IEEE Access
Multidisciplinary | Rapid Review | Open Access Journal

Received April 23, 2021, accepted May 6, 2021, date of publication May 10, 2021, date of current version May 18, 2021.

Digital Object Identifier 10.1109/ACCESS.2021.3078742

Semantic Segmentation Network of Remote Sensing Images With Dynamic Loss Fusion Strategy

WENJIE LIU[1], YONGJUN ZHANG[1], JUN YAN[2], YONGJIE ZOU[1], AND ZHONGWEI CUI[3]

[1]Key Laboratory of Intelligent Medical Image Analysis and Precise Diagnosis of Guizhou Province, College of Computer Science and Technology, Guizhou University, Guiyang 550025, China
[2]Zhuhai Orbita Aerospace Science & Technology Company Ltd., Zhuhai 519000, China
[3]Big Data Science and Intelligent Engineering Research Institute, Guizhou Education University, Guiyang 550018, China

Corresponding authors: Yongjun Zhang (zyj6667@126.com) and Jun Yan (yan@myorbita.net)

This work was supported in part by the construction of automatic and intelligent integrated ground processing system for micro nano hyperspectral satellite data under Grant ZH0405-1900-01PWC, in part by the Research Foundation for Advanced Talents of Guizhou University (2016) under Grant 49, in part by the Key Disciplines of Guizhou Province-Computer Science and Technology under Grant ZDXK[2018]007, in part by the Key Supported Disciplines of Guizhou Province-Computer Application Technology under Grant QianXueWeiHeZi ZDXK [2016]20, in part by the National Natural Science Foundation of China under Grant 61462013 and Grant 61661010, and in part by the Zhuhai Innovation and Entrepreneurship Team (2017) under Grant ZH01110405170027PWC.

ABSTRACT The remote sensing (RS) images are widely used in various industries, among which semantic segmentation of RS images is a common research direction. At the same time, because of the complexity of target information and the high similarity of features between the classes, this task is very challenging. In recent years, semantic segmentation algorithms of RS images have emerged in an endless stream, but most of them are improved around the scale features of the target, and the accuracy has great room for improvement. In this case, we propose a semantic segmentation framework for RS images with dynamic perceptual loss. The framework is improved based on the InceptionV-4 network to form a network that includes contextual semantic fusion and dual-channel atrous spatial pyramid pooling (ASPP). The semantic segmentation network is an encoder-decoder structure. In addition, we design a dynamic perceptual loss module and a dynamic loss fusion strategy by further observing the loss changes of the network, so as to better improve the classified details. Finally, experiment on the ISPRS 2D Semantic Labeling Contest Vaihingen Dataset and Massachusetts Building Dataset. Compared with some segmentation networks, our model has excellent performance.

INDEX TERMS Remote sensing, semantic segmentation, perceptual loss, loss fusion.

I. INTRODUCTION

In recent years, the development of aerospace technology and sensors has provided sufficient conditions for the utilization of RS images. Therefore, the application of RS images in all walks of life has become more and more extensive, and various processing methods have become mature. It is the rapid development of artificial intelligence(AI) which makes AI RS become a hot research direction. In the past, the RS images were mostly 3-channel RGB images. Nowadays, the types of RS images are diversified, such as multispectral images, hyperspectral images, high resolution, and super-high-resolution images, etc. There are more types of data to choose from and meet different needs so that the RS task is simplified and targeted. The research in this paper is based on deep learning [1] to segment RS images.

The semantic segmentation of the RS images has many challenges. On the one hand, the imbalance of the sample size leads to insufficient training of certain object categories. On the other hand, because the RS image is an orthophoto, the target may be obscured by clouds or trees, which makes it difficult to classify targets. Thirdly, the features of the same sample are diverse, such as different materials or colors at the top of the house, which increases the difficulty of segmentation.

The semantic segmentation model based on deep learning solves the above problems to some extent, but there are still many deficiencies. Firstly, for the RS image, some objects to

The associate editor coordinating the review of this manuscript and approving it for publication was Stefania Bonafoni.

be classified are of large size, and the convolution kernel of network is small, which cannot extract the global information of the target well. Secondly, the boundary segmentation of the RS image is not fine enough, and there will be misclassification at the edge of objects. Compared with other segmentation networks, the contribution of this paper is as follows. This work studies these two issues and builds a RS image segmentation framework. Compared with other segmentation networks, the contributions made by this paper are as follows:

(1) Take the InceptionV-4 network [2] as the backbone and construct a dual Atrous Spatial Pyramid Pooling (ASPP) [3] module to form a feature extraction network. The InceptionV-4 network is a newer classification network and we transform it into a fully convolutional neural network for semantic segmentation tasks. In addition, the framework we propose introduces the context feature fusion and the dual ASPP module to solve the first problem mentioned above to the greatest extent.

(2) We analyze the loss function of the network and designed a dynamic loss fusion module based on the perceptual loss [4]. This module uses a pre-trained perceptual loss network to reduce the difference between training features and ground truth, making the edge features of the target closer to the ground truth, thereby improving the segmentation details of the image, and solving the second problem mentioned above.

II. RELATED WORK

A. SEMANTIC SEGMENTATION

The semantic segmentation is a pixel-level image classification task. That is to say, there is a corresponding category for each pixel, and the pixel-level image classification can be obtained by extracting features through training. Semantic segmentation is also a common RS image processing method. Through semantic segmentation, the goal of identifying ground objects is achieved. Traditional semantic segmentation methods include pixel-level threshold segmentation [5], cluster segmentation [6], and decision tree segmentation [7]. In recent years, deep learning has developed rapidly and has made great contributions to semantic segmentation. In particular, Long *et al.* invented the Full Convolutional Neural Network (FCN) [8], which has led to the rapid development of semantic segmentation technology. FCN has demonstrated strong learning ability in the image classification. Later, many researchers began to improve the network based on FCN and obtained more high-precision semantic segmentation methods. The latest semantic segmentation networks are mostly based on codec structures, such as SegNet [9], U-Net [10], Deeplab [11], etc. The segmentation precision of these networks is very high, and it is used in the segmentation task of various scenes.

In addition to these common segmentation networks, some optimized feature extraction modules are also proposed. The atrous convolution [12] greatly improved the effect of semantic segmentation. The atrous convolution has a larger receptive field than ordinary convolution, so that the output of each convolution operation has a larger range of characteristic information. The ASPP [3] module is composed of several convolutions with different sampling rates and pooling layers in parallel. Finally, the feature map of the convolution and pooling output is merged by a 1×1 convolution. Different sampling rates can get more different receptive fields, which can extract more scale information and improve the segmentation accuracy. In this work, we apply the ASPP to the semantic segmentation of RS images, and construct a dual ASPP module for our backbone network, which optimizes the segmentation effect.

B. VERY HIGH RESOLUTION IMAGE SEMANTIC SEGMENTATION

With the development of photogrammetry technology in recent years, the resolution of RS images is getting higher and higher, and image processing methods are also more mature. Cao textitet al. [13] added digital terrestrial model (DSM) as supplementary information to the segmentation task, explored different fusion strategies, and designed an end-to-end segmentation network. Wei et al [14] designed a semantic segmentation network with a codec structure, connected to the CRF module to improve segmentation performance. Mi and Chen [15] proposed a Superpixel-enhanced Deep Neural Forest method based on the DCNN network, and combined with the decision tree to improve the accuracy of segmentation. Jiang *et al.* [16] designed a random walk network based on SegNet (Random-Walk-SegNet), which reduces the effect of blurring on the edge, and the method has lower complexity. Audebert *et al.* [17] studied segmentation of multi-modal RS data, and proposed a method about multi-scale feature extraction, which fused the radar data and multi-spectral data to obtain a powerful segmented method. Most of these methods are based on the deep neural network, and the segmentation accuracy of RS image is high, but there is still a lot of room for improvement, especially the edge segmentation accuracy of the target needs to be improved.

C. PERCEPTUAL LOSS

Perceptual loss [4] was proposed by Johnson, J in 2016. It is used in image style transfer and super-resolution, and the perceptual loss function is used to train feedforward networks for image transformation tasks. In recent years, many researchers have applied perceptual loss to other fields.

In the direction of image reconstruction, Wen *et al.* [18] introduced detailed perceptual loss on the basis of cascaded residual blocks. By reducing detail perceptual loss, the texture details of the reconstructed image and ground truth become more and more similar, and a good single image super-resolution reconstruction effect is obtained on multiple datasets. Later, researchers applied the perceptual loss to the field of medical imaging. Yang, Q *et al.* [19] used a generative adversarial network with Wasserstein Distance [20] and perceptual loss to denoise the Low-Dose CT images, and good results were obtained in clinical CT image tests.

