Multi-Source Fusion Image Semantic Segmentation Model of Generative Adversarial Networks Based on FCN

LIANG ZHAO, YING WANG, ZHONGXING DUAN, DENGFENG CHEN, AND SHIPENG LIU

1College of Information and Control Engineering, Xi’an University of Architecture and Technology, Xi’an, Shaanxi 710055, China
2School of Building Services Science and Engineering, Xi’an University of Architecture and Technology, Xi’an, Shaanxi 710055, China

Corresponding author: Zhongxing Duan (zhaoliange32011@yeah.net)

This work was supported in part by the National Natural Science Foundation of China under Grant 12002251 and Grant 51209167, in part by the Nature Science Foundation of Shaanxi Province under Grant 2019JM-474, and in part by the Science and Technology Project of Xi’an under Grant 2020KJRC0055.

ABSTRACT At present, most of the methods used in the research of image semantic segmentation ignore the low-level feature information of image, such as space, edge, etc., which leads to the problems that the segmentation of edge and small part is not precise enough and the accuracy of segmentation result is not high. To solve this problem, this paper proposes a multi-source fusion image semantic segmentation model of generative adversarial networks based on FCN: SCAGAN. In VGG19 network, add super-pixel and edge detection algorithm, and introduce the efficient spatial pyramid module to reduce the number of parameters while adding the spatial and edge information of image; Adjust the skipping structure to better integrate the low-level features and high-level features; build a generation model DeepLab-SCFCN combining with the atrous spatial pyramid pooling to better capture the feature information of different scales of the target for segmentation; The FCN with five modules is designed as the discrimination model for GAN. It is verified on the data set PASCAL VOC 2012 that the model achieves IoU of 70.1% with a small number of network layers, and the segmentation effect of edge and small part is better at the same time. This technology can be used in image semantic segmentation.

INDEX TERMS Image semantic segmentation, fully convolutional networks, generative adversarial networks, efficient spatial pyramid, atrous spatial pyramid pooling.

I. INTRODUCTION

Image semantic segmentation is the combination of image recognition and image segmentation, and classifies every pixel in the image and gives semantic labels, so as to achieve accurate segmentation. It is a very important technology in computer vision and image recognition [1].

Traditional image semantic segmentation is based on probabilistic graphical models (PGM), which mainly includes generative models, discriminative models and condition random [2]. The results of these models are difficult to achieve satisfactory accuracy and easy to cause error segmentation. In 2015, Shelhamer et al. [3] proposed fully convolutional networks (FCN), which is a kind of network that can achieve pixel to pixel. Liu et al. [4] added the extracted global features to the local feature information for feature fusion, and input it into the network together with the context information. Ghiasi and Fowlkes [5] proposed a new model LRR (Laplacian pyramid reconstruction and refinement model), using Laplacian pyramid algorithm [6] to recombine the feature information extracted from different convolutional layers. After scaling the original images, Chen et al. [7] input these images in FCN, gave the weight of the target in each scale, and fuse multi-scale features to classify pixels. Yu and Koltun [8] introduced dilated convolution to optimize standard convolution. Without loss of resolution and increase of computation, the receptive field and precision are improved. In 2017, Wang et al. [9] introduced hybrid dilated convolution (HDC) and dense up sampling convolution (DUC), which increased the receptive field (RF) and made better use of local information.

