ISBN: 978-981-18-5852-9
WCSE 2022 Spring Event: 2022 9th International Conference on Industrial Engineering and Applications
doi: 10.18178/wcse.2022.04.011

DeepLabv3+ Semantic Segmentation Network Optimized by Double Weighted Feature Fusion

Hong Zhang[1,2], Yue Gao[1,2], Jianchen Miao[1]
[1] Xi'an University of Posts and Telecommunications, Xi'an, China
[2]Automatic Sorting Technology Research Center, State Post Bureau of the People's Republic of China

Abstract. Intelligent identification to prohibited items in lug-gage is of great significance to ensure the safety of passengers. In this paper, we propose CWF-DeepLabv3+ semantic segmentation network to improve the segmentation accuracy to small size objects in X-Ray security images, and reduce the object pixel loss caused by complex background. The CWF-DeepLabv3+ network is based on the DeepLabv3+ network, which improves both the Xception backbone and the Atrous Spatial Pyramid Pooling (ASPP) module. Firstly, we combine the Cross Stage Partial Network (CSPNet) structure to propose a new CSP-Xception backbone network. The CSPNet network is capable of improving the learning ability of the network by using gradient split to build rich gradient combinations. Secondly, we design the Double Weighted Feature Fusion (DWFF) in the ASPP module. Weighted feature fusion by learning the importance of input features, and in this way, the segmentation accuracy of images with small size objects and complex background will be improved. The experimental results of training and testing on X-Ray security images show that the mPA value and mIoU value of CWF-DeepLabv3+ network reach to 90.85% and 81.64% respectively. Furthermore, compared with DeepLabv3+, Unet, and Bisenetv2 semantic segmentation network, the mean pixel segmentation accuracy of our proposed network is improved by at least 3.13%.

Keywords: X-Ray Security Images, Semantic Segmentation, DeepLabv3+, CSPNet, Double Weighted Feature Fusion

1. Introduction

Maintaining urban traffic order and ensuring passenger travel safety are critical for social security. X-Ray security equipment is used to detect and identify the goods in the passengers' luggage, which can effectively avoid the dangerous caused by prohibited items such as knives and guns [1]. At present, the detection of the images obtained by X-ray security machine is still carried out by security inspectors. However, with the higher work intensity of security inspectors, false detection and missed detection are easy to happened. Therefore, it is very important to research automatic identification technology that can assist the security inspector to carry out the object category and shape of the prohibited items. Among many methods, semantic segmentation is a key technology to obtain object information in images.

Traditional segmentation algorithms for X-Ray security images include threshold setting and edge extraction. For example, Wang et al [2] used traditional EM clustering to segment prohibited items, Cao et al [3] used threshold selection and region growth, Gu [4] used the wavelet decomposition and reconstruction method to extract the edge of the image and then realize the segmentation of the security image. These methods are simple and fast, but the recognition results are not fine enough, and they need to be manually set and analyzed. At the same time, they are not universal and applied only to certain specific small areas, so it is difficult to fully meet the needs of security inspection work. With the development of deep learning technology, image semantic segmentation technology based on deep learning has been a breakthrough and applied to automatic driving, medical diagnosis, remote sensing monitoring and other fields.

Semantic segmentation algorithm integrates object recognition technology on the basis of image segmentation, which can not only extract the object, but also recognize the type of the objects. It greatly

Tel.: 18829210411
E-mail address: zhangh@unm.edu, 1983242966@qq.com, 958857480@qq.com