It can be seen that the perceptual loss has not been widely used in the semantic segmentation of RS images. We studied the principle of perceptual loss, which is to calculate the loss function using the pre-trained network, then fuse it with the loss of the feature network, and update the parameters through back propagation to achieve better learning results. Therefore, the performance of perceptual loss applied to semantic segmentation of remote sensing images is explored in this work. At the same time, a dynamic perceptual loss fusion strategy is obtained for training.

III. PROPOSED MODEL

In this work, a deep learning framework for semantic segmentation of RS images is built, which can well solve some problems encountered in the semantic segmentation of the RS images. This framework is roughly divided into four parts: The inceptionV-4 network as the backbone, dual ASPP module, decoder module, and perceptual loss network. Among them, the InceptionV-4 network and dual ASPP modules are used as the encoder, which will be described in detail next.

A. THE INCEPTIONV-4 NETWORK

The InceptionV-4 network [2] is a novel deep learning classification network proposed by Szegedy in 2017. Convolution and pooling are used in parallel in the network to prevent bottleneck problems [21]. At the same time, a 1×1 convolution kernel is also used to prevent such problems. In this paper, the InceptionV-4 network is modified as the backbone. The modified InceptionV-4 network structure is shown in Figure 1. All parts after the Inception-C module are removed and the backbone is connected to the first group of the decoders. In addition, from previous work experience, it can be inferred that the combination of shallow features and deep features will improve the classification accuracy. This is because that as the network deepens, more abstract features can be extracted, but the spatial features of the object are lost a lot. Therefore, this paper merges the output features of the stem module with the output features of the Inception-C module, which reduces the loss of the object spatial features to a certain extent. There are other ways to reduce the loss of the object spatial features, such as atrous convolution, and Atrous Spatial Pyramid Pooling (ASPP) takes advantage of this, which will be described below.

B. THE DUAL ATROUS SPATIAL PYRAMID POOLING

The ASPP [3] was first proposed on DeeplabV-2 and later improved in DeeplabV-3, adding the Batch Normalization (BN) layer compared to the previous. The ASPP uses four atrous convolutions with different sampling rates (or the receptive field), which can effectively extract multi-scale features, but the sampling rate of the convolution kernel cannot be too large. When the receptive field of 3×3 convolution kernels close to the size of the feature maps, the size of actual working filter becomes 1×1.

In this paper, the ASPP is modified to obtain the ASPP-1 and ASPP-2 according to the situation of backbone.

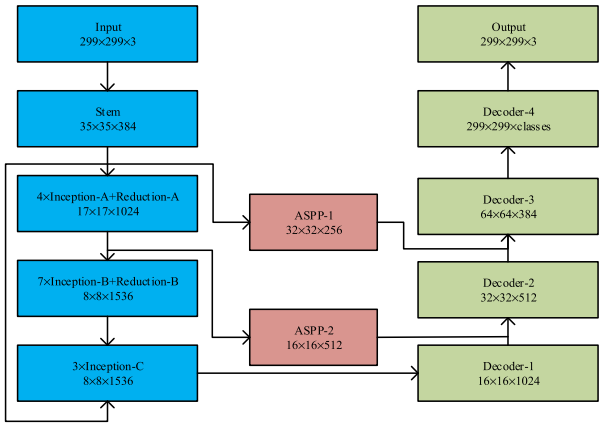


FIGURE 1. The structure diagram of the feature network.

The distinctions between ASPP-1-2 and ASPP are as follow: (1) Before inputting the feature maps into the ASPP, a maximum pooling layer is added to resize the feature maps to the size that can be input into the ASPP; (2) Generally, the ASPP is placed at the end of backbone. In this paper, ASPP-1 and ASPP-2 are embedded in different stages of backbone in order to better extract target location information, as shown in the Figure 1; (3) This work uses dual ASPP to sample on feature maps of different sizes, so the sampling rate of ASPP are also modified.

In Figure 1, the size of the feature map input in the ASPP-1 is 35×35. The sampling rate of atrous convolution in the ASPP-1 is set to 1, 6, 12 and 18, and the size of the output feature maps is $35 \times 35 \times 256$. In addition, there is a pooling layer, and then 35×35 feature maps are obtained. In this way, the ASPP modules with 5 parallel convolution and pooling layers are formed, and finally, the feature maps are obtained by combining the above convolution and pooling through a 1×1 convolution. In order to match the size of the decoder feature maps, a maximum pooling is used to change the feature map size to be $32 \times 32 \times 256$. For the ASPP-2, the size of the input feature maps is 17×17, so it is not advisable to set the sampling rate too large. We remove the atrous convolution with a sampling rate of 18, and the final ASPP-2 contains a pooling layer and atrous convolutions with sampling rates of 1, 6, and 12, respectively. The subsequent operations are the same as the ASPP-1. The structure diagram of the ASPP-1 module used in this paper is shown in Figure 2.

C. THE DECODER

In the previous work, the decoder we designed was too complicated. In this work, we design a relatively simple decoder. As shown in Figure 1, the decoder is mainly divided into four groups, which are composed of upsampling layers and convolutions. Each group of the decoder contains a bilinear upsampling layer. In addition, the first group of the decoder contains three convolutions, the second, third, and fourth groups of the decoder contain two convolutions. The size of the feature maps output by the ASPP-2 is 17×17. Because

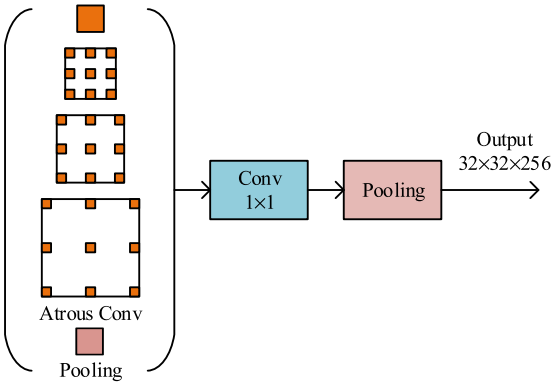


FIGURE 2. Structure diagram of the ASPP-1.

the feature maps need to be connected, the feature size needs to be consistent. Therefore, the size is changed to 16 × 16 by pooling, and then combined with the first group of the decoder to get 1536-dimensional features. Similarly, the size of the feature maps output by the ASPP-1 is 35 × 35. After pooling, the size becomes 32 × 32. Combining it with the second group of decoders to obtain a 768-dimensional feature. In some convolutions, we also use atrous convolution instead of ordinary convolution operations.

The specific parameters of the decoder are shown in Table 1. The stride of each convolution is 1, if not mentioned separately. We study from the decoder design of the SegNet, and simplify it to obtain the decoder. Each decoder group contains an upsampling operation to gradually restore the size of the feature maps. The first group of decoder contains a three convolutions, while the rest of the decoder have two convolutions. Finally, restore the feature maps to the size when the images is input into the network through bilinear upsampling.

By designing the decoder, we construct a U-shaped encoder-decoder structure network, which is called the Feature Network. This segmentation network with a codec structure is a highly recognized segmentation structure. The encoder extracts target features through multiple convolutions while reducing the size of the feature maps. The decoder uses upsampling and convolutions to gradually restore the size of the feature maps, and finally, use the Softmax classifier to obtain the classification result.

D. THE PERCEPTUAL LOSS NETWORK

The perceptual loss was first proposed by Johnson *et al.* to be used in the image style transfer. Later, the method was extended to super-resolution reconstruction and denoising of medical CT images, and achieved good results. Perceptual loss has not been effectively applied in the semantic segmentation of the RS image, so this work will build a semantic segmentation network of RS images with dynamic perceptual loss. The image style transfer network with perceptual loss is mainly divided into two parts, the first is the style transfer network, and the other is the loss network. The style

TABLE 1. Parameters of the decoder network.

Layers	Type	Kernel-Size	Output
Decoder-1	Upsample	\	16×16×2560
	Conv2d	3×3, padding=2, dilate=2	16×16×1280
	Conv2d	3×3, padding=1	16×16×1280
	Conv2d	3×3, padding=1	16×16×1024
Decoder-2	Upsample	\	32×32×1536
	Conv2d	3×3, padding=2, dilate=2	32×32×768
	Conv2d	3×3, padding=1	32×32×512
Decoder-3	Upsample	\	64×64×768
	Conv2d	3×3, padding=2, dilate=2	64×64×384
	Conv2d	3×3, padding=1	64×64×384
Decoder-4	Upsample	\	128×128×384
	Conv2d	3×3, padding=1	128×128×64
	Conv2d	3×3, padding=1	128×128×classes
	Upsample	\	299×299×classes

transfer network is responsible for the feature training of the image transfer. The parameters change with the training of the network, and the parameters of the loss network remain unchanged. The loss network generally uses a pre-trained VGG network, and because the dynamic perceptual loss module is added, two loss functions are defined in the overall framework. The feature maps obtained by the style transfer network is compared with the feature maps obtained by the loss network, while the loss is calculated. Finally, update parameters through back propagation [22].

