Journal of Engineering Mechanics and Machinery (2021) Vol. 6: 85-93
Clausius Scientific Press, Canada
DOI: 10.23977/jemm.2021.060110
ISSN 2371-9133

Surface Defect Detection Model of Motor Commutator Based on Semantic Segmentation

Shenghan Hu

College of Logistic Engineering, Shanghai Maritime University, Shanghai 201306, China

Keywords: Surface defect detection, Computer vision, Deep learning, Semantic segmentation

Abstract: The surface defect detection of industrial parts is very important in industrial automation production, but there are problems with the small number of defect samples and the small-scale defect. To solve the above problems, this paper proposes a surface defect detection model of motor commutator based on semantic segmentation. The model is divided into two parts: segmentation network and classification network. First, the segmentation network uses an encoder-decoder to better capture small targets. The encoder uses an improved lightweight network MobileNet V3 as a feature extractor. Effectively learn the optimal features from a small number of samples, and improve the segmentation accuracy of the network. Then the classification network uses the segmentation results to make predictions, and the segmentation results provide interpretability for the prediction of the classification network. Experiments show that the proposed model has good generalization ability on a small number of samples, can effectively detect small-scale defect, and has high accuracy.

1. Introduction

In modern industrial processes, it is a very important task to check whether the surface quality of the product meets the standard. Therefore, the surface defect detection based on machine vision has very important industrial use and academic research significance. However, traditional machine vision detection methods need to manually extract features, which takes a lot of time, and the quality of feature extraction directly affects the accuracy. Compared with traditional machine vision methods, deep learning can simplify or even omit data preprocessing, and learn abstract and essential features directly from original data, instead of manually extracting features.

At present, deep learning is more and more widely used in the field of defect detection. For example, Xian et al. [1] proposed a novel cascaded autoencoder to accurately locate and classify defects appearing in images captured from real industrial environments. Mei et al. [2] used a convolutional denoising autoencoder network to reconstruct image patches, which can be trained with only a small number of defect-free samples. The above research fully proves the excellent performance of deep learning in detecting surface defects.

Motor is an indispensable equipment in industry, transportation and daily life. As the core component of motor, the quality inspection of commutator is a very important task. There are two problems in this paper. One is that the defects are small-scale, the defects only occupy a small part of

the pixels in the whole high-resolution image, so the detection accuracy of defects is low. The second is that deep convolution neural network needs a large number of samples training to achieve good results. However, the number of defect samples collected in the industrial field is small and the cost of collection is high, which is far from meeting the demand of the number of samples needed for deep learning.

In response to the above problems, this paper divides the model into two parts: segmentation network and classification network. The segmentation network separates anomaly locations from the image texture background, and then uses the segmentation results to assist the classification network to perform classification tasks, which is beneficial to improve the prediction accuracy of the classification network. At the same time, an encoder-decoder is adopted in the segmentation network, which combines high-level semantic information with low-level semantic information, which can effectively improve the segmentation accuracy of small targets. the encoder adopts an improved lightweight network MobileNet V3 to extract features, which greatly reduces the amount of parameters. Through the compact network structure design, the optimal features can be learned from a small number of training samples, and the defect detection can be performed more efficiently and reliably.

2. Defect Detection Model Based on Semantic Segmentation

A two-stage model is designed in this paper. The structure of the model is shown in Figure 1. The first stage is the segmentation network, The segmentation network locates the surface defects at pixel level, provides visualization of defects and interpretability for decision-making of classification network. The second stage is the classification network. The output of the segmented network assists the classification network to make accurate judgment, and outputs the probability value between 0 and 1. The greater the probability, the higher the possibility of defects.

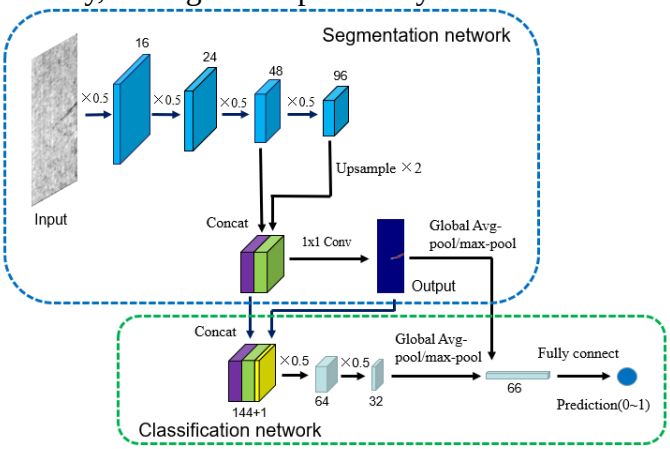


Fig.1 Surface Defect Detection Model of Motor Commutator Based on Semantic Segmentation