improves the efficiency and accuracy of object recognition, and becomes one of the important research fields. Semantic segmentation can realize the inference of each pixel value in the collected images, then put a semantic label on each fine-grained pixel value [5]. The Fully Convolutional Network (FCN) proposed by Long et al [6] replaced the fully connected layer with a fully convolutional layer, which opened up a precedent in the field of semantic segmentation. Subsequently, a series of semantic segmentation networks are designed to improve the feature extraction capability of the network. Such as the Unet [7] network, which employed an encoder-decoder structure and performs better on medical images [8]. The Deeplabv1 [9], which applied conditional random field CRF to improve the ability of the model to capture information. Deeplabv2 [10], consists of multi-scale and multi-atrous rates convolution. And the Deeplabv3 [11] used parallel atrous convolution. All of them are meaningful and innovative methods. In 2018, the DeepLabv3+ [12] network proposed by Chen et al pushed semantic segmentation to a new peak. It adopted Xception [13] as the backbone and used a simple and effective decoder to recover the details of the object boundary. What's more, the DeepLabv3+ network can clearly and intuitively display the object and have strong versatility. In this paper, we use DeepLabv3+ network as the base network to design a new semantic segmentation network to further improve the detection accuracy to dangerous objects in X-Ray security images. In particular, the X-Ray security images are with complex background, multi-class and multi-scales, and more small size objects. For the detection accuracy and less calculation, we propose the CWF-DeepLabv3+ (SCPnet With Weighted Fusion DeepLabv3+) semantic segmentation network. The main contributions of the paper are as follows:

1) Improving the Cross Stage Partial Network (CSPNet) [14] to establish the CSP-Xception backbone network. Utilizing the unique network architecture of CSPNet and the advantages of the rich gradient combination brought by gradient split, its segmentation accuracy can be effectively improved without increasing the amount of calculation.

2) Optimizing the ASPP module. We designed a double weighted feature fusion DWFF structure, which enables the network to focus on important features with different weight factor, solves the lower segmentation accuracy problem of small objects and the object pixel loss caused by complex background.

2. DeepLabv3+ network structure

DeepLabv3+ semantic segmentation network is an improvement on the content of the previous DeepLab series. The structure of DeepLabv3+ has two parts: encoder and decoder, they perform convolution and down-sampling to extract feature information. The network structure is shown in Fig 1. As we can see, the encoder uses the Xception backbone network to extract features, and uses the ASPP module to further get multi-scale context information. Here, the ASPP module includes 1×1 convolution, 3×3 atrous convolution with rates of 6, 12, 18, and global pooling, this module can well capture the multi-scale information brought by different receptive fields. In the decoder, two times up-sampling, two times convolutions and one feature fusion operation are performed to achieve pixel-level prediction.

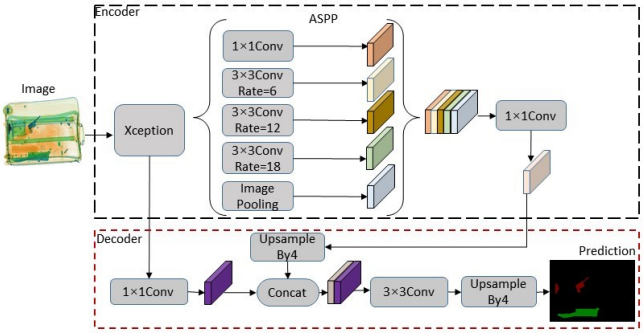

Fig. 1: DeepLabv3 + network structure

3. X-ray security image segmentation network

3.1. CWF-DeepLabV3+ network structure

In order to effectively improve the segmentation performance of the DeepLabv3+ network, we propose the CWF-DeepLabv3+ network, displayed in Fig 2. The CWF-DeepLabv3+ network integrated the CSPNet into the Xception to build the CSP-Xception backbone network. We also designed the DWFF-ASPP module to improve the accuracy. The CWF-DeepLabv3+ network is also divided into two parts: encoder and decoder to fuse multi-scale information.

