Journal of Network Intelligence
Taiwan Ubiquitous Information

©2023 ISSN 2414-8105 (Online)
Volume **8**, Number **4**, November **2023**

Lightweight Semantic Segmentation Algorithm Based on Multi-level Feature Fusion

Guang Huo

Department of Computer Science
Northeast Electric Power University
Jilin, 132012, China
yanhuo1860@qq.com

Yan-Ran Wang*

Department of Computer Science
Northeast Electric Power University
Jilin, 132012, China
2687232208@qq.com

Yan-Chang Liu

Department of Computer Science
Northeast Electric Power University
Jilin, 132012, China
2231606230@qq.com

Ru-Yuan Li

Department of Computer Science
Northeast Electric Power University
Jilin, 132012, China
53921874@qq.com

*Corresponding author: Yan-Ran Wang

Received November 6, 2022, revised January 4, 2023, accepted March 17, 2023.

Abstract. *With the increasing demand for application scenarios such as autonomous driving, robotics, and drones, segmentation networks that require high-performance GPUs to operate cannot satisfy these computationally constrained devices. How to design a lightweight semantic segmentation network with low parameters and fast running speed becomes a challenging problem. Therefore, this paper proposes a lightweight semantic segmentation method based on multi-level feature fusion. The algorithm takes MobileNetv2 as the core to build a lightweight feature extraction network. The feature enhancement module is designed and multi-level features are integrated, enabling the model to obtain more information generated by the backbone network. The combination of CEloss and Diceloss loss function is proposed to reduce the impact of sample imbalance. Experimental results on the PASCAL VOC 2012 dataset show that the segmentation accuracy of the algorithm is 72.61% and the average time of a single picture is 45ms, achieving a good balance between real-time performance and accuracy.*

Keywords: Semantic segmentation; lightweight network; feature fusion; loss function

1. **Introduction.**

Artificial intelligence technology is gradually helping people get rid of pure repetitive work and has been widely used. Relevant research has become a hot spot in recent years [1]. Semantic segmentation is an important image preprocessing method in the field of artificial intelligence. Compared with object detection and image classification, the pixel-by-pixel feature of semantic segmentation makes it suitable for various applications in the field of artificial intelligence, such as autopilot, robot and medical assistant diagnosis.

Since the emergence of deep learning algorithm, CNN (convolutional neural network) has received more and more attention [2]. At present, image semantic segmentation technology is mainly based on depth learning neural network [3]. These models [4, 5, 6, 7, 8, 9] improve network performance by introducing more parameters and various complex operations. For example, the parameter of the classic split network PSPNet network is 65.7M. It takes several seconds to process images on the Titanxp GPU. These high-precision segmentation network models usually have a large number of parameters and require a lot of memory resources. This limits the application of these high-precision networks in equipment with limited computing power, such as autopilot, robot vision and medical assistant diagnosis. In particular, the storage space of smart phones is limited, and it is unlikely to use hundreds of MB of memory to store these segmentation models. With the increasing demand for application scenarios such as autopilots, robots and unmanned aerial vehicles, high-precision segmentation networks that require high-performance GPUs to operate cannot meet these devices with limited computing power. Therefore, many scholars at home and abroad have proposed some effective algorithms to solve this challenging problem. ENet [10] abandoned the final stage of the model in order to pursue an extremely compact framework, resulting in that the acceptance domain of the model is not enough to cover large objects, making the segmentation ability very poor. BiSeNet [11] proposes a new bidirectional split network, which improves the speed of split network. ESPNetV2 [12] uses void convolution and depth-separatable convolution to reduce model parameters. Although it can effectively reduce the number of parameters, the lack of network feature extraction ability leads to low accuracy. CGNet [13] built a lightweight CG module to learn the joint features of local features and surrounding contexts, and further improved the learning of joint features by introducing global context features. Although the above algorithms reduce the number of parameters and improve the real-time performance to a certain extent, the accuracy does not reach the expected effect. Therefore, how to achieve a reasonable balance between the accuracy and speed of semantic segmentation network in the equipment with limited computing power has become a key issue.

