IEEE Access
Multidisciplinary | Rapid Review | Open Access Journal

Received July 3, 2020, accepted July 12, 2020, date of publication July 17, 2020, date of current version July 29, 2020.

Digital Object Identifier 10.1109/ACCESS.2020.3009976

A New Multi-Channel Deep Convolutional Neural Network for Semantic Segmentation of Remote Sensing Image

WENJIE LIU[1], YONGJUN ZHANG[1], HAISHENG FAN[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 and Technology Company Ltd., Oribita Tech Park, 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 Haisheng Fan (fan@myorbita.net)

This work was supported in part by the Research Foundation for Advanced Talents of Guizhou University under Grant (2016) No. 49, in part by the Key Disciplines of Guizhou Province-Computer Science and Technology under Grant ZDXK[2018]007, in part 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 2017 Zhuhai introduces Innovation and Entrepreneurship Team under Grant ZH01110405170027PWC.

ABSTRACT The semantic segmentation of remote sensing (RS) image is a hot research field. With the development of deep learning, the semantic segmentation based on a full convolution neural network greatly improves the segmentation accuracy. The amount of information on the RS image is very large, but the sample size is extremely uneven. Therefore, even the common network can segment RS images to a certain extent, but the segmentation accuracy can still be greatly improved. The common neural network deepens the network to improve the classification accuracy, but it has a lot of loss to the target spatial features and scale features, and the existing common feature fusion methods can only solve some problems. A segmentation network is built to solve the above problems very well. The network employs the InceptionV-4 network as the backbone and improves it. We modify the network structure and introduce the changed Atrous Spatial Pyramid Pooling module to extract the multi-scale features of the target from different training stages. Without losing the depth of the network, using Inception blocks to strengthen the width of the network can obtain more abstract features. At the same time, the backbone network is used for semantic fusion of the context, it can retain more spatial features, then an effective decoder network is designed. Finally, evaluate our model on the ISPRS 2D Semantic Labeling Contest Potsdam and Inria Aerial Image Labeling Dataset. The results show that the network has very superior performance, reaching 89.62% IOU score and 94.49% F1 score on the Potsdam dataset, and the IOU score on the Inria dataset has been greatly improved.

INDEX TERMS Semantic segmentation, neural network, remote sensing, feature fusion.

I. INTRODUCTION

With the development of RS technology, the amount of RS image data is becoming larger and the resolution is higher. RS image contains a lot of information, so there are many aspects in the application of RS image, including target detection, scene classification, semantic segmentation, and so on. The application of RS image tends to be diversified, such as urban planning [1], building extraction [2], road extraction [3], vehicle detection [4], and illegal building extraction [5]. In these fields, high segmentation quality is needed. Although there are many RS image segmentation methods, the segmentation effect still needs to be improved.

Semantic segmentation is the pixel-level classification method, which marks each pixel of an image as a certain kind of object label. There are many challenges in the semantic segmentation task of RS images. First, the RS image contains a large amount of information, but the amount of data in each sample is extremely uneven, and the samples in different scenes are diverse, which puts forward high requirements for segmentation methods. Second, because the RS images are taken vertically from high altitude, some samples will

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

overlap or occlude, such as the occlusion of trees on vehicles, resulting in feature differences in vehicle extraction. Third, the same kind of samples also have different characteristic information, such as the different color trees in the forest, and the top color of buildings will be greatly different, which will bring great challenges to segmentation. Fourth, because of the different angles of the sun, there will be a lot of shadows in the image, which is the noise of the image. Therefore, numerous researchers pay attention to the segmentation of RS images and put forward a lot of segmentation methods.

In recent years, artificial intelligence has made rapid development and shown strong productivity. Deep learning is also increasingly applied to the processing of RS images, especially RS image classification. At the same time, the rapid development of high-performance computing equipment and computer software technology provides a guarantee for the application of deep learning in RS image classification. The application of deep learning in RS images can reduce the dependence of RS image classification on expert knowledge, improve classification accuracy and recognition efficiency. Consequently, the combination of deep learning and RS images is of great significance.

The main purpose of the current study is to establish a deep learning network for semantic segmentation of high-resolution RS images. So, a multi-channel segmentation network (DAPN) is established with dual Atrous Spatial Pyramid Pooling (ASPP) [6] to segment wholesale RS images.

II. RELATED WORK

Over the past few decades, most research in RS has emphasized the employ of machine learning. Firstly, texture [7] and geometry features of ground objects are extracted. Then, for vegetation and water body, it is necessary to further study the reflectivity of ground objects, Normalized Difference Water Index (NDWI) [8], [9].

The features of RS images mainly include spectral, spatial and texture features. Shao *et al.* [10] proposed two improved texture descriptors for RS image classification, which are color Gabor wavelet texture (CGWT) and color Gabor opponent texture (CGOT), and the experimental results show that the performance is better than other texture features. Huaiying [11] adopted a shadow detection method based on a statistical hybrid model to solve the problems of a high reflection area and false positives in the presence of water. These methods are based on machine learning algorithms, but there are still some defects, over-reliance on expert knowledge, recognition efficiency is not high, so it is necessary to study a more efficient and stronger generalization classification algorithm.

In recent years, many researches on deep learning and computer vision have been published, and various high-performance deep learning algorithms have emerged. Image processing based on deep learning has gradually become a trend in the whole field of computer vision. Deep learning was firstly proposed by Hinton and Salakhutdinov [12] in 2006. They built the multi-level structure model of automatic coding, and then extended the depth confidence network based on Restricted Boltzmann Machine (RBM) [13]. Razavian *et al.* [14] applied the convolution neural network algorithm to train pixel feature classifiers, but the accuracy is low because of the shortcomings of traditional segmentation methods. In 2014, GoogleNet [15] won the championship in the ImageNet Large Scale Visual Recognition Challenge (ILSVRC) competition, and VGGNet [16] won the second place in the competition, while made remarkable achievements in image transfer. In 2015, Long *et al.* [17] achieved pixel-level image classification based on a full convolution neural network (FCN), and then many researchers improved it based on FCN. Later, He *et al.* [18] detected objects in the image and generated a segmentation mask for each object, which is called Mask R-CNN. In 2016, Szegedy *et al.* [19] improved the Inception-V3 network to make the network deeper and wider to build the Inception-V4 network. Besides, due to the extension of the codec structure, many FCN-based codec semantic segmentation networks have been built, such as U-Net [20], SegNet [21], Deeplab [22] and so on. These are recently published semantic segmentation networks, and the segmentation effect is much better than the former methods. The application of atrous convolution [23] to FCN enlarges the sampling of the feature image, thus the receptive field is expanded, and the segmentation effect is improved obviously. Lafferty and Mccallum [24] applied conditional random fields to optimize the results of FCN output [25], which became a commonly used method for segmentation post-processing.

Numerous deep learning methods are applied to the semantic segmentation of RS image. Reference [26] used the DCNN framework for semantic segmentation of multi-spectral imagery (MSI) images, which overcomes the label scarcity of MSI data and achieves a good segmentation effect. Reference [27] combining the RGB feature information obtained in the deep learning framework with the optical detection and ranging (LiDAR) features, formed a multi-sensor decision fusion technique, which is applied to mark the LiDAR data and RGB data semantically. Finally, introduce the high-order conditional random field framework to improve the semantic tagging. Reference [28] combined DCNN with decision-making forest, and introduced a super-pixel enhancement region module to further enhance the edge information of the target. Reference [29] proposed a spatial residual module (SRI) to continuously fuse multi-level feature extraction multi-scale information, which shows significant segmentation improvement compared with several latest FCN models. Reference [30] designed the Web-Net which is a layered and densely connected nested network structure. A super-layered sampling block (UHS) is inlaid to integrate the feature map of each layer, and finally identify the building area more accurately.