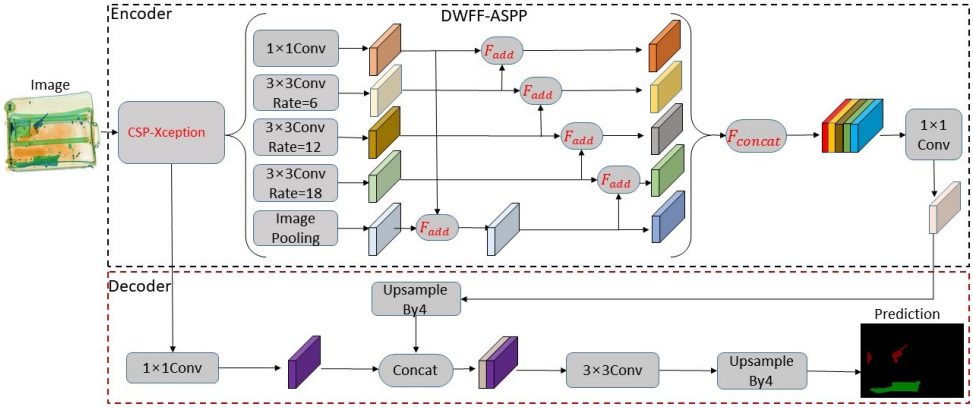

Fig. 2: CWF-DeepLabv3+ network structure

1) In the encoder, the information of two feature layers is extracted, which contains high-level semantic information and low-level semantic information. Then the feature map with higher semantic information enters the DWFF-ASPP module to extract multi-scale context information. The DWFF-ASPP module is composed of ASPP module and DWFF module. First of all, instead of directly performing simple Concat fusion did by original network, the cross-layer and adjacent-layer is combined in weighted fusion style on the five feature layers obtained by the ASPP module. That makes the network not only focus on features of different resolutions, but on important features. Then we perform the second weighted fusion on the five new feature maps obtained, and let feature maps are connected in the channel dimension. Finally, 1×1 convolution is used to reduce the number of feature channels.

2) In the decoder, the features with lower semantic information obtained from the backbone network are compressed by 1×1 convolution. At the same time, the feature map obtained by the encoder is up-sampled by a factor 4 to match the resolution of the low-level features. After that, we connect and fuse the two feature maps and use two times 3×3 convolutions and up-sampled by a factor 4 to refine the features obtained in the previous step. As a result, we can clearly obtain the categories and postures of prohibited items segmented.

3.2. CSP-Xception backbone network

The function of backbone network is to extract the features of multi-scale objects by different convolutions and residuals. In this paper, we based on the original Xception backbone network, incorporated the CSPNet network to propose the CSP-Xception backbone network. The CSPNet network is widely used in object detection, here we set it to CSP_block. Our CSP-Xception backbone network, which fully combines the unique depthwise separable convolution of the Xception network and the rich gradient combinations and residuals mechanism of the CSPNet network, so it can be better used to the task of segmenting X-ray security images. The CSP- Xception network backbone network is shown in Fig 3.

CSP-Xception frame is composed of four CSP_block modules and two Xception_block modules, as shown in Fig 3(a). Specifically, for Xception_block module, displayed in Fig 3(b), we performed three times depthwise separable convolution on the input feature and then connected with the large residual edge to obtain the output feature. Depthwise separable convolution divided the standard convolution into depthwise convolution and pointwise convolution to reduce the parameter of the network and ensure the accuracy segmentation is similar or even higher. In this structure, the 1x1 convolution is pointwise convolution. The structure of CSP_block is shown in Fig 3(c). It is worth noting that the input feature map is split into two paths, one path successively performs 1x1 convolution, N times residual and 1x1 convolution operations. The other path also performs 1x1 convolution. Finally, we concatenate and merge the two branches. This unique network architecture uses split and merge to construct rich gradient combinations, which cross-mixes

gradients at different positions, so segmentation accuracy can be effectively improved without increasing the amount of calculation. Here, a picture with the size of 512×512×3 will pass through the CSP-Xception backbone network to obtain two feature layers, which contains high-level semantic and low-level semantic information with sizes of 32×32×1024 and 128×128×256, respectively. And they will also keep performing the ASPP and Decoder operations for feature extraction. Our new backbone network can effectively enhance the model's learning ability and feature extraction ability. At be the same time, the high-level edge features and texture information obtained by this structure will beneficial to the accurate identification of prohibited items.

