Wafer Defect Map Pattern Recognition Based on Improved ResNet

YANG Yining, *WEI Honglei**

College of Mechanical Engineering and Automation, Dalian Polytechnic University, Dalian 116034, P. R. China

(Received 27 May 2024; revised 15 July 2024; accepted 8 August 2024)

Abstract: The defect detection of wafers is an important part of semiconductor manufacturing. The wafer defect map formed from the defects can be used to trace back the problems in the production process and make improvements in the yield of wafer manufacturing. Therefore, for the pattern recognition of wafer defects, this paper uses an improved ResNet convolutional neural network for automatic pattern recognition of seven common wafer defects. On the basis of the original ResNet, the squeeze-and-excitation (SE) attention mechanism is embedded into the network, through which the feature extraction ability of the network can be improved, key features can be found, and useless features can be suppressed. In addition, the residual structure is improved, and the depth separable convolution is added to replace the traditional convolution to reduce the computational and parametric quantities of the network. In addition, the network structure is improved and the activation function is changed. Comprehensive experiments show that the precision of the improved ResNet in this paper reaches 98.5%, while the number of parameters is greatly reduced compared with the original model, and has well results compared with the common convolutional neural network. Comprehensively, the method in this paper can be very good for pattern recognition of common wafer defect types, and has certain application value.

Key words: ResNet; deep learning; machine vision; wafer defect map pattern recogniton

CLC number: TP391.41; TP18; TN406 **Document code:** A **Article ID:** 1005-1120(2024)S-0081-08

0 Introduction

As semiconductor and information technology continue to advance, the proliferation of chips across various sectors has significantly impacted our lives and production processes. 5G communication systems rely on specialized chips, while the burgeoning field of autonomous driving necessitates sophisticated vehicle-grade chips. The internet of things (IOT) demands chips for control applications, and the aerospace industry imposes even higher standards, requiring chips with enhanced computational capabilities. The demand within the semiconductor market is surging. The semiconductor and integrated circuit industry play a pivotal role in chip manufacturing. Typically, chip design companies initiate the process by completing the integrated circuit design. Subsequently, the manufacturing phase begins with the fabrication of wafers, which serves as the foundation for chip production.

The standard process involves several key steps: Raw materials are cut to form monocrystalline silicon wafers, followed by a series of operations including engraving, cleaning, oxidation, diffusion, chemical vapor deposition, metal sputtering, photoresist coating, lithography, etching, and testing. The fabricated or cut wafers are then forwarded to packaging facilities for pin packaging and final product testing. As in Fig.1, this paper concentrates on the testing phase of the wafer manufacturing process, which is critical for ensuring the quality and reliability of the chips produced. As semiconductor processes advance, the performance of chips increases, as does their complexity, leading to a high-

*Corresponding author, E-mail address: weihl2005@163.com.

How to cite this article: YANG Yining, WEI Honglei. Wafer defect map pattern recognition based on improved ResNet[J]. Transactions of Nanjing University of Aeronautics and Astronautics, 2024, 41(S): 81-88.

http://dx.doi.org/10.16356/j.1005-1120.2024.S.010

er likelihood of defects. Thus, testing wafers to control quality is imperative and timely elimination of unqualified wafer. Generally, using electronic probes for each die for testing, for failure to pass the test grain, will be recorded with different numbers, and ultimately the formation of the wafer test chart, and can be simplified into the wafer defects maps[1-2]. By analyzing various defect patterns, quality engineers are able to identify distinct defect signatures, thereby deducing the underlying causes of these defects. This analysis is crucial for enhancing the production yield rate. As the scale and efficiency of wafer production increase, the timely and rapid detection and inference of wafer defects maps become increasingly imperative. In the production process, sometimes there will be a variety of defect patterns at the same time, so how to efficiently and accurately identify the defect patterns of wafers is very important.

