A prohibited items identification approach based on semantic segmentation

YAO Shao-qing (姚少卿)[1], **SU Zhi-gang** (苏志刚)[1,2]*, **YANG Jin-feng** (杨金锋)[3], **and ZHANG Hai-gang** (张海刚)[3]

1. Tianjin Key Laboratory for Advanced Signal Processing, Civil Aviation University of China, Tianjin 300300, China

2. Sino-European Institute of Aviation Engineering, Civil Aviation University of China, Tianjin 300300, China

3. Institute of Applied Artificial Intelligence of the Guangdong-Hong Kong-Macao Greater Bay Area, Shenzhen Polytechnic, Shenzhen 518055, China

(Received 31 January 2020; Revised 12 June 2020)

Deep learning (DL) based semantic segmentation methods can extract object information including category, location and shape. In this paper, the identification of prohibited items is regarded as a task of semantic segmentation, and proposes a universal model with automatic identification of prohibited items. This model has two improvements based on the general semantic segmentation network. Firstly, the N-type encoding structure is applied to enlarge the receptive field of the network aiming at reducing the misclassification. Secondly, consider the lack of surface texture in X-ray security images. Inspired by feature reuse in Densenet, shallow semantic information is reused to improve the segmentation accuracy. With the use of this model, when using input images of size 512×512, we could achieve 0.783 mean intersection over union (mIoU) for a seven-class object recognition problem.

Document code: A **Article ID:** 1673-1905(2021)04-0247-5

DOI https://doi.org/10.1007/s11801-021-0017-6

Security inspection is an important means to protect public space from various security threats. Modern cities face all kinds of crowded occasions that have put forward higher requirements for security inspection. People want to identify prohibited items quickly, automatically and accurately[1]. But at present, the identification of prohibited items mainly depends on the security personnel's observation of X-ray security images. Security personnel have to work long hours and are under great pressure. Their judgment might be influenced that result in frequent missed inspections, which seriously affects the security of public space. In recent years, the deep learning based convolutional neural network[2] has performed well in image processing and visual understanding. Object recognition capability for conventional images has reached a practical level. Therefore, it is meaningful to use a convolutional neural network to realize the automatic identification of prohibited items to improve the present situation of security inspection.

Object recognition is the core of the automatic identification of the prohibited items model. There are typically three types of object recognition methods based on deep learning. The first one works on the image level which produces a score for each class indicating its presence or absence. The second one instead works on the object level and produces a bounding box as well as a class label for each object individually[3,4]. The third one works on the pixel level, grouping pixels according to the different semantic meanings expressed in the image, such as Fully Convolutional Networks (FCN), U-Net, Pyramid Scene Parsing Network (PSPNet)[3,5,6], etc. At present, the research on prohibited items identification mainly focuses on the second method (object level). However, little research has been done on the subject of prohibited items identification based on semantic segmentation method (pixel level).

In this paper, we design a general semantic segmentation model for prohibited items identification in X-ray security images, which is illustrated in Fig.1. On the one hand, compared with the identification method of object level, our model provides information on the category, location and shape of prohibited items, which constitute the core of what security personnel concerns. On the other hand, our experiments show that the basic semantic segmentation network does not perform well in our X-ray security images datasets so that it cannot be applied directly. Therefore, two modifications are carried out in our model to improve the identification results. (1) This model adopts the N-type encoding structure which is composed of down-sampling in two stages and one-stage up-sampling (Because the model contains two up-sampling, here refers specifically to the up-sampling in the N-type encoding. shown as Fig.2). Down-sampling in two stages will increase the pooling times. However, while pooling

* E-mail: ssrsu@vip.sina.com

enlarges the receptive field of the network, the image details are lost along with the decrease of resolution which is not conducive to the recovery of the feature map. So we insert the one-stage up-sampling in the middle of the down-sampling in two stages. In this way, the model balances the need to enlarge the receptive field and to recover the feature map while reducing misclassification. (2) Because of the lack of surface texture in X-ray security images, the shallow semantic information is more important in the improvement of the segmentation accuracy. Inspired by feature reuse in DenseNet[7], we fuse the shallow semantic information in both up-sampling processes. The experiments below have proved that the two improvements we proposed are feasible and effective.