In order to solve the above problems, this paper proposes a multi-level feature fusion lightweight network. In the coding stage, the whole network uses the lightweight MobileNetv2 network to extract image features, and uses ASPP to capture image multi-scale context content information; In the decoding stage, a feature enhancement FE module is designed to enhance the multi-level features generated by the backbone network, and then merge them with the high-level feature map, so as to make full and effective use of the semantic information of the low-level feature map and solve the problem of rough image segmentation. At the same time, a scheme combining CEloss and Diclose loss functions is proposed to solve the imbalance between positive and negative samples. The experimental results show that our algorithm achieves 72.61% accuracy of MIOU on the dataset, and the average time of a single image is 45ms, and achieves a good balance between real-time and accuracy.

2. **Approach.**

2.1. **The overall structure.** In order to make the algorithm of this paper have real-time segmentation speed and high segmentation accuracy, model of this paper is designed on the basis of the current most advanced semantic segmentation algorithm DeepLabv3. The whole network structure of this paper is shown in Figure 1.

In the encoding stage, the lightweight MobileNetv2 is used as the basic network structure to extract features, and then the extracted feature maps are input to the Atrous Spatial Pyramid Pooling module (ASPP) to capture the multi-scale contextual content information of the image. The size of the convolution kernel limits the range of dependency captures betweendata samples [14]. Therefore, in order to expand the receptive field, the hole convolution sequence with hole ratio of 6, 12 and 18 is used in the ASPP module. The selection of hole ratio here is the same as DeepLabv3+, which makes the segmentation performance better. Besides, a 1×1 convolution and image pooling are parallelized to obtain more features. These feature map obtained by the ASPP module is spliced and fused in the channel dimension, and the dimension of the feature map is reduced by 1×1 convolution.

In the decoding stage, a feature enhancement module FE is designed to enhance the details of the low-level features generated by the backbone network. Then it is spliced and fused with the features output from the up-sampled encoding stage, 3×3 convolution is used to refine the features, and bilinear interpolation is used again for up sampling. Finally, output the final prediction graph of semantic segmentation.