Perceptual loss network is trained by adding a neural network on the basis of feature extraction network, which is different from the ordinary convolutional neural network. This additional network is a pretrained neural network, which can be used for transfer learning, and the remote sensing image dataset in this paper can also be used for training features by transfer learning. A pretrained neural network will also compute a loss function, since the pretrained network has learned features that are easy to generalize, and therefore, the loss will be lower. Combining the two networks, on the one hand, the training features are initialized, and on the other hand, the features that have been generalized are used to obtain better results.

In the proposed method, when feature extraction network is executed to the last layer, it will output feature maps of fixed dimensions. At this time, we input these feature maps into the pre-trained VGG19 network for training. Like the ordinary training network, the loss will be calculated. As shown in Figure 3, the feature network outputs the feature maps $Y^{'}$, which is then input into the perceptual loss VGG19 network along with the ground truth Y. The perceptual loss $L(P)$ is calculated and then fused with the loss $L(F)$ calculated by

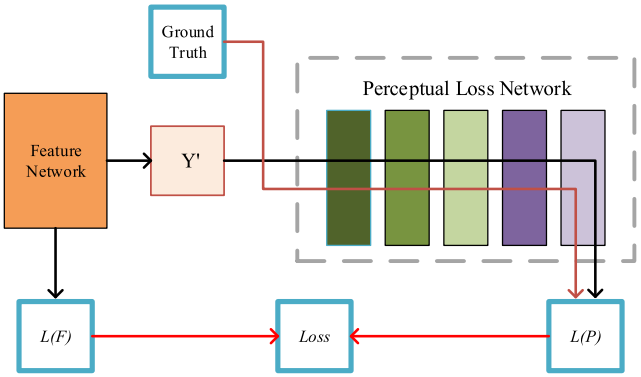


FIGURE 3. Schematic diagram of perceptual loss.

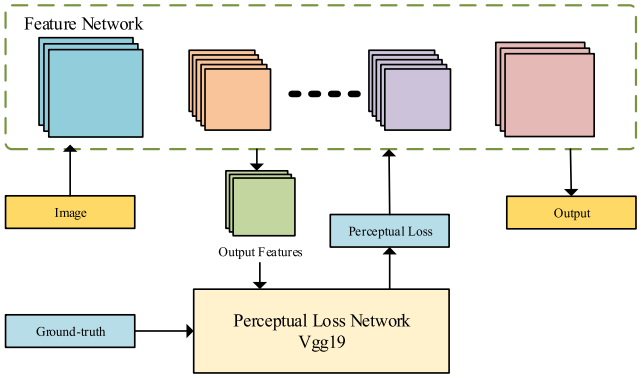


FIGURE 4. The frame structure of semantic segmentation of RS image with dynamic perceptual loss.

the feature extraction network to obtain the loss *Loss* of the whole network. Finally, parameters are updated through back propagation.

Since the pre-trained convolutional neural network has encoded the perceptual and semantic information calculated in the loss function, it has good fitting parameters, so the high-level features are more similar, which makes the training details of the network better and the performance is also improved. In this work, we use the pre-trained VGG19 network as the loss network to build a complete semantic segmentation framework for RS images with perceptual loss. The overall framework is shown in Figure 4. Firstly, input the training set to the feature network to train and extract the features, then calculate the perceptual loss by inputting the data with ground truth and the feature maps obtained by the feature network to the loss network. Finally, the parameters of the feature network are updated through backpropagation.

The main responsibility of the loss network is to calculate the difference between the predicted maps and the ground truth. Here, the mean square error function is used. Supposed that the feature network is f, the loss network is p, and the loss function is *loss*, the loss function of the loss network is:

$$loss(p) = |y - y'|^2 \tag{1}$$

where y is the ground truth and y' is the predicted maps after network p training, that is, $y' = p(x)$. x in the loss network represents the feature graph inputted into it, while in the feature network, represents the original image inputted into the feature network. The loss of feature network adopts the Negative Logarithmic Likelihood Loss Function (NLLLoss), which is often used in multi-classification tasks and is set as $loss(f)$, then the loss function of the whole network is:

$$loss = loss\,(f) + \omega loss\,(p) \tag{2}$$

where ω is the weight. If the loss function is extended to all data, the formula can be derived as follows:

$$loss = \frac{1}{N}\sum_{i=1}^{N}\sum_{j=1}^{C} loss\,(f) + \omega loss\,(p) \tag{3}$$

where N is the number of input samples, and C is the classes of the samples. At this point, the loss function of the overall framework is obtained. A loss is obtained for each training, which is propagated back to the network, and the next training is continued after updating parameters.

Here, we have studied some articles [19], [32], [33] on the application of the perceptual loss. But in these articles, the fusion of loss function is mostly simple sum of two loss functions, and the loss function formula like formula (2) is obtained. In this case, the change of loss in different training stages is not taken into account. In this paper, we take into account the change of loss function in different training stages and improve its formula as follows:

$$loss = \begin{cases} \frac{1}{N}\sum_{i=1}^{N}\sum_{j=1}^{C} loss\,(f) + \alpha \times loss\,(p), \\ (loss\,(f) > \theta_1) \\ \frac{1}{N}\sum_{i=1}^{N}\sum_{j=1}^{C} \beta \times loss\,(f) + \gamma \times loss\,(p), \\ (\theta_2 > loss\,(f) > \theta_3) \\ \frac{1}{N}\sum_{i=1}^{N}\sum_{j=1}^{C} \delta \times loss\,(f) + \varepsilon \times loss\,(p), \\ (loss\,(f) < \theta_4) \end{cases} \tag{4}$$

where $\alpha, \beta, \gamma, \delta$ and ε are the weighting parameters, θ is the threshold of the loss. Then, we can perform loss fusion dynamically at different training stages.

At present, there is no clear theoretical explanation for the weight parameters of the fusion of loss function, because different datasets and different networks have great differences in the calculated loss function, most of which are based on the existing papers and the experience of parameter adjustment. Based on the recently published papers, we expanded the simple loss fusion to carry out loss fusion with different fusion weights at different stages according to the experimental loss changes on the dataset.

IV. EXPERIMENTS

A. DATASET

The ISPRS2D Semantic Labeling Contest Vaihingen dataset [23] is a high-resolution RS dataset with complete

FIGURE 5. Overview of the vaihingen dataset and massachusetts buildings dataset.

semantic labeling. There are six classes of Impervious Surfaces, Building, Low Vegetation, Tree, Car, and background in the dataset. The data types include the True Orthophoto (TOP) and the Digital Surface Models (DSMs). The Vaihingen dataset contains 33 patches of different sizes. Each patch contains a true orthophoto and corresponding semantic annotations. The Orthophoto contains three different bands, which are Near-Infrared, Red and Green (IR-R-G). In this work, we use orthophotos composed of IR-R-G for training.

The Massachusetts Building Dataset [34], labeled buildings and backgrounds, consists of 151 aerial images collected in the Boston area, each 1500 × 1500 pixels in size. Among them, 137 images are training sets, 10 images are test sets, and 4 images are validation sets. The buildings in the dataset are all detached houses and garages. These images are in color and have three RGB bands. Figure 5 shows the overall preview of the two datasets.

B. THE EXPERIMENTAL SETUP

1) DATA EXPANSION

Since the Vaihingen Dataset is small, data expansion is necessary for better training in a deep neural network. At the same time, in order to meet the requirements of the network input for the image size, we cut the original patches into 1891 images with size of 300 × 300. Before the data expansion, about 25% of the images are randomly selected as the test set, about 5% as the validation set, and the rest as the training set. Mirror and rotate the training images randomly to obtain the expanded training set with a total of 11200 images after data partition. Because the image size of these patches is different, the image is resized to the size which can be divisible by 300 before clipping, and then resized to the size of 299 × 299 when inputted into the feature network.

For the Massachusetts Building Dataset, use the same data expansion method as for the Vaihingen Dataset to obtain a total of 13700 training images.