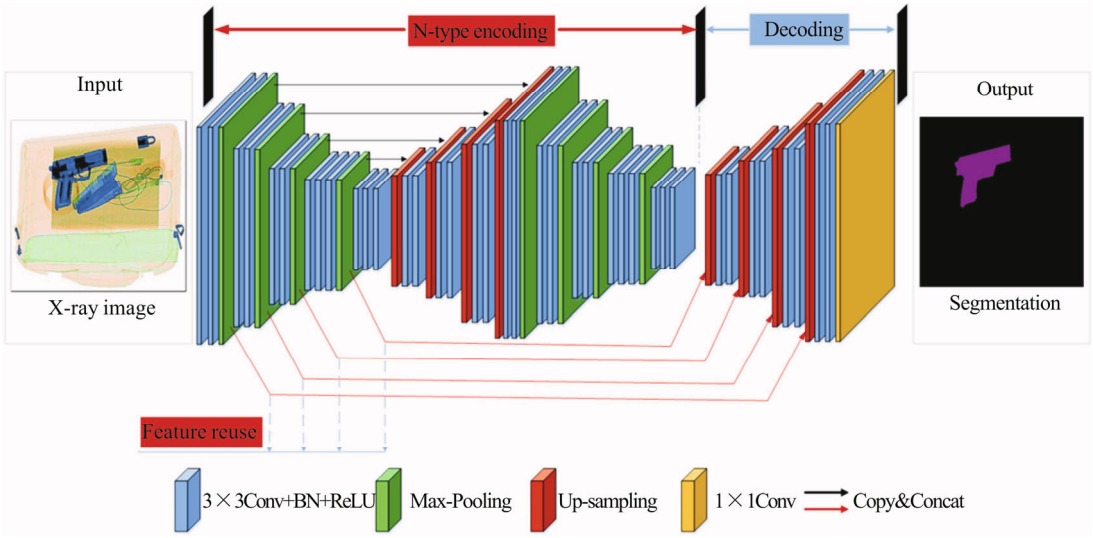

Fig.1 The prohibited items identification model

The main contributions of this paper are as follows: (1) We propose a model for prohibited items identification in X-ray security images based on semantic segmentation technology. (2) N-type encoding structure and feature reuse strategies are introduced to the semantic segmentation to improve the segmentation results. (3) We have achieved state-of-the-art performance on the X-ray security images datasets.

In the past years, deep learning based semantic segmentation methods have provided state-of-the-art performance. Long et al[3] proposed full convolutional neural network for the first time in 2014, which realized the segmentation task of images for any size and improved the segmentation efficiency. In 2015, the birth of U-Net[5] and SegNet[8] marked that encoding and decoding structure became the mainstream of semantic segmentation network. In the same year，dilated convolutions[9] provided a new strategy for the network to enlarge the receptive field, and the performance was excellent. In 2016, Zhao et al[6] proposed the PSPNet, in the following year, DeepLabv3[10] came out. What these models have in common is that they fuse information at different scales to improve the segmentation accuracy of the network. In 2018, Chen et al[11] introduced Decoder and Xception on the basis of DeepLabv3 to improve network performance and proposed DeepLabv3 plus. Since 2019, real -time semantic segmentation has become the focus of research gradually, with Lightweight Encoder-Decoder Network (LEDNet), Efficient Symmetric Network (ESNet)[12,13] and other networks constantly proposed. And these networks aim to balance the accuracy and speed of the network.

In the field of prohibited items identification, most researchers focus on the method of object detection. Wei et al[14] proposed the method of multi-task-transfer learning on the basis of Single Shot MultiBox Detector (SSD) to solve the problem for few positive samples of X-ray security images. Xu et al[15] adopt k-fold cross validation method to preprocess the dataset，and then change the skip structure to dense connection. Meanwhile, replace the Non Maximum Suppression (NMS) algorithm in You Only Look Once (version3) (YOLOv3) with soft-NMS. All kinds of technical improvements are to improve the prohibited items detection accuracy. However, few researchers put their focus on the method about image segmentation. Yang et al[16] introduced the attention mechanism to extract the regions of prohibited items in the image, segmenting the target regions afterwards (in 2018). But the segmentation accuracy of this method is not well. After that, An et al[17] added channel attention module on the basis of DeepLab v3+ to improve the segmentation results, which is the first use case of semantic segmentation in the identification of prohibited items (in 2019).