2.1 Segmentation Network

2.1.1 Proposed Bottleneck

Mobilenet V3 [3] model has high performance, less parameters and fast prediction speed. Therefore, this paper uses MobileNet V3 to extract features, considering that the number of samples used in this paper is small, and the resolution of the input image is high, in order to further reduce the amount of parameters, this paper improves the bottleneck of MobileNet V3 to be suitable for the current task.

Figure 2(a) shows the bottleneck of MobileNet V3, which first performs channel expansion on the input, then uses Depthwise convolution for feature extraction, then uses the SE attention mechanism,

and finally performs channel compression. The reason is that Depthwise convolution cannot change the number of channels, and the ability of feature extraction is limited by the number of input channels, so the number of channels is increased first, and then the Depthwise convolution is performed to extract features.

For the current task, this paper improves the bottleneck structure of MobileNet V3. We use multiple parallel depthwise convolution that will generate feature maps in different receptive fields, which is also equivalent to expanding channels to the next layer. as shown in Figure 2(b). Research [4] shows that 1×k, k×1 convolutions can provide a larger receptive field, and using a large convolution kernel is better than stacking multiple small convolution kernels. At the same time, considering that a large convolution kernel will bring more parameters and calculations, a separate convolution is used. Given that the current task is to detect small-scale defects in high-resolution images, a larger convolution kernel is used. Specifically, the size of the convolution kernel is 3×3, 5×5, 1×7, 7 ×1, 1×9, 9×1, 1×11, 11×1. Each convolution kernel generates a depthwise convolution path, then the feature maps of the paths are concatenated. Afterwards, a pointwise convolution is used to shrink the channels. In this paper, the improved lightweight bottleneck is used for feature extraction, which effectively reduces the number of parameters of the segmentation network.

The bottleneck expansion rate in this paper is 8, and the dimension reduction rate in the SE attention mechanism is 4. In the improved lightweight bottleneck, the number of weights is:

$$\begin{aligned} W_1 &= (9+25+7+7+9+9+11+11)\times C_{\text{in}} \\ &\quad + 8\times C_{\text{in}} \times 1\times 1\times C_{out} + 8\times C_{\text{in}} \times C_{\text{in}} \times 8\times C_{\text{in}} \\ &= (88+8\times C_{out} + 64C_{in}{}^2)\times C_{in} \end{aligned} \tag{1}$$

and a MobileNet V3 bottleneck is:

$$\begin{aligned} W_2 &= C_{in} \times 1\times 1\times 8\times C_{in} + 8\times C_{in} \times 3\times 3\times 1 \\ &\quad +8\times C_{in} \times 1\times 1\times C_{out} + 8\times C_{in} \times C_{in} \times 8\times C_{in} \\ &= (54+8\times C_{in} + 8\times C_{out} + 64C_{in}{}^2)\times C_{in} \end{aligned} \tag{2}$$

The value range of c_{in} is 16, 24, 48, 96. It can be seen that the weights of bottleneck improved in this paper is less under the same conditions.

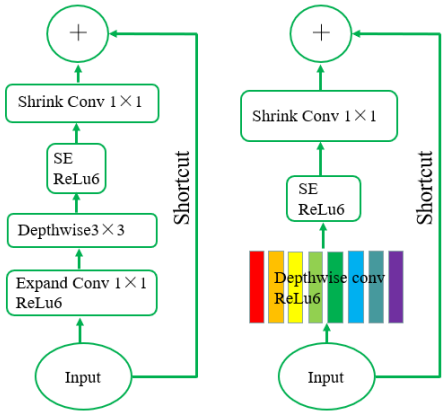


Fig.2 Mobilenet V3 Bottleneck on the Left and Our Bottleneck Structure on the Right

2.1.2 Encoder-Decoder Network

Table 1 describes the structure of the encoder in detail. A block is the basic convolution unit. It can either be a standard convolution or a bottleneck. In the table, N represents that the block are repeated N times, S is the stride. S is used in the first depthwise convolution when the bottleneck blocks are stacked repetitively. We use stride rather than pooling in the depthwise convolution to enlarge the

receptive field and reduce the feature dimension, since it is proved that fewer computational cost while no impact on the accuracy in using stride than pooling[5]. The input of the encoder is a gray image, which is subsampled by four times. In the decoder, the output of the 10th block is up-sampled by two times to the size of the 7th block output, and the result is concatenated with the output feature map of the 7th block, and finally the channel is reduced by pointwise convolution to obtain the final single channel segmentation output map. Considering the actual task requirements, only 2 times of up sampling is used in the decoder, which means that the final segmentation output map is 1 / 8 of the original image resolution.

