Accuracy (OA) and the mean Intersection-over-Union (mIoU) using the average of 5 runs.

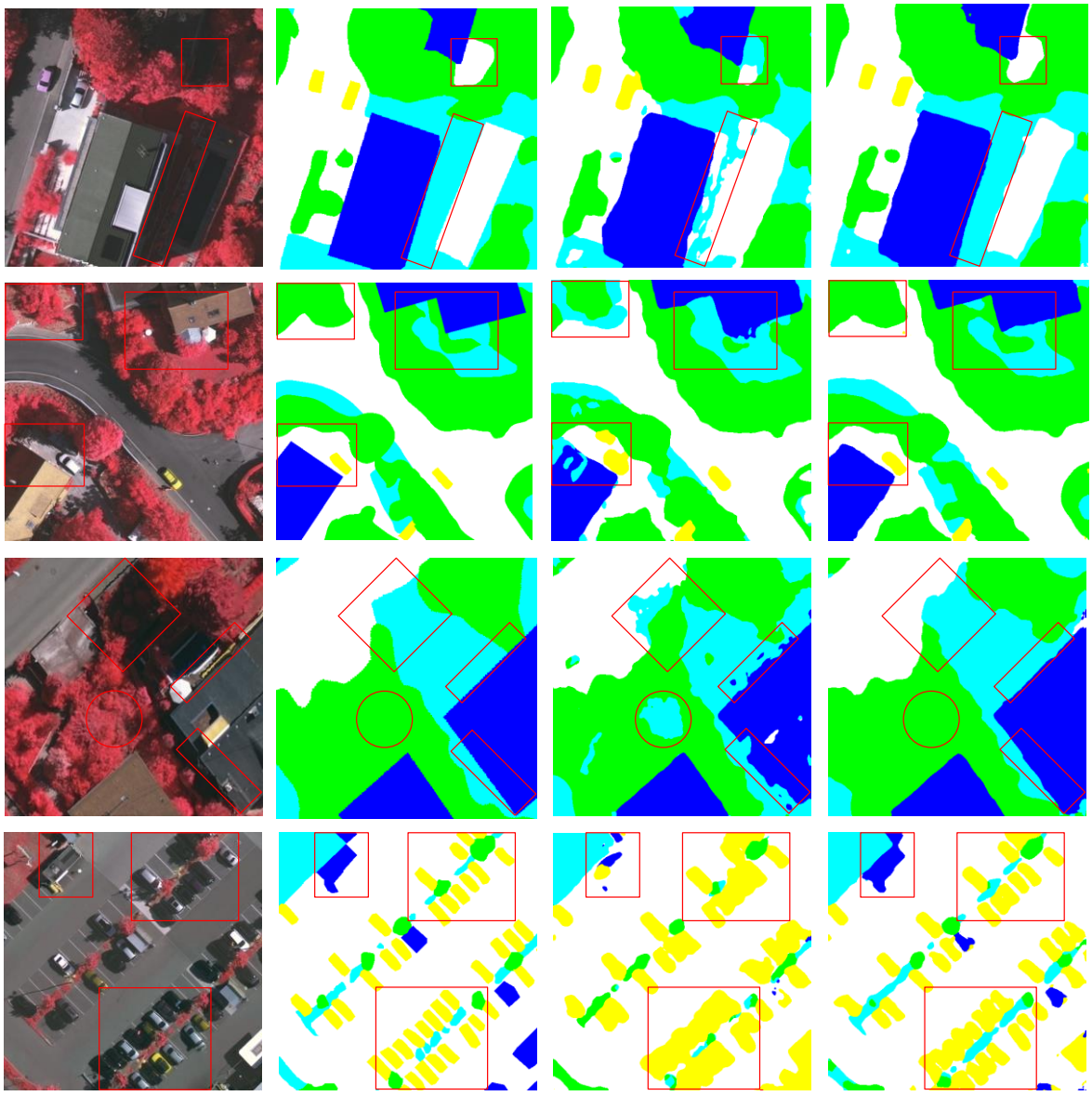
Authorized licensed use limited to: Indraprastha Institute of Information Technology. Downloaded on July 24,2024 at 06:16:39 UTC from IEEE Xplore. Restrictions apply.