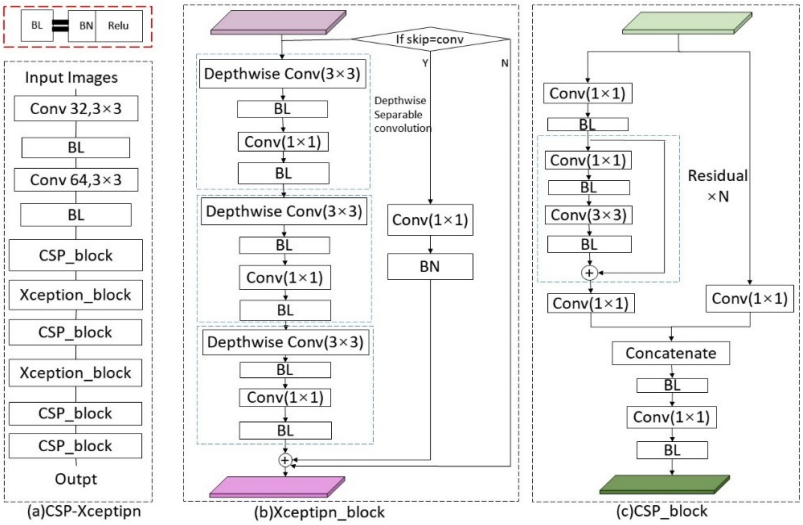

Fig. 3: CSP-Xception backbone network

3.3. Double Weighted Feature Fusion DWFF module

Through the ASPP model, we can obtain features with different information. However, the DeepLabv3+ network performs a simple Concat operation on these features to fuse in the channel. This fusion does not consider the importance of different features and the contribution to the output feature map. And important information are missing. Therefore, in order to obtain an effective feature layer with high-level semantic information, we propose a Double Weighted Feature Fusion DWFF. As shown in Fig 4, which includes F_{sum} and F_{concat} fusion methods.

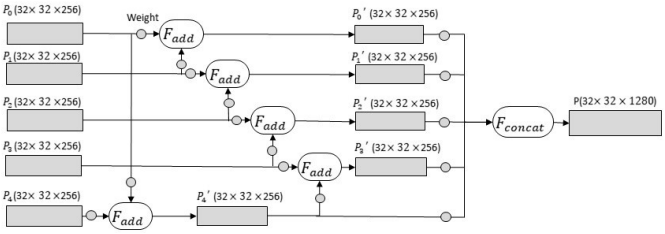

Fig. 4: DWFF module

In detail, five feature layers perform five groups F_{sum} weighted fusion to form the first weighted fusion block, and the five feature layers perform one group F_{concat} F_{concat} weighted fusion to perform the second weighted fusion block. The weight of the feature layer is formula (1):

$$W = \frac{w_i}{\sum_j w_j + \varepsilon} \tag{1}$$

After obtaining the weight of each feature layer, we adopt F_{add} and F_{concat} fusion methods. As shown in formula (2) and (3):

$$F_{add} = \sigma\left(\sum_i W_i P_i\right) \tag{2}$$

$$F_{concat} = \sigma\left(\phi_{i=0}^{4} W_i P_i\right) \tag{3}$$

Where w is the learned importance parameter, W_i is the weight of the i layer feature map, and ε is a small value to ensure numerical stability, σ is the swish activation function, and ϕ is the Concat fusion function.

As shown in Fig 4, the feature maps obtained by the ASPP module are P_0, P_1, P_2, P_3, P_4. The size of these five feature maps with different receptive fields is 32×32×256. Then we perform DWFF operation on these five features. Taking the first group of the first weighted fusion block P_0 and P_4 as an example. According to the importance parameters w_0 and w_4 of P_0 and P_4 learned by the network, we can get the weights W_0 and W_4. Further, we combine the σ swish activation function and adopt the F_{add} fusion method to obtain the fused feature map $P_4^{'}$. The size of $P_4^{'}$ is 32×32×256. Similarly, P_3 and $P_4^{'}$, P_2 and P_3, P_1 and P_2, P_0 and P_1 were fused by F_{add} to obtain $P_3^{'}$, $P_2^{'}$, $P_1^{'}$, $P_0^{'}$. The second weighted block is based on the five feature maps: $P_4^{'}$, $P_3^{'}$, $P_2^{'}$, $P_1^{'}$, $P_0^{'}$. In the same way, according to the learned importance of them, we redistribute the new weights are W_4, W_3, W_2, W_1, W_0. We also combine the $\sigma\sigma$swish activation function and adopt the F_{concat} fusion method, that is, the Concat operation is performed by the ϕ fusion function. The number of channels of the feature map P is 1280. Because the DWFF module gives different weights to each layer for fusion, so the processed feature map can better extract object information.