2) EXPERIMENTAL SETTING

After data expansion, the batch size of training data is set to 32 and inputted to the feature network. The network is implemented using the Pytorch framework and deployed on the NVIDIA Tesla V100 (32GB RAM) server with CUDA10.0. The Adam is used to train the model and realize the decay of learning rate. The initial learning rate is set to 0.0001, and L2 regularization is used to prevent overfitting. After the training, the model with the best performance was used for validation. The specific parameters are shown in Table 2.

TABLE 2. Network training settings.

Items	Value
Batch size	32
Epoch	200
Initial learning rate	0.0001
num-workers	8
parameters	133883558

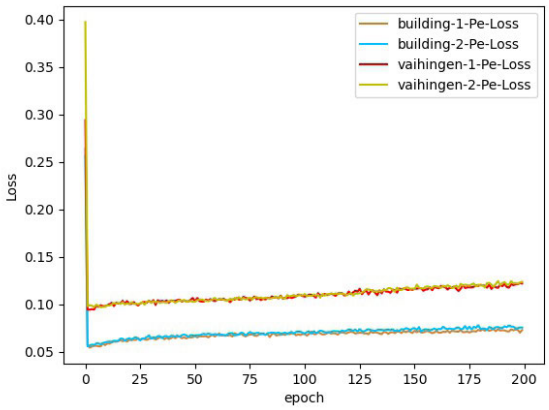


FIGURE 6. Comparison curve of perceptual loss.

As shown in Figure 6, we conduct two groups of pre-experiments on the Vaihingen dataset and Massachusetts Buildings dataset under the same conditions to obtain the perceptual loss curves of 200 epochs. Except for the large error in the first epoch, which can be ignored, the perceptual loss basically fluctuates slightly around 0.10. As the fusion of loss functions is needed, the loss calculated by the feature network needs to be observed. Figure 7 is the loss changing curve of the feature network and the loss network. It can be seen that the feature loss and the perceptual loss are getting closer and closer with the training. With the approach of feature loss and perceptual loss, simply adding the two loss functions will lead to excessive loss and fail to achieve the fitting effect. Therefore, we refer to the weight setting in literature [19], [32], [33], as well as the relative changes of two kinds of losses and the experience of parameter tuning, to determine the weighting parameters of different data sets to control the tradeoff between the feature loss and perceptual loss.

For the Vaihingen dataset, when the loss of the feature network $loss\,(f) > 0.15$, α is set as 0.1 in the formula (4);

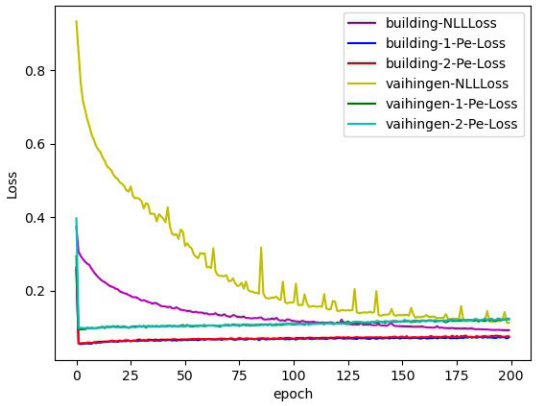


FIGURE 7. Loss comparison between the loss network and the feature network.

when $0.15 > loss(f) > 0.12$, γ is set as 0.05 and β is set as 0.95; when $loss(f) < 0.12$, ε is set as 0.01 and δ is set as 0.98. For the Massachusetts Buildings dataset, because the losses are different from the Vaihingen dataset, the weighting parameters are also different. When the loss of the feature network $loss(f) > 0.15$, α is set as 0.15 in the formula (4); when $0.15 > loss(f) > 0.10$, γ is set as 0.10 and β is set as 0.90; when $loss(f) < 0.10$, ε is set as 0.05 and δ is set as 0.95. In addition, in the later stage of the training, the loss reduction become more and more subtle. Even if the weight of the perceptual loss becomes smaller, the fusion of the two losses will make the overall loss larger. Therefore, we add a weight to the feature loss to dilute the impact of the loss fusion.

C. EVALUATION CRITERIA

In order to comprehensively evaluate the performance of the model, this work uses three benchmark indicators, namely Intersection Over Union (IOU), Overall Accuracy (OA) and F1 score (F1). The indicators are calculated as follows:

$$IOU = \frac{TP}{TP + FP + FN} \tag{5}$$

$$OA = \frac{TP + TN}{TP + TN + FP + FN} \tag{6}$$

$$1 = 2 \times \frac{P \times R}{P + R}, \quad P = \frac{TP}{TP + FP}, \quad R = \frac{TP}{TP + FN} \tag{7}$$

where TP, FP, FN, TN represent the number of true positives, false positives, false negatives and true negatives, respectively. P represents Precision, R represents Recall.

D. EXPERIMENT ANALYSIS

In this section, the performance of our method, some common semantic segmentation networks and the latest researches are compared on the Vaihingen test set and Massachusetts Buildings test set with three different indicators. The results are shown in the below tables. The compared networks are the U-Net [10], ERFNet [24], DABNet [25], SegNet [9] and so on. These types of networks are relatively mainstream semantic segmentation networks. In addition, some other comparative experiments are conducted for two datasets, which are described later.

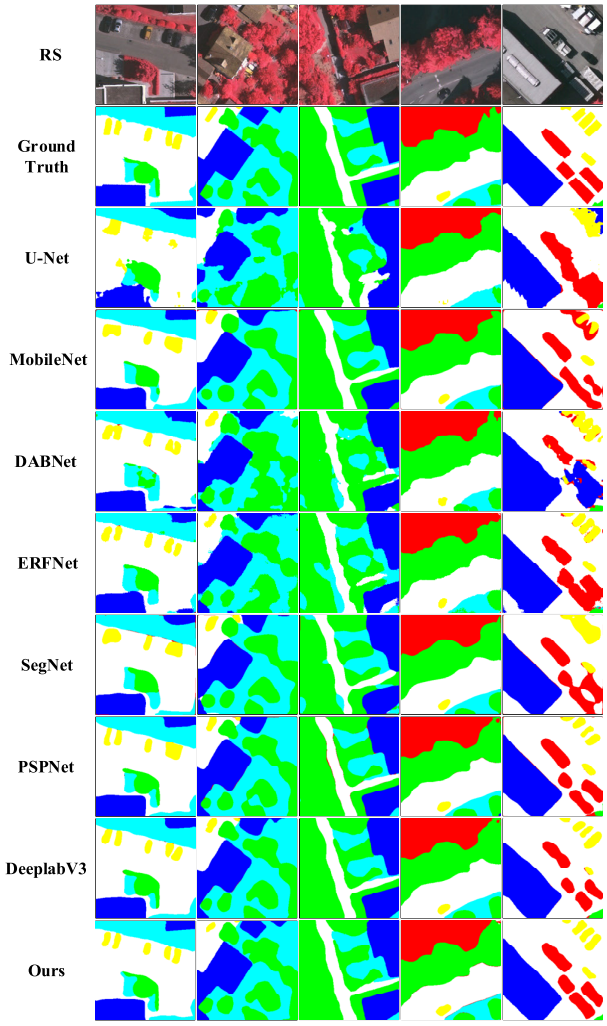


FIGURE 8. Comparison of the results on the vaihingen validation set.

1) EXPERIMENTS ON THE VAIHINGEN DATASET

In common segmentation network, the performance of DeeplabV3 is relatively excellent, with the IOU, OA and F1 scores achieving 88.31%, 93.33% and 91.92%, respectively, which meets the requirements of practical application. It can generally make segmentation for all kinds of ground targets, but the segmentation of edges is still not accurate enough, and the segmentation accuracy still has great room for improvement. The method proposed in this paper achieves extremely excellent performance on the Vaihingen dataset, with the highest IOU value of 90.30%, and the best results are also achieved on the two indicators of OA and F1, which are 92.63% and 93.48% respectively. The method in this paper greatly improves the accuracy of segmentation.

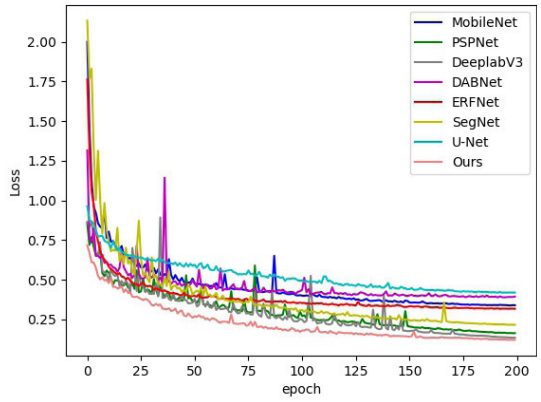


FIGURE 9. Comparison of test-loss changing curves on the vaihingen test set.