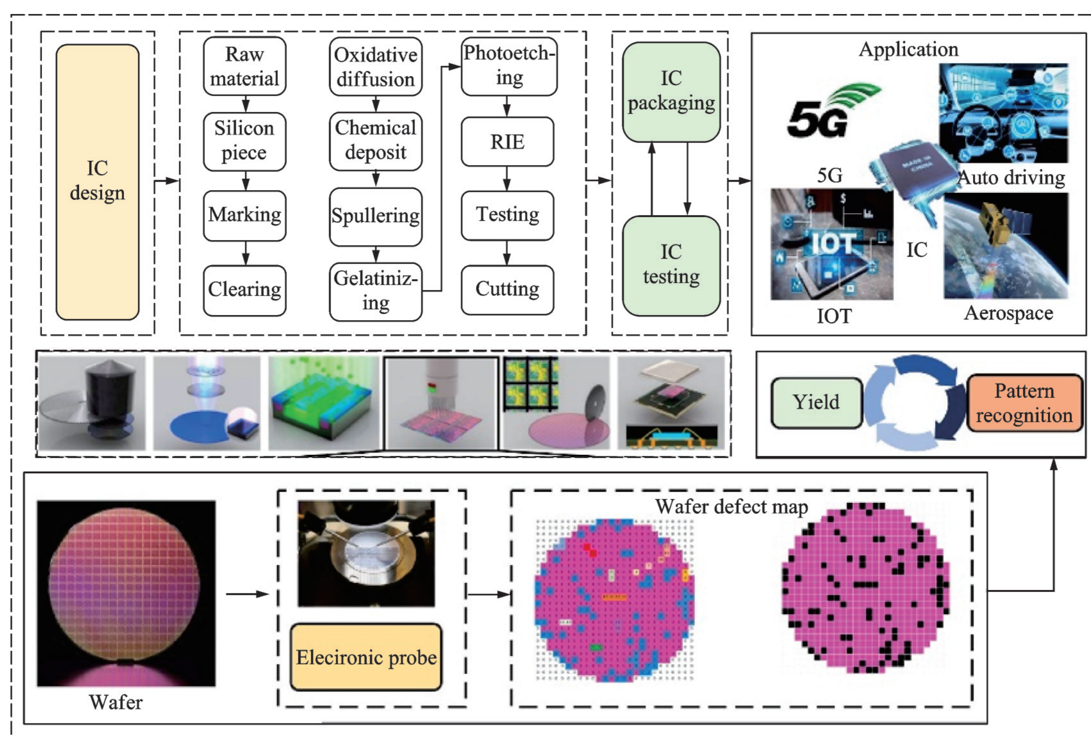

Fig.1 Wafer test framework

1 Related Work

Traditional methods for identifying wafer defect maps typically require engineers to have relevant knowledge and experience, followed by manual analysis[3]. However, the manual identification process is slow and inefficient, which is incompatible with the rapid and voluminous production and inspection demands of the modern integrated circuit industry. Consequently, machine vision technology has been increasingly employed to automate the detection of wafer defect maps. It can be divided into two main parts: traditional machine vision methods and machine learning clustering methods. Common approaches for machine vision involve Hough[4] transform to detect defects in the contour, the template matching method to match defects in the graph, HU invariant, Fisher discriminant[5], regional feature modeling[6], and support vector[7]. Such as this kind of method can detect simple defects to a certain extent in the contour, but the accuracy of detection is poor and manual feature extraction is required, which is difficult to use in practice and inefficient. Meanwhile, the clustering method often exhibits limitations when applied to practical scenarios due to its complex and time-consuming, which render it unsuitable for industrial-scale productions[8]. Besides, clustering paramenters often need to be set manually. With the development of deep learning, there are many scholars trying to use deep learning methods

