IEEE Access
Multidisciplinary ⋮ Rapid Review ⋮ Open Access Journal

SPECIAL SECTION ON ADVANCED OPTICAL IMAGING FOR EXTREME ENVIRONMENTS

Received April 21, 2019, accepted May 15, 2019, date of publication May 24, 2019, date of current version June 7, 2019.

Digital Object Identifier 10.1109/ACCESS.2019.2918700

Scene-Aware Deep Networks for Semantic Segmentation of Images

ZHIKE YI[1], TAO CHANG[1], SHUAI LI[1,4], RUIJUN LIU[2], JING ZHANG[3], AND AIMIN HAO[1]

[1]State Key Laboratory of Virtual Reality Technology and Systems, Beihang University, Beijing 100191, China
[2]Beijing Key Laboratory of Big Data Technology for Food Safety, Beijing Technology and Business University, Beijing 100048, China
[3]College of Software, Beihang University, Beijing 100191, China
[4]Shandong Key Laboratory of Digital Medicine and Computer Assisted Surgery, The Affiliated Hospital of Qingdao University, Qingdao 266003, China

Corresponding authors: Shuai Li (lishuai@buaa.edu.cn) and Ruijun Liu (liuruijun@btbu.edu.cn)

This work was supported in part by the Foundation of Shandong Provincial Key Laboratory of Digital Medicine and Computer assisted Surgery under Grant SDKL-DMCAS-2018-01, in part by the National Natural Science Foundation of China under Grant 61672077 and Grant 61532002, in part by the Applied Basic Research Program of Qingdao under Grant 161013xx, and in part by the Beijing Natural Science Foundation-Haidian Primitive Innovation Joint Fund under Grant L182016.

ABSTRACT Scene classification and semantic segmentation are two important research directions in computer vision. They are widely used in the research of automatic driving and human–computer interaction. The purpose of the scene classification is to use the image classification to determine the category of the scene in an image by analyzing the background and the target object, while semantic segmentation aims to classify the image at the pixel level and mark the position and semantic information of the scene unit. In this paper, we aimed to train the semantic segmentation neural network in different scenarios to obtain the models with the same number of scene categories, which they are used to process the images. During the process of the actual test, the semantic segmentation dataset was firstly divided into three categories based on the scene classification algorithm. Then the semantic segmentation neural network is trained under three scenarios, and three semantic segmentation network models are obtained accordingly. To test the property of our methods, the semantic segmentation models we got were selected to treat other pictures, and the results obtained from the performance of scene-aware semantic segmentation were much better than semantic segmentation without considering categories. Our study provided an essential improvement of semantic segmentation by adding category information into consideration, which will be helpful to obtain more precise models for further picture analysis.

INDEX TERMS Semantic segmentation, scene classification, convolutional neural network.

I. INTRODUCTION

Scene classification and semantic segmentation are two essential conceptions in computer vision, for they are widely used in the research of automatic driving, human-computer interaction and augmented reality. The purpose of scene classification is to use the image classification to determine the category of the scene in an image by analyzing the background and the target object. The goal of semantic segmentation is to classify images at the pixel level, and to mark the position and semantic information of the scene unit.

Semantic segmentation involves two aspects: segmentation and classification. In order to determine the corresponding category of a pixel, we use not only the information of the pixel near the point, i.e the local information, but also the information of the pixel far away from the distance, i.e the global information, should be considered.

Many of the current semantic segmentation methods are developed on the basis of the Fully Convolutional Network (FCN), which usually takes the semantic information of long distance into consideration, leading to the mismatches in the analysis results. [30]. However, if a priori can be added to the semantic segmentation, the errors developed from FCN could be greatly reduced, and the accuracy could be improved [18]. Thus, many studies of existing semantic segmentation expanded the receptive field with the convolution kernel and feature fusion, which aimed to let the convolution kernel in the network structure get the semantic information of the image from a distance [17], [26]. During this process, once the scene category information is determined, more priors of the ‘‘farthest distance’’ (ie the whole picture) semantic information that can be used may also increase

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

and the accuracy may also be improved. If training multiple scenarios separately without changing the network structure, the results of all the scenarios can only achieve this level, however, if adding the scene category information, the results of the accuracy will be definitely improved. Moreover, this adding information will also possess instructive significance for the design of the network structure and the training of the network model.

In this paper, we managed to improve the scene-aware semantic segmentation methods by combining scene category information into semantic segmentation. Specifically, we trained the semantic segmentation neural network in different scenarios to obtain the semantic segmentation neural network model with the same number of scene categories. In the actual test, the images were classified by scene classification, and then processed with the semantic segmentation neural network model of the corresponding scene category. Moreover, we also explored a main question whether the use of scene category information could improve the accuracy of semantic segmentation by setting control experiments. For the improvement of accuracy of image analysis, our results can be used in related researches, such as automatic driving, human-computer interaction, and augmented reality.

II. RELATED

A. SEMANTIC SEGMENTATION ALGORITHM

There are two important aims of the researches of semantic segmentation algorithms: firstly, to continuously improve the ability of abstracting features from neural networks, including the expansion of receptive fields and the fusion of multiple ranges of features, such as PSPNet [30], DeepLab series algorithms [1]; secondly, to try to restore the reduced feature map to a larger spatial resolution by different methods. Many algorithms use feature fusion to merge the shallower feature maps during recovery, such as FCN [14], U-Net [24], DeconvNet [21], DeepLab v3+ [3], because the lower level feature map has more spatial resolution and contains more position information, which can make the segmentation more accurate [13]. However, recent artificial intelligence technologies have many limitations [16], [19], for example, the classification of FCN is too rough. The main reason is that when the resolution of feature map is enlarged by linear interpolation, the magnification is too large, and even if a skip structure is used, the accuracy will only be increased limitedly, and as mentioned above, FCN does not combine information over long distances, which can lead to some relationship mismatch errors.

Compared with FCN, DeepLab v2 algorithm [2] makes better use of remote semantic information by expanding the receptive field and feature fusion, which greatly improves the accuracy of semantic segmentation. These improvements make the network structure more complex, thus, compared with FCN, more computation is required in both training and testing. Besides, DeepLab v3+ [3] is the latest semantic segmentation neural network structure of DeepLab series. The main feature of DeepLab v3+ is to propose an Encoder-Decoder structure to implement semantic segmentation algorithm. Regarding to the inconsistency of categories, DeepLab v2 has been greatly improved by expanding convolution and space pyramid pooling, while DeepLab v3+ has been improved by spatial pyramid pooling, which makes the results are slightly better than those of DeepLab v2. The obvious alternation in DeepLab v3+ is that the boundaries become more precise, which attributed to the fact that the decoder part fuses the shallow feature maps and scales up on the feature map in two steps instead of scaling to the ultimate goal once directly like DeepLab v2. As a result, the results of the semantic segmentation algorithm have been greatly improved by using the methods of expanding convolution and spatial pyramid pooling to merge the remote information, revealing the essentiality of long-distance semantic information. As a kind of global information, scene category information is regarded as a kind of special long-distance information, which is considered in semantic segmentation, and receive a positive result.