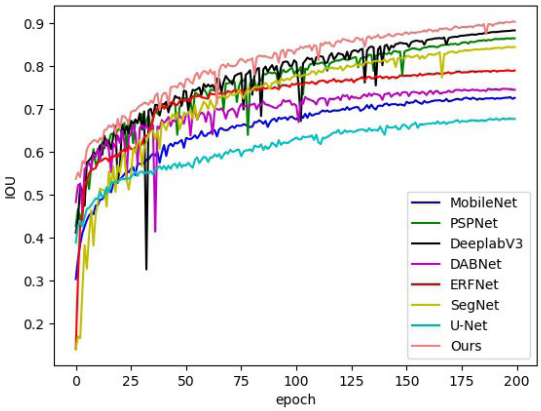


FIGURE 10. Changing curves of test IOU on the vaihingen test set.

The performance is rendered on the validation set and compared with the performance of other networks. As shown in Figure 8, we randomly selected five semantic images in the validation set for comparison. The other networks have obvious incorrect regions, while the method in this paper basically has no incorrect regions, and the image segmentation of other networks is relatively rough. Our method is more precise and accurate than other networks, and is closer to ground truth. In addition, we compare the time when loading the model and render the validation image. Because the model in this work takes up more memory than other models, the time required to call the model and render is not the shortest, but the absolute speed is not too long. It takes an average of 0.124s to render a picture, which can also meet basic needs in practical applications.

We compare the loss changes in this method during training. Figure 9 shows the test loss curve obtained by training 200 epochs in the same situation. It can be seen that the loss convergence of this method on the test set is faster than that of other networks, and the loss is also lower, which means that the resulting classification results are more similar to the ground truth. In addition, Figure 10 shows the changing curve of the IOU of each method on the test set. Similarly,

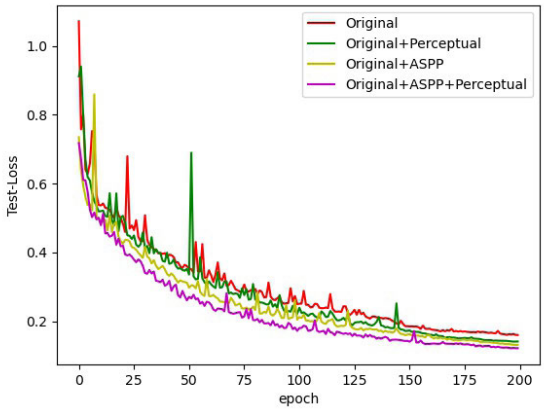


FIGURE 11. Comparison of the loss changing curves on the vaihingen test set. The original+ASPP+perceptual is the test loss when all modules are included. The Original+ASPP is the test loss when only the dual ASPP module is included. The original+perceptual is the test loss when only dynamic perceptual loss module is included. The Original is the test loss when the two modules are not included.

the method in this work achieves the highest accuracy and achieves higher values faster.

In addition, in order to prove the effectiveness of the proposed method, experiments were carried out without the dual ASPP and dynamic perceptual loss. As shown in Figure 11, after adding the dynamic perceptual loss module, the distance between the predicted maps and the ground truth becomes smaller, and the test loss also converges faster than the network without the perceptual loss module. This fully proves that our dynamic perceptual loss module is beneficial to feature extraction. In the early stage of training, since the perceptual loss module has encoded information, the distance between the predicted maps and the ground truth is smaller. Through backpropagation, the feature network can extract features more accurately, so the convergence of loss is faster. With the deepening of training, the learning of the target feature by the feature network has approached the ability of the perceptual loss network, and the rate of loss reduction has also changed. Therefore, we change the weight of each part of the loss function to achieve better training results, and the overall accuracy is also improved.

The experiment was carried out without dual ASPP module. It can be seen from Figure 11 and 12 that, if the framework does not contain dual ASPP module, the test loss is higher than that contains the module, while the accuracy is lower. It can be seen that the overall training effect will be reduced, which proves that the dual ASPP module will also have a good effect on training.

In conclusion, it can be concluded that both the dual ASPP module and Dynamic Perceptual Loss module have a positive impact on the semantic segmentation of RS images. The experimental results are shown in Table 4. The IOU, OA, and F1 scores of the network with the Dynamic Perceptual Loss module and dual ASPP module are 3.23%, 2.68%, and 1.72% higher than that of the original network, respectively.

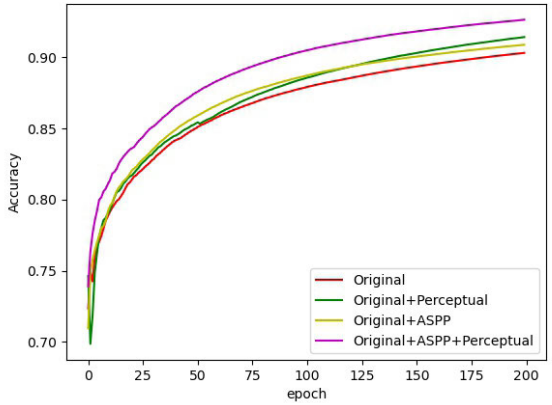


FIGURE 12. Comparison of the accuracy curves on the vaihingen test set. The original+ASPP+perceptual is the accuracy when all modules are included. The Original+ASPP is the accuracy when only the dual ASPP module is included. The original+perceptual is the accuracy when only dynamic perceptual loss module is included. The original is the accuracy when the two modules are not included.

TABLE 3. Comparison of experimental results of different networks on the vaihingen dataset.

Model	IOU(%)	OA(%)	F1(%)	Time(s)
UNet	67.64	86.46	90.52	0.103
MobileNet	72.57	86.94	91.02	**0.065**
DABNet	74.45	88.12	87.92	0.104
ERFNet	78.96	88.90	88.35	0.117
SegNet	84.42	89.58	91.81	0.127
PSPNet	86.45	91.88	92.57	0.203
DeeplabV3	88.31	**93.33**	91.92	0.170
ours	**90.30**	92.63	**93.48**	0.124

TABLE 4. Comparison of network results with or without dynamic perceptual loss on the vaihingen test set.

Dual ASPP	Dynamic Perceptual Loss	IOU (%)	OA (%)	F1 (%)
×	×	87.07	89.95	91.76
✓	×	88.91	90.52	93.07
×	✓	89.71	91.06	93.37
✓	✓	**90.30**	**92.63**	**93.48**

Finally, we compare our network with the latest research results. Most of these methods are based on mainstream segmentation frameworks, and improvements have been made on them, and they have achieved good segmentation results.

Table 5 lists the F1, overall F1 and OA scores of the five classes in the Vaihingen dataset except for the background. In this work, the F1 scores of Imp, Low and Tree classes reach the maximum value, which are 94.86%, 89.15% and 91.17% respectively. The classes of Low and Tree have similar features in the image, so it is not easy to distinguish. However, the F1 score of our method is higher than that of other methods for these two classes, which shows that the sensitivity of our method to the target boundary is higher, and the score of the class Car is also higher, so the segmentation of small targets is also very effective. The overall F1 and OA scores are also higher than other methods. Table 5 also lists the IOU scores of some methods. It can be seen that our method is better than other methods in the IOU score. By comparing the results, it is obvious that the proposed method is superior to these recently proposed algorithms, and has better performance for RS image segmentation tasks.

2) EXPERIMENTS ON THE MASSACHUSETTS BUILDINGS DATASET

After finishing the experiment on the Vaihigen dataset, in order to further verify the effectiveness of the proposed method, we conducted experiments on the Massachusetts Buildings dataset, and compared the results with some models on the dataset, as well as with the latest research results.

Table 6 records the evaluation scores obtained by each mainstream model on the Massachusetts Buildings dataset. Among them, the Deeplab-V3 has relatively high IOU, OA and F1 scores, which are 72.50%, 91.06% and 89.64% respectively, but it takes a long time to process each image. The method proposed in this paper has achieved the best performance among many methods, with the highest scores of IOU, OA and F1 (75.61%, 94.57% and 92.08%, respectively). Meanwhile, the processing time of each image is also faster than that of the Deeplab-V3, and the processing time of each image is 0.128s. Among all the methods, the time is relatively fast. Considering the accuracy and processing efficiency, the proposed method is undoubtedly the best.

In Figure 13, we randomly selected the predicted images on part of the validation set, so that the predicted effect of each method can be seen more intuitively. It can be seen that the predicted effect of the method in this paper is closest to the ground truth. Compared with other models, the result is more accurate for both small buildings and large buildings.