The prohibited items identification network proposed in this paper is based on an encoder-decoder style architecture commonly used in image segmentation. For the convenience of training, we choose the general semantic segmentation network with fewer parameters as the basic framework of the model. At the same time, in view of the

misclassification and poor accuracy of segmentation network, we made the following improvements in encoding and decoding.

In this paper, we propose an N-type encoding structure, which is different from traditional encoders. It is composed of down-sampling in two stages and one-stage up-sampling. The encoder is shown in Fig.2. Here we use Visual Geometry Group Network (16 layers) (VGG16) as the framework to complete the task of down-sampling at single stage, which realize extracting semantic information and shrinking the size of feature map. As a result, down-sampling of two stages consist eight pooling operations totally. Pooling enlarges the receptive field of network, but it is not conducive to the size recovery of feature map.

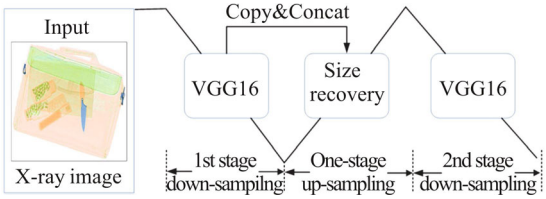

Fig.2 N-type encoding structure

We insert the one-stage up-sampling in the middle of the down-sampling in two stages. Meanwhile, inspired by the U-Net structure, we adopt the step by step up-sampling to fuse the semantic information extracted from first stage down-sampling. Ablation experiments prove all of these strategies help the feature map recover better ((U-Net & Deep-unet) vs OurNet). After the one-stage up-sampling, we carry out the second stage of down-sampling. In this stage, we retain the same convolution, pooling, and nonlinearization operations as the first stage down-sampling (VGG16).

At the end of encoding, the decoder generates a segmentation image of the same size as the input image by up-sampling. In this part, we focus on the selection of up-sampling methods and the reuse of shallow semantic information.

In FCN[3], authors use deconvolution to complete the task of up-sampling, but later research shows that this method is easy to produce checkerboard artifacts. We adopt bilinear interpolation combined with convolution to restore the size of feature maps. Compared with deconvolution, bilinear interpolation does not need to be learned. It runs fast and has simple operation. Just set the fixed parameter value, which is the coefficient that the center value needs to be multiplied.

In addition, there are two main characteristics of X-ray security images. The first is that the image color is monotonous. The second is the lack of surface texture. When judging the content of X-ray security images, the security personnel mainly depends on the contour information and color information of the object. What means that the low level semantic information plays a more important role in image recognition. For neural networks, the main task of shallow networks is to extract the contour of the object and some shallow semantic information. In the decoder, we use the shallow semantic information for the second time (feature reuse). Copying these and splice with the feature maps of the up-sampling to form thicker features. From the test results, it can be seen that the reuse of shallow semantic information can improve the segmentation accuracy of X-ray security images. Analysis of the reasons, in our network, each layer of the shallow network not only accepts the supervision from loss function in the original network, but also because there are connections between shallow network and decoder (shown as Fig.3), the supervision of the network is diverse. The advantages of deep supervision have been proved in Densenet[7].

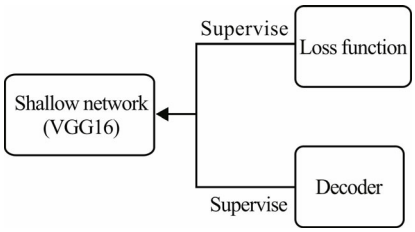

Fig.3 Double supervision

In this section, our datasets built by the laboratory are first introduced. After that, we perform a series of contrast experiments to prove the effectiveness of this model. Finally, we select the best model through carrying out some ablation studies. The experimental results indicate that our model could achieve state-of-the-art performance.