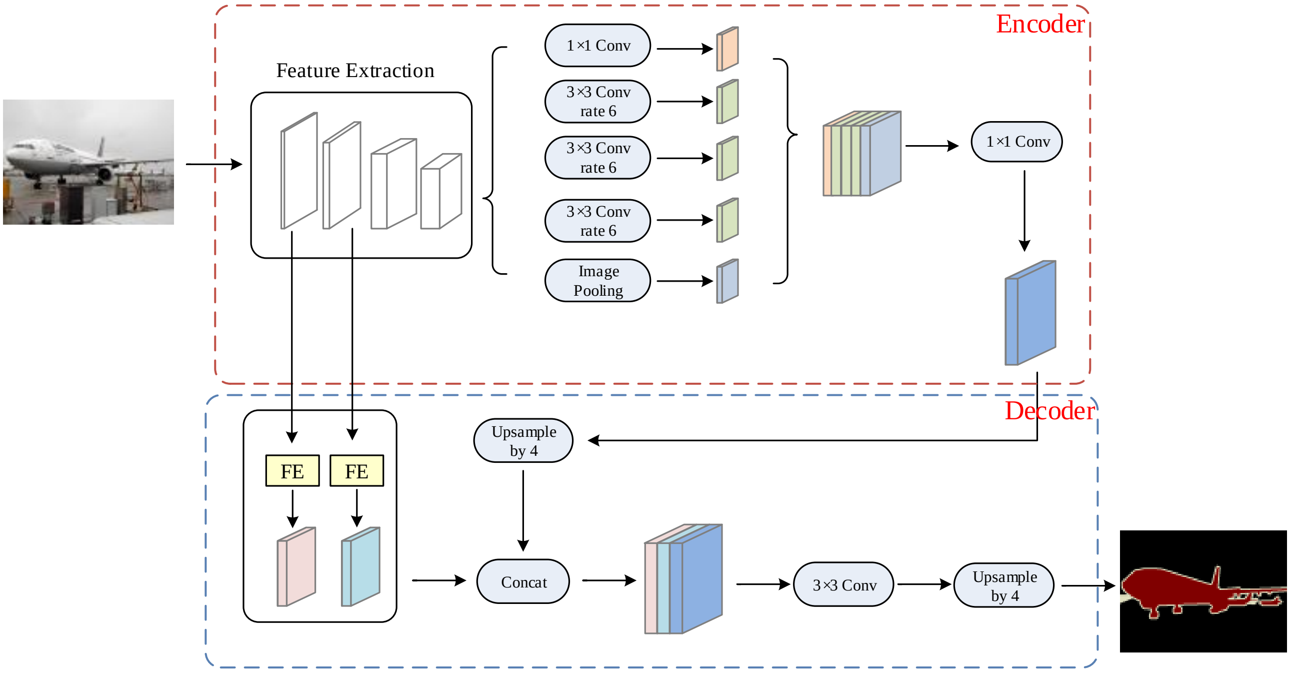


Figure 1. The overall structure of the algorithm in this paper

2.2. **Backbone network MobileNetv2.** MobileNetv2 is a lightweight neural network focused on mobile terminals or embedded devices proposed by the Google team in 2018. In standard convolution, each convolution kernel performs convolution operations simultaneously on all channels of the input [15]. While MobileNetv2 replaces ordinary convolution with deep separable convolution, which can reduce the number of parameters and speed up the calculation. At the same time, the inverted residual structure is used to further improve the performance of the network. The above structure can make MobileNetv2 greatly reduce the amount of parameters and computation of the model when the accuracy rate only decreases by a small margin. The structure of MobileNetv2 is shown in Table 1. Where t is the expansion factor, c is the output channel, n is the number of repetitions, and s represents the step size. In this model, the last pooling layer and full connection layer are deleted from the original MobileNetv2 structure. Because the original MobileNetv2 directly classifies the target after extracting the feature, and the model in this paper also needs to further process the extracted feature, so the final full connection layer is not required.

Table 1. Backbone Extraction Network Structure

input	Operator	t	c	n	s
$512^2 \times 3$	conv2d	-	32	1	2
$256^2 \times 32$	conv2d	1	16	1	1
$256^2 \times 16$	bottleneck	6	24	2	2
$128^2 \times 24$	bottleneck	6	32	3	2
$64^2 \times 32$	bottleneck	6	64	4	2
$32^2 \times 64$	bottleneck	6	96	3	1
$32^2 \times 96$	bottleneck	6	160	3	2
$16^2 \times 160$	bottleneck	6	320	1	1
$16^2 \times 320$	conv2d 1×1	-	1280	1	1

It should be noted that there is residual connection if and only if the step size is 1, as shown in Figure 2(a), and if the step size is 2, there is no skip connection, as shown in Figure 2(b). MobileNetv2 uses the inverted residual structure to increase the dimension first and then reduce the dimension, so that the network allows smaller input and output dimensions, thereby reducing the calculation cost and parameter amount of the network.