B. SCENE CLASSIFICATION ALGORITHM

Traditional artificial design features include underlying features extracted based on pixel points, for instance GIST features [22], [23], SIFT features [15], HOG [5] features, and advanced features which are similar with those including Object banks, and features [10], such as Latent pyramidal regions [11], [25], Bag of parts features [8] etc. containing more semantics. After the emergence of convolutional neural networks, the potential features of images are extracted well when dealing with large-scale image data sets and widely used in scene classification. Traditional classifiers in machine learning research include Support Vector Machine (SVM), Bayesian classifier, K-nearest neighbor classifier, and BP neural network classifier. With the rapid development of deep learning, convolutional neural network has become the most commonly used classification method. After AlexNet [9] achieving great success in image classification, deep learning is now a widely used method in scene classification, and more valuable methods have been proposed, such as VGGNet [27], Inception Network [28], [29] and ResNet [7].

C. SEMANTIC SEGMENTATION DATA SET

Common scene analysis data sets include LMO [12], PASCAL VOC and PASCAL context [6], [20], Cityscapes [4] and ADE20K [31], among which ADE20K is the latest data set. Compared with the previous data sets, both scene category and object category are more enormous, which become the most used data sets in the recent semantic segmentation research.

III. METHOD OVERVIEW

A. SCENE-AWARE SEMANTIC SEGMENTATION ALGORITHM DESIGN

In this paper, we used the flow chart shown in Figure 1 to show the scene-aware semantic segmentation algorithm. Specifically, there are three steps:

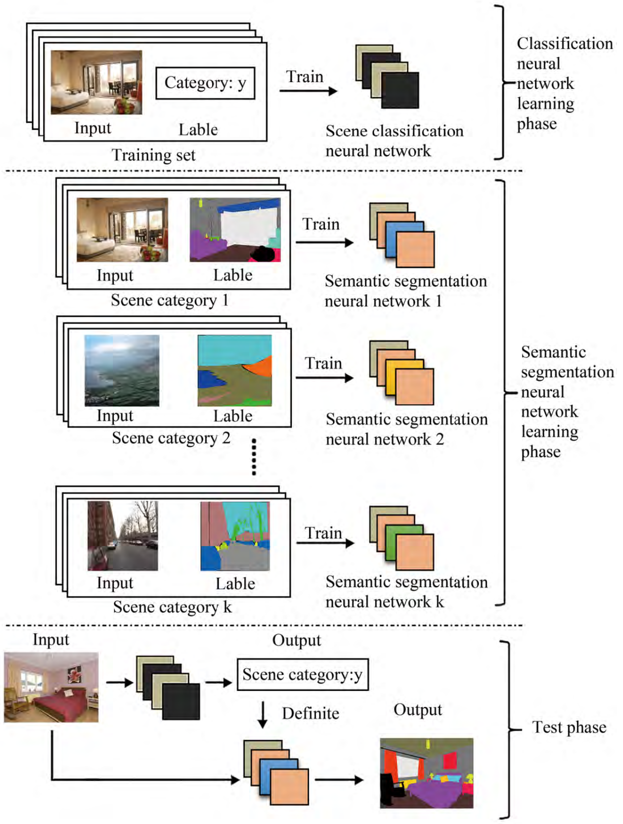

Category: y
Input
Lable
Training set
Train
Scene classification neural network
Classification neural network learning phase
Input
Lable
Scene category 1
Train
Semantic segmentation neural network 1
Input
Lable
Scene category 2
Train
Semantic segmentation neural network 2
Input
Lable
Scene category k
Train
Semantic segmentation neural network k
Semantic segmentation neural network learning phase
Input
Output
Scene category:y
Definite
Output
Test phase

FIGURE 1. **Flowchart of scene-aware semantic segmentation algorithm.**

1) Classification neural network learning phase: In order to classify the scenes of the pictures in the test set, it is necessary to train the convolutional neural network of the scene classification, in which the images are input and the scenes are labeled.

2) Semantic segmentation neural network learning phase: Semantic segmentation neural network is trained in each scene category, and a priori information in different scenarios is learned to improve the final result. The picture is input and the semantic segmentation result is labeled.

3) Test phase: The final result is obtained by processing the test set using the scene classification neural network and semantic segmentation neural network trained in step 1 and step 2. In the flow chart, the scene classification neural network and the semantic segmentation nerve are distinguished by different colors. After performing the test phase, we classify the pictures to obtain the number of scene category: y, and then select the corresponding segmentation network for further processing to obtain the final semantic segmentation result.

The results of experiment, which use the above-mentioned sub-scenario training semantic segmentation neural network combined with the scene classification test method, show that when training only in a single category of data set, the accuracy rate of the experimental semantic segmentation is dropped compared to the entire training set. Further analysis of the experimental results revealed that it is due to the

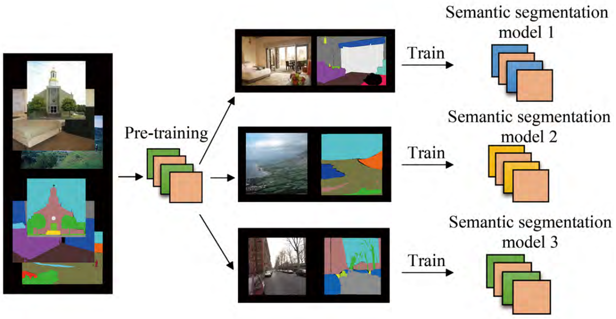

Pre-training
Train
Semantic segmentation model 1
Train
Semantic segmentation model 2
Train
Semantic segmentation model 3

FIGURE 2. **Schematic diagram of adding the training phase after pre-training.**

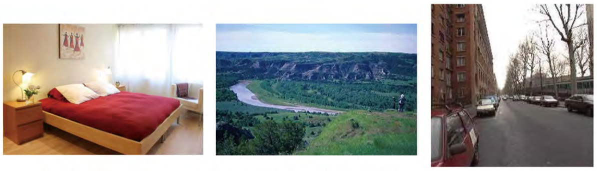
reduction of target resolution accuracy, which rarely occurs in the scene, when considering that the time cost of every single iteration during training is roughly the same. For a more reasonable comparison, the experimental comparison is based on the equalization of the total time cost (the number of iterations) of the semantic segmentation training. At the same time, in order to avoid the problem of the accuracy of the training in the scene, the pre-training step needs to be added, as shown in Figure 2.