Although the above segmentation methods are recently published and introduced some methods to improve the

performance of segmentation, it cannot extract enough features to classify, so the segmentation effect still needs to be improved. Given this situation, the DAPN is established, which takes into account the depth training of the network and the preservation of shallow features, on the basis of which more multi-scale features are extracted. As a result, they obtain higher segmentation accuracy.

III. PROPOSED MODEL

The DAPN is a codec structure, in which the encoder module downsampling and extracts the abstract features of the image through the multi-layer convolution module. Then the upsampling is gradually restored to the original size of the image by the decoder module. It will be covered in detail in this section.

A. MODULE OF THE ENCODER

1) BACKONE

The DAPN takes the InceptionV-4 network as the backbone and adds dual ASPP modules as the encoder of the network. The InceptionV-4 is a popular deep learning network recently, which comes from the improved InceptionV-3 network [32]. Over the past few years, there have been many popular full convolution neural networks, such as VGG, ResNet and YOLO. Compared with these mainstream networks, the InceptionV-4 has higher classification accuracy and less model memory. In the test set of the ImageNet Classification Challenge, the Top-5 error rate is 3.08%, and the network is deeper. Our RS images have more information, it is necessary to extract more abstract features for learning and classification. Therefore, the InceptionV-4 network is used as the network backbone. Compared with the InceptionV-3 network, the conspicuous difference of InceptionV-4 is the Stem module and the Reduction-B module, which adopts more skills to reduce the calculation of the model. In order to obtain feature maps, it modifies the Stem module to make the structure more complex and the network level deeper. The Reduction module changes the width and depth of the network, and improves the bottleneck problem without adding too much network depth. Because using the parallelism of convolution and pooling to prevent bottleneck problems has been mentioned in the InceptionV-3, convolution and pooling parallelism is used again in the InceptionV-4.

Figure 1 is the model structure of the InceptionV-4 network. Firstly, input the image with 299 × 299 sizes into the Stem module, and in this module, through parallel groups of convolution layers and pooling layers, the feature map of 35 × 35×384 is passed into the Inception module, that is, to achieve the purpose of preprocessing. The Inception-A, Inception-B, Inception-C module have the same structure as in the InceptionV-3. The Reduction-A and Reduction-B modules reduce the size of the feature graph. In Reduction-B, asymmetric convolution and pooling parallel strategies are applied to reduce the calculated amount. The structure of the Reduction module is shown in Figure 2.

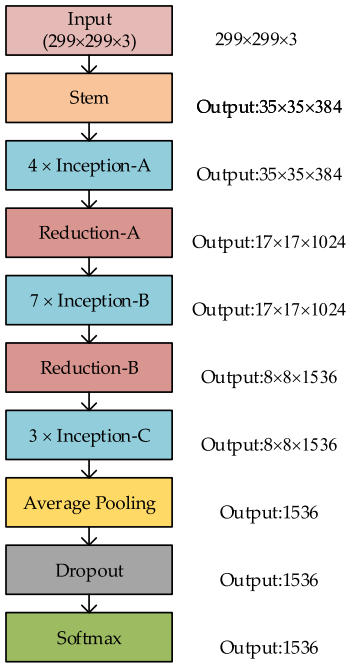

FIGURE 1. The InceptionV-4 architecture diagram.

2) ATROUS SPATIAL PYRAMID POOLING

In addition to apply the Inception network to extract features, the ASPP module is applied to further extract multi-scale features of images [31]. To put it simply, the ASPP is composed of multiple parallel atrous convolutions [23], and fuses the convolution blocks of the feature maps, that is, a spatial pyramid structure with atrous convolution. The ASPP proposed firstly in the deeplabv2 that the mapping at the top of the feature uses four kinds of atrous convolution with different dilated rates. The verification results show that convolution with different dilated rates is effective. Compared with ordinary convolution, atrous convolution enlarges the area of the receptive field due to the difference of dilated rate, and a larger range of information can be obtained in each convolution, while parallel convolution with dilated rates of 6, 12 and 18 are used in the ASPP, so more scale features can be obtained. As shown in Figure 3, the upper feature maps are inputted into a module containing parallel operations of four convolutions and a pooling and extract the multi-scale features through the characteristics of four convolutions (where the dilated rates of the four parallel convolutions are 1, 6, 12, 18 respectively). At the same time, the global average pooling operation is performed on the feature maps, and then combine the outputs of the five parallel operations, and finally through a 1 × 1 convolution. This is the overall flow of the ASPP module. By introducing this module, our network can extract sufficient multi-scale features, and strengthen the whole network training. In our work, a dual ASPP module is built, which will be described in detail in Section 3.3.

B. DECODER

In order to restore the feature size without losing the local information of the image, a simple decoder mod-

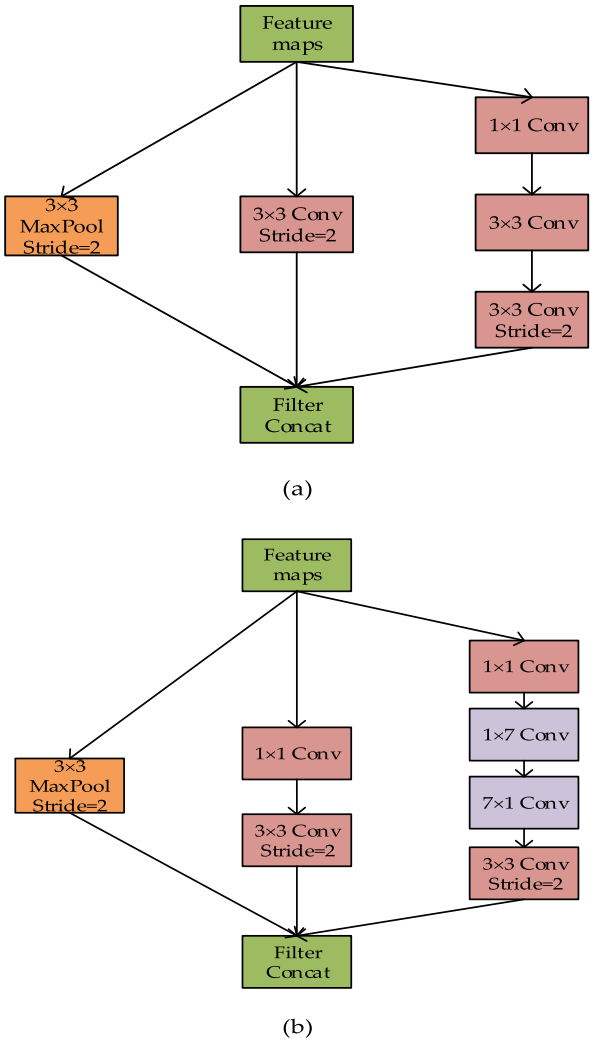

FIGURE 2. Reduction module structure diagram, (a) is Reduction-A structure diagram, (b) is Reduction-B structure diagram.

ule is designed, which receives the feature maps from the encoder, and then through four sets of convolution modules. Finally, the image is restored to its original size by a bilinear upsampling. Each set of convolution modules contains three complete convolution operations, and the last decoding block includes an upsampling, a convolution, and a bilinear upsampling.