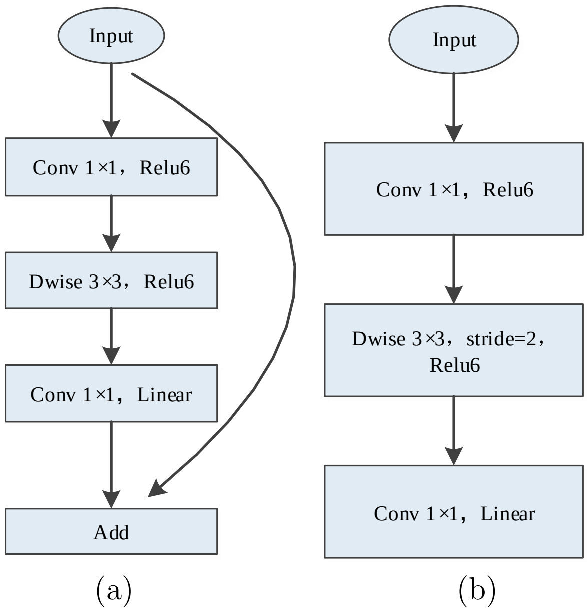


Figure 2. Bottleneck structure of MobileNetv2

2.3. **FE module design.** In encoder-decoder based semantic segmentation network structure, the low-level features generated by the backbone network have rich spatial and detailed information, such as the primary edge and spatial bit information of the image.

However, due to less convolution and more noise, these edge and position information are relatively blurred. In order to obtain a more accurate segmentation effect, these low-level features must be further processed and utilized. Therefore, this paper proposes an FE module, as shown in Figure 3. The FE module is used to enhance the local edge contour information and global position information. This module contains two branches. One branch uses 3×3 convolution further extract local edge detail information. In addition, considering that when low-level features and high-level features are fused, the low-level features should have a lower dimension ratio, so adding a 1×1 convolution reduces dimensionality of low-level features. Another branch uses the global average pooling operation to integrate the global spatial information to achieve a better effect of strengthening the global location information. Through the above processing, the two branches contain enhanced global position information and enhanced local contour detail information respectively. The features obtained from the two branches are fused to enhance the overall information of low-level features. The high-level semantic features obtained through the deep backbone network contain rich semantic information. The fusion of low-level features and high-level features enhanced by FE module will greatly help the decoder to recover image resolution. In this paper, we first upsampling the deep features obtained at the coding end to obtain a feature map with the same size as the low-level features. Then connect the high-level semantic features with the low-level spatial features in the two stages of the backbone network generation on the channel dimension, and then use 3×3 to refine features and further enhance the effectiveness of features. Through multi-level feature fusion, the whole model can make full use of the feature map generated in each stage of the network, and improve the accuracy of segmentation.

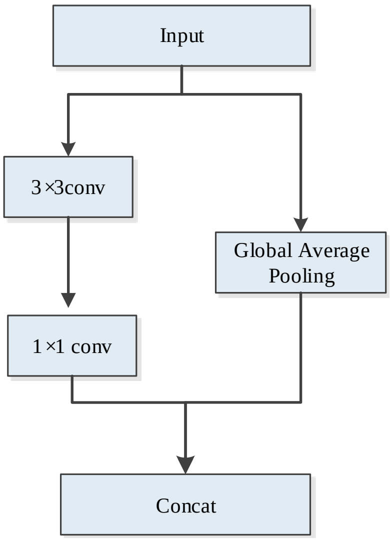


Figure 3. FE module structure diagram

2.4. **Loss function.** Designing a loss function that represents the learning goal more clearly based on the data and characteristics of a task often brings some improvements to the task [16]. Cross entropy loss function (CEloss) is a classical loss function in pixel-wise semantic segmentation. CEloss calculates the distance between the predicted probability distribution and the actual output probability distribution, and its formula is shown in Equation(1). It's often used in the field of semantic segmentation mainly because of its excellent convergence speed. However, it is not suitable for solving the problem of unbalanced of samples, and data imbalance will make the model learning bias and make the training process fall into the local minimum of the final loss function. But Diceloss is a common approach to solve this problem. Dice loss was proposed by Fausto Milletari in V-Net [17], which takes the evaluation metric of semantic segmentation as Loss. Dice coefficient is a measure function used to evaluate the similarity between two samples, and its value ranges from [0,1]. When the coefficient is 1, it means complete overlap, and its formula is shown in Equation(2). The larger the value of s in Equation(2), the more similar the predicted value and the true value, so the closer the Dice coefficient is to 1, the better. If used as a Loss the smaller the better, so Diceloss is shown in Equation(3). Although Diceloss can have good segmentation performance for scenarios with unbalanced data, the training loss tends to be unstable, especially when the target is very small. In addition, gradient saturation may occur in extreme cases. Therefore, in this paper, a scheme combining CEloss and Diceloss loss function is proposed, and the specific formula is shown in Equation(4).