Trying to use the above experimental scheme incorporate scene category information (section III-B) into the semantic segmentation algorithm, there are two notable factors.

Firstly, under such experimental settings, how to compare with the control semantic segmentation algorithm (section III-C) (training on the entire training set) is more reasonable. Considering that in the absence of over-fitting, the general training time is more costly and the result is better. In this paper, we use the same semantic separation training, compare the pixel accuracy and average cross-comparison ratio of the test results. Secondly, in this experimental scheme, the semantic segmentation algorithm between our experiment and the control group are different in terms of the complexity of the algorithm flow and the number of models. Even if the results of the former are better, they are not determined because of the addition of scene category information. In order to do further explore, we design another set of controlled trials for comparison, and find it mainly affects by the scenario class information rather than algorithm flow changes, which is called the algorithm flow control experiment (section III-D). After that, we mainly detect the effect of scene classification accuracy on scene-aware semantic segmentation algorithm (section III-E).

B. SCENE CLASSIFICATION DESIGN

In order to classify the scenes of the pictures in the test set, we use the deep convolutional neural network to implement the scene classification algorithm, and select the more commonly used VGGNet [27] network structure. The algorithm will be trained on the semantic segmentation dataset to determine the category of the scene in which the image to be processed belongs to. Since the subsequent algorithm requires to train the semantic segmentation neural network



Indoor scene Outdoor nature scene Outdoor artificial scene

FIGURE 3. Pictures of three different scenes.

under each scene, the division of the scene category should be reasonably selected, that is, all the pictures are divided into several categories. Firstly, the scene category should not be divided too much, for the following reasons:

1) The follow-up algorithm should train the semantic segmentation neural network in each scene. If there are too many scene categories, the workload is too large;

2) The total number of training pictures is determined. The more categories, the fewer the training pictures under each scene on average, which may lead to over-fitting;

3) When the scene is divided into many parts, the difference of images between different scenes is small, and the difficulty of scene classification increases, which is not conducive to the subsequent experiments.

After analyzing, our work classifies all the pictures into three categories according to the scene: indoor scene pictures, outdoor natural scene pictures and outdoor artificial scene pictures which were shown in Figure 3.

Dividing pictures into these categories is mainly based on the following considerations:

1) The total number of categories is three, which is not large, and there will be no disadvantages caused by too many classifications;

2) The differences between the pictures of the same category are small, and the pictures of different categories are very different, thus, the image classification will reach a high accuracy rate;

3) The types of objects in different categories of pictures vary greatly. For example, beds and carts do not appear in outdoor natural scenes, mountains and airplanes do not appear in indoor scenes, etc. These differences make the scene category containing more information and may play a greater role in the semantic segmentation algorithm.

C. SEMANTIC SEGMENTATION ALGORITHM COMPARISON DESIGN

For the scene-dependent semantic segmentation algorithm, we chose the deep convolutional neural network to implement the semantic segmentation algorithm, and select the two network structures DeepStudio v2 [2] and DeepLab v3+ [3] to avoid accidental results that lead to the wrong conclusion. We also set up a group of contrast experiment in the following experiments.

In order to compare whether the addition of scene category information improves the results of semantic segmentation, a set of semantic segmentation experiments needs to be set for comparison. For the semantic segmentation algorithm is implemented in this study, which is trained on the whole data set without considering the scene category, the semantic segmentation model will be used to compare with the following experiments. The convolutional neural network can efficiently extract potential features in the image through the continuous stacking of multiple network layers. The feature map is transmitted in the form of a three-dimensional matrix from the front to the back in the neural network. In order to continuously extract more abstract and advanced information, the width and height of the feature map are often reduced. Semantic segmentation requires the annotation of all the pixels, and the reduced feature map needs to be restored to larger resolution.

The existing semantic segmentation algorithms basically use the convolutional neural network to extract the feature map with reduced spatial resolution, and then restore the feature map to a larger spatial resolution by means of deconvolution and bilinear interpolation, and finally get the result of semantic segmentation. Most of the results do not reach the size of the original image, and need to be enlarged to get the final result.

To train semantic segmentation using the training set, we need to label the dataset carefully in advance, and mark the category of each pixel in each image. The categories are indicated by numbers, and generally have a special number such as 0 or 255, which means that the pixel is ignored and can be ignored during training, testing, and evaluating. During training, it is necessary to calculate the differences between the output of the neural network and the actual annotation, then to adjust the weight of the neural network model according to the differences. By using the adjusted model to carry out the next training and cycle until the differences between the output of the neural network and the actual results is small, then the training can be ended.

D. CONTROL EXPERIMENT DESIGN

In the control experiment, the algorithm flow is similar to the scene-aware semantic segmentation algorithm experiment. We divide the pictures into three categories and train each of them to obtain a semantic segmentation neural network model. In test stage we first determine the category, then select the corresponding model. Instead of classifying categories based on the scene category, we select the picture randomly. The proportion of images in different categories is still consistent with the previous ones. The flow chart of the training is shown in Figure 4. The semantic segmentation models in different categories are distinguished by different colors. The flow chart of the test is shown in Figure 5.

Using the algorithm flow similar to Figure 1. In addition to the image classification method is not in accordance with the scene category, the experiment controls other variables and the scene-aware semantic segmentation experiment is the same. The specific settings are as follows:

1) Pre-processing the training set;

2) The DeepLab v3+ network model is also used during training;

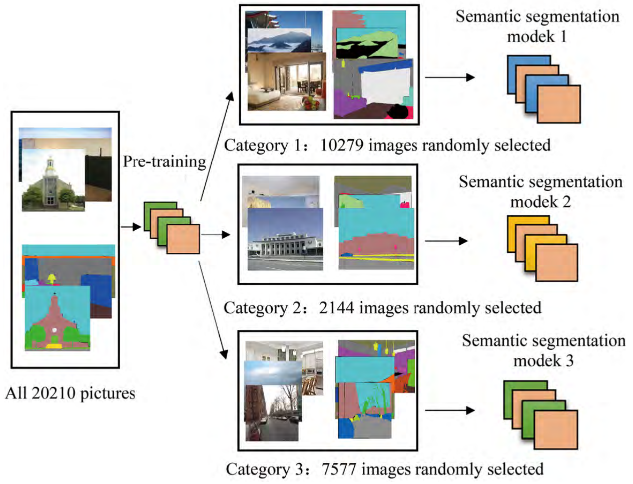

Semantic segmentation modek 1
Pre-training
Category 1：10279 images randomly selected
Semantic segmentation modek 2
Category 2：2144 images randomly selected
Semantic segmentation modek 3
All 20210 pictures
Category 3：7577 images randomly selected

FIGURE 4. Schematic diagram of the algorithm flow control experiment training stage.

