Wuhan University
Journal of Natural Sciences

2020, Vol.25 No.5, 461-468

Article ID 1007-1202(2020)05-0461-08

DOI 10.19823/j.cnki.1007-1202.2020.0058

A Lightweight Improved U-Net with Shallow Features Combination and Its Application to Defect Detection

□ **WU Hong, SUN Xiankun†, XIONG Yujie**

School of Electronic and Electrical Engineering, Shanghai University of Engineering Science, Shanghai 201620, China

Abstract: In order to solve the problems of shallow features loss and high computation cost of U-Net, we propose a lightweight with shallow features combination (IU-Net). IU-Net adds several convolution layers and short links to the skip path to extract more shallow features. At the same time, the original convolution is replaced by the depth-wise separable convolution to reduce the calculation cost and the number of parameters. IU-Net is applied to detecting small metal industrial products defects. It is evaluated on our own SUES-Washer dataset to verify the effectiveness. Experimental results demonstrate that our proposed method outperforms the original U-Net, and it has 1.73%, 2.08% and 11.2% improvement in the intersection over union, accuracy, and detection time, respectively, which satisfies the requirements of industrial detection.

Key words: U-Net; depth-wise separable convolution; shallow features combination; defect detection

CLC number: TP 391

Received date: 2020-03-10

Foundation item: Supported by the Youth Fund of National Natural Science Foundation of China (61801286, 62006150), Shanghai Young Science and Technology Talents Sailing Program (19YF1418400), and Fund Project of Shanghai Science and Technology Commission (16dz1206002)

Biography: WU Hong, male, Master candidate, research direction: computer vision. E-mail: whaaa2018@126.com

† To whom correspondence should be addressed. E-mail: xksun@sues.edu.cn

0 Introduction

Small metal industrial products are widely used in engineering fields such as ship, machinery, and automobile manufacturing. In the production process, due to the influence of the production environment, various types of defects are generated on the surface of the small metal industrial products, such as edge cracking, oxidation, holes, and scratches. These defects have a great impact on the appearance and performance of the products, so it is important to detect the surface defects effectively.

At present, surface defect detection is performed mainly by deep learning methods based on convolutional neural network (CNN). For example, CNN has been successfully applied to polishing metal shaft surface defect detection, gear surface defect detection, wood surface defect detection and solar cells surface defect detection[1-5], etc. Cha *et al* [6] detected defects by combining CNN with the sliding window technology. CNN can only find the massive defects, but cannot completely detect the weak defects. Long *et al* [7] proposed fully convolutional neural network (FCN), which changes the fully connection layer of CNN into the convolutional layers. FCN classifies each pixel to realize semantic segmentation. However, FCN does not consider the correlation between pixels. The image details are easily lost after segmentation, which reduces the detection accuracy. Badrinarayanan *et al* [8] proposed SegNet to realize boundary segmentation by pooling index. SegNet can accurately detect large-size objects. However, neighboring information is ignored in the process of de-pooling, so it is impossible to accurately detect small defects. Chen *et al* [9-11] proposed a series of Deeplab networks, which use atrous convolutions to increase the size of receptive field. However, the use of atrous convolutions might lead to the loss of local details, which makes it impossible to accurately locate defects. Hu *et al* [12] proposed ACNet, which uses attention assistant module to segment images. ACNet can be used for semantic segmentation of indoor scene images with uneven brightness, but it is complicated and cannot satisfy real-time requirement. Olaf *et al* [13] proposed the U-Net architecture, which can use the encoder and decoder structure to segment small objects and retain location information. In recent years, U-Net has also been successfully applied to road detection [14-17], water body extraction [18], identification of ore minerals [19] and other fields. However, the feature extraction of U-Net mainly extracts the semantic features in the deep layer of the image, while the shallow layer texture information is lost. The loss of shallow layer features reduces the accuracy of segmentation and affects the final detection results. Moreover, the U-Net has many parameters and high calculation cost, so that the U-Net cannot satisfy the real-time requirement. In order to solve the above problems, we propose the IU-Net. IU-Net uses depth-wise separable convolution to reduce the number of parameters. At the same time, IU-Net adds several convolution layers to the skip path to combine more shallow layer features. And we use the IU-Net to detect the surface defects of small metal industrial products successfully.