$$\text{Ł}_{CE} = -\frac{1}{n}\sum_{x}[y\ln a + (1-y)\ln(1-a)] \tag{1}$$

$$s = \frac{2|X \cap Y|}{|X| + |Y|} \tag{2}$$

$$\text{Ł}_{DL} = 1 - s = 1 - \frac{2|X \cap Y|}{|X| + |Y|} \tag{3}$$

$$L = L_{CE}(a) + L_{DL} \tag{4}$$

When there is an extreme imbalance between the front and back scenes, CEloss cannot solve this extreme imbalance. However, Diceloss is not affected by the size of the foreground and can deal with this problem well. In some extreme cases, Diceloss cannot learn the correct direction of gradient descent. Combined with CEloss, it can give the network a learning direction. Therefore, this paper mainly combines Dicelos with CEloss to train the network.

3. **Experiment.**

3.1. **Dataset.**

The model will be trained and tested on Pascal VOC 2012 data set. Pascal VOC 2012 is a classic dataset in the field of semantic segmentation. Pascal VOC 2012 dataset plus background has 21 categories, including 1 background class and 20 target classes. For the segmentation task, the data set has 1464 training, 1449 verification and 1456 test images.

3.2. **Implementation details.**

The hardware and software environment configurations of all experiments are shown in Table 2.

TABLE 2. Experimental software and hardware environment configuration

Configuration	version
OS	64 Windows10
processor	Intel(R) Xeon(R)Silver 4114 CPU @2.20GHZ 2.19GHZ
GPU	NVIDIA Quadro P4000
Cuda	Cuda 9.2
Python	Python 3.6
Deep Learning Framework	Pytorch 1.2.0
development tools	PyCharm Community Edition 2020.3.3 x64

3.3. **Result analysis.**

3.3.1. *Evaluation indicators.* For evaluation, we uses Mean Intersection Over Union (MIOU), Time, pixel accuracy (PA) and model size as evaluation metrics to evaluate model performance. MIOU refers to the degree of coincidence between the segmentation result and its true value. It is currently the most commonly used in the field of image semantic segmentation, and is also the main evaluation metric used in this experiment. The equation is as follows:

$$MIOU = \frac{\sum_i n_{ii}}{n[t_i + \sum_j (n_{ji} - n_{ii})]} \tag{5}$$

Among them, n is the number of categories, n_{ij} is the number of pixels of the ith category that are predicted to be the jth category, t_i is the number of all the ith category pixels, $t_i = \sum_j n_{ij}$.

3.3.2. *Ablation experiment.* In this subsection, four different ablation experiments are designed to verify the effectiveness of each module. The performance of the proposed model is compared with that of the DeepLabv3+ benchmark model using MobileNetv2 as the feature extraction network. Table 3 shows the results of whether the module is effective.

Compared with the DeepLabv3+benchmark model with MobileNetv2 as the feature extraction network, the performance effect of the model proposed in this study is shown in Table 3. The second line of Table 3 shows the segmentation effect of combining CELoss and DiceLoss loss functions. Combining the two loss functions can effectively solve the problem of sample imbalance, enhance the robustness of the model, and increase the MIOU by 0.8%. It can be seen from the third line that the MIOU of the improved model is 0.84% higher than that of the benchmark model after using the FE module and multi-level features proposed in this study in parallel. From the fourth line, it can be seen that the MIOU of the PASCAL VOC 2012 test set can be increased from 71.66% to 72.61% by using the feature enhancement fusion structure and combining the loss function. The experimental results show that the proposed module can effectively improve the recognition accuracy of the network.