4. Experiments and result analysis

4.1. Experimental environment and related parameters

The network adopts cuda 9.0 and python3.6 operating environment. Tensorflow deep learning framework and Windows10 operating system. The hardware platform is Intel (R) Xeon (R) CPU E5-2640 @ 2.60 GHz. GPU: 4 GTX1080 Ti graphics GPU calculators, optimization strategy for Adam, complete training and testing. The network training parameters shown in Table 1.

Table 1: Network training parameters

Classes	Value
Learning rate	0.0001
Training Epoch	100
Batch-size	4

4.2. X-ray security image dataset

In this paper, we constructed the X-Ray security image dataset, and use the pictures collected by the X-Ray security inspection machine. The image has a complex background and different image quality, it includes five categories of prohibited items: Gun, Knife, Scissors, Pliers, and Wrench. More importantly, we modeled on the VOC2012 dataset format and used Labelme software to manually label these images as semantic segmentation ground truth. The semantic annotations of X-Ray security images are shown in Fig 5. To prevent data imbalance, we use data augmentation for categories with a small number of images, as shown in Fig 6. First, we transform the image randomly, such as crop, revolve and sharpening enhancement to increase the diversity of the data in terms of shape and size. At the same time, in order to further expand the sample size, CycleGAN [15] is used to transform the image from the source domain X to the object domain Y. That is, we collect regular images on the network, and use the cyclic generation GAN network to extract the underlying features of prohibited items, then we can transform its style into the X-Ray security image style. In the end, the total number of dataset is 4030, which is divided into train set and test set according to the ratio of 7:3, and we compressed images to 512 × 512 pixels as the inputs of network.

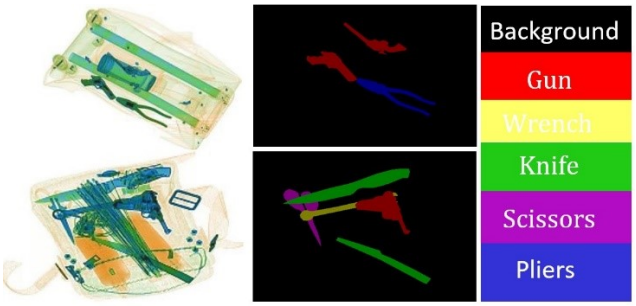

Fig. 5: Semantic annotation of X-Ray security images

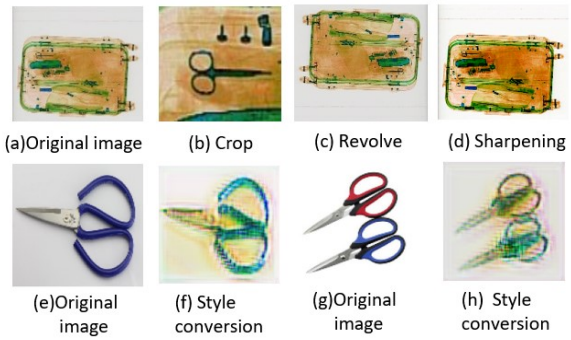

Fig. 6: Data augmentation

4.3. Results and analysis

On the self-made X-Ray security image dataset, in order to contrast with our network, we experiment with several existing methods. They are classic Unet, state-of-the-art Bisenetv2 [17], original DeepLabv3+ and the improved DeepLabv3+ network based on Mobilenetv2 [16]. And in order to fairly evaluate the effectiveness of the proposed CWF-DeepLabv3+ network, all experiments in this paper are based on the same benchmark and setting, that is, we did not use the pre-trained weight model obtained on large data sets such as VOC2012 and Cityscapes.