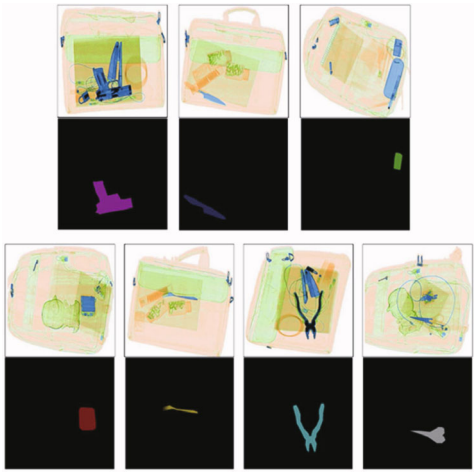

Fig.4 Examples of images in datasets

At present, there is no public datasets of X-ray security images based on semantic segmentation task. Using the equipment in the laboratory, we collected the X-ray security images and labeled all of them. There are 1 883 images in total. The size of each image is 512×512, before the experiment, we set the proportion of the training set to the test set to be 8:2 (Empirical reference value). In our datasets, there are 7 pixel-level labels, which include

gun, knife, power bank, lighter, fork, pliers and scissors. The examples of the datasets are shown in Fig.4. In the process of image collection, we randomly stack items in the baggage to simulate the actual security inspection scene as much as possible. In addition, because the scale of the datasets is still small. We do some data augmentation operations, such as flip, rotate, scale, Gaussian Blur and so on. (The following experiments were performed on the same dataset.)

In order to test the segmentation performance of this model, we do three groups of contrast experiments. Firstly, our model is compared with FCN8s (model published in 2014). Secondly, this model compared with PSPNet (model published in 2016). Thirdly, our network is compared with An's[17] model (model published in 2019). At the same time, all models are retrained in our datasets for a fair comparison. And these models are trained on a single NVIDIA 1080Ti GPU. We set the initial learning rate to 0.000 1. The epoch is set to 100. Using the pixel accuracy (PA) and the mean intersection over union (mIoU) as the measurement to evaluate the identification performance. Our experimental results are shown in Tab.1.

Tab.1 The accuracies of different methods

Models	molU	Time	Fps
ENet	51.131	**7 ms**	**128.7**
FCN8s	60.521	60 ms	16.6
PSPNet	70.400	76 ms	13.1
DeepLabv3	71.087	81 ms	12.3
An's model	74.500	280 ms	3.6
OurNet	**78.267**	39 ms	25.3

Combined with Tab.1, it can be seen that on the index of mean intersection over union (mIoU), there are large differences among several models. The score of Enet is 51.131, FCN8s is 61.521, PSPNet is 70.400, DeepLabv3 is 71.087, and An's model is 74.500. Compared to these models, our network get the best score of 78.267. Observing the final result (Fig.5), PSPNet and An's model are also good at the identification of power bank and gun. However, they have poor performance in the identification of prohibited items with small size, such as lighter and fork. Compared with all above models, our model could get more accurate segmentation results whatever the object is big or small. In addition, to show the speed of these models. We added two parameters, Time to process one picture (Time) and Frame per second (Fps). The results of Tab.1 show that real-time semantic segmentation network (ENet) get the best performance. It only takes 7 ms to process one picture, our network needs 39 ms to do the same operation. However, the segmentation result of ENet is poor. The score of mIoU is only 51.131, while ours is 72.267.

In this section, we test the different structures of the network to verify the effectiveness of the two modifications for this model. At the same time, in order to have a fair comparison. These ablation experiments performed the same training strategies as the contrast experiments (single NVIDIA 1080Ti GPU, lr= 0.0001, epochs=100). Our experimental results are listed in Tab.2.