Table 3. Evaluating the effectiveness of each module

CELoss	DiceLoss	FE module	PA%	MIOU%
✓			81.43	71.66
✓	✓		81.3	72.46
✓		✓	81.41	72.5
✓	✓	✓	81.89	72.61

3.3.3. *Comparative Experiment.* In order to verify the effectiveness of the algorithm, the experiment compares the segmentation accuracy and speed of this algorithm with several mainstream methods in the field of semantic segmentation on Pascal VOC 2012 dataset. During the test, the experimental hardware environment is the same. MIOU is adopted as the main measurement standard.

Comparing the algorithm in this study with the current mainstream high-precision semantic segmentation algorithm, it can be seen from Table 5 that the network proposed in this study achieves 72.61% of MIOU on PASCAL VOC 2012 dataset. In terms of segmentation accuracy, compared with the classical segmentation network FCN-8s and UNet, the algorithm in this study improves the MIOU on the dataset by 13.02% and 17.61% respectively under the condition of lighter weight and higher efficiency. Compared with the current representative high-performance network PSPNet, the precision has decreased by 0.63%, but the reasoning speed has been greatly shortened by 75 ms. The experimental results show that the network proposed in this study can maintain a high segmentation accuracy under the premise of less parameters, less memory occupation and faster operation speed.

Table 4. Performance comparison of different algorithms

Model	MIOU(%)	Time(ms)	weights
FCN-8s	59.59	160	91.3MB
UNet [18]	55	146	94.9MB
PSPNet	**73.24**	120	178MB
Ours	72.61	**45**	**23.2MB**

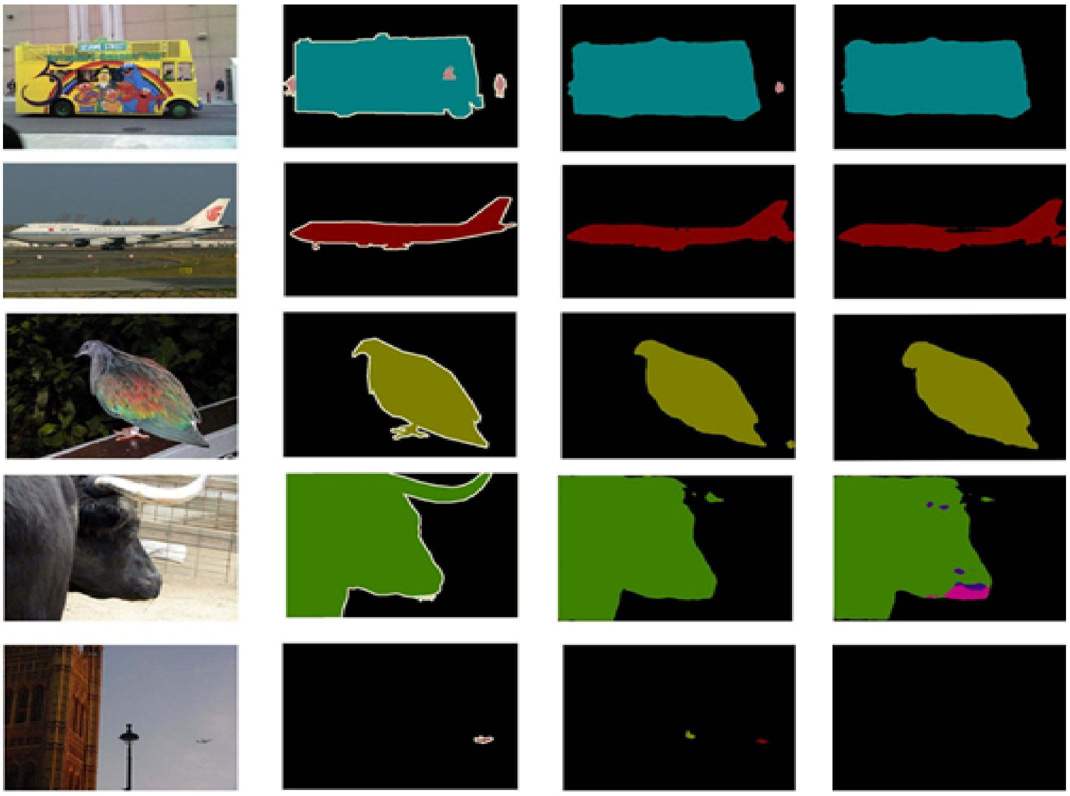
Compare this research algorithm with the current mainstream lightweight semantic segmentation algorithm, as shown in Table 6. It can be seen that ESPNetv2 is one of the fastest real-time networks. It processes a single image faster than the network in this study, but it only achieves 68.0% MIOU on the PASCAL VOC 2012 dataset, which is nearly 5 percentage points less than the network in this study. Compared with other lightweight segmentation networks, the network in this paper not only processes a single image faster, but also has higher segmentation accuracy. The above experimental results show that the network in this study achieves the best balance between accuracy and efficiency, and is superior to the current lightweight network structures.