We compare the loss changes during training. Figure 14 shows the test loss curve obtained by training 200 epochs in the same situation on the Massachusetts Buildings test set. It can be seen that the loss convergence of our method on the test set is faster and the loss is lower. Figure 15 shows the changing curve of the IOU of each method on the test set. Similarly, the method in this work achieves the highest value and achieves the value faster. It can be seen that, compared with Vaihingen dataset, the oscillation of IOU curve on the Massachusetts Buildings dataset is more obvious, which is caused by the differences of the dataset. The number of samples in the Massachusetts Buildings dataset is small, and the training is not as sufficient as the Vaihingen dataset. However, in Figure 15, the oscillation of our method is more slight than other methods, which also reflects that the segmentation of buildings in this method is more stable than other methods.

TABLE 5. Comparison of the latest methods on the vaihingen test set (%).

Method	Imp	Building	Low	Tree	Car	F1	OA	IOU
Cao[13]	93.60	96.00	85.50	90.30	88.70	\	91.50	\
Wei[14]	92.20	93.70	86.70	89.80	79.90	89.70	90.40	\
Mi[15]	93.40	**97.60**	87.40	91.10	85.30	90.00	92.20	89.50
Jiang[16]	89.70	94.30	77.90	88.20	80.40	\	88.64	\
Nicolas [17]	91.00	96.30	87.30	88.50	**95.40**	90.60	91.10	\
Liu[26]	90.10	93.20	81.40	87.20	72.00	84.80	87.80	\
Sina [27]	90.80	94.60	81.50	88.70	83.00	\	89.00	\
Shang[28]	90.20	94.10	80.90	87.00	81.20	86.70	88.20	\
Chai[29]	90.40	94.00	80.00	87.30	72.30	\	87.90	\
Chen[30]	\	\	\	\	\	\	\	62.38
Guo[31]	\	\	\	\	\	\	\	76.15
Ours	**94.86**	96.90	**89.15**	**91.17**	85.61	**93.48**	**92.63**	**90.30**

TABLE 6. Comparison of experimental results of different networks on the massachusetts buildings dataset.

Model	IOU(%)	OA(%)	F1(%)	Time(s)
UNet	53.28	72.70	74.21	0.157
MobileNet	61.23	80.95	78.44	**0.084**
DABNet	55.37	73.40	72.61	0.213
ERFNet	62.89	80.04	80.36	0.113
SegNet	70.76	93.16	86.65	0.109
PSPNet	61.17	80.31	73.79	0.088
DeeplabV3	72.50	91.06	89.64	0.134
ours	**75.61**	**94.57**	**92.08**	0.128

TABLE 7. Comparison of the latest methods on the massachusetts buildings test set.

Method	IOU (%)	F1 (%)	OA (%)
Liu[35]	**81.80**	90.00	91.30
Shao[36]	74.46	85.36	\
Zhu[37]	71.19	84.75	\
Chen[38]	66.00	79.50	\
Kang[39]	73.93	85.01	\
He[40]	69.23	81.75	\
ours	75.61	**92.08**	**94.57**

TABLE 8. Comparison of network results with or without dynamic perceptual loss on the massachusetts buildings dataset.

Dual ASPP	Dynamic Perceptual Loss	IOU (%)	OA (%)	F1 (%)
×	×	67.42	93.42	88.73
✓	×	71.82	93.85	91.32
×	✓	72.94	94.21	90.70
✓	✓	**75.61**	**94.57**	**92.08**

We also make a comparison with the latest research. As shown in Table 7, the F1 score of our method reaches the highest 92.08%, the OA is the highest 94.57%, and the IOU also has a good result, with the value of 75.61%, which is only a little lower than [35]. In general, the method in this paper also has excellent performance.

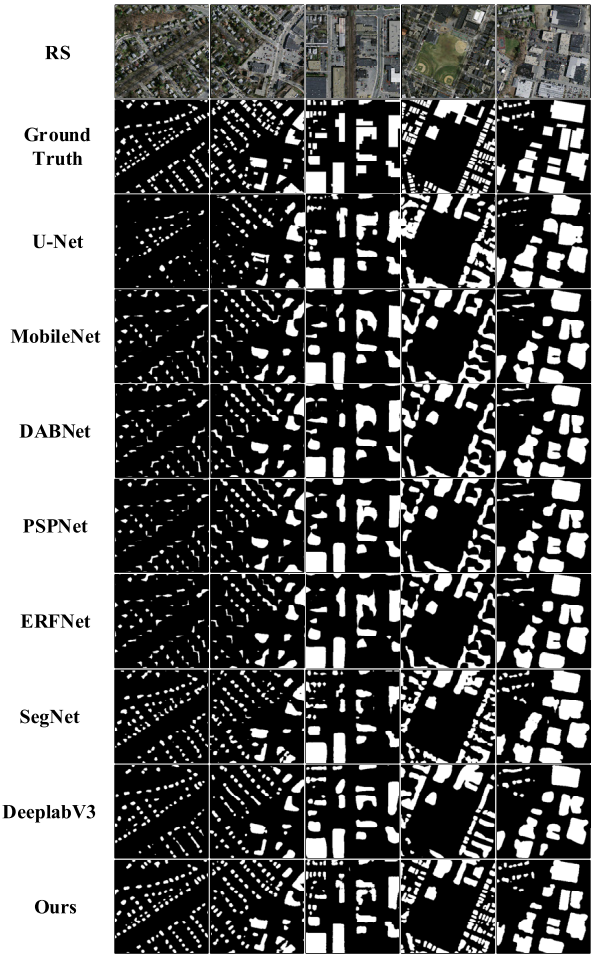


FIGURE 13. Comparison of the results on the massachusetts buildings validation set.

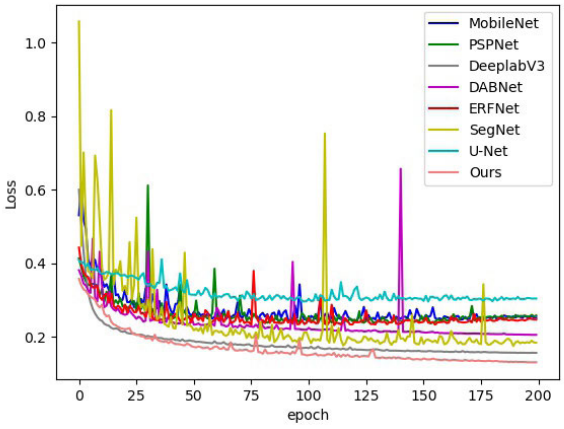


FIGURE 14. Comparison of test-loss changing curves on the massachusetts buildings test set.

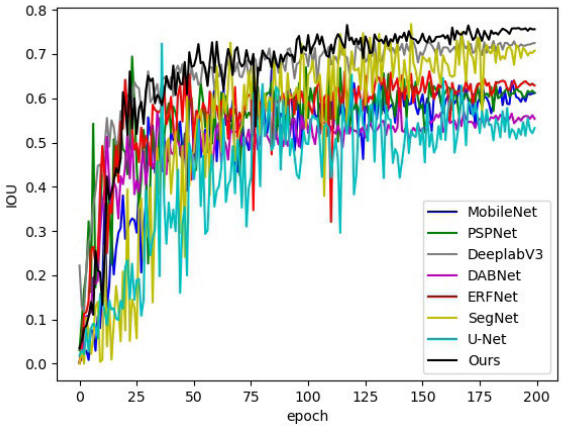


FIGURE 15. Changing curves of test IOU on the massachusetts buildings test set.

TABLE 9. Comparison of experimental results with different backbones on the vaihingen and massachusetts buildings dataset (%).

Backbone	Vaihingen			Massachusetts		
	IOU	OA	F1	IOU	OA	F1
EfficientNet	70.17	80.41	87.66	65.23	83.97	84.45
ResNet	86.19	90.82	90.56	71.52	91.73	90.86
InceptionV-4	90.30	92.63	93.48	75.61	94.57	92.08

Finally, we further verified the superiority of our method, experiments are carried out with or without Dual ASPP module and Perceptual Loss module. It can be seen from Figures 16 and 17 that the presence or absence of the Dual ASPP module and Perceptual Loss module will have an impact on the experimental results. The loss of the proposed method is lower than that of other cases. In other words, the predicted images of the proposed method are closer to the ground truth, so the OA in Figure 17 is also higher. In short, it can be concluded that the Dual ASPP module and Perceptual Loss module have a positive impact on the

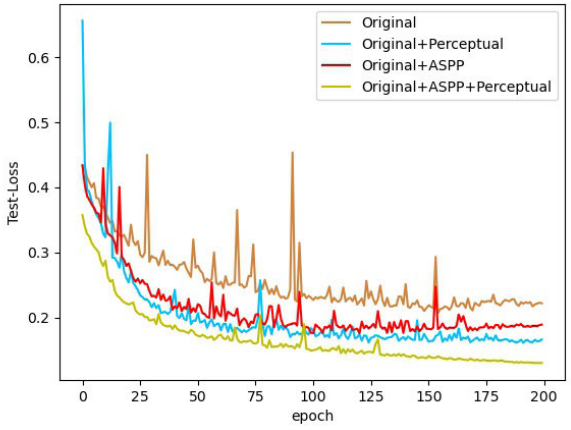


FIGURE 16. Comparison of the loss changing curves on the Massachusetts Buildings test set. The original+ASPP+perceptual is the test loss when all modules are included. The Original+ASPP is the test loss when only the dual ASPP module is included. The original+perceptual is the test loss when only dynamic perceptual loss module is included. The Original is the test loss when the two modules are not included.