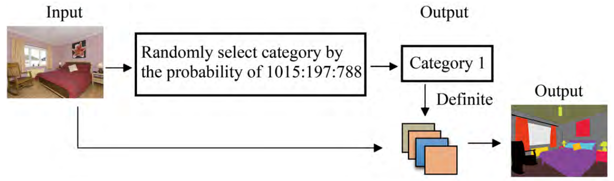


FIGURE 5. Schematic diagram of the algorithm flow control experiment test phase.

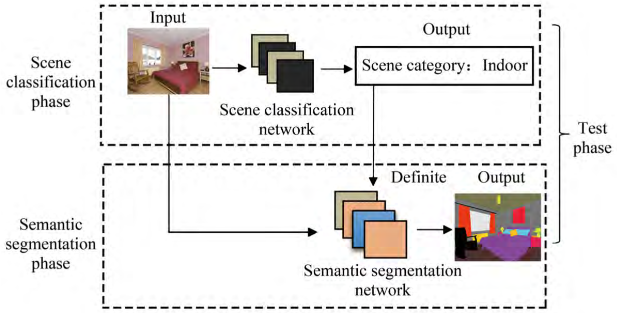

Input
Output
Scene classification phase
Scene category：Indoor
Scene classification network
Test phase
Definite
Output
Semantic segmentation phase
Semantic segmentation network

FIGURE 6. Scene-aware semantic segmentation algorithm test phase process decomposition.

3) The total number of training iterations is 480K, and the two distribution modes given in section III-C are also allocated according to the proportion of the picture and the average distribution.

E. DESIGN OF THE METHOD FOR EVALUATING THE IMPACT OF THE ACCURACY OF THE SCENE CLASSIFICATION ALGORITHM ON THE RESULTS

This section mainly studies the impact of the accuracy of scene classification on the scene-aware semantic segmentation algorithm. As shown in Figure 6, the algorithm can be divided into scene classification stage and semantic segmentation stage while testing on the test set and in current section, the scene classification stage was the emphases. The effect of the accuracy of the scene classification on the final result is explored by adjusting the results of the scene classification.

IV. RESULTS

A. EVALUATION CRITERIA

In order to evaluate the accuracy of the scene classification algorithm, we select accuracy as the evaluation standard, and set the total number of pictures as m, while Q is set as the prediction result of all pictures. Q_{ij} indicates the number of pictures with real category i and prediction category j, then:

$$accuracy = \frac{\sum_i Q_{ij}}{m} \tag{1}$$

In this experiment, m is 1611, the value of the i and j is indoor scene, outdoor natural scene, outdoor artificial scene.

We also use VGGNet as the network structure for scene classification. The final full connection layer and output number of the network are modified according to the number of categories.

According to the above implementation details and evaluation criteria, our work carries out an experiment on the ADE20K semantic segmentation dataset. The experimental results are as follows:

1) 16326 training pictures, 1611 test pictures, 3 category, accuracy rate is 96.46%;

2) Use the VGGNet model obtained to classify pictures that are not labeled in the training set, and then be trained. There are 20210 training pictures, 1611 test pictures and.3 categories,Accuracy rate is 96.90

Among the 20210 training pictures, there are 10279 for indoor scene pictures, 2144 for outdoor natural scene pictures and 7757 for outdoor artificial scene pictures. The test set was classified using the 2) trained VGGNet From the experimental results, the accuracy of the scene classification is still relatively high, which indicates that there will be little influence in the subsequent experiments.

B. SEMANTIC SEGMENTATION ALGORITHM COMPARISON EXPERIMENTAL RESULTS

We use the ADE20K data set as the data set of the semantic segmentation experiment part. The training data set has all 20210 pictures, and the test data contains 2000 pictures. Each picture has a corresponding label, and the label data is a single-channel image that has the same size as the original image. The value on the image is the category label of the pixel in the corresponding original image, and the category is represented by a number. There are 150 types of target objects in the ADE20K dataset, numbered from 1 to 150, and points marked 0 are ignored.

The preparation of the training data is mainly the variation of the annotation. Most of the datasets are labeled from 0, and 255 is used as the annotation of the ignored pixels. In order to be consistent with other datasets, the preprocessing is done in the experiment.

After pre-processing the experimental data according to the above method, the Deeplab v2 network structure was used to train on the data set, and the test results of the model on the test set were evaluated with Mean IoU and Pixel Accuracy respectively.

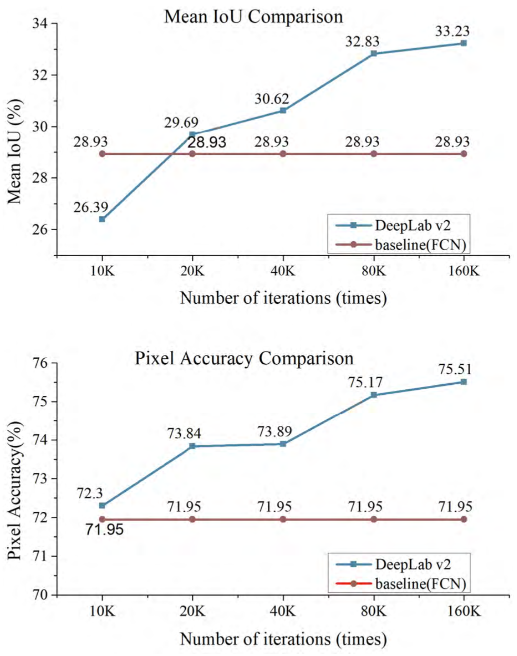


FIGURE 7. Semantic segmentation algorithm experimental test results - DeepLab v2.