The effect of network segmentation depends on the similarity between the segmentation result obtained by the network and the semantic ground truth of its corresponding image. Here, we select five representative pictures from the dataset, as shown in (a) of Fig 7. They include multi-objects, multi-shapes, occlusions, complex background, small object and other difficult segmentation samples.

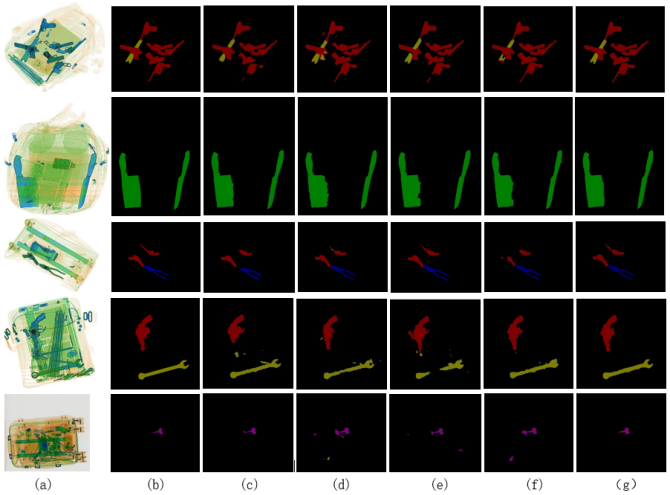

Fig.7: Results of different Semantic segmentation networks. (a) Selected pictures. (b) Ground truth. (c) Segmentation results on Bisenetv2. (d) Segmentation results on Unet. (e) Segmentation results on DeepLabv3+. (f) Segmentation results on improved DeepLabv3+ based on Mobilenetv2. (g) Segmentation results on CWF-DeepLabv3+.

Semantic segmentation results of selected pictures are shown in Fig 7. For the first picture with more objects, the CWF-DeepLabv3+ network can clearly and completely segment the contour of each object, the Bisenetv2 and the DeepLabv3+ have obvious barrel pixel loss, the Unet and the improved DeepLabv3+ based on Mobilenetv2 [16] have problems with wrench, guns, and background misclassifi-cations. For the second picture, the knife is multi-forms and close to the background color, other networks not only segment rough edges, but appear superabundance segmentation and missing segmentation. The CWF-DeepLabv3+ can process the edge of the knife more finely and smoothly. For the third picture, we compared with Bisenetv2, Unet and the improved DeepLabv3+ based on Mobilenetv2, although the original DeepLabv3+ model can successfully segment the gun that covered by cup, there are still a few pixel misclassification. The result of CWF- DeepLabv3+ network segmentation retains more detail and is closer to the ground truth. The fourth picture shows a gun and wrench with complicated background, our algorithm can comprehensively predict the geometric shapes of the gun and wrench, but the boundary still needs to be smooth. In the fifth picture, scissors as a small size object, in contrast, other networks predict clothes hangers with similar colors as a part of scissors, but our method can more accurately segment the scissors, and better handle the details

of small objects. In summary, for the segmentation of prohibited items in X-Ray security images, the segmentation effect of the CWF-DeepLabv3+ network is closer to the corresponding ground truth.

In addition, in order to further evaluate the advantages of our network, we use the specific evaluation metrics of semantic segmentation network to measure these networks. These four metrics are: Pixel Accuracy (PA), Mean Pixel Accuracy (mPA), Intersection over Union (IoU), Mean intersection over union (mIoU). They are very authoritative to verify the performance of the network. The calculation formulas are as follows:

$$\mathrm{PA} = \frac{\sum_{i=0}^{k} p_{ii}}{\sum_{i=0}^{k}\sum_{i=0}^{k} p_{ij}} \tag{4}$$

$$\mathrm{mPA} = \frac{1}{k+1}\frac{\sum_{i=0}^{k} p_{ii}}{\sum_{i=0}^{k}\sum_{i=0}^{k} p_{ij}} \tag{5}$$

$$\mathrm{IoU} = \sum_{i=0}^{k} \frac{p_{ii}}{\sum_{j=0}^{k} p_{ij} + \sum_{j=0}^{k} p_{ji} - p_{ii}} \tag{6}$$

$$\mathrm{mIoU} = \frac{1}{k+1}\sum_{i=0}^{k} \frac{p_{ii}}{\sum_{j=0}^{k} p_{ij} + \sum_{j=0}^{k} p_{ji} - p_{ii}} \tag{7}$$