The first application of the generative adversarial networks (GAN) [10] in image is image generation. GAN is also used in image editing, malicious attack detection [11], data
generation [12], attention prediction [13], three-dimensional structure generation, causal reasoning [14]. Afterword, it was introduced into semantic segmentation to optimize and train the model, and got better performance. In the GAN, there are two main parts, generator and the discriminator. The two parts are trained by the way of confrontation. Isola et al. [15] first applied GAN to image semantic segmentation. This method can make the network generate samples closer to the target data without increasing the complexity of post-processing. The main image semantic segmentation network is used as the generator, and a new neural network is designed as discriminator. This structure has become a general framework for image semantic segmentation. However, it cannot be used for all data sets and has no significant performance improvement. Souly et al. [16] first used a full convolutional network discriminator as a segmentation network. On the data set side, it can be supplemented by samples generated by the generator. This is a semi-supervised method, and there is a big difference between the samples produced by generator and the real samples, so the effect of improving network accuracy is not obvious. Hung et al. [17] used the mainstream image semantic segmentation network as the generator, and the discriminator used the FCN. There are two advantages of designing the discriminator in this way. On the one hand, it can accurately determine each input source pixel (network segmentation result graph or standard segmentation graph). On the other hand, it can generate the confidence graph of input image without ground truth, which can be used as additional judgment information of semi-supervised learning. However, the image semantic segmentation network used in this method is too single, which failed to make effective use of the rich semantic feature information in the lower layer, and does not use multi-scale information fusion. In this paper, we propose a multi-source fusion generative adversarial network image semantic segmentation model, named SCAGAN, which is based on FCN. Where, S stands for super pixel SLIC algorithm [18]. C is the Canny edge detection algorithm. A represents the Atrous Spatial Pyramid Pooling (ASPP) module. The last 3 letters GAN is the generative adversarial networks mentioned above. This method combines the traditional image segmentation method with the deep learning method. This method makes full use of the edge and spatial information of the image to do the fusion of low-level and high-level features, and improves the segmentation accuracy by generative adversarial network.

II. METHODS
A. FULLY CONVOLUTIONAL NETWORKS
Full convolutional networks (FCN) uses an end-to-end network structure to solve the image semantic segmentation. That is, after the image is input into the network, the pixel class prediction is directly output. FIGURE 1 shows an end-to-end network structure, with its convolution and deconvolution on the left.

The whole structure of FCN is divided into full convolutional part and deconvolution part. Among them, the convolutional part is the classic convolution neural network [19], [20] (CNN). On the basis of CNN, the last full connection layer is replaced by convolution, which is used to extract features to generate a hot spot image. The deconvolution part is sampling the small hot spot image to get the original size of the semantic segmentation image. The structure of FCN is shown in Figure 2.

After many convolutions and five pooling operations, the original image features are extracted. After each pooling operation, the size of the image is reduced to 1/2 of the original image, and after five pooling operations, the image size changed to 1/32, with the resolution is gradually reduced. In order to obtain the resolution of the original image from the low-resolution image, and save the spatial information of the image to realize the classification by pixel. This paper adopts up-sampling to restore the feature map to the original image size.

Since the image obtained by convolution, pooling and up-sampling is relatively rough, this paper uses the skipping connection in FCN. As shown in FIGURE 2, the third and fourth feature maps after pooling were taken out and then a $1 \times 1$ convolutional layer is adopted. The convolutional layer outputs 21 channels, and then reduces the number of channels to 20 through the $1 \times 1$ convolutional layer. A convolutional layer is added here for prediction, and the prediction results are directly obtained in the third and fourth pooling layers. In the process of convolutional, the feature information obtained from the low-level layers is more precise and precise. The feature information obtained from the high-level layers is more abstract and robust, but the location information is not clear. While the skipping connection can make a fusion of the low-level information and the high-level information, and obtain more accurate results.

B. EFFICIENT SPATIAL PYRAMID NETWORKS
Efficient spatial pyramid module (ESP) [21] is a new convolution module, which is efficient in calculation, storage and
FIGURE 3. Factorize standard convolution to point convolution and spatial pyramid dilated convolution.

FIGURE 4. Schematic diagram of ESP module.

power, which separated standard convolution into point convolution and spatial pyramid dilated convolution. As shown in Figure 3. The point convolution in the ESP module uses the $1 \times 1$ convolution to project the high-dimensional feature map into the low-dimensional space. The spatial pyramid dilated convolution uses $K$ and $n \times n$ dilated convolution to resample these low dimensional feature maps at the same time. The dilation rate of each feature map is $2^{k-1}$, $k = \{1, 2, \ldots, K\}$. This factorization greatly reduces the number of parameters and memory required by ESP, while retaining a large effective receiving range of $((n-1)2K-1+1)^2$.