Table 1 Segmentation Network Encoder

Input	operator	Output	S	N	SE	Activation	Block
1408×512	5×5	16	2	1	False	h-Swish	1
704×256	bottleneck	16	1	1	False	ReLU6	2
704×256	bottleneck	24	2	3	False	ReLU6	3-4
352×128	bottleneck	48	2	3	True	ReLU6	5-7
178×64	bottleneck	96	2	4	True	h-Swish	8-10

2.2 Classification Network

Classification network uses multiple convolution layer and down sampling layer operation, so that the network has enough depth to learn the characteristics of defects, and the down sampling operation enables it to capture larger global features and capture the overall shape of the defect, and distinguish noise and defect features, which is helpful to improve the performance of classification network. Among them, compared with the traditional convolution, the convolution operation adopts deep separable convolution, which has lower parameters and operation cost, and increases the nonlinear of the network and improves the learning ability of the network. The classification network uses 5 × 5 kernel sizes, down sampling twice, using stride of 2 to down sample.

The classification network not only uses the concatenated feature maps (144 channels) in the segmentation network, but also uses 1x1 convolution to obtain the final segmented feature maps (1 channel) after channel reduction. The two feature maps are concatenated together (145 channels) as the input of the classification network, so that the network can make full use of beneficial features, simplify the structural design of the classification network, and improve the prediction performance of the classification network.

Finally, the global maximum pooling and global average pooling are performed on the output feature maps (32 channels) of the classification network output, and 64 output neurons are generated, In addition, the final output feature map (1 channel) of the segmented network is also processed into global maximum pooling and global average pooling, and two output neurons are obtained. The 64 output neurons are connected with the two output neurons as the input of the full connection layer, and the final output probability is between 0 and 1, According to the fixed threshold, judge whether it is a defect.

3. Experiments

The model is implemented using the pytorch, and Adam is used for training in both classification and segmentation tasks. The loss function is both cross-entropy loss and the learning rate is 0.01. Considering the high resolution and GPU memory limitations, batch size is set to 1.

3.1 The Datasets

The KolektorSDD surface defect dataset [6] consists of 50 defective motor commutators. Specifically, microscopic fractions or cracks were observed on the surface of the plastic embedding in motor commutator. There are 399 images in total, 52 of which have visible defects (positive samples), for each image a detailed pixel-wise annotation mask is provided. The remaining 347 images have no defects (negative samples). The size of each picture is scaled to 1408×512. Annotation accuracy is particularly important in industrial settings since it is fairly time consuming. Considering that the labor spent on annotations should be minimized, four more annotation types were generated by dilating the original annotations with the morphological operation using different kernel sizes. as shown in Figure 3. The proposed model will be evaluated under different annotation accuracy. Some examples are shown in Figure 4.

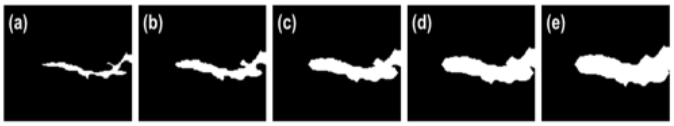


Fig.3 (a) the Original Annotation, (B) (C) (d) (e) with Different Morphological Kernel Sizes, I.e. 5, 9, 13, 15

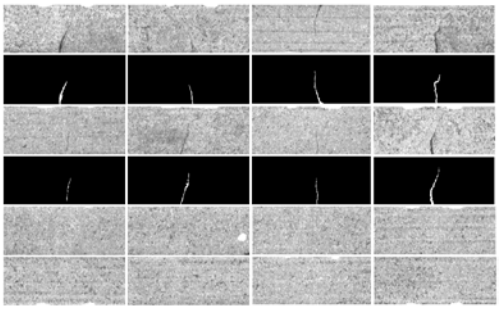


Fig.4 The Top is the Defective Sample and Their Annotation Masks, and the Bottom is the Defect-Free Sample

3.2 Results and Discussion

In actual industrial scene applications, it is usually more important to accurately classify whether there are anomalies in each image than to accurately locate defects in the image. Therefore, this paper does not evaluate the accuracy of the segmentation network, only evaluate the accuracy of the classification network. The output of the segmentation network is only used to visualize defects, which provides interpretability for the model. The dataset is divided into three subsets for 3-fold cross validation.