for the detection of scratches on the wafer surface, the classification of defects. Nakazawa and Kang et al.[9-10] have successfully used the power of convolutional neural networks (CNNs) for the recognition of wafer defects maps, as well as for wafer defects maps retrieval tasks. Their approach achieved a well classification accuracy, demonstrating the efficacy of deep learning techniques in the domain of semiconductor manufacturing quality control. Yu et al.[11] proposed an 8-layer structure convolutional neural network to detect whether a wafer has a defect, and built a 13-layer model to classify wafer defect maps. Wang et al.[12] used the polar mapping to convert the wafer map into a matrix and then used CNN to classify the matrix data to obtain the better performance. Jin et al.[13] used CNN to extract features from wafer defects maps, which were subsequently fed into a support vector machine incorporating an error-correcting output code to classify them. Shim et al.[14] have developed a cost-effective convolutional neural network (CNN) classification model that not only exhibits well performance but also incurs minimal labeling cost. To augment the training efficiency of this model, they integrated active learning techniques to selectively filter unlabeled wafer defect maps, thereby prioritizing them for labeling. However, the above deep learning methods still have low accuracy and poor adaptation to wafer defect classification. So there is still room for improving the model effect. At the same time, most of the model parameter counts are large, which is not conducive to subsequent deployment. Therefore, this paper improves the ResNet network. Firstly, by adding the squeez-and-excitation (SE) attention mechanism to improve the feature extraction ability of the network, the useless information is suppressed. Then the activation function is replaced to improve the nonlinear expression ability of the model. And the depth separation convolution is used to lighten the model. Finally, good results are achieved.

2 Theories and Methods

As a class of feed-forward neural networks that contain convolutional computation and have a deep structure, CNN is one of the representative algorithms for deep learning. ResNet[15] is a classical convolutional neural network. The network pioneers the residual architecture, which improves the information propagation efficiency by adding directly connected edges to the nonlinear convolutional layers, and enables the gradient at the higher level to be directly transmitted back to solve the problem of gradient vanishing in deeper networks. The ResNet network is scalable and can be directly used in other networks, and the network training is fast and easy to optimize. In this paper, ResNet34 is used, in which the number 34 represents the number of layers of the network. The structure of ResNet34 is shown in Table 1.

Table 1 Structure of ResNet34

Layer name	Output size	Structure	Number
Conv 1	112 × 112	7 × 7 ,64,stride 2	1
Conv 2_x	56×56	3×3,64 3×3,64	3
Conv 3_x	28×28	3×3,128 3×3,128	4
Conv 4_x	14×14	3×3,256 3×3,256	6
Conv 5_x	7×7	3×3,512 3×3,512	3
Output	1×1	Average pooling Full connection	1
—	—	Softmax	1

In order to make the ResNet network better meet the requirements of pattern recognition of wafer defects maps, this paper improves it in several aspects: changing the activation function, embedding the SE attention mechanism, and using depth-separated convolution.

2.1 Changing the activation function

Activation functions enhance model nonlinearity. The ReLU activation in CNNs mitigates the vanishing gradient problem but risks dead neuron issues, potentially reducing training efficacy. In this paper, we use Leaky ReLU with leakage linear rectification function as the activation function, which can avoid the above problems. The mathematical expression of Leaky ReLU function is shown as

$$\text{Leaky ReLU}(x)=\begin{cases} x & x>0 \\ ax & x\leqslant 0 \end{cases} \tag{1}$$

Leaky ReLU function in $x<0$ can still calculate the gradient, so set x equal to 0.01.

2.2 Embedding the SE attention mechanism

Squeeze-and-excitation network (SE-net)[16] is a channel attention network. Hence SE-net is embedded into ResNet. Its function is to enhance critical features and suppress generic features. SE-net can learn the importance of feature channels independently without changing the original spatial dimension, and improve the model training by boosting the weights of more effective feature information based on the importance of feature channels. The SE-net model structure consists of three parts: Squeeze, excitation, and reweight, as shown in Fig.2. The process of SE-net function is as follows.

(1) Squeeze operation

The feature graph with the size of $C\times W\times H$ is compressed into a feature graph with the size of $1\times 1\times C$ by using average pooling, as F_{sq} shown in Eq.(2).

$$F_{\text{sq}}(C)=\frac{1}{HW}\sum_{i=1}^{H}\sum_{j=1}^{W}C(i,j) \tag{2}$$

where H and W are the hight and width of the feature map, respectively; C is the channel of the feature map; i and j are the pixel setting variables.