The rest of this paper is organized as follows. Section 1 provides a detailed description of the IU-Net architecture. Section 2 gives the experiment and the results, including the dataset and the implementation details of the experiment. The paper ends with a conclusion of the major findings in Section 3.

1 IU-Net

1.1 The Original U-Net Architecture

The U-Net consists of a contracting path (Encoder) for feature extraction and an expanding path (Decoder) for precise positioning. Encoder consists of repeated application of two 3×3 convolution operations, each followed by a rectified linear unit (ReLU) for activation. ReLU function can be expressed as:

$$f(x)=\max\,(0,x) \tag{1}$$

where x is the input value and $f(x)$ is the output value. After the ReLU layer, the 2×2 max-pooling layer with step size of 2 is used for down-sampling. In order to reduce the loss of features in the down-sampling process, the number of feature channels is doubled in each down-sampling process.

Decoder consists of repeated application of two 3×3 convolution operations, a concatenation with the correspondingly cropped feature map from the encoder and a rectified linear unit for activation. After the ReLU layer, the 2×2 convolution (“up-convolution”) is used for up-sampling. The size of the feature map is doubled and the number of feature channels is halved in each up-sampling.

The decoder enlarges the size of the feature maps to ensure that the output image has the same resolution as the input image. The pooling process can be expressed as:

$$o=\text{ceil}[(\frac{i+2\times p-k}{s})]+1 \tag{2}$$

where o is the size of the output image, ceil is the return of the smallest integer, i is the size of the input image, p is the parameter of padding setting, k is the size of the convolution kernel, and s is the step size.

Accordingly, de-convolution operation can be expressed as:

$$o=s\times(i-1)+k-2\times p \tag{3}$$

Pooling and de-convolution operation are opposite. The pooling operation of the encoder reduces the resolution of the image, while de-convolution operation improves the resolution to restore the image size.

The distance between the expected value and the predicted value is calculated by the binary cross-entropy loss function. The loss function can be calculated as:

$$\text{Loss}=-\frac{1}{n}\sum_{i}^{n}(y_i\log(\hat{y}_i)-(1-y_i)\log(1-\hat{y}_i)) \tag{4}$$

where y_i is the expected value and $\hat{y}_i$ is the predicted value. Adam optimizer is used to accelerate the convergence speed. The step size update function can be defined as:

$$\theta_t=\theta_{t-1}-\alpha\times\hat{m}_t/(\sqrt{\hat{v}_t}+\varepsilon) \tag{5}$$

where α is the initial learning rate, $\hat{m}_t$ is the average value of the gradient, $\hat{v}_t$ is the variance of the gradient, and ε is an infinitesimal positive number to ensure that the denominator is non-zero.

1.2 The Improved U-Net Architecture

IU-Net architecture is shown in Fig. 1. Convolution operation is added in the skip path to extract more shallow layer features. Normal convolution is replaced by depth-wise separable convolution, and batch normalization (BN) layers are added to reduce parameters. Those changes reduce the calculation cost and improve the network generalization capability. To keep the dimensions of the input and the output images consistent, a padding layer is added to each convolution layer.