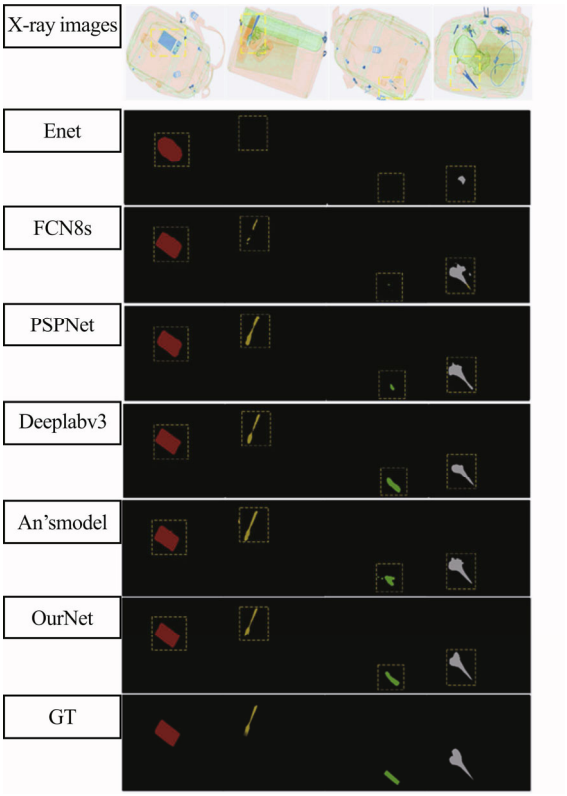

Fig.5 Results of contrast experiment

Tab.2 The accuracies of different structures

Models	molU	Time	Fps
U-Net	64.886	**28 ms**	**34.8**
LadderNet+	65.920	39 ms	25.3
Deep-unet	67.780	48 ms	20.7
OurNet	**78.267**	39 ms	25.3

First of all, our encoding structure is similar to U-Net. We did the first group of ablation experiment. As can be seen from Tab.2, compared with U-Net (mIoU:64.886), our network could improve performance by 20.62%. Moreover, Adding operation of up-sampling during encoding process，which is our first strategy to improve the performance of identification. To verify it, we cancel the one-stage up-sampling but keep the down-sampling of two stages, so that the network depth reaches 30 layers (call it Deep-unet (mIoU:67.780)). Experimental results show that our strategy can improve performance by 15.47%. Finally, in order to prove that reusing the shallow features of this network can improve the segmentation accuracy. We change the feature layer of fusion in decoding, and turn to fuse the deep semantic information of this network. It's similar to LadderNet after the structural change (call it LadderNet+ (mIoU:65.920)). Compared with it, the results

in the Tab.2 show that our network can increase performance by 18.73%.

Our network takes 39 ms to process one picture, U-Net only need 28 ms to do the same operation. It can be seen from Tab.2 that our network sacrifices a certain running speed to obtain a high segmentation accuracy. The above experimental results are shown in Fig.6.

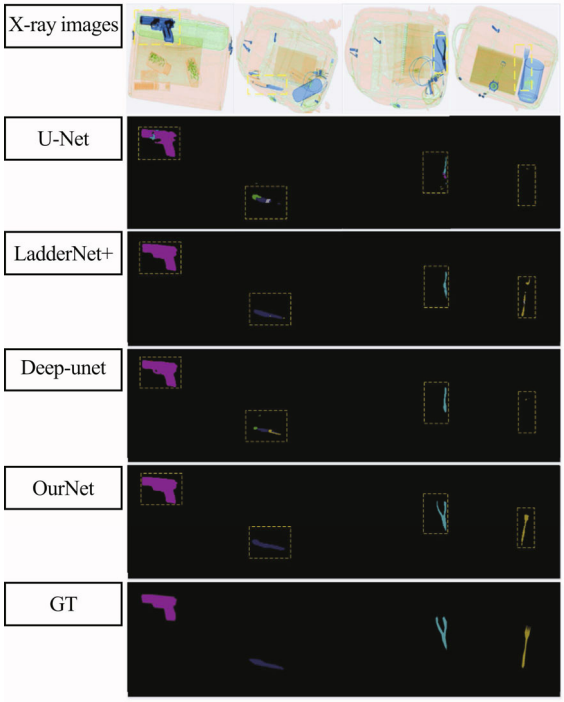

Fig.6 Results of ablation experiment

In this paper, we propose a semantic segmentation method, for the identification of prohibited items in the X-ray security images. However, semantic segmentation technology is facing two main problems: misclassification and poor segmentation accuracy. For this reason, we design an N-type encoder to enlarge the receptive field of network to reduce misclassification. Combined with the characteristic of the lack of surface texture in X-ray security images. We adopt a strategy for reusing shallow features to improve the accuracy of model segmentation.