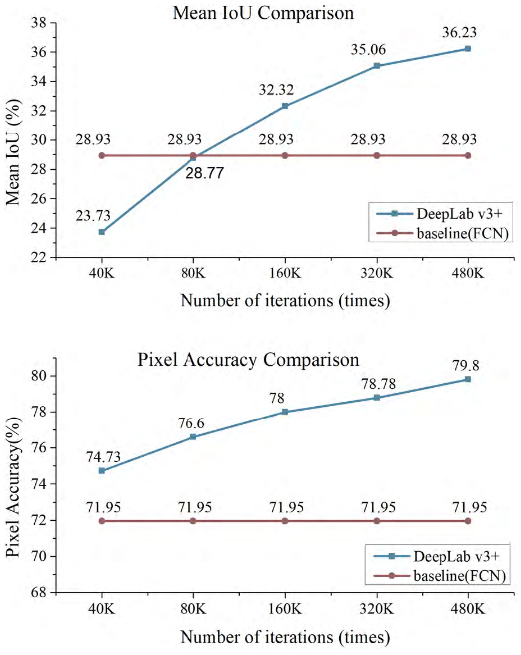


FIGURE 8. Semantic segmentation algorithm experimental test results - DeepLab v3+.

The red line in Figure 7 is the result of a test using the FCN model as the baseline published by the ADE20K Dataset. We can find that when the number of training iterations exceeds 20K, the result of DeepLab v2 is significantly better than FCN.

According to the same pre-processing, we also use the Deeplab v3+ network structure to train the data set. The results on the test set are shown in Figure 8.

Compared with DeepLab v2, DeepLab v3+ requires more resources during training. Due to equipment limitations, the number of pictures input in one iteration is less than DeepLab v2 when training DeepLab v3+, so more iterations are needed. From the results, after a certain training stage, the results of DeepLab v3+ are much better than FCN and DeepLab v2.

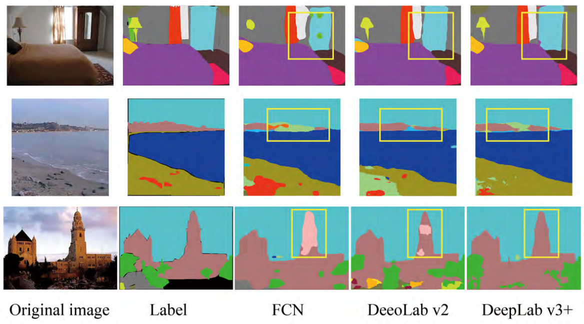


FIGURE 9. Comparison of semantic segmentation effects of different network structures.

In order to make the comparison more intuitively, we selects several pictures from the test set, processes them with different semantic segmentation algorithms, and visualizes the category numbers with different colors. The comparison results are shown in Figure 9. The first column is the original image, the second column is the image after the visualization, and each column is the result of different semantic segmentation algorithms. Each row represents the same graph. For the convenience of analysis, the following will be called Picture 1, Picture 2, and Picture 3 from top to bottom.

From the Figure 9 we find that in the results obtained by the FCN, there are some inconsistent results in the continuous area. For example, there is incorrect color appears on the door of the Picture1, which also happens on the parts of the yellow frame in the Figure1 and Picture2. It is probably that we don't take the semantic information of the far range (a small pixel that was misjudged does not take into account points outside this local area) into account in the feature extraction phase.

C. SCENE-AWARE SEMANTIC SEGMENTATION ALGORITHM EXPERIMENTAL RESULTS

This experiment is also carried out on the ADE20K dataset. The preparation of the experimental data mainly involves two aspects. On the one hand, this experiment divides the training set into three parts according to the scene, and the classification standard according to the results of the classification algorithm(section III-B). There are 10279 pictures of indoor scene, 2144 pictures of outdoor natural scene and 7757 pictures of outdoor artificial scene. On the other hand, similar to the section IV-B, the labels are also processed in this experiment as we have mentioned above.

The results of this experiment were compared with the experimental results of the semantic segmentation algorithm (section IV-B) to analyze whether the addition of scene category information can improve the results of semantic segmentation. In order to make the comparison more reasonable, the experimental comparison is based on the equalization of the total time cost of the semantic segmentation training. Considering that the same network structure is used to train on the same data set on the same device, the time cost of a single iteration is roughly the same, and the experimental comparison will be based on the same semantic iteration.

We use the network structure of DeepLab v2 and DeepLab v3+ which are modified slightly at the output layers, as in the experiments performed while implementing the semantic segmentation algorithm.

The number of training iterations for pre-training is also counted in the total number of times. There are three scenarios in total. The number of pre-trainings is one-fourth of the total for simplicity. There are two ways to allocate time in three scenarios: average allocation and distribution according to the training picture proportion. The training scenes pictures of the three scenes are 10279 indoor scene pictures, 2144 outdoor natural scene pictures and 7757 outdoor artificial scene pictures, the ratio is about 5:1:4.

The essential reason why we want to allocate the training time according to the proportion of the picture is that the pictures proportions in the three scenes are too different. If the distribution is average simply, the semantic segmentation network model may be over-fitting in the scene with few pictures with the increase of training time. However, the semantic segmentation model with a lot of pictures is still in an underfitting state.

Taken the number of iterations as 40K as an example: in the first scheme, 10K pre-training and 10K in three scenarios, which means that in order to obtain the semantic segmentation model of the indoor scene, the selected network model is trained and iterated 10K times on all the training images firstly, and then trained and iterated 10K times on the indoor training images. The other two scenes are similar: in the second scheme, pre-training 10K times, indoor scene model allocation 15K times; outdoor nature scene is allocated 3K times, and the outdoor artificial scene is allocated 12K times.

In the end, the two schemes get three semantic segmentation neural network models, corresponding to three scenarios. The test is still using the method shown in Figure 1. Firstly, we use the scene classification algorithm to decide which model to use, and secondly we use a specific model for semantic segmentation.

We use DeepLab v2 network structure in above comparative experimental scheme, and carry out experiments under the condition that the total training amount is 40K, 80K, 160K iterations. The test results are shown in Figure 10.

The scene-aware semantic segmentation of the two distribution methods is better than that of the control group, and the experimental group that allocates the training time proportionally is better than the average distribution result. We conduct the experiment using DeepLabv3+ in the same comparative experimental scheme with a total training of 160K, 320K, and 480K iterations. The results on the test set are shown in Figure 11.