Considering the definition of average precision (AP):

$$AP = \sum_{n} (R_{\mathrm{n}} - R_{n-1}) P_{\mathrm{n}} \quad (3)$$

Where R_{n} and P_{n} are the recall and precision at the n-th threshold. It can be seen that the average precision (AP), as a comprehensive index under different thresholds, is inconsistent with the reality of industrial production. In fact, F-Measure under a certain fixed threshold for the whole dataset is more suitable than the average precision. In addition, false negative (FN) and false positive (FP) can more intuitively reflect the precision and recall. Therefore, this paper chooses F-Measure (F1), false negative (FN), false positive (FP), precision, recall as the evaluation indicators of the model [7]. Choose 0.9 as the threshold for judging positive and negative samples.

$$Precision = \frac{TP}{TP + FP} \quad (4)$$

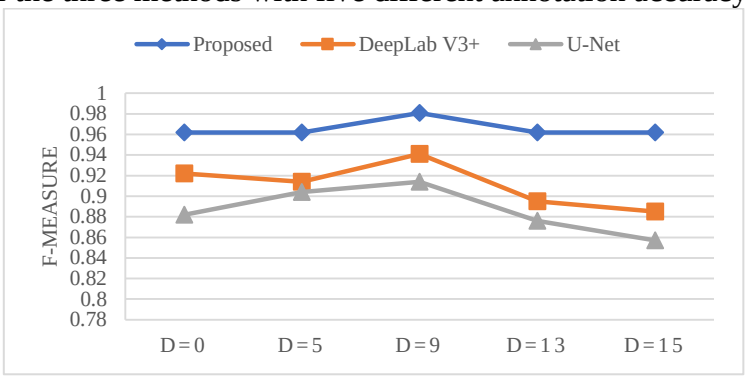
$$Recall = \frac{TP}{TP + TN} \quad (5)$$

$$F_\beta = (1+\beta^2)\frac{Precision \times Recall}{\beta^2(Precision + Recall)} \quad (6)$$

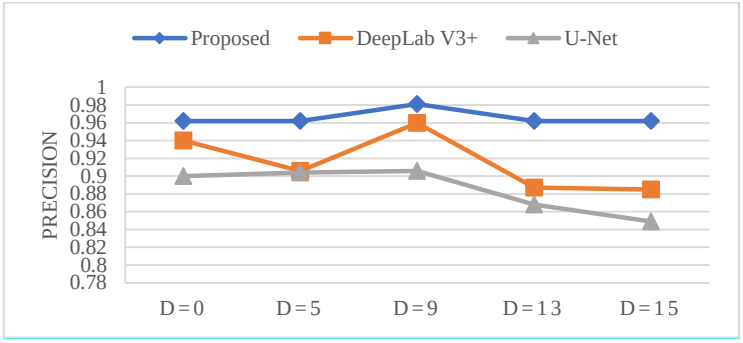
Where F_β is a weighted harmonic average value of precision and recall, and β=1 means that precision and recall are equally important in this paper.

The training of the classification network and the training of the segmentation network are carried out separately. First, the segmentation network is trained separately, and then the weights of the segmentation network are frozen, and only the classification network is trained. By fine-tuning the classification network, the over-fitting problem of the segmentation network with a large number of weights is avoided. In the process of training, the model adopts a resampling strategy to alternately train defective and non-defective samples, so as to ensure that the network observes the same number of defective and non-defective images, and prevent the training effect of the model is not good due to the imbalance of positive and negative samples.