The strategy of ESP module is divided into four parts: reduction, split, transformation and concatenation. As shown in Figure 4. The total parameters are $\frac{MN}{K} + (N^2K^2)$. Where the $M \times \frac{N}{K}$ is reduction, and $K \times n^2 \times (\frac{N}{K})^2$ is the parameters of split.

In this paper, three ESP modules were used to replace the three full connection layers of the original VGG19, and a $1 \times 1$ convolution was attached behind the ESP module.

C. SCFCN

FCN can only obtain a limited local pixel spatial relationship. Because of the calculation of the convolutional layer. The segmentation results lack spatial consistency and are not sensitive to details. While the super pixel SLIC algorithm can provide more detailed information, the Canny algorithm can enhance the edge information of the image. Therefore, we combine these three methods and propose a pyramid context fuses image semantic segmentation model, SCFCN. Among them, S represents the Super pixel SLIC algorithm, and C is the Canny operator. Super pixel SLIC algorithm provides more detailed segmentation information, and the segmentation of Canny [22] algorithm increases the information of image edge. In training, the network fuses the extracted low-level information and the high-level information to complete the end-to-end semantic segmentation of the image. The schematic diagram of SCFCN algorithm is shown in Figure 5.

The steps that the algorithm follows:

1. The original images is processed by SLIC algorithm to get the new image data set of super-pixel segmentation.
2. The original images is processed by Canny algorithm to get the image data set of image edge segmentation.
3. The original data set and the new data set processed in the first two steps were used together as the input images of SCFCN network to train a new image semantic segmentation model.
4. The validation set image was input in SCFCN, and the segmentation result obtained by Canny algorithm was corrected by coincidence with the edge segmentation image obtained by Canny algorithm, and the final segmentation result was output.

Shelhamer et al. [3] mentioned that the fusion of more layers of information did not significantly change the segmentation results, so they did not study the fusion of lower layers of information. However, the SCFCN in this paper add the images processed by SLIC algorithm and Canny algorithm as the input. The low-level information is particularly important, including the super-pixel segmentation information and edge segmentation information.

After the first and the second pooling layer, we design the skipping connection, so that the low-level features obtained...
by SLIC algorithm and Canny algorithm can be more effectively preserved. And more spatial information and edge information are added in FCN. The low-level features from the first and second pooling are fused with the high-level features from the network output, and output the segmentation results by up-sampling to the original image. Figure 6 shows the network structure of SCFCN.

**D. ATROUS SPATIAL PYRAMID POOLING**

Atrous Spatial Pyramid Pooling (ASPP) [23] module which was proposed from DeepLabv2 uses a multi-scale context capture method for images, and parallel sampling with dilated convolution with different sampling rates is a significant feature. In order to obtain multi-scale information of the target, we applied ASPP to our proposed algorithm. In the ASPP module, the interval between adjacent weights of dilated convolutional rate is \( r - 1 \), and the default value of ordinary convolution rate is 1.

So the actual size of the dilated convolutional is \( k + (k - 1)(r - 1) \). Where \( k \) is the size of the original convolutional kernel.

Figure 8 shows the dilated convolution with \( 3 \times 3 \) kernel size in the ASPP module. The expansion rate is 6, 12, 18, 24, and the step size is 1. The corresponding filling rate \( p \) of input feature can be calculated by Eq (1), and \( p \) is set to 6, 12, 18, 24, respectively.

\[
W_{out} = W_{in} + 2p - [k + (k - 1) \times (d - 1)] / s + 1 \tag{1}
\]

Parameter setting will affect the input and output characteristics of dilated convolution. Such parameter setting has two advantages: the first is to get the same size of output, and the second is that the resolution of input and output feature is the same. In this paper, we make some adjustments to ASPP, and set the number of channels of each dilated convolution as the total number of categories instead of the original 1 \( \times \) 1 convolution, which can simplify ASPP. After fusing the features extracted from different scales, the segmentation results are obtained by bilinear interpolation method and Softmax classifier.