(2) Excitation operation

Through two fully connected layers, the weight and channel correlation of different feature channels are represented by the parameter W, shown as

$$F_{\text{ex}}(Z,W)=\sigma(W_2\delta(W_1,Z)) \tag{3}$$

where Z is the output of the squeeze operation, W_1 and W_2 are channel weights, $\sigma(\cdot)$ is the activation function and $\delta(\cdot)$ the normalization function.

(3) Reweight operation

The weight vector just generated is multiplied with the feature graph of $C\times H\times W$, and the weight is assigned to it to get a new feature graph. The resulting feature map is exactly the same size as before shown as

$$F_{\text{scale}}(C,S)=SC \tag{4}$$

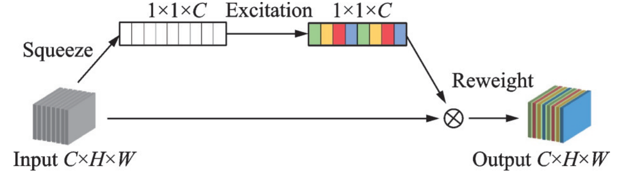

Fig.2 SE attention mechanism

2.3 Depth-separated convolution

Because the ResNet guarantees that the quality of network training will not be degraded by continuously deepening the number of network layers and using short-connection, it eventually achieves good results. However, as the number of network layers increases, the complexity of the network and the number of parameters also increase, ultimately leading to a too large model, which is not conducive to the subsequent industrial deployment. Therefore, it is very important to lighten the model. Therefore, in this paper, we use depth separable convolution to replace the original convolution.

The deep separable convolution consists of two parts, depth convolution and point-by-point convolution, as shown in Fig.3. As can be seen from the figure, the number of convolution kernels used for depth convolution is the same as that of channels of the input feature matrix, and the number of channels of the convolution kernel is 1. This operation can reduce a large number of parameters, but the depth convolution is only a convolution operation for each channel of the input layer independently, which cuts off the connection between different channels of the input feature map in the same spatial position. Following this, the newly generated output results are produced. Therefore, the newly generated output feature matrix needs to be convolved point-by-point to be used as the input layer, which finally realizes the weighting operation of the feature map in the depth direction. Point-by-point convolution is similar to the traditional convolution, and the difference is that the convolution kernel size used in point-by-point convolution is 1×1 and multiply the channel value. The depth-separable convolution used reduces the number of parameters used in the model compared to the traditional convolution, greatly reducing the complexity of the model.

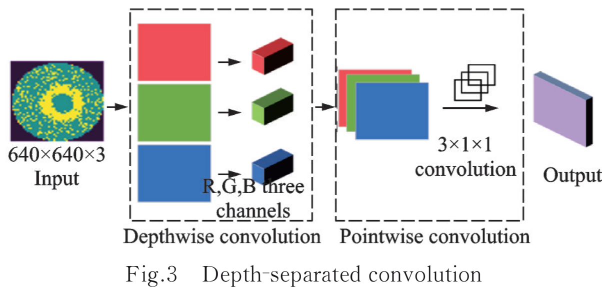

Fig.3 Depth-separated convolution

Finally, the depth-separable convolution is incorporated into the residual block and replaces part of the traditional convolution, while the SE attention mechanism is added to the residual block, and the improved residual block is shown in Fig.4.