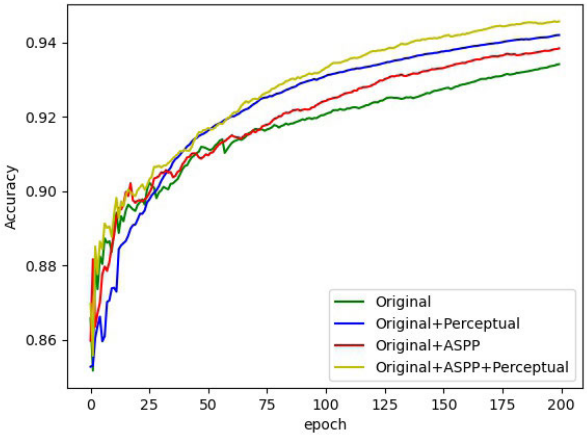


FIGURE 17. Comparison of the accuracy curves on the massachusetts buildings test set. The original+ASPP+perceptual is the accuracy when all modules are included. The Original+ASPP is the accuracy when only the dual ASPP module is included. The original+perceptual is the accuracy when only dynamic perceptual loss module is included. The original is the accuracy when the two modules are not included.

semantic segmentation of RS image. The results are shown in Table 8, the IOU, OA and F1 scores of the network with Dual ASPP module and Perceptual Loss module are 8.19%, 1.15% and 3.35% higher than those of the original network, respectively.

Experiments on two datasets show that the proposed method has good segmentation effect on remote sensing images. There are three main reasons for the effectiveness of our approach. The first is the superiority of the selected backbone. The unique design of the InceptionV-4 not only contributes to computational efficiency, but also improves classification accuracy. Secondly, the introduction of the ASPP module. Although some papers also have the ASPP, we deform the ASPP module and design a dual ASPP to extract multi-scale features according to different training

stages. Thirdly, the pre-trained VGG network is used to calculate the perceptual loss, and the Feature Loss is combined with the Perceptual Loss, which helps to improve the classification accuracy.

Finally, we used different backbones for experiment comparison. As shown in Table 9, we respectively used EfficientNet and ResNet for comparison. It can be seen that InceptionV-4 works better on both datasets. At the same time, the OA value of multi-classification is lower than that of two-classification, which is caused by the different calculation methods. The background category in the building dataset accounts for most of the sample size, the judgment accuracy of the background is higher, so its OA value is also higher. In addition, EfficientNet has a much worse performance than the others, as its network is obtained by the NAS technology. Its disadvantage is that it has good performance only for individual datasets, but its generalization ability is poor, so its performance in remote sensing datasets is not as good as other networks.

V. CONCLUSION

In this paper, we design a semantic segmentation network for RS images, aiming at the problem that the segmentation edges of RS images are not fine enough and are misclassified due to the complex ground information. This network adopts the common codec structure. We take the InceptionV-4 network as the backbone, and introduce a dual-channel atrous pyramid pooling module, using different sampling rates of atrous convolution to fully extract the multi-scale features of the target. Then a simple and effective decoder is designed, which contains four groups of convolution and upsampling to gradually restore the image size. Finally, we study the change of network loss. Using the pre-trained VGG19 network as the perceptual loss network, we design the dynamic perceptual loss module, which inputs the feature maps of the feature network to calculate the perceptual loss and then propagate back to the feature network. Experiments on the Vaihingen dataset and Massachusetts Buildings dataset show that our method has good segmentation performance. However, due to introducing the perceptual loss network, the training time will be longer, which is a disadvantage of the method in this paper. In the follow-up research, we can start from the training efficiency.

REFERENCES

[1] G. E. Hinton, "Reducing the dimensionality of data with neural networks," *Science*, vol. 313, no. 5786, pp. 504–507, Jul. 2006.

[2] C. Szegedy and S. V. A. I. V. Alemi, "Inception-v4, Inception-ResNet and the impact of residual connections on learning," *Comput. Sci.*, vol. 3, no. 6, pp. 105–112, 2016.

[3] L.-C. Chen, G. Papandreou, I. Kokkinos, K. Murphy, and A. L. Yuille, "DeepLab: Semantic image segmentation with deep convolutional nets, atrous convolution, and fully connected CRFs," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 40, no. 4, pp. 834–848, Apr. 2018.

[4] J. Johnson and A. F. A. Li, "Perceptual losses for real-time style transfer and super-resolution," in *Proc. 14th Eur. Conf. Comput. Vis. (ECCV)*, vol. 9906, Oct. 2016, pp. 694–711.

[5] H. Y. Wang and D. L. Pan, "A fast algorithm for two-dimensional Otsu adaptive threshold algorithm," *Acta Automatica Sinica*, vol. 33, no. 9, pp. 969–970, 2005.

[6] A. Coates and A. Y. Ng, "Learning feature representations with K-means," in *Neural Networks: Tricks of the Trade* (Lecture Notes in Computer Science), vol. 7700. 2012, pp. 561–580.

[7] J. Shotton and P. Kohli, "Semantic image segmentation," in *Computer Vision*, K. Ikeuchi, Ed. Boston, MA, USA: Springer, doi: 10.1007/978-0-387-31439-6_251.

[8] E. Shelhamer, J. Long, and T. Darrell, "Fully convolutional networks for semantic segmentation," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 39, no. 4, pp. 640–651, Apr. 2017.

[9] V. Badrinarayanan, A. Kendall, and R. Cipolla, "SegNet: A deep convolutional encoder-decoder architecture for image segmentation," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 39, no. 12, pp. 2481–2495, Dec. 2017.

[10] O. Ronneberger, P. Fischer, and T. Brox, "U-Net: Convolutional networks for biomedical image segmentation," in *Proc. Int. Conf. Med. Image Comput. Comput.-Assist. Intervent. (MICCAI)*, vol. 9351. Cham, Switzerland: Springer, Nov. 2015, pp. 234–241.

[11] L. C. Chen, G. Papandreou, I. Kokkinos, K. Murphy, and A. L. Yuille, "Semantic image segmentation with deep convolutional nets and fully connected CRFs," *Comput. Sci.*, vol. 2014, no. 4, pp. 357–361, 2015.

[12] F. Yu and V. Koltun, "Multi-scale context aggregation by dilated convolutions," in *Proc. 4th Int. Conf. Learn. Represent.*, 2016, pp. 2–14.

[13] Z. Cao, K. Fu, X. Lu, W. Diao, H. Sun, M. Yan, H. Yu, and X. Sun, "End-to-end DSM fusion networks for semantic segmentation in high-resolution aerial images," *IEEE Geosci. Remote Sens. Lett.*, vol. 16, no. 11, pp. 1766–1770, Nov. 2019.

[14] X. Wei, K. Fu, X. Gao, M. Yan, X. Sun, K. Chen, and H. Sun, "Semantic pixel labelling in remote sensing images using a deep convolutional encoder-decoder model," *Remote Sens. Lett.*, vol. 9, no. 3, pp. 199–208, Mar. 2018.

[15] L. Mi and Z. Chen, "Superpixel-enhanced deep neural forest for remote sensing image semantic segmentation," *ISPRS J. Photogramm. Remote Sens.*, vol. 159, pp. 140–152, Jan. 2020.

[16] J. Jiang, C. Lyu, S. Liu, Y. He, and X. Hao, "RWSNet: A semantic segmentation network based on SegNet combined with random walk for remote sensing," *Int. J. Remote Sens.*, vol. 41, no. 2, pp. 487–505, Jan. 2020.

[17] N. Audebert, B. Le Saux, and S. Lefèvre, "Beyond RGB: Very high resolution urban remote sensing with multimodal deep networks," *ISPRS J. Photogramm. Remote Sens.*, vol. 140, pp. 20–32, Jun. 2018.