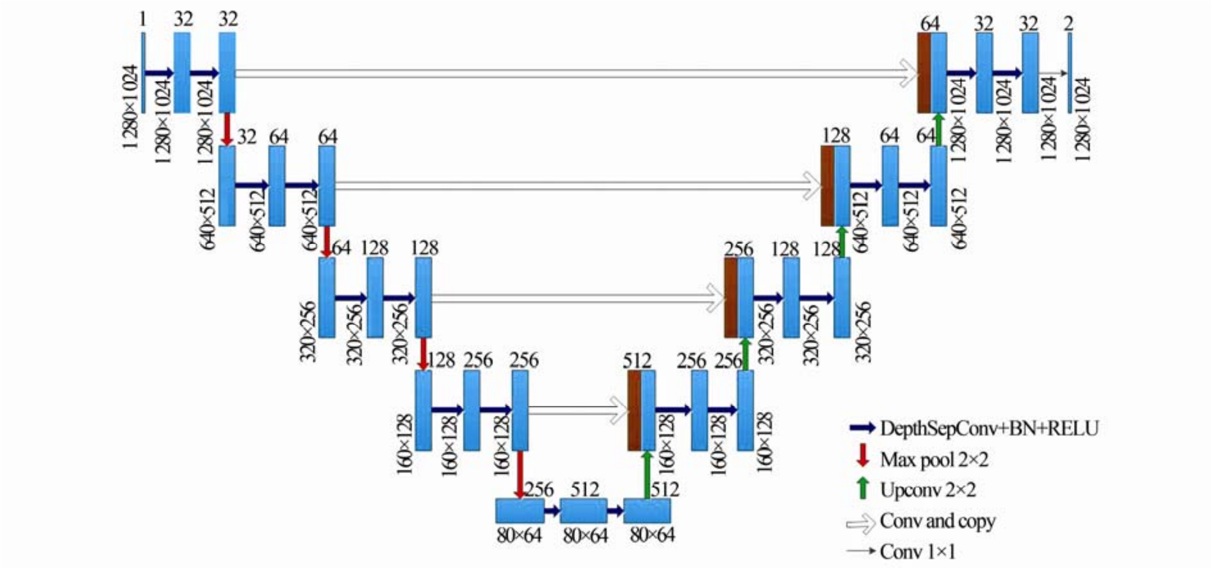

Fig. 1 IU-Net architecture

In the U-Net, there is a simple crop and copy operation to combine more information. The location and other information of the shallow layers are combined into the high-dimensional feature maps. However, due to the high complexity of the defect pictures, this simple concatenating operation leads to the loss of shallow layer features. The lost features are rich in texture and location information. Therefore, only the central part of the defects can be extracted. It is impossible to segment defects from the background accurately.

In order to combine the shallow layer features into the deep feature maps better, IU-Net adds the convolution operations and short links into the skip path. The output of node $X^{i,j}$ is represented by $x^{i,j}$, where i is the down-sampling layer, and j is the convolution layer added in the skip path. Then the output feature map can be expressed as:

$$x^{i,j}=\begin{cases}H(x^{i-1,j}), & j=0\\ H([[x^{i,k}]_{k=0}^{j-1},U(x^{i+1,j-1})]), & j>0\end{cases} \tag{6}$$

where $H(\cdot)$ is the convolution operation after the activation function, $U(\cdot)$ is the up-sampling layer, $[\cdot]$ represents the concatenation layer.

The visualization of the feature map of each layer in the down-sampling process is shown in Fig. 2. The shallow layer features are extracted through the convolution operation. The extracted shallow layer features are combined with the high-dimensional features to make up for the semantic gap between the feature maps of encoder and decoder. The skip path allows high-dimensional feature maps to be combined with more shallow layer features. And it ensures that the extracted features contain not only semantic information that can distinguish the defects from the background, but also effective texture features and location information that can accurately segment the scratches from the background. Therefore, the defects can be accurately located to the corresponding position of the original pictures.