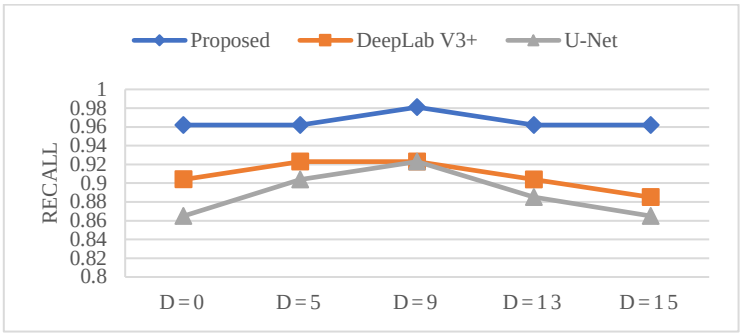
The proposed method is compared with two standard semantic segmentation networks DeepLabv3+[8] and U-Net[9]. The DeepLabv3+ network has been pre-trained before the evaluation, and the U-Net network are initialized randomly with normal distribution as in this paper. The output of these two kinds of segmentation networks are input into the proposed classification network for prediction. Because the output resolution of this paper is reduced by 8 times, the two semantic segmentation networks are only up-sampled to 1/8 of the original image. Both networks are trained with the same training configuration as the proposed model. Figure 5 and Figure 6 show the classification results of the three methods with five different annotation accuracy.



(a)



(b)



(c)

Fig.5 Performance Comparison under Different Annotation Precision. (a)f-Measure;(B)Precision;(C)Recall

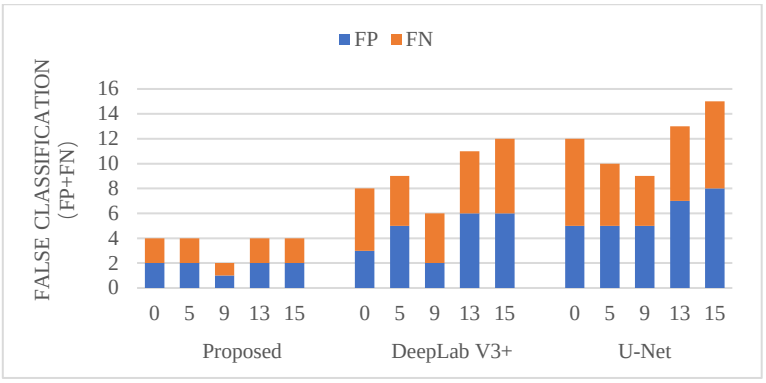


Fig.6 Comparison of Fp and Fn under Different Annotation Accuracy

As shown in Figure 5 and Figure 6, when the annotation accuracy is 9, DeepLabV3+, U-Net and the proposed method all achieve the best results. the proposed method is superior to the other in all annotation accuracy, and DeepLabV3+ performed the second best. Experiments have proved that coarse annotations can be sufficient to achieve a performance similar to the one with finer annotations, or even better performance than using fine annotations. This conclusion can reduce the labor cost of manual annotation and improve the flexibility of the production line. Observing the number of miss-classifications under the best results, it can be seen that the proposed method has only one false positive and one false negative. DeepLabv3+ has 2 false positives and 4 false negatives, while U-Net is worse, reaching 5 false positives and 4 false negatives.

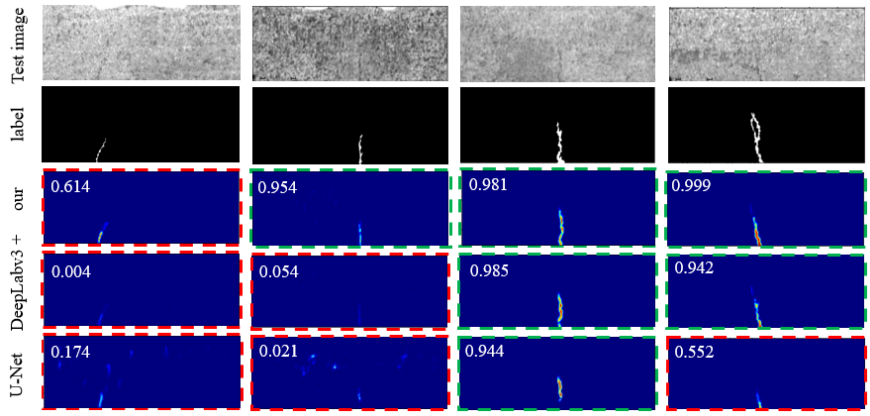


Fig.7 Examples of True-Positive (Green Border) and False-Negative (Red Border)