C. MODEL STRUCTURE

The above two sections describe the basic modules, and this section will describe the architecture of the overall model in detail. The DAPN architecture is shown in Figure 4, which takes the InceptionV-4 network as the backbone, abandons its final Average Pooling, Dropout and Softmax, while introducing the dual-ASPP module, and then combine the features to form the encoder module. After the Stem module, establish the first ASPP module to form a parallel training network. In the first ASPP module, introduce 35 × 35 × 384 feature maps, and fully extract the multi-scale features of the first

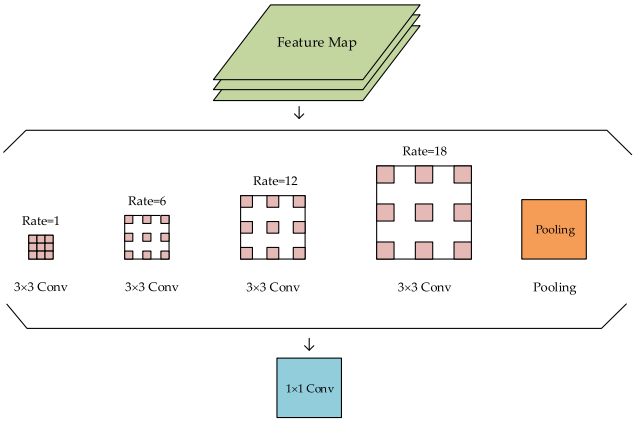

FIGURE 3. The ASPP structure diagram.

stage through the atrous convolutions with the dilated rate of 1, 6, 12 and 18 respectively. Then combine the feature maps of the five branches of the first ASPP module to form the 35 × 35 × 256 feature maps, and the combined feature maps are convoluted through a 1 × 1 convolution. In order to match the size of the feature maps to be fused, add the max pool operation of 4 × 4 after the 1 × 1 convolution, and the size of the feature maps is reduced to 32 × 32. The second ASPP module is established after the Reduction-A module of the InceptionV-4 network, and it receives 17 × 17 × 1024 feature maps. To match the size of the feature maps, the dilated rate of convolution in the second ASPP module is respectively set to 1, 6, 8, 12, and then combine the four branches of the ASPP module. Through a 1 × 1 convolution layer, the feature maps of 17 × 17 × 512 is outputted, and to match the size of the fused feature map. The max pool layer of 2 × 2 is added after the convolution layer, obtain the feature maps of 16 × 16 × 512. After completing the training of the dual ASPP module, fuse the feature images of the corresponding size in the decoder. In addition, another branch of the network is constructed. Considering that the degree of fusion training and the learned features in the earlier stage may not be enough, combine the feature graph of Inception-A module with the output feature graph of Inception-C to form an encoder with multi-channel training branch, which can fully extract the context information of the network. The decoder module is divided into four convolution modules, each convolution block has an upsampling operation, and finally, the image size is restored by bilinear upsampling. The parameter configuration of the overall network is shown in Table 1,2,3,4.

Compared with the traditional InceptionV-4 network, the DAPN uses this network as the backbone and modifies it to make the network more complex, while improving the training quality. On the basic of the InceptionV-4 network, remove the last two layers, combine the contextual semantic information of the network, embed the dual ASPP module into the backbone, and finally establish the decoder module.

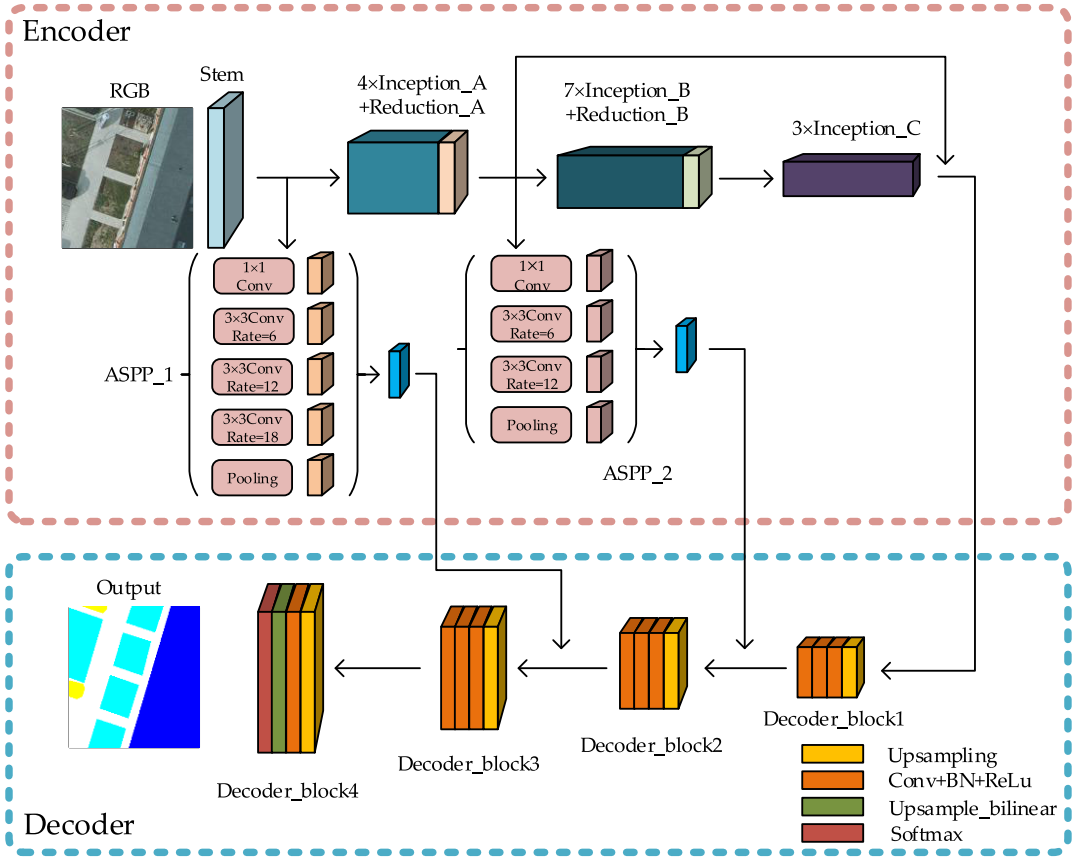

FIGURE 4. The architecture of DAPN.

TABLE 1. The configuration parameters of backbone.

Layer	Type	Output Shape	Connect to
Block_1	Stem	35×35,384	Block_2, ASPP1,
Block_2	Inception-A×4+Reduction-A	17×17,1024	Block_3, ASPP2, Decoder_1
Block_3	Inception-B×7+Reduction-B	8×8,1536	Block_4
Block_4	Inception-C×3	8×8,1536	Decoder_1

TABLE 2. The configuration parameters of the first ASPP.

Layer	Type	Kernel Size	Output Shape	Connect to
Branch_1	Conv	1×1	35×35,256	Conv_1
Branch_2	Atrous_Conv	3×3, rate=6	35×35,256	Conv_1
Branch_3	Atrous_Conv	3×3, rate=12	35×35,256	Conv_1
Branch_4	Atrous_Conv	3×3, rate=18	35×35,256	Conv_1
Branch_5	Pooling	1×1	35×35,256	Conv_1
Conv_1	Conv	1×1	35×35,256	Pooling_1
Pooling_1	Pooling	4×4	32×32,256	Decoder_3

The common ASPP module at the end of the encoder, which only expands the receptive field to extract multi-scale features, but in RS images with a lot of information, this operation is not enough to extract more features. Due to the deepening of the network, the shallow features of the target are seriously lost. That is, the location information of the target is seriously lost, so the prediction effect of the model still needs to be improved. By adding a dual ASPP module, a new semantic segmentation network is constructed. The DAPN adds dual ASPP in the first two stages of the network so that the location information of the data will not be lost, and multi-scale features can be extracted from feature maps of different sizes through different receptive fields. Then combine with the corresponding modules of the decoder. So, on the one hand, the context information of the network is fused together to ensure that the shallow features of the network will not lose too much. On the other hand, more multi-scale features can be obtained through the dual ASPP module. Finally, to strengthen the learning ability of the network, fuse the backbone network. In order to verify the good segmentation performance of the DAPN, we have carried out experiments on ISPRS 2D Semantic Labeling Contest Potsdam and Inria Aerial Image Labeling Dataset.

TABLE 3. The configuration parameters the second ASPP.