[18] Z. Wen, J. Guan, T. Zeng, and Y. Li, "Residual network with detail perception loss for single image super-resolution," *Comput. Vis. Image Understand.*, vol. 199, Oct. 2020, Art. no. 103007.

[19] Q. Yang, P. Yan, Y. Zhang, H. Yu, Y. Shi, X. Mou, M. K. Kalra, Y. Zhang, L. Sun, and G. Wang, "Low-dose CT image denoising using a generative adversarial network with Wasserstein distance and perceptual loss," *IEEE Trans. Med. Imag.*, vol. 37, no. 6, pp. 1348–1357, Jun. 2018.

[20] I. Gulrajani, F. Ahmed, M. Arjovsky, V. Dumoulin, and A. Courville, "Improved training of Wasserstein GANs," in *Proc. 31st Annu. Conf. Neural Inf. Process. Syst.*, Dec. 2017, pp. 5768–5778.

[21] C. Szegedy, W. Liu, Y. Jia, P. Sermanet, S. Reed, D. Anguelov, D. Erhan, V. Vanhoucke, and A. Rabinovich, "Going deeper with convolutions," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2015, pp. 1–9.

[22] D. E. Rumelhart, G. E. Hinton, and R. J. Williams, "Learning representations by back-propagating errors," *Nature*, vol. 323, no. 6088, pp. 533–536, Oct. 1986.

[23] *International Society for Photogrammetry and Remote Sensing. 2D Semantic Labeling Contest*. Accessed: Sep. 20, 2020. [Online]. Available: http://www2.isprs.org/commissions/comm3/wg4/semantic-labeling.htm

[24] E. Romera, J. M. Alvarez, L. M. Bergasa, and R. Arroyo, "ERFNet: Efficient residual factorized ConvNet for real-time semantic segmentation," *IEEE Trans. Intell. Transp. Syst.*, vol. 19, no. 1, pp. 263–272, Jan. 2018.

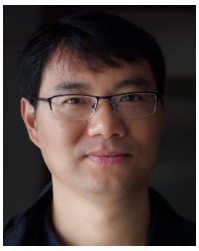
[25] G. Li and J. Kim, "DABNet: Depth-wise asymmetric bottleneck for real-time semantic segmentation," in *Proc. 30th Brit. Mach. Vis. Conf.*, Sep. 2019, pp. 1–12.

[26] Y. Liu, S. Piramanayagam, S. T. Monteiro, and E. Saber, "Semantic segmentation of multisensor remote sensing imagery with deep ConvNets and higher-order conditional random fields," *J. Appl. Remote Sens.*, vol. 13, no. 1, Jan. 2019, Art. no. 016501.

[27] S. Ghassemi, A. Fiandrotti, G. Francini, and E. Magli, "Learning and adapting robust features for satellite image segmentation on heterogeneous data sets," *IEEE Trans. Geosci. Remote Sens.*, vol. 57, no. 9, pp. 6517–6529, Sep. 2019.

[28] R. Shang, J. Zhang, L. Jiao, Y. Li, N. Marturi, and R. Stolkin, "Multi-scale adaptive feature fusion network for semantic segmentation in remote sensing images," *Remote Sens.*, vol. 12, no. 5, p. 872, Mar. 2020.
[29] D. Chai, S. Newsam, and J. Huang, "Aerial image semantic segmentation using DCNN predicted distance maps," *ISPRS J. Photogramm. Remote Sens.*, vol. 161, pp. 309–322, Mar. 2020.
[30] G. Chen, X. Zhang, Q. Wang, F. Dai, Y. Gong, and K. Zhu, "Symmetrical dense-shortcut deep fully convolutional networks for semantic segmentation of very-high-resolution remote sensing images," *IEEE J. Sel. Topics Appl. Earth Observ. Remote Sens.*, vol. 11, no. 5, pp. 1633–1644, May 2018.
[31] S. Guo, Q. Jin, H. Wang, X. Wang, Y. Wang, and S. Xiang, "Learnable gated convolutional neural network for semantic segmentation in remote-sensing images," *Remote Sens.*, vol. 11, no. 16, p. 1922, Aug. 2019.
[32] K. Xia, H. Yin, P. Qian, Y. Jiang, and S. Wang, "Liver semantic segmentation algorithm based on improved deep adversarial networks in combination of weighted loss function on abdominal CT images," *IEEE Access*, vol. 7, pp. 96349–96358, 2019.
[33] M. Li, W. Hsu, X. Xie, J. Cong, and W. Gao, "SACNN: Self-attention convolutional neural network for low-dose CT denoising with self-supervised perceptual loss network," *IEEE Trans. Med. Imag.*, vol. 39, no. 7, pp. 2289–2301, Jul. 2020.
[34] V. Mnih, *Machine Learning for Aerial Image Labeling*. Toronto, ON, Canada: Univ. Toronto, 2013.
[35] Y. Liu, L. Gross, Z. Li, X. Li, X. Fan, and W. Qi, "Automatic building extraction on high-resolution remote sensing imagery using deep convolutional encoder-decoder with spatial pyramid pooling," *IEEE Access*, vol. 7, pp. 128774–128786, 2019.
[36] Z. Shao, P. Tang, Z. Wang, N. Saleem, S. Yam, and C. Sommai, "BRRNet: A fully convolutional neural network for automatic building extraction from high-resolution remote sensing images," *Remote Sens.*, vol. 12, no. 6, p. 1050, Mar. 2020.
[37] Q. Zhu, Z. Li, Y. Zhang, and Q. Guan, "Building extraction from high spatial resolution remote sensing images via multiscale-aware and segmentation-prior conditional random fields," *Remote Sens.*, vol. 12, no. 23, pp. 1–18, Dec. 2020.
[38] M. Chen, J. Wu, L. Liu, W. Zhao, F. Tian, Q. Shen, B. Zhao, and R. Du, "DR-Net: An improved network for building extraction from high resolution remote sensing image," *Remote Sens.*, vol. 13, no. 2, pp. 1–19, Jan. 2021.
[39] W. Kang, Y. Xiang, F. Wang, and H. You, "EU-Net: An efficient fully convolutional network for building extraction from optical remote sensing images," *Remote Sens.*, vol. 11, no. 23, p. 2813, Nov. 2019.
[40] N. He, L. Fang, and A. Plaza, "Hybrid first and second order attention Unet for building segmentation in remote sensing images," *Sci. China Inf. Sci.*, vol. 63, no. 4, pp. 1–12, Apr. 2020.

WENJIE LIU received the B.S. degree in information and computing science, in 2018. He majored in software engineering with the School of Computer Science and Technology, Guizhou University. His research interests include machine learning, computer vision, and remote sensing image processing based on deep learning.



YONGJUN ZHANG received the master's and Doctoral degrees in software engineering from Guizhou University, Guiyang, China, in 2010 and 2015, respectively. He is currently an Associate Professor with Guizhou University. From 2012 to 2015, he is a joint Training Doctoral Student with Peking University and Guizhou University. He is also studying at the Key Laboratory of Integrated Microsystems, Shenzhen Graduate School, Peking University. His research interest includes intelligence image algorithms of computer vision, such as scene target detection, extraction, tracking, recognition, and behavior analysis.

JUN YAN was born in Gaomi, Shandong, China, in October 1962. He received the Ph.D. degree from Dublin City University. He is currently the Founder and the Chairman of Zhuhai Orbita Aerospace Science & Technology Company Ltd. He is also the Standing Director of Guangdong Western Returned Scholars Association, a Part-Time Professor with the Harbin Institute of Technology, the Honorary President of Bigdata College, Qingdao University of Science and Technology, and the Chief Designer of the Zhuhai No. 1 satellite constellation, which consists of 34 satellites. Since 2000, he has participated in and directed the research and development of S698 series of rad-hardened chips based on SPARC architecture, SIP cubic package modules, satellite spatial information platform, remote sensing satellite constellation, and satellite big data projects.

YONGJIE ZOU received the B.S. degree in computer science and technology from Guizhou Education University, in 2018. He majored in computer technology with the School of Computer Science and Technology, Guizhou University. His research interests include machine learning, computer vision, and remote sensing image processing based on deep learning.

ZHONGWEI CUI received the master's degree in computer application technology from Guizhou University, Guiyang, in 2008. He is currently pursuing the Ph.D. degree with Guizhou University. He has been an Associate Professor with the School of Mathematics and Big Data, Guizhou Education University, Guiyang, China, since December 2013. He has 11 years of teaching experience. His research interests include machine vision and wireless networks.