To sum up, on the one hand, our network could get the state-of-the-art performance, which we can get this conclusion from the above experimental results. On the other hand, from the perspective of application field. little research has been done on the subject of prohibited items identification based on semantic segmentation. We hope our work can promote the development of intelligent security inspection, so as to improve the status quo of security inspection.

References

[1] Caijing Miao, Lingxi Xie, Fang Wan, Chi Su, Hongye Liu, Jianbin Jiao and Qixiang Ye, SIXray: A Large-Scale Security Inspection X-Ray Benchmark for Prohibited Item Discovery in Overlapping Images, CVPR, 2119 (2019).

[2] LeCun Y and Bengio Y, Nature **521**, 436 (2015).

[3] Long J, Shelhamer E and Darrell T, Fully Convolutional Networks for Semantic Segmentation, CVPR, 3431 (2015).

[4] Bolei Zhou, Aditya Khosla, Agata Lapedriza, Aude Oliva and Antonio Torralba, Learning Deep Features for Discriminative Localization, CVPR, 2921 (2016).

[5] Ronneberger O, Fischer P and Brox T, U-net: Convolutional Networks for Biomedical Image Segmentation, Medical Image Computing and Computer Assisted Intervention, 234 (2015).

[6] Hengshuang Zhao, Jianping Shi, Xiaojuan Qi, Xiaogang Wang and Jiaya Jia, Pyramid Scene Parsing Network, CVPR, 2881 (2017).

[7] Gao Huang, Zhuang Liu, Laurens van der Maaten and Kilian Q, Weinberger: Densely Connected Convolutional Networks, CVPR, 4700 (2017).

[8] Badrinarayanan V, Kendall A and Cipolla R, Segnet: A Deep Convolutional Encoder-Decoder Architecture for Image Segmentation, Pattern Analysis and Machine Intelligence **39**, 2481 (2017).

[9] Yu F, Koltun V and Funkhouser T, Dilated Residual Networks, CVPR, 472 (2017).

[10] Liang-Chieh Chen, George Papandreou, Florian Schroff and Hartwig Adam, Rethinking Atrous Convolution for Semantic Image Segmentation, arXiv preprint arXiv 1706.05587, 2017.

[11] Liang-Chieh Chen, Yukun Zhu, George Papandreou, Florian Schroff and Hartwig Adam, Encoder-Decoder with Atrous Separable Convolution for Semantic Image Segmentation, ECCV, 801 (2018).

[12] Yu Wang, Quan Zhou, Jia Liu, Jian Xiong, Guangwei Gao, Xiaofu Wu and Longin Jan Latecki, LEDNet: a Lightweight Encoder-Decoder Network for Real-Time Semantic Segmentation, arXiv preprint arXiv, 1905.02423 (2019).

[13] Yu Wang, Quan Zhou and Xiaofu Wu, ESNet: An Efficient Symmetric Network for Real-Time Semantic Segmentation, PRCV, 41 (2019).

[14] Wei Y and Liu X, Dangerous Goods Detection Based on Transfer Learning in X-Ray Images, Neural Computing and Applications, 1 (2019).

[15] Xu C, Han N and Li H, A Dangerous Goods Detection Approach Based on YOLOv3, Computer Science and Artificial Intelligence, 600 (2018).

[16] Xu M, Zhang H and Yang J, Prohibited Item Detection in Airport X-Ray Security Images via Attention Mechanism Based CNN, PRCV, 429 (2018).

[17] Jiuyuan An, Haigang Zhang, Yue Zhu and Jinfeng Yang, Semantic Segmentation for Prohibited Items in Baggage Inspection, International Conference on Intelligent Science and Big Data Engineering, 495 (2019).