Where, k represents the category of semantic segmentation, p_{ii} represents the number of pixels correctly predicted, p_{ij} represents the number of pixels belonging to category i recognized as category j. p_{ji} represents the number of pixels belonging to category j recognized as category i.

Based on the above evaluation metrics, we train and test the CWF-DeepLabv3+ network and the DeepLabv3+ network. The comparison of the IoU and PA values of the two networks for background and five categories of object is shown in Table 2.

Table 2: Comparison of IoU and PA values between two networks

Category	CWF-DeepLabv3+		DeepLabv3+	
	IoU(%)	PA(%)	IoU(%)	PA(%)
Background	99.62	99.81	99.4	99.7
Gun	85.96	91.06	83.29	89.98
Knife	88.59	93.21	83.96	90.0
Wrench	71.69	88.78	52.56	85.68
Pliers	71.42	86.52	62.42	79.6
Scissors	72.58	85.7	62.68	79.14

It can be seen from the Table 2, the CWF-DeepLabv3+ network has the best prediction results for knife, and the PA and IoU reach to 93.21% and 88.59%, which are improved by 3.21% and 4.63% than the DeepLabv3+ network. At the same time, for images containing wrenches with more complex background, our algorithm's IoU and PA values are increased by 19.13% and 3.1%. In addition, our network boosts the IoU and PA values of the scissors, which are mostly small objects, by 9.9% and 6.56%, respectively. Therefore, CWF-DeepLav3+, which is improved by integrating CSPNet and DWFF modules, significantly improves the segmentation metric value of each category of prohibited items.

Simultaneously, according to the evaluation metrics of semantic segmentation, we also obtain the mIoU, mPA and FPS values of these networks on the X-Ray security image test set. The mIoU, mPA and FPS values of each network are shown in the Table 3.

Table 3 : Comparison of metric values of networks

Network	mIoU(%)	mPA(%)	FPS
Bisenetv2	78.32	85.47	13.62
Unet	77.91	87.72	11.50
DeepLabv3+	74.06	87.35	11.93
Mobilenetv2-DeepLabv3+	72.07	86.8	12.08
CWF-DeepLabv3+	81.64	90.85	12.12

From the Table 3, we can clearly see that the mIoU value of the CWF-DeepLabv3+ network can reach to 81.64%, and the mPA value can reach to 90.85%. Compared with the original DeepLabv3+ network and improved DeepLabv3+ based on Mobilenetv2, the mIoU value has increased by 7.04% and 9.57%, and the mPA value has increased by 3.5% and 4.05%, respectively. Furthermore, compared with the Unet network and the latest Bisenetv2 network, the mIoU value has increased by 3.73% and 3.32%, and the mPA value has increased by 3.13% and 5.38%. At the same time, our CWF-DeepLabv3+ network has good FPS value. Obviously, our network as a whole has also been improved.

5. Conclusions

In this paper, we proposed CWF-DeepLabv3+ semantic segmentation network. This network is not only integrated the CSPNet structure to improve the learning ability of model, but also designed a double weighted feature fusion DWFF structure to calulate the weight of input features according to their importance factors. The experimental results in the X-ray security image dataset show that our proposed network can segment prohibited items more accurately and completely than Unet, Bisenetv2, DeepLabv3+, improved DeepLabv3+ based on Mobilenetv2, and the segmentation effect is closer to the ground truth. The performance of PA, mPA, IoU, mIoU metric is significantly improved. In particular, compared with the original DeepLabv3+ network, the mIoU value and mPA value were increased by 7.04% and 3.5%, respectively. It is shown that the CWF-DeepLabv3+ network can effectively enhance the segmentation performance and has application feasibility in X-Ray security image segmentation.