Layer	Type	Kernel Size	Output Shape	Connect to
Branch_5	Conv	1×1	17×17,512	Conv_2
Branch_6	Atrous_Conv	3×3, rate=6	17×17,512	Conv_2
Branch_7	Atrous_Conv	3×3, rate=12	17×17,512	Conv_2
Branch_8	Pooling	1×1	17×17,512	Conv_2
Conv_2	Conv	1×1	17×17,512	Pooling_2
Pooling_2	Pooling	2×2	16×16,512	Decoder_2

TABLE 4. The configuration parameters of the decoder.

Layer	Type	Kernel Size	Output Shape	Connect to
Decoder_1	Upsampling	\	16×16,2560	\
	Conv	3×3, padding=1	16×16,1280	\
	Conv	3×3, padding=1	16×16,1280	\
	Conv	3×3, padding=1	16×16,1024	Decoder_2
Decoder_2	Upsampling	\	32×32,1536	\
	Conv	3×3, padding=1	32×32,768	\
	Conv	3×3, padding=1	32×32,768	\
	Conv	3×3, padding=1	32×32,512	Decoder_3
Decoder_3	Upsampling	\	64×64,768	\
	Conv	3×3, padding=1	64×64,384	\
	Conv	3×3, padding=1	64×64,384	\
	Conv	3×3, padding=1	64×64,384	Decoder_4
Decoder_4	Upsampling	\	128×128,384	\
	Conv	3×3, padding=1	128×128,64	\
	Conv	3×3, padding=1	128×128, num	\
	up_bilinear	\	299×299	softmax

IV. EXPERIMENTS

A. DATASET AND PREPROCESSING

1) ISPRS 2D SEMANTIC LABELING CONTEST POTSDAM DATASET

ISPRS 2D Semantic Labeling Contest Potsdam dataset [32] is a high-resolution aerial image dataset with complete semantic markings in the International Society for Photogrammetry and RS, including high-resolution true orthophoto (TOP) and digital surface model (DSM). Image files are composed of different channels, there are IRRG (IR-R-G,3 channels), RGB (R-G-B, 3 channels) and RGBIR (R-G-B-IR, 4 channels) three kinds of image format respectively. In this section, only use TOP RGB images for training. The dataset contains 38 RS patches (6000 × 6000), and each patch is extracted from orthophoto images, of which 24 images have the corresponding semantic label. Dataset labels are divided into six categories, including Impervious Surfaces, Building, Low Vegetation, Tree, Car, and background.

Cut 24 images into 299 × 299 size images, due to the depth of the network is deep, the data volume is too small to get enough features. Therefore, data augmentation is carried out to reduce the impact on training. After cropping, flip the image horizontally and then vertically, obtain 76800 images of 299 × 299 sizes by rotation, which ensures sufficient training data. Select randomly 75% of the total samples as a training set, 20% as the test set, and the rest is used as a validation set. The effect prediction is made on the validation set after deriving the model.

2) INRIA AERIAL IMAGE LABELLING DATASET

The Inria Aerial Image Labeling Dataset [33] contains high-resolution RS data of five areas. There are Austin, Chicago, Kitsap County, Western Tyrol, Vienna, and each with 36 orthophoto images of 5000 × 5000 sizes, the image band combination is RGB. The semantic labels of the dataset can be divided into architectural and non-architectural (background) categories.

Cut each image into 400 images size of 250 × 250, and then use bilinear interpolation to resize the image into 299 × 299. All the images are flipped horizontally and vertically, and then rotated to get the final training set, each area contains 115200 RS images. Selected randomly 75% of the total samples as trainset, 20% as a test set, and the rest is used as the validation set. The effect prediction is made on the validation set after deriving the model.

B. EVALUATION FUNCTION

In order to comprehensively evaluate the performance of the proposed model, use Intersection over Union (IOU), Overall Accuracy (OA), F1, Precision and Recall to evaluate the experimental results. The above evaluation indicators are frequently used in previous papers, and they are compared with the recognized semantic segmentation evaluation indicators. The calculation formulas of each evaluation index are as follows:

$$IOU = \frac{TP}{TP + FP + FN} \tag{1}$$

$$OA = \frac{TP + TN}{P + N} \tag{2}$$

$$Precision = \frac{TP}{TP + FP} \tag{3}$$

$$Recall = \frac{TP}{TP + FN} \tag{4}$$

$$F1 = 2 \times \frac{Precision \times Recall}{Precision + Recall} \tag{5}$$

where P is the number of positive samples, N is the number of negative samples, TP is the number of positive samples that predict correctly, FP is the number of positive samples that predict falsely, TN is the number of negative samples that predict correctly, FN is the number of negative samples that

predict falsely, and the number of samples is the number of pixels in each picture.

C. TRAINING OPTIMIZATION

In the present study, training samples have multiple classifications and binary classifications, so use the Negative Log-Likelihood Loss function (NLLLoss). It is applicable to both multi-class and second class.NLLLoss is similar to the Cross-Entropy Loss function, but the Cross-Entropy Loss function is more widely used. In the classification task, predicted label y is discrete categories, and the output target $f(x, \delta)$ of the model is the conditional probability for each category. Suppose $y \in \{0, 1, \ldots, N-1\}$, the conditional probability of the *ith* class predicted by the model is the formula (6):

$$P(y = i|x) = f_i(x, \delta) \tag{6}$$

then $f(x, \delta)$ meet the formula (7):

$$f_i(x, \delta) \in [0, 1], \quad \sum_{i=1}^{N-1} f_i(x, \delta) = 1 \tag{7}$$

So $f_y(x, \delta)$ can be seen as a likelihood function of category y, and take the negative logarithm to get the negative log-likelihood loss function:

$$L(y, f(x, \delta)) = -\log f_y(x, \delta) \tag{8}$$

that is

$$L(y, f(x, \delta)) = -\sum_{i=1}^{N-1} y_i \log f_i(x, \delta) \tag{9}$$

Due to the large training data and the limited computing power of the computer, use the Adam (Adam optimization algorithm) algorithm for training optimization. It combines the optimal performance of the two optimization algorithms AdaGrad and RMSProp, and it is easier to adjust the parameters, while it has high computational efficiency and can adapt to large datasets. The training acceleration effect on a large amount of training data is obvious.

D. EXPERIMENTAL RESULTS

1) IMPLEMENTATION SETTINGS

The DAPN is implemented using the Pytorch framework. Train our models using the adaptive learning rate algorithm Adam with a learning rate of 0.0005 to converge the model quickly, while with a momentum of 0.9. In addition, use L2 regularization with a weight decay of 0.0001 to avoid over-fitting. The model is deployed on NVIDIA Tesla V100 (32GBRAM) server with CUDA10.0, and train 200 epochs with a batch size of 32. After training, the model with the best evaluation index is selected for testing.

2) COMPARISON OF COMMON MODELS

Compare the performance of four common models on two datasets. These models are common codec structures and fusion network for segmentation. Because the number of samples of each category is not balanced, the IOU score and the F1 score are selected to compare. Table 5 and 6 show the IOU, F1 scores of all kinds of samples on the Potsdam dataset. The results show that the segmentation accuracy of the DAPN is much better than that of other models, with an overall IOU score of 89.62% and a total F1 score of 94.49%. The Inria dataset has only two categories, in which the segmentation target is the building. Experiments are carried out on the dataset of five regions, and only the IOU score is calculated. The table shows that the performance of our model is the best in each regional test set. Then predict on the validation set and compare it with the prediction results of several mainstream models. Figure 5 and Figure 6 show the comparison of the marked effects of various common segmentation models on the two datasets. The marked effect of the DAPN is better than that of other models.