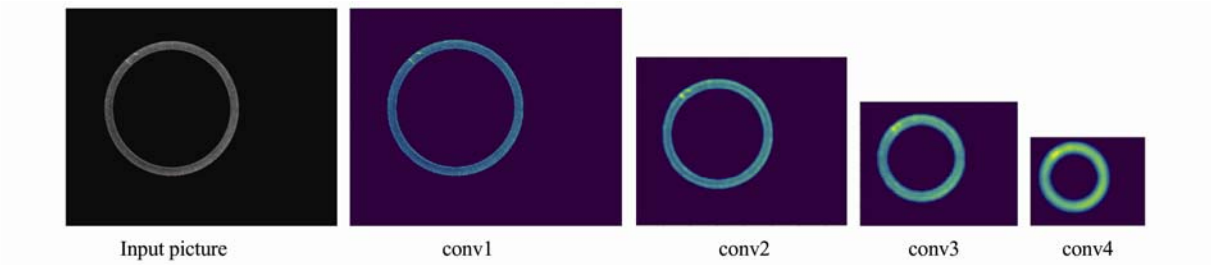

Fig. 2 Visualization of down-sampling feature map

In order to reduce the network parameters and calculation cost, and improve the generalization capability of the network, the depth-wise separable convolution is applied to IU-Net. The ordinary convolution is divided into the depth-wise convolution and a 1×1 point convolution, followed by the BN layer. Normal convolution operates on all input channels simultaneously. For each additional feature to be extracted, a convolution kernel must be added. The depth-wise separable convolution operates on different input channels respectively, and extracts the spatial characteristics of each input channel. Then the feature of each point is extracted by 1×1 convolution operation. If more features need to be extracted, only more 1×1 convolutions need to be designed. The ordinary convolution calculation cost is:

$$C_1 = D_K \times D_K \times M \times N \times D_F \times D_F \tag{7}$$

The total computing cost of the depth-wise separable convolution is:

$$C_2 = D_K \times D_K \times M \times D_F \times D_F + M \times N \times D_F \times D_F \tag{8}$$

where C_1 represents the ordinary convolution calculation cost, C_2 represents the depth-wise separable convolution cost, $D_K \times \mathrm{D}_K$ represents the size of the convolution kernel, $D_F \times D_F$ represents the size of the feature map, M represents the number of input channels, and N represents the number of output channels.

Their ratio is:

$$\frac{D_K \times D_K \times M \times D_F \times D_F + M \times N \times D_F \times D_F}{D_K \times D_K \times N \times M \times D_F \times D_F} = \frac{1}{N} + \frac{1}{D_K^2} \tag{9}$$

In order to prevent the disappearance of the gradient, BN layer and ReLU activation function are added after each depth-wise convolution and point convolution respectively, as shown in Fig. 3.

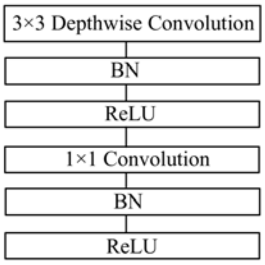

Fig. 3 Depth-wise separable convolution structure

The BN layer normalizes the output of the previous layer, speeds up the training speed and prevents gradient explosion. When the convolution layer outputs m feature maps of $p \times q$, the BN layer firstly calculates the mean value μ_B of all pixel points in the feature maps, μ_B is calculated as:

$$\mu_B = \frac{1}{M}\sum_{i=1}^{M} x_i \tag{10}$$

where $M = m \times p \times q$ and x_i is the pixel point. And then the variance σ_B^2 of all points is calculated as:

$$\sigma_B^2 = \frac{1}{M}\sum_{i=1}^{M}(x_i - \mu_B)^2 \tag{11}$$

Then the normalization operation as follow is carried out:

$$\hat{x}_i = \frac{x_i - \mu_B}{\sqrt{\sigma_B^2 + \varepsilon}} \tag{12}$$

where ε is an infinitesimal positive number.

Finally, the normalized data is reconstructed as follows:

$$y_i = \gamma \hat{x}_i + \beta \tag{13}$$

where γ is the scale factor and β is the offset factor.

The BN layer normalizes the output of the previous layer, which further speeds up the training speed and reduces the difficulty of parameter optimization.

2 Experiment

2.1 Experimental Environment and Parameter

The proposed methods are all performed on a workstation with i7-8700k CPU @ 3.6 GHz, 128GB RAM and an NVIDIA GTX1080 GPU with 8GB GPU memory. The code is implemented with PyTorch.

In this paper, Adam algorithm is used to optimize the model. Adam algorithm combines the Momentum algorithm and the RMSprop algorithm. It adaptively adjusts the learning rate and tremendously speeds up the training. The learning rate of the Adam algorithm is set to 0.000 1, the exponential decay rate of the first moment estimate is set to 0.9, and the exponential decay rate of the second moment estimate is 0.999. Batch size is the number of random samples per training step. An epoch refers to one full iteration of the training set. The main parameter settings of the networks are shown in Table 1.