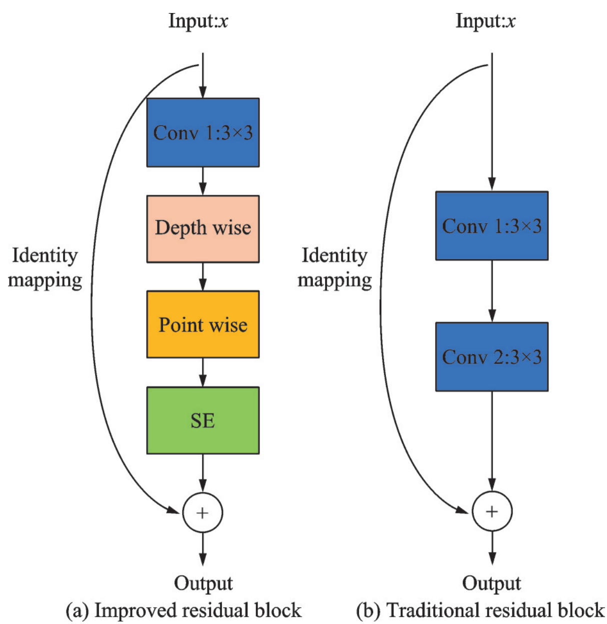

Fig.4 Residual block comparison

3 Experiments

To test the effectiveness of the proposed method, comparative tests are conducted. The operating system used is Windows 10, the CPU model is AMD Ryzen 7 4800 H, 16 GB of RAM, and the CUDA11.1.114 parallel computing framework with NVIDIA GTX1650 graphics card is used.

The experimental part is mainly divided into experimental data, model training, and comparison experiments.

3.1 Experimental data

A total of 6 149 wafer defect map data are used from the international public dataset, containing a total of 7 categories shown in Fig.5. The image size is 640×640.

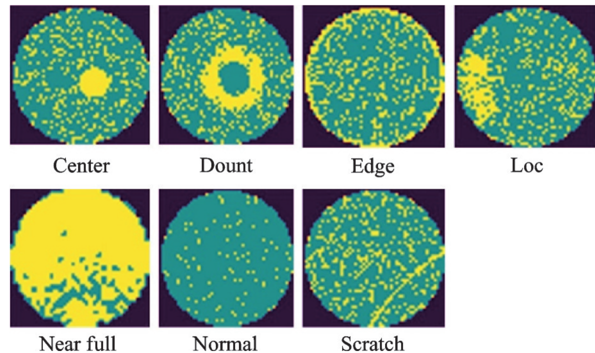

Fig.5 Wafer map sample

3.2 Model training

The number of training rounds, Epoch, is set to 100; the learning rate is set to 0.001, which is a hyperparameter determining the required step size for optimization; and the momentum is 0.9, which is also a hyperparameter in deep learning and is used to update the weight parameters of the model. The batch size is set to 10 to use the Adam optimizer.

Precision (P), Recall (R), Accuracy (Acc) and F_1 point are used to evaluate the detection results, which can be respectively expressed as

$$P = \frac{\mathrm{TP}}{\mathrm{TP}+\mathrm{FP}} \tag{5}$$

$$R = \frac{\mathrm{TP}}{\mathrm{TP}+\mathrm{FN}} \tag{6}$$

$$\mathrm{Acc} = \frac{\mathrm{TP}+\mathrm{TN}}{\mathrm{TP}+\mathrm{FP}+\mathrm{TN}+\mathrm{FN}} \tag{7}$$

$$F_1 = \frac{2\times P}{P+R} \tag{8}$$

where TP is the number of true classes, FP the number of false positive classes, FN the number of false negative classes, and TN the number of true negative classes. The train result of the improved ResNet is shown in Fig.6.

In Fig.6, the red line represents the loss value of training, the loss is the comparison between the prediction and the real label, and the smaller the loss value represents the better the training effect. From Fig.6, it can be seen that the loss value of the improved ResNet model in this paper decreases and converges rapidly, and finally stays below 0.1, with a better training effect. The blue line represents the

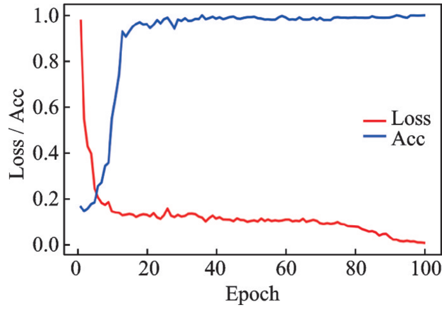

Fig.6 Train loss curves

accuracy value in training. As the number of training rounds increases, the accuracy rate increases rapidly and finally reaches about 98%. Considering the aforementioned results, the model improved in this paper shows good training performance.