**E. GENERATOR OF GAN**

Inspired by the characteristics of each version of DeepLab, a new encoder-decoder model DeepLab-SCFCN is designed as a generator model of GAN. DeepLab-SCFCN combines DeepLabv2 and SCFCN. The structure diagram is shown in Figure 8.

In order to capture the context information of different scales of the target and improve the segmentation accuracy and edge accuracy of the network, the ASPP in DeepLabv2 is introduced between the encoder and decode. Encoder is the convolutional part of SCFCN, and decoder obtains the same size of the original image through ASPP and up sampling.

**F. DISCRIMINATOR OF GAN**

According to the relationship between the discriminator and generator, a full convolutional network is designed as the discriminator. The structure of the full convolutional network is shown in the Figure 9.

The discrimination network is composed of five convolutional layers, \( 3 \times 3 \) kernel size, step size 1. The number of channels in each convolutional layer is 256, 128, 64, 32, 16, 1, respectively. An up-sampling layer is added at the end of the discriminator, so that the output image with the same size as the original image can be obtained. Figure 9 shows the structure of the discriminator.

The discriminator can judge whether the input image is the result image of model segmentation or the ground truth. So the input of discriminator should have two sets of images. When the input image is the original image and the ground
image. The loss function is defined as shown in Eq (2).

\[ l_{mce}(y, y') = \sum_{i=1}^{H\times W} \sum_{m=1}^{M} y_{im} \ln y'_{im} \]

where \( l_{mce}(y, y') \) represents the multi classificaon cross entropy, the number of train set is \( N \), the segmentation label is \( y_{im} \), the segmentation network is recorded as \( s() \), \( y \) represents the segmentation label using one-hot coding, \( s(x_{n}) \) and \( y' \) represent the prediction label of the segmentation network and the segmentation model, respectively.

Using the difference between segmentation results and ground truth, the discrimination model can judge whether the input image is the segmentation result of the generated model or ground truth, and can’t define the specific loss function, which simplifies the discrimination process. The discriminator can continuously optimize the generated model by judging the true and false (1,0) effect of the input image, and finally get the optimal result. If the discriminant model is set as \( d() \), the form of its loss function is shown in Eq (3).

\[ l_d = l_{bce}(d(s(x_{n})), 0) + l_{bce}(d(y_{n}), 1) \]  

where, \( l_{bce}(z, z') = -[z \ln z' + (1-z)\ln(1-z')] \) represents the cross entropy of binary classificaon.

The discriminator is a FCN model. After multi-layer convolutional operation, the features are extracted and classified. The objecvon function of the model is as follows Eq (4):

\[ L(s, d) = l_s + l_d \]

III. EXPERIMENTAL RESULTS AND ANALYSIS

A. EXPERIMENTAL ENVIRONMENT AND DATA

In this paper, the hardware of the experiment is that with Nvidia GeForce RTX 2080Ti as the graphics card, and Intel Core i7-9700k@3.60GHz as the processor eight core, Asustek PRIME Z390-A (Z390 chipset group) as main board. The software uses Pycharm as compiler and python as programmg language. The specific experimental environment configuration is shown in Table 1.

PASCAL VOC 2012 [24] is adopted for the data set. The data set includes 11530 images of train set and validation set, of which 27450 are manually marked. There are 20 semantic categories such as human, cat, car and dining table in the tag, and 21 categories in total with background. PASCAL VOC 2012 shows the 20 categories, each with assigned colors and numbers, with black background.
FIGURE 11. The basic structure of the algorithm in this chapter Optim
means optimizer.

B. EXPERIMENT
In this paper, we use Tensorflow framework to improve the
algorithm on Linux. Based on SCFCN network, we add the
ASPP in DeepLab-v2 to transform it into encoder-decoder
structure as generator of GAN, and design 5 modules of FCN
as the discriminator. The basic structure of the algorithm
is shown in Figure 11. The super parameter settings in the
experiment are shown in Table 2.