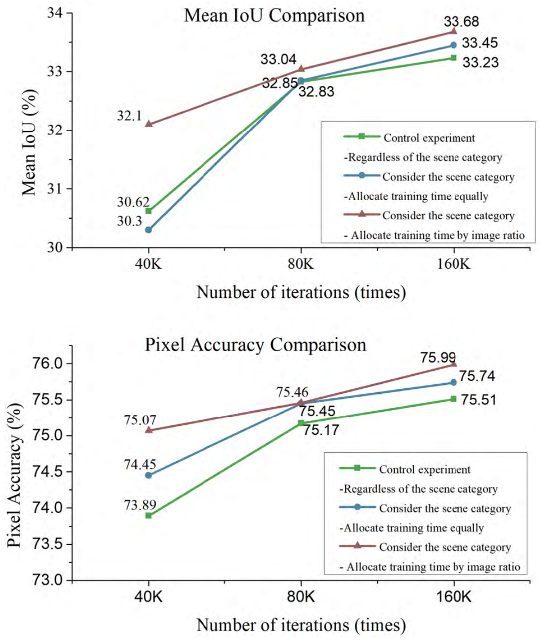


FIGURE 10. Scene-aware semantic segmentation test results - DeepLab v2.

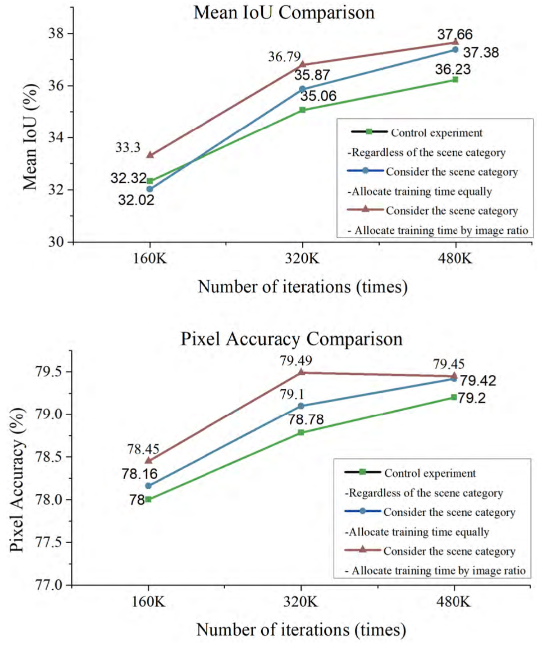


FIGURE 11. Scene-aware semantic segmentation test results - DeepLab v3+.

Similar to the experiment using DeepLab v2 network structure, the results of scene-aware semantic segmentation of the two distribution methods are better than those of the control group. The results of experimental group with proportional training time are better than those of average distribution. We mainly summarize three conclusions from experimental results on DeepLab v2 and DeepLab v3+:

After a certain number of training iterations, compared with the control experiment without considering the scene category, the experimental results which considered the

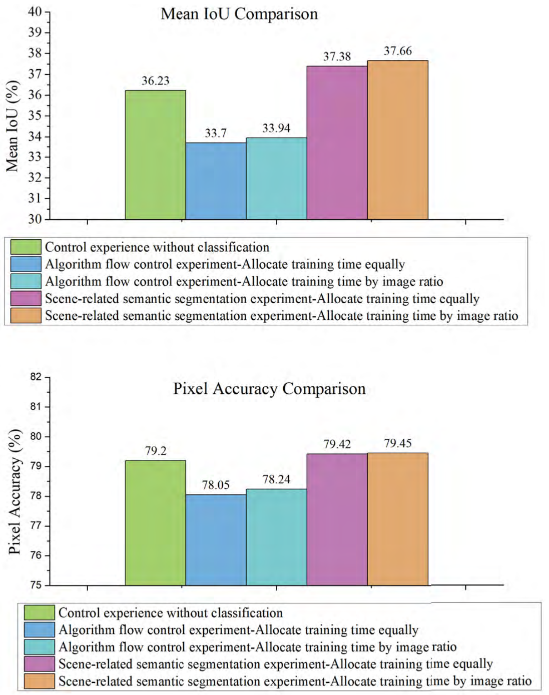


FIGURE 12. Algorithm flow comparison experiment results.

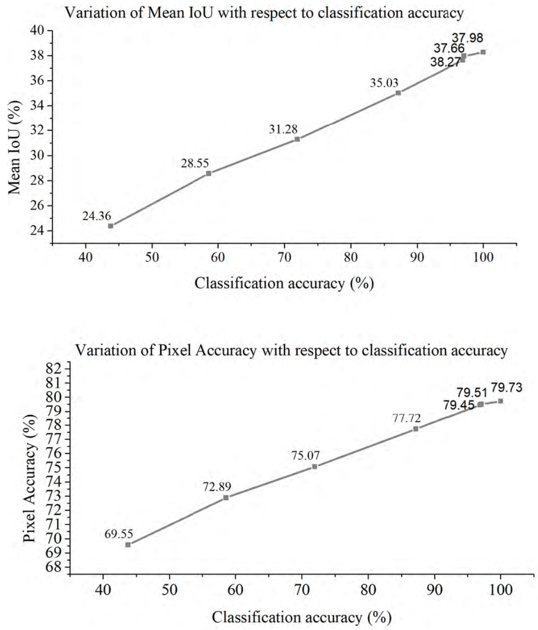


FIGURE 13. Effect of classification accuracy on scene-aware semantic segmentation.

scene category obviously had higher pixel accuracy (Pixel Accuracy) and average cross ratio (Mean IoU). It verifies the hypothesis mentioned above that the addition of scene category information does make the semantic segmentation algorithm get better results. However, the premise is the scene category information, rather than the process itself, which is much more complex than the control group.

In the two ways of allocating training time, the proportion of distribution according to the picture is better than the average allocation. For the ratio of the pictures in the three scenes is too different, the training time should be divided according to the proportion of the picture. If the distribution is simply average, the semantic segmentation network model will be over-fitting in the scene with especially few pictures with the increase of training time, but the semantic segmentation model is still in an under-fitting state in the scene with a lot of pictures. The experimental results also prove that the distribution according to the picture is a more reasonable choice.

In comparison with the final results, the experimental results using the Deep Lab v3+ network structure are much better than those using DeepLab v2. It is mainly due to the advantages of DeepLab v3+ network structure, including more efficient decoders, stronger feature extraction capabilities, etc., which is consistent with the results of section IV-B.

D. ALGORITHM FLOW CONTROL EXPERIMENT RESULTS

After obtaining the test results, we use Pixel Accuracy and Mean IoU as the quantitative evaluation criteria. By using DeepLab V3+ network structure with 480K total training, the experiments of scene-aware semantic segmentation in section 4.2 and non-categorized control experiments in section 4.1 are compared. The comparison results are shown in Figure 12.

When compared with the control experiment in which the category is not divided in section IV-C, the accuracy of the algorithmic flow using randomized image category does not improve, but decreases, which shows that the improvement of accuracy in the semantic segmentation experiment related to the section IV-B scenario is not from the algorithm flow, but the addition of the scene category information.