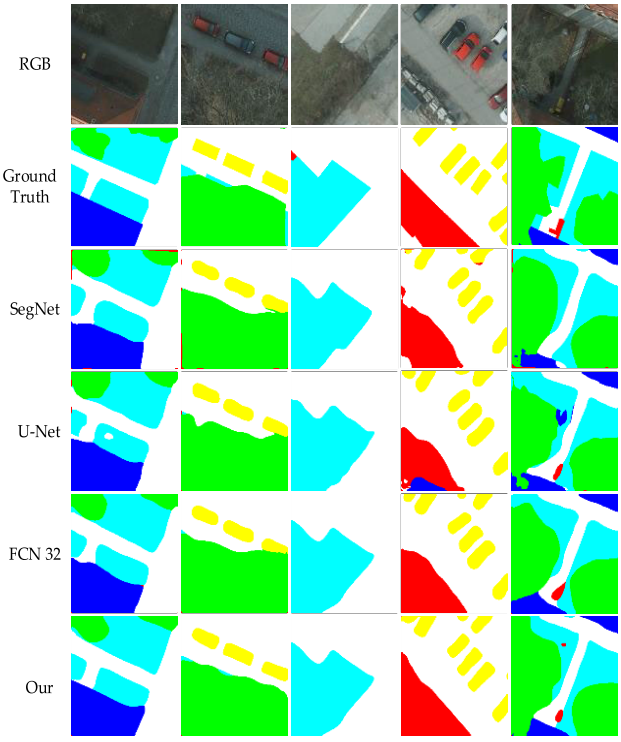

FIGURE 5. Comparison of marked effects with common models on Potsdam dataset.

As shown in Table 8, we also calculate the FLOPs of the DAPN compared to common classification networks with the same inputs. The FLOPs is the floating point operations which can be comprehended the calculated amount. Networks with a similar FLOPs do not necessarily perform at the same speed, so it is not an absolute measure, but because it measures the complexity of the network, it is also a reference. The DAPN have the highest params and FLOPs, among them, params of the VGG and the DAPN are similar. In addition, the processing time of each image is calculated in the test set and validation set. On the test set, many evaluation indicators

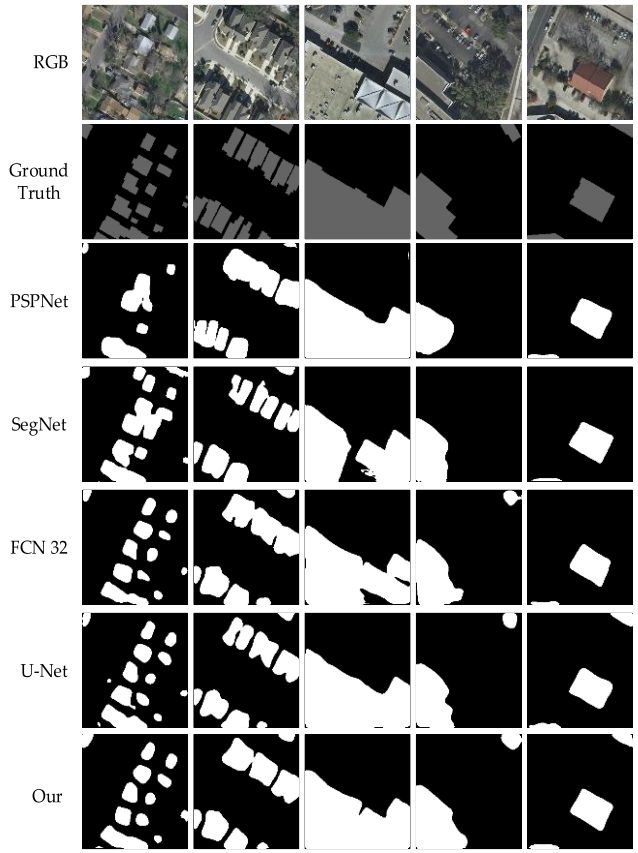

FIGURE 6. Comparison of marked effects with common models on Inria dataset.

TABLE 5. The IOU scores compared with common models on Potsdam test set.

	Imp.S.	Build.	Low.V.	Tree	Car	Overall
FCN32	89.62	92.14	73.96	64.36	68.63	78.53
SegNet	87.07	91.33	69.10	60.25	60.79	73.71
PspNet	64.09	68.57	46.86	38.03	39.41	50.60
U-Net	89.79	91.07	72.95	61.62	62.97	75.12
Our	97.23	97.26	84.42	84.71	84.47	89.62

TABLE 6. The F1 scores compared with common models on Potsdam test set.

	Imp.S.	Build.	Low.V.	Tree	Car	Overall
FCN32	85.15	94.42	79.26	76.51	80.16	84.25
SegNet	82.65	89.76	76.04	73.74	71.63	82.69
PspNet	71.39	75.58	65.71	45.32	40.61	63.36
U-Net	83.49	90.14	78.36	75.41	73.10	83.20
Our	96.97	97.25	89.97	86.32	93.74	94.49

need to be calculated, while on the validation set, only classification prediction is needed, so the runtime of test is slightly longer than that of prediction. In general, the predicted time is very short, and the predicted efficiency of remote sensing images is appreciable.

TABLE 7. The IOU scores compared with common models on Inria test set.

	Austin	Chicago	Kitsap	Tyrol	Vienna
FCN32	83.10	89.15	73.90	87.07	90.55
SegNet	80.83	86.79	72.31	83.74	88.43
PspNet	79.03	84.92	70.96	82.03	86.43
U-Net	85.94	91.06	78.24	87.36	92.64
Our	91.83	92.38	83.06	87.94	94.96

TABLE 8. The comparison of params and FLOPs for networks.

	Params	FLOPs(G)
AlexNet	61100840	1.20
Vgg16	138357544	26.99
ResNet101	44549160	14.54
InceptionV-3	27161264	5.73
DenseNet161	28681000	13.49
Our	133883558	70.57

3) EXPERIMENTAL COMPARISON OF THE LATEST SEGMENTATION METHODS

With the rapid development of deep learning, there are more and more methods to combine RS image classification with artificial intelligence, and the segmentation effect is greatly improved. Compare it with some RS image semantic segmentation methods. Some of the methods are as follows:

Audebert *et al.* [34] proposed a three-stage segment-before-detect method. Firstly, the full winder neural network is used to infer the semantic segmentation of the pixel-level classification mask, then the boundary box of the connection part is used for vehicle detection, and finally, the traditional convolution neural network is used for target-level classification.

Wang *et al.* [35] used the deep residual network as the encoder and combines two proportional high-level features and the corresponding low-level features into a decoder to further develop the multi-scale loss function and enhance the learning process. finally, the final segmentation post-processing technique of conditional random field based on superpixel is added to improve the segmentation effect.

Zhang *et al.* [36] studied the role of each feature layer in FCN, proposes an effective fusion strategy, quantifies the sensitivity of multimodal data through recall rate and recall decline rate in the multi-resolution model, analyzes the influence of different modes on pixel prediction, and explains the reasons for poor performance caused by common fusion. Finally, propose an optimization scheme of fusion elevation information.

Yu *et al.* [37] proposed an incremental learning method, which makes the network suitable for learning the previously learned features on the new training data, retains the previous features, and minimizes the loss function of the network.

Guo *et al.* [38] designed a gated convolution (L-GCNN). Firstly, design a parameterized gate module (PGM) to generate pixel-level weights. Then, embed a single PGM and its connected extension units into different levels of encoders in

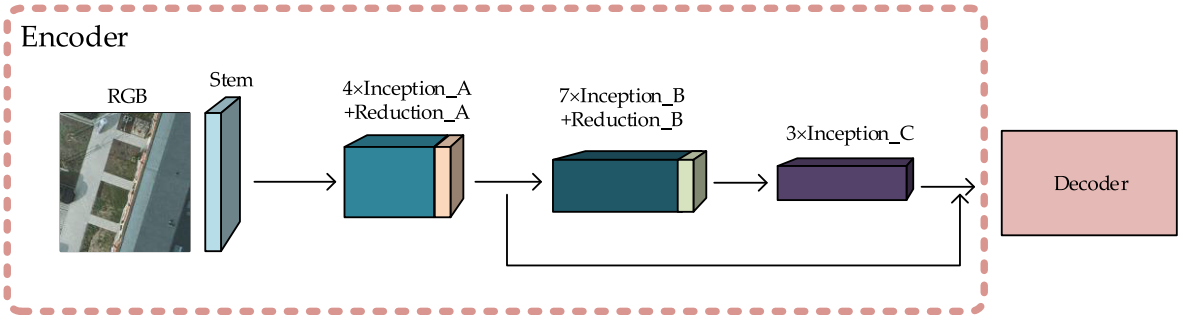

FIGURE 7. The first fusion convergence solution network diagram.