C. EXPERIMENTAL RESULTS AND ANALYSIS
Mean Intersection over Union (MIoU): a standard measure of
semantic segmentation. It calculates the ratio of intersection
and union of two sets. In the problem of semantic segmenta-
tion, the two sets are real value and predicted value.

In order to reduce the memory consumption, the original
image is resize to $224 \times 224$. Let batch_size = 10, the
initial learning rate is set to 1e-4, momentum is 0.95, and the
maximum number of iterations is set to 20000.

The training process uses the train set of PASCAL VOC
2012 for training and test set for validate. Finally, the IoU of
each category in the model is shown in Table 3. The MIoU of
the whole model is in the last row of the Table 3. The MIoU
is the operation of summing up and then averaging the values
of all 21 categories. The final MIoU of this model is 70.10%.

It can be seen from the IoU data of the test results of the
model in Table 3.

(1) The model is the best for background segmentation.
In addition to the 20 types of target in the data set, all parts
are divided into background, so the area of background is
relatively large. The background is relatively complete, so it
is easy to train and segment.

(2) The results of Aeroplane, bird, bus, cat, motorbike
and sheep are also good. Most of these types have relatively
fixed contour, the overall contour area is relatively large
and concentrated, and the difference between them and the
background is relatively large. The area of bird in the image
is generally small, and the parts such as bird legs and beak
are even smaller, but the structure is relatively simple, so the
segmentation effect is better in this model.

(3) The second effect are the bottle, car, cow, dog, horse,
person, sofa, train and TV monitor. These kinds of structures
are relatively single. In the process of training, the number of
each category on the image are more, which is conducive to
feature extraction and learning. Therefore, a better segmenta-
tion effect is achieved.

(4) Bicycle, boat, dining table and potted plant are four
kinds of segmentation methods with poor effect. The features
of boat and dining table are not obvious, and they are easy to
be segmented by mistake with other categories when extract-
ing the features. The structure of bicycle is complex, and the
spokes on tire cross and block each other, and the number
is large, which affects the segmentation results of the model.
It is also difficult to divide plants in potted plants.

(5) Chair is the worst in all 21 categories. For the four
legs of the chair are relatively thin, the area occupied in
the image is relatively small. At the same time, due to the
small space between the legs of the chair, the probability of
mutual occlusion and overlap between the legs of the chair
obtained from different angles is high, and there is a clear
space between the legs of the chair, which corresponds to the
background pixel mark. If the space is too small, it is not easy
to divide the model in detail. So the chair is the worst one.

FCN, SCFCN and SCAGAN algorithms are tested in the
same experimental environment and data set. Table 4 is the
comparison of the segmentation results of the three algo-
rithms. The segmentation results are specific to the IoU value
of each category, and finally use the MIoU value to measure
the segmentation accuracy of the model.

It can be seen from Table 4 that MIoU of SCFCN is 6.9%.
higher than FCN, SCAGAN is 3.8% higher than SCFCN and
TABLE 4. IoU comparison of three algorithms.

| Class        | FCN     | SCFCN   | SCAGAN  |
|--------------|---------|---------|---------|
| background   | 84.0%   | 92.1%   | 94.5%   |
| Aeroplane    | 74.7%   | 81.6%   | 87.2%   |
| bicycle      | 29.6%   | 37.5%   | 40.6%   |
| bird         | 65.4%   | 74.3%   | 84.2%   |
| boat         | 48.9%   | 54.2%   | 56.7%   |
| bottle       | 57.0%   | 63.9%   | 65.8%   |
| bus          | 74.9%   | 85.8%   | 87.8%   |
| car          | 69.3%   | 76.2%   | 76.5%   |
| cat          | 70.9%   | 78.8%   | 80.1%   |
| chair        | 20.0%   | 25.4%   | 35.4%   |
| cow          | 63.1%   | 70.0%   | 69.3%   |
| dining table | 45.4%   | 52.3%   | 56.1%   |
| dog          | 68.7%   | 75.4%   | 74.3%   |
| horse        | 64.8%   | 76.7%   | 79.1%   |
| motorbike    | 71.4%   | 78.3%   | 83.2%   |
| person       | 69.0%   | 70.7%   | 76.3%   |
| pottedplant  | 42.5%   | 47.1%   | 48.1%   |
| sheep        | 71.3%   | 79.2%   | 83.2%   |
| sofa         | 37.1%   | 42.3%   | 60.3%   |
| train        | 66.8%   | 72.7%   | 70.0%   |
| tvmonitor    | 53.5%   | 58.4%   | 63.5%   |
| MIoU         | 59.40%  | 66.30%  | 70.10%  |