Table 1 Parameters settings of training

Parameter	Value
Loss	Cross Entropy Loss
Batch size	2
Epoch	1 100
Optimizer	Adam
Momentum	0.9
Learning rate	0.000 1

2.2 Dataset

We used our own SUES-Washer dataset to train and validate our models. Totally 500 defects images of metal washers (including 300 images for training the model and 200 for testing the model) were taken in the factory with a resolution of 1 280×1 024. In order to ensure better generalization of the trained models, these images were taken from different conditions. Each image is manually labeled with the defects for the training and verification process.

In order to improve the generalization capacity of the model, we adopted the mix-up method [20] to augment the dataset, and finally augmented the training dataset to 500 images.

The examples of data augmentation are shown in Fig. 4.

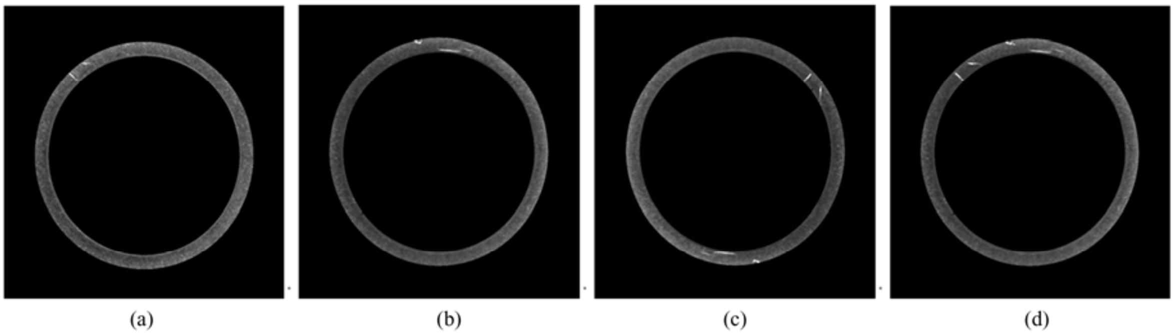

Fig. 4 Examples of data augmentation

(a) and (b) are original images, while (c) and (d) are generated images

2.3 Evaluation Criteria

In this paper, the performance of IU-Net is objectively evaluated by the mean intersection over union (MIoU) and the accuracy of detection. These two evaluation criteria are common in the field of image segmentation. MIoU can be expressed as:

$$\mathrm{MIoU} = \frac{1}{k+1}\sum_{i=0}^{k}\frac{\mathrm{TP}}{\mathrm{FN}+\mathrm{FP}+\mathrm{TP}} \tag{14}$$

where k is the number of classification, TP is true positive, FN is false negative and FP is false negative.

Accuracy can be defined as:

$$\mathrm{Acc} = \frac{N_{\mathrm{Correct}}}{N_{\mathrm{Sum}}} \tag{15}$$

where N_{Sum} is the total number of samples, and N_{Correct} is the number of samples segmented accurately.

2.4 Experiment and Analysis

IU-Net changes 3×3 ordinary convolution into depth-wise separable convolution. Table 2 shows the comparison between IU-Net and U-Net in the number of parameters and average time consumption of detecting a picture. The application of depth-wise separable convolution greatly reduces the number of parameters, reduces the calculation cost and improves the detection efficiency.

Table 2 Comparison between IU-Net and U-Net

Network	Number of parameters	Time / ms
IU-NET	**3 061 680**	**717**
U-Net[13]	3 836 213	807

The comparison networks in the experiment are the popular semantic segmentation network FCN8S [7] and SegNet [8]. The mean value of test results is shown in Table 3, and the accuracy and MIoU of the four network training processes are shown in Fig. 5. The accuracy of IU-Net is higher than that of other networks. Compared with the original U-Net, the accuracy is improved by 2.08% and MIoU by 1.73%. The IU-Net can detect surface defects more accurately.

Table 3 Comparison of evaluation criteria

%

Method	Accuracy	MIoU
FCN8S[7]	90.46	83.60
SegNet[8]	86.21	71.77
U-Net[13]	93.76	90.60
IU-Net	**95.84**	**92.33**

In order to verify the robustness of the IU-Net in the metal washer defect detection, four models are used to detect the defects. The IU-Net, original U-Net [13], FCN8S [7] and SegNet [8] are used to detect the defects images respectively, and the detection results are shown in Fig. 6. It can be seen intuitively from the figure that IU-Net can detect the edges of the defects accurately, and the defects can be precisely located to the corresponding position of the original pictures. However, the original U-Net is easy to make defects incomplete, which results in much error detection, and error detection is difficult to remove by morphological operation. FCN8S and SegNet are not accurate in the location of defects and result in much error detection. Moreover, it can be seen from Fig. 6 (5) (d) and (5) (e) that, FCN8S and SegNet fail to detect the defects.