Table 5. Comparison between the model in this article and the lightweight model

Model	MIOU (%)	Time(ms)
DeepSN [19]	64.94	54.6
ESPNetv2	68.0	**16.2**
CRF-RNN [20]	72.0	56
Ours	**72.61**	45

We illustrate visual results of our approach in Figure 4. From the figure, we can see that for the bus, plane, bird and cow targets in the first 4 rows, the segmentation results of this method have smoother edges and clearer outlines of details. The experiments demonstrate that the FE module and the multi-level feature fusion designed in this paper have indeed improved the segmentation capability of the model. Simultaneously, the last line of Figure 4 is shown, the size of the aircraft and the background is seriously unbalanced. The baseline model is unable to identify the aircraft target. Our proposed model can correctly classify this image. This example shows that the combined loss function can deal with samples with unbalanced data.

4. **Conclusion.** We have proposed a new lightweight network architecture. The network reduces the amount of parameters and computation by introducing MobileNetv2 as the feature extraction network. The feature enhancement module is designed to enhance the detailed information in the low-level features of multiple stages, making the target edge segmentation clearer. The combination of CEloss and Diceloss function improves the



(a) Original image (b) Groud Truth (c) ours (d) mobilenetv2_deeplabv3+

FIGURE 4. Visualize the results

segmentation ability of the network to unbalanced samples. Our work has been verified and compared on Pascal VOC 2012 dataset, and the results show that our network can obtain relatively good segmentation results with less computational cost. In addition, the algorithm we proposed still has some limitations. The structure of this paper is lightweight structure, which lacks the overall segmentation accuracy and is lower than the current best segmentation model. Therefore, in the next research work, how to improve the accuracy of lightweight networks will become our focus.

Acknowledgment. This research is funded by Science and technology project of the Jilin Provincial Education Department, Grant no. JJKH20220118KJ.

REFERENCES

[1] F. Zhang, T.-Y. Wu, and G. Zheng, "Video salient region detection model based on wavelet transform and feature comparison," *EURASIP Journal on Image and Video Processing*, vol. 58, 2019.

[2] K.-K. Tseng, J. Lin, C.-M. Chen, and M.-M. Hassan, "A fast instance segmentation with one-stage multi-task deep neural network for autonomous driving," *Computers & Electrical Engineering*, vol. 93, 107194, 2021.

[3] X. Tian, L. Wang, and Q. Ding, "Research review of image semantic segmentation technology based on deep learning," *Journal of Software*, vol. 30, no. 2, pp. 440–468, 2019.

[4] J. Long, E. Shelhamer, and T. Darrell, "Fully convolutional networks for semantic segmentation," in *2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. IEEE, 2015, pp. 3431–3440.