Since the categories of prohibited items in the dataset used in this article cannot represent all categories, we will further collect X-Ray security images containing more categories to expand the scope of application of the model. At the same time, it is necessary to propose an effective solution to the situation that similar objects cannot be segmented due to serious overlap.

6. Acknowledgements

This research is supported in part by Shaanxi Provincial Natural Science Foundation of China (Grant No.2021 SF-478).

7. References

[1] Yuxiao Wang, Liang Zhang. Multi-scale feature fusion security image dangerous goods detection[J]. *Laser and optoelectronics progress,* 2021,58 (08): 152-159.

[2] Huaiying Wang, Li Chen. Segmentation method of contraband in X-Ray backscatter image based on EM clustering[J]. *CT theory and application research*, 2013,22 (03): 463-468.

[3] Jinbo Cao. X-Ray machine security image recognition of prohibited items[D]. *Northwest Normal University*, 2017.

[4] Lexu Gu. Design of tool identification system based on X-Ray security image[D]. *Northeast Electric Power University*, 2018.

[5] Longkang Peng, Licong Liu, Xuehong Chen, Jin Chen, Xin Cao and Yuean Qiu. Research on generalization performance of cloud detection network for remote sensing images: a case study of DeepLabv3 + [J].

[6] LONG J, SHELHAMER E, DARRELL T. Fully convolutional networks for semantic segmentation[C]// *Proceedings of the 2015 IEEE Confe-rence on Computer Vision and Pattern Recognition. Piscataway: IEEE,* 2015: 3431-3400.

[7] Ronnebrger O, Fischer P, BROX T.U-Net: Convolutional Networks for Biomedical Image Segmentation[J]. 2015, 9(20):1-8

[8] Xiliang Zhu, et al."Coronary angiography image segmentation based on PSPNet." *Computer Methods and Programs in Biomedicine*. Prepublish (2020): doi:10.1016/J.CMPB.2020.105897.

[9] Chen L C, Papandreou G, Kokkinos I，et al.Semantic image segmentat-ion with deep convolutional nets and fully connected CRF[J]. *Computer Science*, 2014(4) :357- 361.

[10] Chen L C, Papandreou, G, Kokkinos, I., Murphy, K., Yuille, A.L.: Deeplab: Semantic image segmentation with deep convolutional nets, atrous convolution, and fully connected crfs. arXiv (2016).

[11] Chen L C, Papandreou, G, Schroff F, et al. Rethinking Atrous Convolution for Semantic Image Segmentation, arXiv: 1706.05587v117 Jun 2017:3-4.

[12] Chen L C, ZHU Y, Papandreou, G, et al. Encoder-Decoder with Atrous Separable Convolution for Semantic Image Segmentation [C]// Proce-edings of the 2018 European Conference on Computer Vision. Switzerland：Springer, 2018: 801-818.

[13] Chollet, F.: Xception: Deep learning with depthwise separable convolutions. *In: CVPR.* (2017).

[14] Wang C Y, Liao H Y M, Wu Y H, et al. CSPNet: A new backbone that can enhance learning capability of CNN[C] //2020 *IEEE/CVF Conference on Computer Vision and Pattern Recognition workshops*, June 14-19.

[15] ZHU J Y, ISOLA P, ZHOU T, et al. Image-to-image translation with conditional adversarial networks[C]// *Proceedings of IEEE conference on computer vision and pattern recognition . Honolulu, HI, USA:IEEE,* 2017:5967－5976.

[16] Jianfang Cao, Xiaodong Tian, Yiming Jia, Minmin Yan. Application of Improved DeepLabv3+ Model in Mural Segmentation [J]. *Computer Application*, 2021,41 (05): 1471-1476.

[17] Changqian Yu, Changxin Gao, Jingbo Wang, Gang Yu, Chunhua Shen, Nong Sang. BiSeNet V2: Bilateral Network with Guided Aggregation for Real-Time Semantic Segmentation[J]. *International Journal of Computer Vision,*2021,129(11):3051-3068.