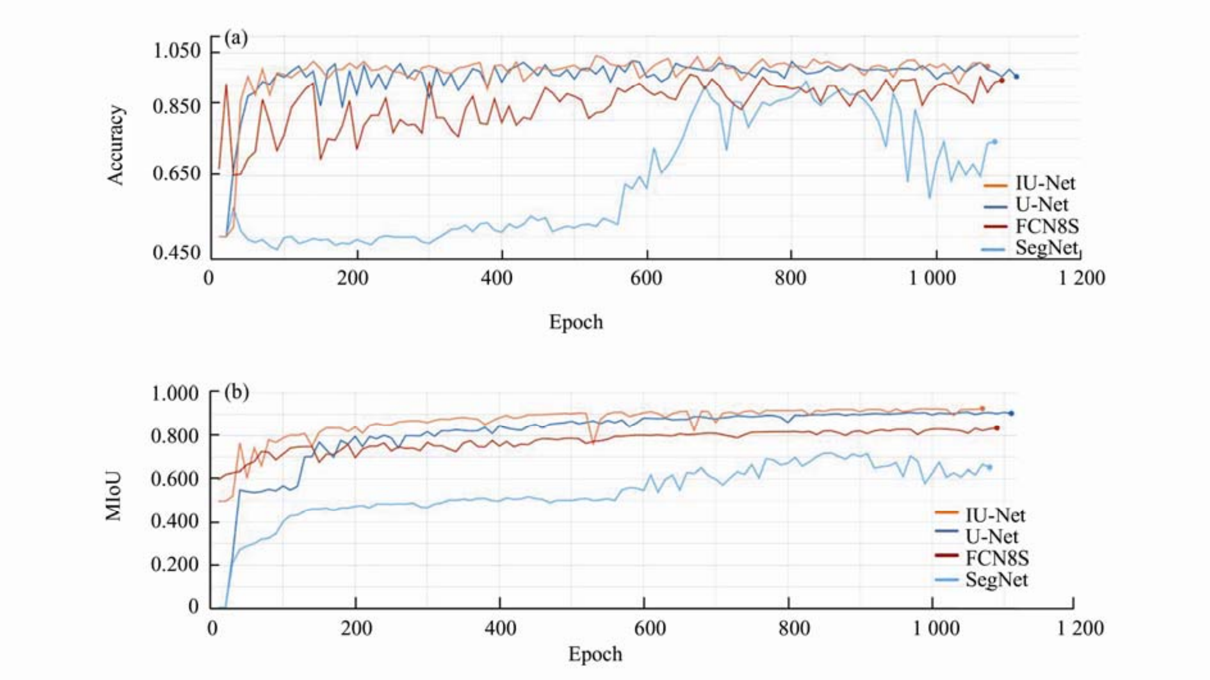

Fig. 5 The comparison of training accuracy (a) and MIoU (b)