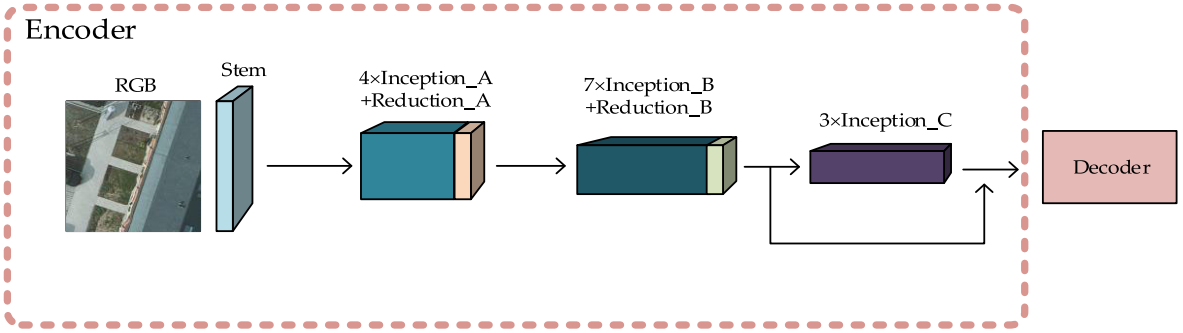

FIGURE 8. The second fusion solution network diagram.

TABLE 9. Processing time of each image for the DAPN.

	Test time(s)	Predicted time(s)
Potsdam	0.4736	0.2934
Inria	0.5253	0.2605

the L-GCNN, resulting in a fine segmentation framework that aggregates context information.

Some of the methods mentioned above are the recently published RS image segmentation methods. The experiment is carried out on two datasets and compared with the above methods.

a: Comparison of experiments on Potsdam datasets

Apply the IOU, F1 and OA scores to evaluate our model on the Potsdam test set. Table 10 and Table 11 show the experimental results on the Potsdam dataset. The method proposed achieved the highest IOU, F1 and OA average scores. Table 11 provides the comparative results of per class, the DAPN can obtain the best performance by and large. The F1 score of the class Imp.Surf achieves the highest 96.97%, which is higher than the [28] about 3%. The F1 score of the [27] for class Car can reach a maximum of 96.40%, while the score of our model is near. Expect the class Tree, other classes of F1 scores also achieved the highest results by and large.

b: Comparison of experiments on Inria datasets

Table 12 is the experimental results on the Inria dataset. Through experiments on the Inria five regional datasets,

TABLE 10. Comparison of different methods on Potsdam test set.

	IOU	F1	OA
[27]	\	92.90	91.50
[28]	84.60	92.60	\
[35]	\	88.80	88.30
[36]	\	81.80	80.00
[37]	\	\	89.42
[38]	84.12	91.24	\
[39]	\	90.60	\
[40]	\	\	90.10
[41]	\	84.25	\
Our	89.62	94.49	96.13

TABLE 11. Comparison of F1 scores on Potsdam test set.

	Imp.Surf.	Building	Low.Veg.	Tree	Car
[27]	93.50	97.20	88.20	89.20	96.40
[28]	94.10	97.80	89.40	90.30	95.10
[35]	90.20	95.90	83.90	84.30	89.60
[36]	83.00	90.00	77.00	72.00	74.00
[37]	91.77	95.71	84.40	79.56	88.25
[38]	92.84	96.17	87.78	85.74	93.65
[39]	92.70	96.30	87.30	88.50	95.40
[40]	92.30	97.00	86.80	86.90	94.50
[41]	92.10	96.86	85.28	92.10	94.43
Our	96.97	97.25	89.97	86.32	93.74

obtain better segmentation results than the other two methods, in which the IOU score has been greatly improved. Combined with the results on the two datasets, our model performs excellently in the segmentation of the building.

In addition, verify the segmentation effect of three cases which fusing context information without adding a dual ASPP module on two datasets. In the first case,

the InpectionV-4 network is used as the encoder and then directly connected to the decoder used in this paper without any other processing. In the second case, the InpectionV-4 network is used as the encoder, in which the characteristic image of the Reduction-A module is fused to the end of the encoder and then connected to the same decoder. In the third case, the trunk of the encoder is the same as above, and then the feature image output by the Reduction-B module is fused to the end of the encoder, and finally connected to the same decoder module. Among them, two kinds of fusion features use common fusion techniques to extract features from different layers, and then fuse the feature with the feature map of the deeper network, so that the context information of the neural network is simply fused. To achieve the purpose of improving the accuracy and effect of segmentation. The two merged network architectures are shown in Figure 7 and Figure 8.

Experiment on two datasets with the methods of three cases. On the Potsdam dataset, calculated five evaluation indicators of various situations, and more comprehensively showed the segmentation effect of different fusion methods. On the Inria dataset, use only one fusion strategy, and then obtain the IOU scores after training in five areas respectively. As can be seen from Table 13 and Table 14, the evaluation indicators obtained from the training on the two datasets are significantly higher than those obtained by other strategies, because the feature map of the moderate training stage is fused to the end of the encoder. The loss of target location information can be greatly reduced, and enough abstract features can be obtained. After merging the feature map of the Reduction-B output with the feature map output at the end of the encoder, the evaluation indicator is significantly reduced. The result shows that the performance of segmentation can be improved by using the fusion strategy. After adding the dual ASPP module, the evaluation indicators have been significantly improved. Without losing too many shallow features of the network, the dual ASPP module has a wider receptive field through atrous convolution with different dilated rates. Features are extracted from different depths of the convolution network to obtain more location information, the final segmentation accuracy is also higher. Figure 9 and 10 shows the comparison of the prediction results of different strategies on the validation set, and the prediction effect of the DAPN proposed in the present study is better.

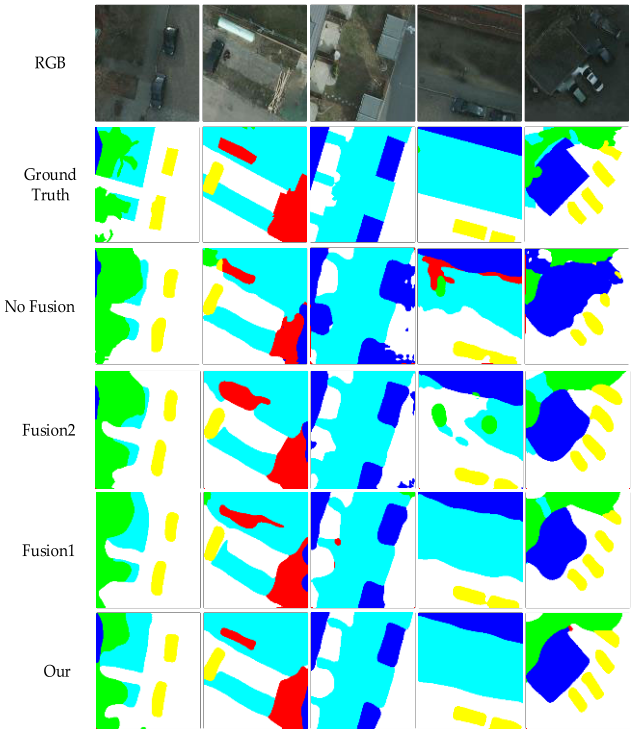

FIGURE 9. Comparison of marked effects with different fusion schemes on Potsdam datasets.