In addition, the deep learning typically employs a comparison between the true and predicted values for each class to construct a confusion matrix, thereby evaluating the performance of the model in recognizing various categories within the test dataset. The confusion matrix plot for the ResNet model modified in this paper is shown in Fig.7.

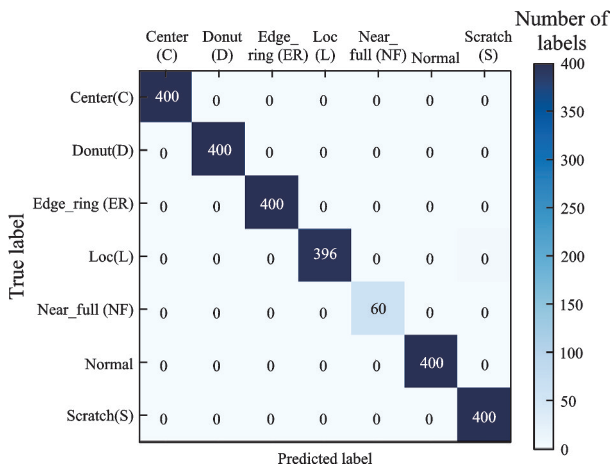

Fig.7 Confusion matrix

From Fig.7, it can be seen that the improved ResNet model can effectively identify most of the wafer defect types, with only a few errors, proving that the proposed model has good identification results.

In addition, to verify the robustness and generalization of the improved model presented in this paper, validation experiments are conducted on the trained model using a validation set. Experimental results indicate that the improved ResNet network achieves a precision value of 98.5%, an accuracy of 99.0%, a recall of 99.5%, and an F_1 point of 99.5%. The validation set shows that the improved model in this paper can well accomplish the pattern recognition of seven common wafer defects maps.

3.3 Comparison test

In order to verify the effectiveness of the improved model in this paper, comparison tests with other common convolutional neural networks (MobileNet, Alexnet, VGG) are also conducted. The test results are shown in Table 2.

Table 2 Comparison results %

Model	Precision	Accuracy	Recall
MobileNet	88.4	87.5	86.5
Alexnet	93.8	92.7	94.0
VGG	92.5	91.6	90.4
Our method	98.5	99.0	99.5

From the experimental results in Table 2, the performance of the improved ResNet in this paper is optimal compared to the other convolutional neural networks with the same test set, and the precision value is 10.1% higher than that of the worst-performing Mobilenet. Taken together, the improvement of ResNet in this paper achieves good results in the wafer defect map pattern recognition task. Finally, in order to verify the effectiveness of the lightening effect by using depth-separable convolution, the improved model is compared with the original ResNet34, whose results are shown in Table 3.

Table 3 Parameter comparison

Model	Number of total parameters	Parameter size/MB	Precision/%
ResNet34	21 797 672	83.15	95.5
Our method	13 707 176	52.29	98.5

From the results in Table 3, it can be clearly seen that improved ResNet variant exhibits a reduction in the number of parameters, amounting to a mere 62.8% of the parameter count in the original ResNet, and the network size and complexity are reduced significantly, but at the same time the accuracy is higher than that of the original ResNet by 3%. It shows that the improved network achieves a bal-

ance between lightweight and performance, showing good performance.

4 Conclusions

A deep learning convolutional neural network approach is used for automated pattern recognition of wafer defect maps, while the original ResNet is improved in order to increase the accuracy of the recognition and lighten the network. The feature extraction capability of the network is increased by embedding the SE attention mechanism, the activation function is changed to improve the nonlinear capability of the network, and finally a deep separable network is used to make the network lightweight for deployment. The final experiments a precision show that the improved ResNet in this paper achieves good results, with a precision of 98.5%, an accuracy of 99%, and a recall of 99.5%. Comprehensively, it is better than similar methods, and the amounts of parameters is reduced to 62.8% of the original ResNet, achieving the balance of network lightweight and performance, and has good application effect.