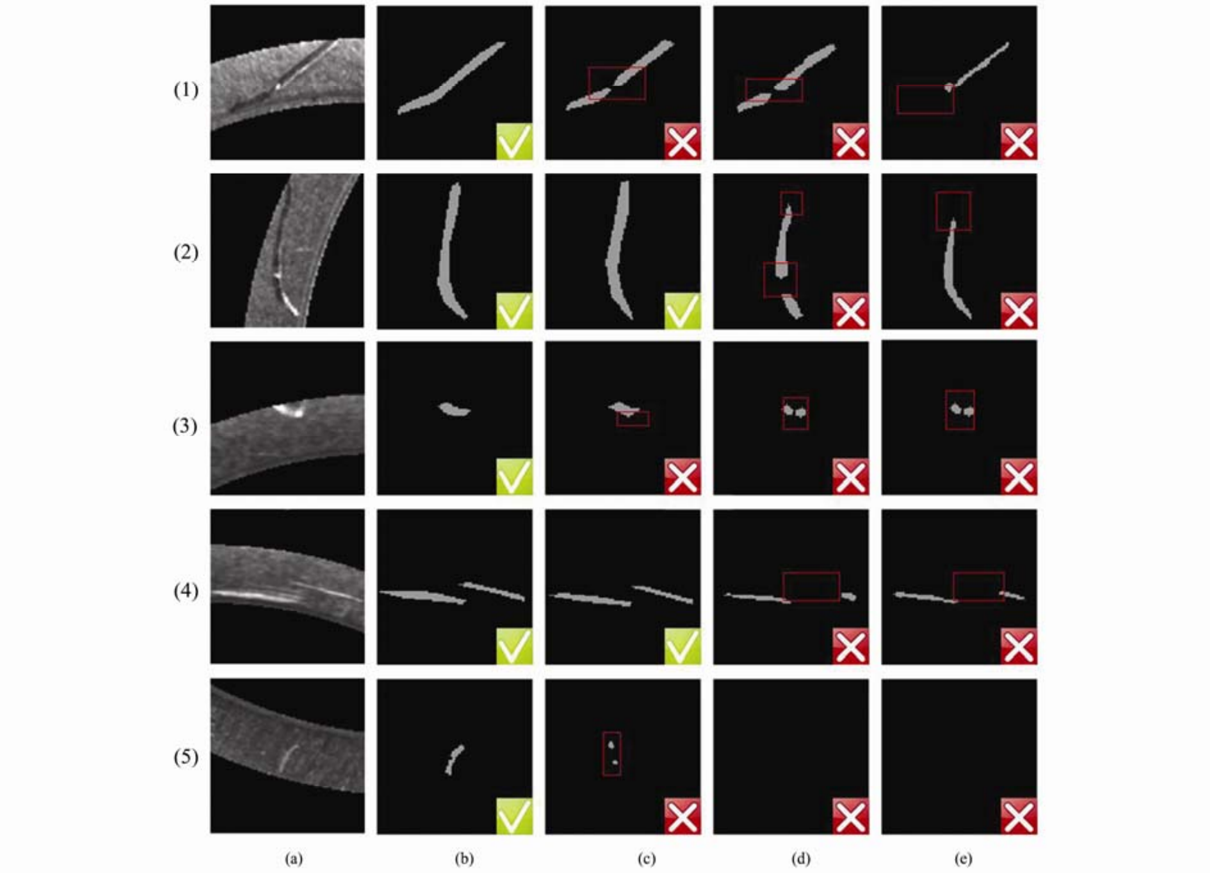

Fig. 6 Detection results

(a) Original pictures; (b) Our method; (c) U-Net; (d) FCN8S; (e) SegNet

In order to further verify the effectiveness of IU-Net in defect detection, 10 groups of four networks detection results are selected for the test, and the defects length is measured by the minimum rotating rectangle method. The error analysis is shown in Table 4. It can be seen from the results that the shape and size of defects obtained by the IU-Net are basically consistent with the ground truth. The maximum error is 1.32 mm, which satisfies the requirements of industrial detection.

Table 4 Measurement error comparison

mm

Method	Maximum error	Average error
FCN8S[7]	1.66	0.68
SegNet[8]	1.72	0.70
U-Net[13]	1.56	0.62
IU-Net	**1.32**	**0.53**

3 Conclusion

According to the features of the small metal industrial products, a novel defect detection method of IU-Net is proposed and tested. The major findings of this paper can be concluded as follows.

We propose a lightweight improved U-Net with shallow features combination. In the IU-Net, ordinary convolution is changed into depth-wise separable convolution to reduce the number of parameters and calculation cost. The convolution layer is added to the skip path to combine more shallow layer features to prevent the loss of defects details. And IU-Net is used to detect the surface defects of small metal industrial products. The results of the experiment show that IU-Net proposed in this paper can be used to detect the defects accurately, which satisfies the requirements of industrial detection.

References

[1] Yann L, Yoshua B, Geoffrey H. Deep learning [J]. *Nature*, 2015, **521**(1): 436-444.