[5] H. Zhao, J. Shi, X. Qi, X. Wang, and J. Jia, "Pyramid scene parsing network," in *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. IEEE, 2017, pp. 6230–6239.

[6] G. Lin, A. Milan, C. Shen, and I. Reid, "RefineNet: Multi-path refinement networks for high-resolution semantic segmentation," in *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. IEEE, 2017, pp. 5168–5177.

[7] 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 Transactions on Pattern Analysis and Machine Intelligence*, vol. 40, no. 4, pp. 834–848, 2018.

[8] L.-C. Chen, Y. Zhu, G. Papandreou, F. Schroff, and H. Adam, "Encoder-decoder with atrous separable convolution for semantic image segmentation," in *Proceedings of the European Conference on Computer Vision*. ECCV, 2018, pp. 833–851.

[9] Y. Zhang, Z. Qiu, T. Yao, C.-W. Ngo, D. Liu, and T. Mei, "Transferring and regularizing prediction for semantic segmentation," in *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. IEEE, 2020, pp. 9618–9627.

[10] A. Paszke, A. Chaurasia, S. Kim, and E. Culurciello, "ENet: A deep neural network architecture for real-time semantic segmentation," 2016. [Online]. Available: https://doi.org/10.48550/arXiv.1606.02147

[11] C. Yu, J. Wang, C. Peng, C. Gao, G. Yu, and N. Sang, "BiSeNet; Bilateral segmentation network for real-time semantic segmentation," in *Proceedings of the European Conference on Computer Vision*. ECCV, 2018, pp. 334–349.

[12] S. Mehta, M. Rastegari, L. Shapiro, and H. Hajishirzi, "ESPNetv2: A light-weight, power efficient, and general purpose convolutional neural network," in *2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. IEEE, 2019, pp. 9182–9192.

[13] T. Wu, S. Tang, R Zhang, J. Cao, and Y. Zhang, "CGNet: A light-weightcontext guided network Ior semantic segmentation," in *IEEE Transactions on Image Processing*, vol. 30, pp. 1169–1179, 2021.

[14] F. Zhang, T.-Y. Wu, J. Pan, G. Ding, and Z. Li, "Human Motion Recognition Based on SVM in VR Art Media Interaction Environment," *Human-centric Computing and Information Sciences*, vol. 9, pp. 1–15, 2019.

[15] W. Liu, H.-R. Wang, and B.-J. Zhou, "DeepLabv3plus-IRCNet: An image semantic segmentation method for small target feature extraction," *Journal of Image and Graphics*, vol. 26, no. 2, pp. 391–401, 2021.

[16] K.-K. Tseng, R. Zhang, C.-M. Chen, and M.-M. Hassan, "DNetUnet: A semi-supervised CNN of medical image segmentation for super-computing AI service," *The Journal of Supercomputing*, vol. 77, no. 4, pp. 3594–3615, 2021.

[17] F. Milletari, N. Navab, and S.-A. Ahmadi, "V-Net: Fully convolutional neural networks for volumetric medical image segmentation," in *2016 Fourth International Conference on 3D Vision (3DV)*. 3DV, 2016, pp. 565-571.

[18] O. Ronneberger, P. Fischer, and T. Brox, "U-Net: Convolutional networks for biomedical image segmentation," 2015. [Online]. Available: https://arxiv.org/abs/1505.04597v1

[19] D. Lai, Y. Deng, and L. Chen, "DeepSqueezeNet-CRF: A Lightweight Deep Model for Semantic Image Segmentation," in *2019 International Joint Conference on Neural Networks (IJCNN)*. IEEE, 2019, pp. 1–8.

[20] S. Zheng, S. Jayasumana, B. RomeraParedes, V. Vineet, Z. Su, D. Du, C. Huang, and P.-H. Torr, "Conditional random fields as recurrent neural networks," in *2015 IEEE International Conference on Computer Vision (ICCV)*. IEEE, 2015, pp. 1529–1537.