10.7% higher than FCN. Based on the above data, the model SCAGAN has a corresponding accuracy improvement in the segmentation of each category. The IoU obtained by the algorithm is expressed in a more intuitive way as shown in Figure 12. Where the horizontal axis is the categories and the vertical axis is the value of the IoU.

From the bar chart in Figure 12, it can be seen that the segmentation effect of the algorithm proposed in this chapter on sofa is the most improved, reaching 60.3% IOU, and 23.2% higher than FCN. There are many kinds of birds, the size of birds is generally small. The beak and claw are even smaller, which increases the difficulty of segmentation. The algorithm in this chapter also has a significant improvement in bird segmentation, which is 18.8% higher than FCN. In addition, in the segmentation effect in FCN, only 20.0% of chairs have reached 35.4% in the algorithm in this chapter and increase 15.4%. Horse, bus, aeroplane, sheep, motorcycle, bicycle, dining table, background, TV monitor and other types all increase more than 10% compared with FCN. Boat, bottle, truck, cat, cow, dog, person and potted plant all increased by more than 5%. The IoU of train has the least improvement, which is only 3.2%. The total average accuracy of this algorithm is 10.7% higher than FCN.

In this paper, we compared the visualization of the results of three algorithms. Each category is shown in a different color. The visualization results are shown in Figure 13.

It can be seen from Figure 13 that the segmentation results of horse legs, aircraft landing gears and people have clear edges and obvious continuity. Among them, the landing gear part of the aircraft is very small, but the segmentation effect is significantly improved. Therefore, the algorithm in this paper has better effect on edge, small part segmentation and so on.

In this paper, we first get the inspiration from feature fusion and heteromorphic convolution, and made the improvement based on the VGG19 network in CNN. We introduced super-pixel feature (SLIC) and edge detection feature (Canny) and made these two features as low-level features and high-level features for fusion. At the same time, we introduce ESP module on the network and improve the skipping connection to get the model of SCFCN. SCAGAN model is designed with DeepLab-SCFCN as the generator and with five modules of full convolution network as the discriminator. The comparison of SCAGAN algorithm with FCN and DeepLab is shown in Table 5. It can be seen that the MIoU of SCAGAN algorithm is better than that of FCN and DeepLabv1, which verifies its effectiveness. However, the segmentation effect is not better than DeepLabv2 and DeepLabv3, because our SCAGAN algorithm has fewer
network layers than DeepLab algorithm of each version. Increasing the number of layers of the segmented network is one of the future directions of this algorithm.

IV. CONCLUSION

Image semantic segmentation is a very important subject in computer vision. The algorithm proposed in this paper has two main contributions.

(1) The combination of traditional image segmentation method and the depth learning method. This paper introduced the space efficient pyramid module into VGG19, and added image super-pixel information and edge information. A new image semantic segmentation algorithm SCFCN is proposed, and which have achieved good results.

(2) The GAN is applied in the field of image semantic segmentation. In SCFCN, the pyramid pooling of dilated space is added as the generation model, and a full convolution network including five modules is designed as the discrimination model. A new generation antagonism network model SCAGAN for image semantic segmentation is proposed, which achieves a MIoU of 70.1% with a small number of network layers, and has a better segmentation effect on the edge and the small part at the same time.

Our future work will focus on improving its efficiency while maintaining its performance. The algorithm needs high real-time performance in practical application.