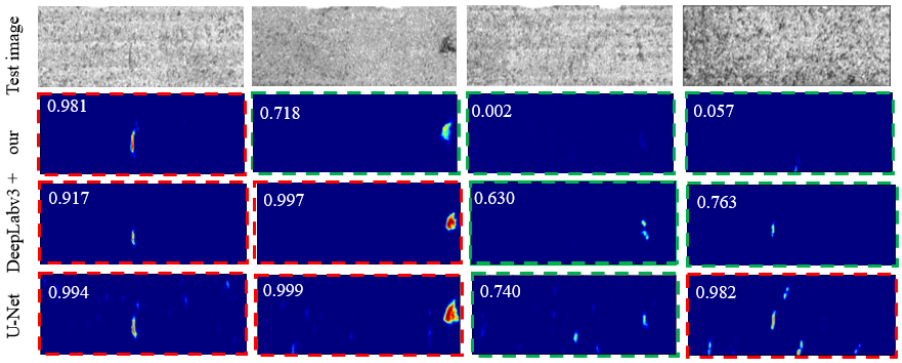


Fig.8 Examples of True-Negative (Green Border) and False-Positive (Red Border)

Several miss-classified images for all methods under the best results are presented in Figures 9 and 10. As shown in Figure 9, the proposed method has one false negative, This sample contains a small defect that is difficult to detect and was not detected with any of the remaining methods as well. For the remaining samples, the proposed method can correctly predict the existence of defects. The proposed method can also locate defects with excellent accuracy. Good positioning can also be observed in the other two methods. However, the segmentation effect of these methods is not good, and the prediction of the existence of defects is poor. True negative and false positive detection, as shown in Figure 10, Among the three methods, the proposed method has only one false positive, and other methods have observed more false alarms. In particular, the output of U-Net contains a lot of noise, and the classification network cannot completely distinguish defects from noise. The proposed method can correctly predict whether there are defects in the image.

4. Conclusion

This paper discusses the effect of using segmentation networks to detect surface defects from the perspective of specific industrial production. A two-stage detection method is proposed. The first stage uses pixel-wise labels of defect to train the segmentation network, and the second stage adds a classification network on top of the segmentation network to predict the presence of anomaly for the whole image. The proposed method is compared with two classic semantic segmentation methods on the Kolektor dataset. Experiments show that the proposed model achieves better results, while the other have a lot of miss-classified images. This is due to the two-stage design of segmentation and classification, and the lightweight bottleneck of the segmentation network, which reduces the amount of parameters and achieves high accuracy. In addition, the segmentation network uses an encoder-decoder and a large convolution kernel, which enhances the network's ability to capture small targets in high-resolution images.

References

[1] TAO X, ZHANG D, MA W, et al. Automatic Metallic Surface Defect Detection and Recognition with Convolutional Neural Networks [J]. Applied Sciences, 2018, 8(9):

[2] MEI S, WANG Y, WEN G. Automatic Fabric Defect Detection with a Multi-Scale Convolutional Denoising Autoencoder Network Model [J]. Sensors, 2018, 18(4):

[3] HOWARD A, SANDLER M, CHEN B, et al. Searching for MobileNetV3; proceedings of the 2019 IEEE/CVF International Conference on Computer Vision (ICCV), F, 2020 [C].

[4] CHAO P, ZHANG X, GANG Y, et al. Large Kernel Matters - Improve Semantic Segmentation by Global Convolutional Network; proceedings of the 2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), F, 2017 [C].

[5] SPRINGENBERG J T, DOSOVITSKIY A, BROX T, et al. Striving for Simplicity: The All Convolutional Net [J/OL]

2014, arXiv:1412.6806[https://ui.adsabs.harvard.edu/abs/2014arXiv1412.6806S.

[6] TABERNIK D, SELA S, SKVARC J, et al. Segmentation-based deep-learning approach for surface-defect detection [J]. Journal of Intelligent Manufacturing, 2020, 31(759-776).

[7] LIU G, YANG N, GUO L, et al. A One-Stage Approach for Surface Anomaly Detection with Background Suppression Strategies [J]. Sensors (Basel), 2020, 20(7):

[8] CHEN L-C, ZHU Y, PAPANDREOU G, et al. DeepLabv3+:Encoder-Decoder with Atrous Separable Convolution for Semantic Image Segmentation [M]//FERRARI V, HEBERT M, SMINCHISESCU C, et al. ECCV (7). Springer. 2018: 833-851.

[9] RONNEBERGER O, FISCHER P, BROX T. U-Net: Convolutional Networks for Biomedical Image Segmentation, Cham, F, 2015 [C]. Springer International Publishing.