E. THE ACCURACY OF THE SCENE CLASSIFICATION ALGORITHM AFFECTS THE RESULTS

We also try to improve the scene classification algorithm implemented in section III-B by changing the parameters and increasing the number of training iterations. The classification accuracy rate on the test machine is increased from 96.90% to 97.08%. The scene correlation semantic segmentation algorithm was re-evaluated by using the neural network model after the replacement scenario. The pixel accuracy rate (Pixel Accuracy) increased from 79.45% to 79.51%, and the average cross-over ratio (Mean IoU) increased from 37.66% to 37.98%. Mean IoU and Pixel Accuracy were further improved to 38.27% and 79.73% when directly using scene category annotations (equivalent to a classification accuracy of 100%).

In addition, we also attempt to scramble some of the scene category annotations to obtain a lower classification accuracy rate for comparison, and obtain the variation of Mean IoU and Pixel Accuracy with classification accuracy as shown in Figure 13.

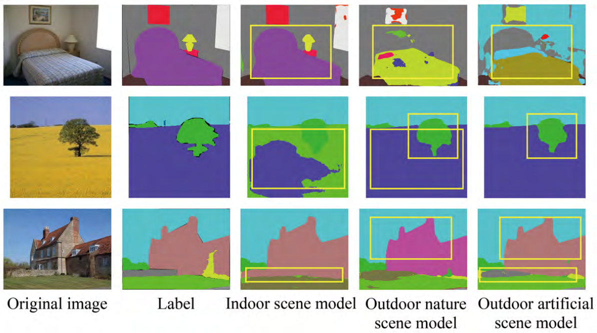


FIGURE 14. Scene-aware semantic segmentation algorithm Comparison of network models in different scenarios.

In the figure, the results of the scene-aware semantic segmentation algorithm are increased with the classification accuracy rate, which shows that better results will be obtained if process the corresponding scene image. For example, the results of indoor image segmentation by indoor semantic segmentation model are better than those by natural segmentation semantic segmentation model. For visual comparison, the comparison of sample images results is shown in Figure 14.

From the first row to the third row in Figure 14 are examples of indoor scenes, outdoor natural scenes, and outdoor artificial scenes. The first column is the original image, the second is the label, and the last three columns are the results of the indoor semantic segmentation model, outdoor natural scene semantic segmentation model and the outdoor artificial scene semantic segmentation model. It can be intuitively found that the image is processed best by using corresponding model, and the semantic segmentation neural network in each scenario learns the scenario category information can improve the results of semantic segmentation, which led to a better final result when compared with the semantic segmentation algorithm without the combination of scene categories.

V. CONCLUSION

Our work include below key points:

1.For the semantic segmentation algorithm, we choose to use the deep convolutional neural network to implement the semantic segmentation algorithm, and select DeepLab v2 and DeepLab v3+ networks. The obtained algorithm was trained on the ADE20K data set, whose result was used as a control to compare with the scene-related semantic segmentation algorithm.

2.In order to classify the scenes of the pictures in the test set, we choose to use the deep convolutional neural network to implement the scene classification algorithm, and select the VGGNet network structure. The training set pictures of ADE20K data set are divided into three categories: indoor scene pictures, outdoor natural scene pictures and outdoor artificial scene pictures. We trained the obtained algorithm on the ADE20K data set to determine the scene category that distinguish the pending image in the test set.

3.For the scene-aware semantic segmentation algorithm, there are 10279, 2144 and 7757 images of indoor scene pictures, outdoor natural scene pictures and outdoor artificial scene pictures respectively according to the results of the scene classification algorithm. After the data set was classified, we trained the semantic segmentation neural network in three scenarios, and obtained three semantic segmentation neural network models. The test used the scene classification algorithm to determine the result of the image to be tested, and selected the semantic segmentation network model of the corresponding scene for processing. Comparing with the results of the semantic segmentation algorithm, the accuracy of the scene-aware semantic segmentation algorithm is higher than that without considering the scene category, which indicates that the addition of the scene category information can indeed improve the results of the semantic segmentation algorithm.

The shortcoming of this paper is that we use the existing semantic segmentation algorithm instead of proposing a new semantic segmentation network model. The training sample is limited to the picture of the specific scene to combine the scene category information, and the scene category information is combined into the semantic segmentation to improve the result.

Adding category information is just one way to increase a priori to improve the results. The experimental results showed that the result of semantic segmentation can be improved by adding a priori, and thus a new thought can be drawn. The next step is to try more ways to increase the a priori for the preprocessing of semantic segmentation to increase the accuracy of semantic segmentation in the future.

REFERENCES

[1] L.-C. Chen, G. Papandreou, I. Kokkinos, K. Murphy, and A. L. Yuille, "Semantic image segmentation with deep convolutional nets and fully connected CRFs," 2014, *arXiv:1412.7062*. [Online]. Available: https://arxiv.org/abs/1412.7062

[2] 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. 2017.

[3] L.-C. Chen, Y. Zhu, G. Papandreou, F. Schroff, and H. Adam, "Encoder-decoder with atrous separable convolution for semantic image segmentation," in *Proc. Eur. Conf. Comput. Vis. (ECCV)*, Jun. 2018, pp. 801–818.

[4] M. Cordts, M. Omran, S. Ramos, T. Rehfeld, M. Enzweiler, R. Benenson, U. Franke, S. Roth, and B. Schiele, "The cityscapes dataset for semantic urban scene understanding," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2016, pp. 3213–3223.

[5] N. Dalal and B. Triggs, "Histograms of oriented gradients for human detection," in *Proc. Int. Conf. Comput. Vis. Pattern Recognit. (CVPR)*, vol. 1, Jun. 2005, pp. 886–893.

[6] M. Everingham, L. Van Gool, C. K. I. Williams, J. Winn, and A. Zisserman, "The PASCAL visual object classes (VOC) challenge," *Int. J. Comput. Vis.*, vol. 88, no. 2, pp. 303–338, 2010.

[7] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2016, pp. 770–778.

[8] M. Juneja, A. Vedaldi, C. V. Jawahar, and A. Zisserman, "Blocks that shout: Distinctive parts for scene classification," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2013, pp. 923–930.

[9] A. Krizhevsky, I. Sutskever, and G. E. Hinton, "ImageNet classification with deep convolutional neural networks," in *Proc. Adv. Neural Inf. Process. Syst.*, 2012, pp. 1097–1105.

[10] L.-J. Li, H. Su, L. Fei-Fei, and E. P. Xing, "Object bank: A high-level image representation for scene classification & semantic feature sparsification," in *Proc. Adv. Neural Inf. Process. Syst.*, 2010, pp. 1378–1386.