[2] Jiang Q S, Tan D P, Li Y B, *et al*. Object detection and classification of metal polishing shaft surface defects based on convolutional neural network deep learning [J]. *Applied Sciences*, 2019, **10**(1): 121-127.

[3] Yu L Y, Wang Z, Duan Z J, *et al*. Detecting gear surface defects using background-weakening method and convolutional neural network [J]. *Journal of Sensors*, 2019, **1**(1): 221-237.

[4] Augustas U, Vidas R, Rytis M, *et al*. Automated identification of wood veneer surface defects using faster region-based convolutional neural network with data augmentation and transfer learning [J]. *Applied Sciences*, 2019, **9**(22): 223-229.

[5] Zhang X, Hao Y W, Shangguan H, *et al*. Detection of surface defects on solar cells by fusing multi-channel convolution neural networks [J]. *Infrared Physics and Technology*, 2020, **108**(1): 11-18.

[6] Cha Y J, Choi W. Deep learning-based crack damage detection using convolutional neural networks [J]. *Computer-aided Civil and Infrastructure Engineering*, 2017, **32**(5): 361- 378.

[7] Long J, Shelhamer E, Darrell T. Fully convolutional networks for semantic segmentation [C] // *Proceedings of the* 2015 *IEEE Conference on Computer Vision and Pattern Recognition*. Washington D C: IEEE, 2015: 3431-3440.

[8] Badrinarayanan V, Kendall A, Cipolla R. SegNet: A deep convolutional encoder-decoder architecture for image segmentation [J]. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2017, **39** (12): 2481-2459.

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

[10] Chen L C, Papandreou G, Kokkinos I, *et al*. Deeplab: Semantic image segmentation with deep convolutional nets, atrous convolution, and fully connected CRFs [J]. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2018, **40**(4): 834-848.

[11] Chen L C, Papandreou G, Kokkinos I, *et al*. Rethinking atrous convolution for semantic image segmentation [EB/OL]. [2017-01-20]. *http//www.arXiv preprint arXiv*: 1706.05587.

[12] Hu X, Yang K, Fei L, *et al*. AcNet: Attention based network to exploit complementary features for rgbd semantic segmentation [EB/OL]. [2019-08-25]. *http//www. arXiv preprint arXiv*: 1905. 10089.

[13] Olaf R, Philipp F, Thomas B. U-Net: Convolutional networks for biomedical image segmentation [C]//*International Conference on Medical Image Computing and Computer-Assisted Intervention*. Washington D C: IEEE, 2015: 234-241.

[14] Zhang Z, Liu Q, Wang Y. Road extraction by deep residual U-Net [J]. *IEEE Geoscience and Remote Sensing Letters*, 2018, **15**(5): 749-753.

[15] Liu L Z, Zhao Y. A closer look at U-Net for road detection [C] // *International Conference on Digital Image Processing*. Washington D C: IEEE, 2018: 1080-1086.

[16] Constantin A, Ding J J, Lee Y C. Accurate road detection from satellite images using modified U-Net [C] // *IEEE Asia Pacific Conference on Circuits and Systems*. Washington D C: IEEE, 2018: 423-426.

[17] Yang X F, Li X T, Ye Y M, *et al*. Road detection and centerline extraction via deep recurrent convolutional neural network U-Net [J]. *IEEE Transactions on Geoscience and Remote Sensing*, 2019, **57**(9): 7209-7220.

[18] Feng W, Sui H, Huang W, *et al*. Water body extraction from very high-resolution remote sensing imagery using deep U-Net and a superpixel-based conditional random field model [J]. *IEEE Geoscience and Remote Sensing Letters*, 2019, **16**(4): 618-622.

[19] Xu S T, Zhou Y Z. Artificial intelligence identification of ore minerals under microscope based on deep learning algorithm [J]. *Acta Petrologica Sinica*, 2018, **34**(11): 3244-3252.

[20] Zhang H, Moustapha C, Yann N, *et al*. Mixup: Beyond empirical risk minimization [EB/OL]. [2018-05-16]. *http//www.arXiv preprint arXiv*: 1710.09412.

□