REFERENCES

[1] Y. Chen, J. Tao, L. Liu, J. Xiong, R. Xia, J. Xie, Q. Zhang, and K. Yang, “Research of improving semantic image segmentation based on a feature fusion model,” J. Ambient Intell. Hum. Comput., vol. 623, pp. 1–13, May 2020.

[2] K. Hu, S. Zhang, and X. Zhao, “Context-based conditional random fields as recurrent neural networks for image labeling,” Multimedia Tools Appl., vol. 79, nos. 23–24, pp. 17135–17145, Jun. 2020.

[3] E. Sheheramer, 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, doi: 10.1109/TPAMI.2016.2572683.

[4] D. Wu, A. Rabinovich, and A. C. Berg, “ParseNet: Looking wider to see better,” 2015, arXiv:1506.04579. [Online]. Available: https://arxiv.org/abs/1506.04579

[5] G. Ghiasi and C. C. Fowlkes, “Laplacian pyramid reconstruction and refinement for semantic segmentation,” in Computer Vision (Lecture Notes in Computer Science), vol. 9907, B. Leibe, J. Matas, N. Sebe, and M. Welling, Eds. Cham, Switzerland: Springer, 2016, doi: 10.1007/978-3-319-46847-9_32.

[6] P. J. Burt and E. H. Adelson, “The Laplacian pyramid as a compact image code,” IEEE Trans. Commun., vol. COM-31, no. 4, pp. 532–540, Apr. 1983, doi: 10.1109/TCOM.1983.1095851.
L. C. Chen, Y. Zhu, G. Papandreou, F. Schroff, and H. Adam, “Encoder-decoder with atrous separable convolution for semantic image segmentation,” in Computer Vision (Lecture Notes in Computer Science), vol. 11211, V. Ferrari, M. Hebert, C. Sminchisescu, and Y. Weiss, Eds. Cham, Switzerland: Springer, 2018, doi: 10.1007/978-3-030-01234-2_49.

LIANG ZHAO received the B.S. degree in electronic information engineering and the M.S. degree in control science and engineering from the Xi’an University of Technology, in 2003 and 2006, respectively, and the Ph.D. degree in intelligent building information technology from the Xi’an University of Architecture and Technology, in 2012. From 2014 to 2015, he was a Visiting Scholar with the Department of Electrical and Computer Engineering, Virginia Polytechnic Institute and State University. He is currently an Associate Professor with the Department of Information and Control Engineering, Xi’an University of Architecture and Technology. His main research interests include intelligent building information processing, image processing, and pattern recognition.

YING WANG is currently pursuing the M.S. degree major in intelligent building with the Department of Information and Control Engineering, Xi’an University of Architecture and Technology. Her research interests include the fields of computer control, image processing technique, intelligent building information processing, artificial intelligence, and computer vision.

ZHONGXING DUAN received the B.S. degree major in industrial electric automation and the M.S. degree major in computer application from the Xi’an University of Architecture and Technology, in 1992 and 1999, respectively, and the Ph.D. degree major in computer architecture from Xi’an Jiaotong University, in 2006. He is currently a Professor with the Department of Information and Control Engineering, Xi’an University of Architecture and Technology. His main research interests include intelligent detection and intelligent control, and building environment optimization control.

DENGFENG CHEN received the B.S. degree in industrial automation from the Shaanxi Institute of Technology, in 1997, and the M.S. degree in control theory and control engineering and the Ph.D. degree in control science and engineering from Northwestern Polytechnical University, in 2004 and 2011, respectively. From 2012 to 2013, he was a Visiting Scholar with the Department of Electrical and Computer Engineering, The Ohio State University. He is currently an Associate Professor with the School of Building Services Science and Engineering, Xi’an University of Architecture and Technology. His main research interest includes building intelligent system and technology.

SHIPENG LIU received the B.S. degree major in automation from Zhoukou Normal University, in 2019. He is currently pursuing the M.S. degree major in intelligent building with the College of Information and Control Engineering, Xi’an University of Architecture and Technology. His research interests include the fields of deep learning, image processing technique, artificial intelligence, and computer vision.

* * *