[11] L.-J. Li, H. Su, Y. Lim, and L. Fei-Fei, "Objects as attributes for scene classification," in *Proc. Eur. Conf. Comput. Vis.* Berlin, Germany: Springer, 2010, pp. 57–69.

[12] C. Liu, J. Yuen, and A. Torralba, "Nonparametric scene parsing: Label transfer via dense scene alignment," Inst. Elect. Electron. Eng., Piscataway, NJ, USA, Tech. Rep., 2009. doi: 10.1109/CVPR.2009.5206536.

[13] R. Liu, Y. Chen, X. Zhu, and K. Hou, "Image classification using label constrained sparse coding," *Multimedia Tools Appl.*, vol. 75, no. 23, pp. 15619–15633, 2016.

[14] J. Long, E. Shelhamer, and T. Darrell, "Fully convolutional networks for semantic segmentation," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2015, pp. 3431–3440.

[15] D. G. Lowe, "Distinctive image features from scale-invariant keypoints," *Int. J. Comput. Vis.*, vol. 60, no. 2, pp. 91–110, 2004.

[16] H. Lu, Y. Li, M. Chen, H. Kim, and S. Serikawa, "Brain intelligence: Go beyond artificial intelligence," *Mobile Netw. Appl.*, vol. 23, no. 2, pp. 368–375, Apr. 2018.

[17] H. Lu, Y. Li, S. Mu, D. Wang, H. Kim, and S. Serikawa, "Motor anomaly detection for unmanned aerial vehicles using reinforcement learning," *IEEE Internet Things J.*, vol. 5, no. 4, pp. 2315–2322, Aug. 2018.

[18] H. Lu, Y. Li, T. Uemura, H. Kim, and S. Serikawa, "Low illumination underwater light field images reconstruction using deep convolutional neural networks," *Future Gener. Comput. Syst.*, vol. 82, pp. 142–148, May 2018.

[19] H. Lu, D. Wang, Y. Li, J. Li, X. Li, H. Kim, S. Serikawa, and I. Humar, "CONet: A cognitive ocean network," 2019, *arXiv:1901.06253*. [Online]. Available: https://arxiv.org/abs/1901.06253

[20] R. Mottaghi, X. Chen, X. Liu, N.-G. Cho, S.-W. Lee, S. Fidler, R. Urtasun, and A. Yuille, "The role of context for object detection and semantic segmentation in the wild," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2014, pp. 891–898.

[21] H. Noh, S. Hong, and B. Han, "Learning deconvolution network for semantic segmentation," in *Proc. IEEE Int. Conf. Comput. Vis.*, Dec. 2015, pp. 1520–1528.

[22] A. Oliva and A. Torralba, "Modeling the shape of the scene: A holistic representation of the spatial envelope," *Int. J. Comput. Vis.*, vol. 42, no. 3, pp. 145–175, 2001.

[23] A. Oliva and A. Torralba, "Building the gist of a scene: The role of global image features in recognition," *Prog. Brain Res.*, vol. 155, pp. 23–36, Oct. 2006.

[24] O. Ronneberger, P. Fischer, T. Brox, "U-net: Convolutional networks for biomedical image segmentation," in *Proc. Int. Conf. Med. Image Comput. Comput.-Assist. Intervent.* Cham, Switzerland: Springer, 2015, pp. 234–241.

[25] F. Sadeghi and M. F. Tappen, "Latent pyramidal regions for recognizing scenes," in *Proc. Eur. Conf. Comput. Vis.* Berlin, Germany: Springer, 2012, pp. 228–241.

[26] S. Serikawa and H. Lu, "Underwater image dehazing using joint trilateral filter," *Comput. Elect. Eng.*, vol. 40, no. 1, pp. 41–50, 2014.

[27] K. Simonyan and A. Zisserman, "Very deep convolutional networks for large-scale image recognition," 2014, *arXiv:1409.1556*. [Online]. Available: https://arxiv.org/abs/1409.1556

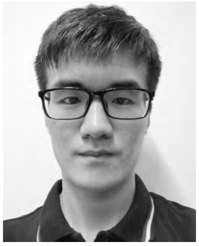
[28] C. Szegedy, W. Liu, Y. Jia, P. Sermanet, S. Reed, D. Anguelov, D. Erhan, V. Vanhoucke, and A. Rabinovich, "Going deeper with convolutions," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2015, pp. 1–9.

[29] C. Szegedy, V. Vanhoucke, S. Ioffe, J. Shlens, and Z. Wojna, "Rethinking the inception architecture for computer vision," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2016, pp. 2818–2826.

[30] H. Zhao, J. Shi, X. Qi, X. Wang, and J. Jia, "Pyramid scene parsing network," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jul. 2017, pp. 2881–2890.

[31] B. Zhou, H. Zhao, X. Puig, S. Fidler, A. Barriuso, and A. Torralba, "Scene parsing through ADE20K dataset," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jul. 2017, pp. 633–641.

ZHIKE YI received the M.S. degree in computer science from Beihang University, where he is currently pursuing the Ph.D. degree with the State Key Laboratory of Virtual Reality Technology and Systems. His research interests include computer graphics, computer vision, machine learning, and image processing.



TAO CHANG received the B.E. degree in computer science from Beihang University, Beijing, China, in 2018, where he is currently pursuing the M.S. degree in computer science and technology. His research interests include computer vision and machine learning.

SHUAI LI received the Ph.D. degree in computer science from Beihang University, where he is currently an Associate Professor with the State Key Laboratory of Virtual Reality Technology and Systems. His research interests include computer vision, image processing, computer graphics, physics-based modeling and simulation, and virtual surgery simulation.

RUIJUN LIU received the Ph.D. degree from Ecole Centrale de Nantes, France, in 2013, and the M.S. degree from Beihang University, in 2009. He is currently with the Beijing Technology and Business University. His current research interests include machine learning, virtual reality, and 3D reconstruction.

JING ZHANG received the B.E. degree in control technology and instruments from Yanshan University, Hebei, China, in 2017. She is currently pursuing the M.S. degree in software engineering with Beihang University, Beijing, China. Her research interests include computer graphics and machine learning.

AIMIN HAO received the B.S., M.S., and Ph.D. degrees in computer science from Beihang University, where he is currently a Professor with the Computer Science School and the Associate Director of the State Key Laboratory of Virtual Reality Technology and Systems. His research interests include virtual reality, computer simulation, computer graphics, geometric modeling, image processing, and computer vision.