C. Results

The experimental results of different methods on the Vaihingen dataset are demonstrated in TABLE I. The performance of the proposed E-Unet++ is significantly improved. For the Vaihingen dataset, the improvements are more than 3% in OA and 4.2% in mIoU compared with UNet++. Some visual results generated by our method and UNet++ are provided in Fig. 3. which shows that the proposed network can achieve more accurate segmentation, especially the boundaries of objects are clearer, and densely distributed small-size objects can also be distinguished from each other.

TABLE I. THE EXPERIMENTAL RESULTS ON THE VAIHINGEN DATASET.

Method	OA	mIoU
FCN	87.80	72.44
UNet++	91.23	80.12
E-Unet++(Ours)	94.29	84.36



(a) Original image (b) Ground truth (c) UNet++ (d) E-Unet++(Ours)

Fig. 3. Visualization of results on the Vaihingen dataset.

Authorized licensed use limited to: Indraprastha Institute of Information Technology. Downloaded on July 24,2024 at 06:16:39 UTC from IEEE Xplore. Restrictions apply.

IV. CONCLUSIONS

In this paper, in order to improve the accuracy of semantic segmentation of remote sensing images, we combined multi-scale fusion of U-Net++ and capture feature of EfficientNet to design E-Unet++ for remote sensing image semantic segmentation. Using weighted cross-entropy loss function to deal with the uneven distribution of data. Moreover, cosine annealing decay improves the model's ability to find the optimal solution. Experiments on the data set show that E-Net++ is significantly improved in the segmentation of object boundaries and small-size objects.

ACKNOWLEDGMENT

This work was supported by the National Natural Science Foundation of China under Grant No.41901378.

REFERENCES

[1] X. Yuan, J. Shi, and L. Gu, "A review of deep learning methods for semantic segmentation of remote sensing imagery." Expert Systems with Applications, vol.169, pp. 114417, 2021.

[2] M. Alam, J.F. Wang, G. Cong, L.V. Yunrong, and Y. Chen, "Convolutional Neural Network for the Semantic Segmentation of Remote Sensing Images." Mobile Networks and Applications, vol.4, pp.1-16, 2021.

[3] R. Li, C. Duan, S. Zheng, C. Zhang, and P. M. Atkinson, "MACU-Net for Semantic Segmentation of Fine-Resolution Remotely Sensed Images," IEEE Geoscience and Remote Sensing Letters, pp. 1-5, 2021

[4] B. Chen, M. Xia, and J. Huang, "MFANet: A Multi-Level Feature Aggregation Network for Semantic Segmentation of Land Cover" Remote Sens, vol. 13, issue 4, pp. 731, 2021.

[5] Y. Feng, W Diao, X. Sun, J. Li, K. Chen, K. Fu, and Gao X, NPALOSS: NEIGHBORING PIXEL AFFINITY LOSS FOR SEMANTIC SEGMENTATION IN HIGH-RESOLUTION AERIAL IMAGERY. ISPRS Annals of Photogrammetry, Remote Sensing and Spatial Information Sciences. Vol.2, pp. 475-482, 2020.

[6] Z. Zhou, MMR Siddiquee, N. Tajbakhsh, and J. Liang, "UNet++: A Nested U-Net Architecture for Medical Image Segmentation," Deep Learn Med Image Anal Multimodal Learn Clin Decis, vol. 11, pp.3-11, 2019.

[7] M. Tan, Q. Le, "EfficientNet: Rethinking Model Scaling for Convolutional Neural Networks," Proceedings of Machine Learning Research, vol.97, pp.6105-6114, 2019.

[8] Y.Q. Jiang, Automatic change detection method of island based on deep learning, Zhe Jiang University, 2020.

Authorized licensed use limited to: Indraprastha Institute of Information Technology. Downloaded on July 24,2024 at 06:16:39 UTC from IEEE Xplore. Restrictions apply.