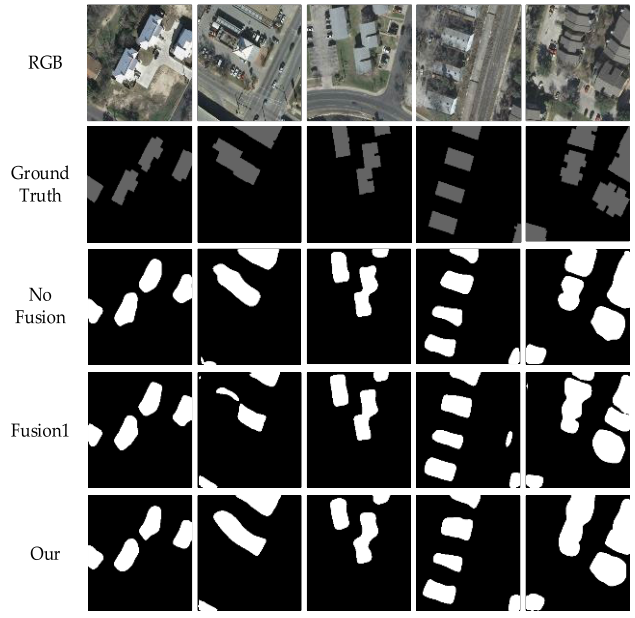

FIGURE 10. Comparison of marked effects with different fusion schemes on Inria datasets.

V. DISCUSSION

After completing the basic experimental comparison, have a more in-depth discussion of the generalization ability of the DAPN, so a transfer learning experiment is carried out with Potsdam dataset and Vaihingen dataset [32]. The original image of the Vaihingen dataset is composed of IR-R-G three channels. Although the categories of the two datasets are the same, the different channel combinations can result in significant color differences in the image. Therefore, the prediction results are very bad by using the model trained on Potsdam R-G-B dataset and Vaihingen IR-R-G dataset to carry out transfer learning experiment. To avoid this problem, the training data of Potsdam is also used IR-R-G images in this section.

The preprocessing of the Potsdam IR-R-G dataset is the same as the Potsdam R-G-B dataset, which obtain 60800 Potsdam IR-R-G images to train. Since each patch size of the Vaihingen IR-R-G dataset is different, the Vaihingen IR-R-G dataset is cut according to the corresponding size and the data

TABLE 12. Experimental comparison of various methods on Inria test set.

	Austin		Chicago		Kitsap		Tyrol		Vienna	
	IOU	OA	IOU	OA	IOU	OA	IOU	OA	IOU	OA
[30]	82.49	97.47	73.90	93.90	72.45	99.35	70.71	98.73	83.72	95.35
[42]	73.09	96.43	70.38	92.92	72.45	99.43	76.40	98.12	78.88	93.98
Ours	91.83	97.03	92.38	94.48	83.06	98.36	87.94	97.05	94.96	96.44

TABLE 13. Comparison of different fusion schemes on Potsdam test set.

	None Fusion	Fusion1	Fusion2	ours
IOU	86.42	87.38	83.68	89.62
F1	88.65	92.70	90.12	94.49
Precision	89.14	90.93	82.20	95.06
Recall	89.77	95.14	87.14	94.29
OA	93.89	95.58	91.06	96.13

TABLE 14. Comparison of IOU scores with different Fusion schemes on Inria test set.

	None Fusion	Fusion1	ours
Austin	85.80	87.24	91.83
Chicago	91.10	91.65	92.38
Kitsap	77.66	80.40	83.06
Tyrol	88.06	86.62	87.94
Vienna	93.41	94.76	94.96

TABLE 15. Comparison of IOU scores with other models on Vaihingen IR-R-G dataset.

Networks	IOU
FCN32	63.74
SegNet	58.19
U-Net	60.30
Our	71.85

is expanded by mirroring. Finally, 5559 images with the size of 299 × 299 are obtained. The training parameters are the same as the previous section, and the Vaihingen dataset is tested using the model obtained after training with Potsdam IR-R-G datasets.

In addition, use several common segmentation networks for comparison experiments, and the IOU scores are shown in Table 15. As can be seen from the table, the DAPN has a stronger generalization ability than other models. However, compared with the performance on the Potsdam dataset, there is a significant decrease in the IOU scores on the Vaihingen IR-R-G dataset due to the differences between the two datasets. Althoug h the categories of the two datasets are the same, there is a lot of variability in ground targets because of the images collected in different regions, resulting in the decline of the prediction ability of the model.

VI. CONCLUSION

In the current study, a network with the multi-channel convolutions and dual ASPP modules is proposed, which fully extract the multi-scale features of the image and retain the spatial features of the object at the same time, while taking into account the loss of network features in many aspects. The network is a segmentation network with codec structure, which extracts features by downsampling and restores the original resolution by upsampling, and it is a well-recognized structure in the task of semantic segmentation. Through the method of data augmentation, the problem of sample imbalance is solved to a certain extent. In addition, experiments are carried out on Potsdam data sets and Inria datasets, and the results are compared with a variety of mainstream segmentation models and the newly proposed segmentation algorithms. The calculation of a variety of evaluation indicators shows that the network has a significant improvement effect compared with other methods. Finally, explore further the generalization ability of the DAPN and predict on Vaihingen IR-R-G datasets. The code of the DAPN is in https://github.com/Udellliu/InceptionV4-ASPP-semantic.

REFERENCES

[1] Q. Weng, "Modeling urban growth effects on surface runoff with the integration of remote sensing and GIS," *Environ. Manage.*, vol. 28, no. 6, pp. 737–748, Dec. 2001.
[2] W. Zhai, H. Shen, C. Huang, and W. Pei, "Fusion of polarimetric and texture information for urban building extraction from fully polarimetric SAR imagery," *Remote Sens. Lett.*, vol. 7, no. 1, pp. 31–40, Jan. 2016.
[3] M. Herold and D. Roberts, "Spectral characteristics of asphalt road aging and deterioration: Implications for remote-sensing applications," *Appl. Opt.*, vol. 44, no. 20, pp. 4327–4334, Jul. 2005.
[4] X. Yu and Z. Shi, "Vehicle detection in remote sensing imagery based on salient information and local shape feature," *Optik*, vol. 126, no. 20, pp. 2485–2490, Oct. 2015.
[5] S. Pang, X. Hu, Z. Wang, and Y. Lu, "Object-based analysis of airborne LiDAR data for building change detection," *Remote Sens.*, vol. 6, no. 11, pp. 10733–10749, Nov. 2014.
[6] 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.
[7] R. M. Haralick, "Textural features for image classification," *IEEE Trans. Syst., Man, Cybern.*, vol. SMC-3, no. 6, pp. 610–621, Nov. 1973, doi: 10.1109/TSMC.1973.4309314.
[8] F. T. Ulaby, M. C. Dobson, and G. A. Bradley, "Radar reflectivity of bare and vegetation-covered soil," *Adv. Space Res.*, vol. 1, no. 10, pp. 91–104, Jan. 1981.
[9] B.-C. Gao, "NDWI—A normalized difference water index for remote sensing of vegetation liquid water from space," *Remote Sens. Environ.*, vol. 58, no. 3, pp. 257–266, Dec. 1996.
[10] Z. Shao, W. Zhou, L. Zhang, and J. Hou, "Improved color texture descriptors for remote sensing image retrieval," *J. Appl. Remote Sens.*, vol. 8, no. 1, Jan. 2014, Art. no. 083584.
[11] X. Huaiying, "A shadow detection of remote sensing images based on statistical texture features," *J. remote Sens.*, vol. 15, no. 4, pp. 778–791, 2011.
[12] G. E. Hinton, "Reducing the dimensionality of data with neural networks," *Science*, vol. 313, no. 5786, pp. 504–507, Jul. 2006.
[13] C. L. P. Chen, C.-Y. Zhang, L. Chen, and M. Gan, "Fuzzy restricted Boltzmann machine for the enhancement of deep learning," *IEEE Trans. Fuzzy Syst.*, vol. 23, no. 6, pp. 2163–2173, Dec. 2015.
[14] A. S. Razavian, H. Azizpour, J. Sullivan, and S. Carlsson, "CNN features Off-the-shelf: An astounding baseline for recognition," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. Workshops*, Jun. 2014, pp. 512–519.
[15] K. Simonyan and A. Zisserman, "Very deep convolutional networks for large-scale image recognition," 2014, *arXiv:1409.1556*. [Online]. Available: http://arxiv.org/abs/1409.1556
[16] C. Szegedy, W. Liu, Y. Jia, P. Sermanet, S. Reed, D. Anguelov, D. Erhan, V. Vanhoucke, and A. Rabinovich, "Going deeper with convolutions," 2014, *arXiv:1409.4842*. [Online]. Available: http://arxiv.org/abs/1409.4842

[17] 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.

[18] K. He, G. Gkioxari, P. Dollár, and R. Girshick, "Mask R-CNN," 2017, *arXiv:1703.06870*. [Online]. Available: http://arxiv.org/abs/1703.06870

[19] C. Szegedy, S. Ioffe, and V. Vanhoucke, "Inception-v4, Inception-ResNet and the Impact of Residual Connections on Learning," *Comput. Sci.*, vol. 3, no. 6, pp. 105–112, 2016.

[20] O. Ronneberger, P. Fischer, and T. Brox, "U-net: Convolutional networks for biomedical image segmentation," 2015, *arXiv:1505.04597*. [Online]. Available: http://arxiv.org/abs/1505.04597

[21] V. Badrinarayanan, A. Kendall, and R. Cipolla, "SegNet: A deep convolutional encoder-decoder architecture for image segmentation," 2015, *arXiv:1511.00561*. [Online]. Available: http://arxiv.org/abs/1511.00561

[22] L.-C. Chen, G. Papandreou, F. Schroff, and H. Adam, "Rethinking atrous convolution for semantic image segmentation," 2017, *arXiv:1706.05587*. [Online]. Available: http://arxiv.org/abs/1706.05587

[23] F. Yu and V. Koltun, "Multi-scale context aggregation by dilated convolutions," 2015, *arXiv:1511.07122*. [Online]. Available: http://arxiv.org/abs/1511.07122

[24] J. Lafferty and A. Mccallum, "Pereira FCN Conditional random fields: Probabilistic models for segmenting and labeling swquence data," in *Prec. ICML*, 2001, vol. 3, no. 2, pp. 282–289.

[25] S. Zheng, S. Jayasumana, B. Romera-Paredes, V. Vineet, Z. Su, D. Du, C. Huang, and P. H. S. Torr, "Conditional random fields as recurrent neural networks," 2015, *arXiv:1502.03240*. [Online]. Available: http://arxiv.org/abs/1502.03240

[26] C. Michaelis, I. Ustyuzhaninov, M. Bethge, and A. S. Ecker, "One-shot instance segmentation," 2018, *arXiv:1811.11507*. [Online]. Available: http://arxiv.org/abs/1811.11507

[27] Diakogiannis F I, Waldner, Franáois, Caccetta P, "ResUNet-a: A deep learning framework for semantic segmentation of remotely sensed data," *ISPRS J. Photogramm. Remote Sens.*, vol. 162, pp. 94–114, Apr. 2020.

[28] 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.

[29] P. Liu, X. Liu, M. Liu, Q. Shi, J. Yang, X. Xu, and Y. Zhang, "Building footprint extraction from high-resolution images via spatial residual inception convolutional neural network," *Remote Sens.*, vol. 11, no. 7, p. 830, Apr. 2019.

[30] Y. Zhang, W. Gong, J. Sun, and W. Li, "Web-net: A novel nest networks with ultra-hierarchical sampling for building extraction from aerial imageries," *Remote Sens.*, vol. 11, no. 16, p. 1897, Aug. 2019.

[31] S. Ioffe and C. Szegedy, "Batch normalization: Accelerating deep network training by reducing internal covariate shift," 2015, *arXiv:1502.03167*. [Online]. Available: http://arxiv.org/abs/1502.03167

[32] International Society for Photogrammetry and Remote Sensing. *2D semantic labeling contest*. Accessed: Mar. 20, 2020. [Online]. Available: http://www2.isprs.org/commissions/comm3/wg4/semantic-labeling.html

[33] E. Maggiori, Y. Tarabalka, G. Charpiat, and P. Alliez, "Can semantic labeling methods generalize to any city? The Inria aerial image labeling benchmark," in *Proc. IEEE Int. Geosci. Remote Sens. Symp. (IGARSS)*, Jul. 2017, pp. 3226–3229.

[34] N. Audebert, B. Le Saux, and S. Lefávre, "Segment-before-detect: Vehicle detection and classification through semantic segmentation of aerial images," *Remote Sens.*, vol. 9, no. 4, p. 368, Apr. 2017.

[35] Y. Wang, B. Liang, M. Ding, and J. Li, "Dense semantic labeling with atrous spatial pyramid pooling and decoder for high-resolution remote sensing imagery," *Remote Sens.*, vol. 11, no. 1, p. 20, Dec. 2018.

[36] W. Zhang, H. Huang, M. Schmitz, X. Sun, H. Wang, and H. Mayer, "Effective fusion of multi-modal remote sensing data in a fully convolutional network for semantic labeling," *Remote Sens.*, vol. 10, no. 2, p. 52, Dec. 2017.

[37] Y. Liu, D. Minh Nguyen, N. Deligiannis, W. Ding, and A. Munteanu, "Hourglass-ShapeNetwork based semantic segmentation for high resolution aerial imagery," *Remote Sens.*, vol. 9, no. 6, p. 522, May 2017.

[38] 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.

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

[40] 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.

[41] O. Tasar, Y. Tarabalka, and P. Alliez, "Incremental learning for semantic segmentation of large-scale remote sensing data," *IEEE J. Sel. Topics Appl. Earth Observ. Remote Sens.*, vol. 12, no. 9, pp. 3524–3537, Sep. 2019.

[42] X. Li, Y. Jiang, H. Peng, and S. Yin, "An aerial image segmentation approach based on enhanced multi-scale convolutional neural network," in *Proc. IEEE Int. Conf. Ind. Cyber Phys. Syst. (ICPS)*, May 2019, pp. 47–52.

WENJIE LIU received the B.S. degree in information and computing science, in 2018. He is currently pursuing the degree 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 Ph.D. degrees in software engineering from Guizhou University, Guiyang, China, in 2010 and 2015, respectively. From 2012 to 2015, he was a Joint Training Ph.D. Student with Peking University and Guizhou University. He was also with the Key Laboratory of Integrated Microsystems, Shenzhen Graduate School, Peking University. He is currently an Associate Professor with Guizhou University. His research interests include intelligence image algorithms of computer vision, such as scene target detection, extraction, tracking, recognition, and behavior analysis.

HAISHENG FAN is currently pursuing the Ph.D. degree. He is also a President with the Division of Satellite Big Data Solutions. Expert of GIS & RS Software Design and Development. He is also a Researcher on image processing methods and manager of remote sensing application projects. He is one of the major pioneers of first domestic commercial RS service platform based on cloud-computing system. He is also a Designer and an Executor of Orbita Satellite Big Data Service Platform which is under construction.

YONGJIE ZOU was born in Qianjiang, Hubei, China, in 1993. He received the B.S. degree in computer science and technology from Guizhou Education University, in 2018. 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, China, in 2008, where he is currently pursuing the Ph.D. degree. Since December 2013, he has been an Associate Professor with the School of Mathematics and Big Data, Guizhou Education University, Guiyang. He has 11 years of teaching experience. His research interests include machine vision and wireless networks.