References

[1] EBAYYEH A A R M A, MOUSAVI A. A review and analysis of automatic optical inspection and quality monitoring methods in electronics industry[J]. IEEE Access, 2020, 8: 183192-183271.

[2] YU J. Enhanced stacked denoising autoencoder-based feature learning for recognition of wafer map defects[J]. IEEE Transactions on Semiconductor Manufacturing, 2019, 32(4): 613-624.

[3] EBAYYEH A A R M A, DANISHVAR S, MOUSAVI A. An improved capsule network (WaferCaps) for wafer bin map classification based on DCGAN data upsampling[J]. IEEE Transactions on Semiconductor Manufacturing, 2021, 35(1): 50-59.

[4] CHANG C W, CHAO T M, HORNG J T, et al. Development pattern recognition model for the classification of circuit probe wafer maps on semiconductors[J]. IEEE Transactions on Components, Packaging and Manufacturing Technology, 2012, 2(12): 2089-2097.

[5] YU J, LIU J. Two-dimensional principal component analysis-based convolutional autoencoder for wafer map defect detection[J]. IEEE Transactions on Industrial Electronics, 2020, 68(9): 8789-8797.

[6] YUAN T, KUO W, BAE S J. Detection of spatial defect patterns generated in semiconductor fabrication processes[J]. IEEE Transactions on Semiconductor Manufacturing, 2011, 24(3): 392-403.

[7] YUAN T, BAE S J, PARK J I. Bayesian spatial defect pattern recognition in semiconductor fabrication using support vector clustering[J]. The International Journal of Advanced Manufacturing Technology, 2010, 51(5): 671-683.

[8] CHEN S, YI M, ZHANG Y, et al. A self-adaptive DBSCAN-based method for wafer bin map defect pattern classification[J]. Microelectronics Reliability, 2021, 123: 114183.

[9] KANG S. Rotation-invariant wafer map pattern classification with convolutional neural networks[J]. IEEE Access, 2020, 8: 170650-170658.

[10] NAKAZAWA T, KULKARNI D V. Wafer map defect pattern classification and image retrieval using convolutional neural network[J]. IEEE Transactions on Semiconductor Manufacturing, 2018, 31(2): 309-314.

[11] YU N, XU Q, WANG H. Wafer defect pattern recognition and analysis based on convolutional neural network[J]. IEEE Transactions on Semiconductor Manufacturing, 2019, 32(4): 566-573.

[12] WANG R, CHEN N. Defect pattern recognition on wafers using convolutional neural networks[J]. Quality and Reliability Engineering International, 2020, 36(4): 1245-1257.

[13] JIN C H, KIM H J, PIAO Y, et al. Wafer map defect pattern classification based on convolutional neural network features and error-correcting output codes[J]. Journal of Intelligent Manufacturing, 2020, 31(8): 1861-1875.

[14] SHIM J, KANG S, CHO S. Active learning of convolutional neural network for cost-effective wafer map pattern classification[J]. IEEE Transactions on Semiconductor Manufacturing, 2020, 33(2): 258-266.

[15] HE K, ZHANG X, REN S, et al. Deep residual learning for image recognition[C]//Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. [S.l.]: IEEE, 2016: 770-778.

[16] HU J, SHEN L, ALBANIE S, et al. Squeeze-and-excitation networks[J]. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2019, 42(8): 2011-2023.

Acknowledgements This work was supported by the 2021 Annual Scientific Research Funding Project of Liaoning Provincial Department of Education (Nos. LJKZ0535,

LJKZ0526) and the Natural Science Foundation of Liaoning Province(No. 2021-MS-300).

Authors Mr. **YANG Yining** is now a postgraduate student in mechanical engineering at Dalian Polytechnic University. His research interests include machine vision and deep learning applications.

Dr. **WEI Honglei** received his Ph.D. degree from Dalian University of Technology in 2007. He is currently an associate professor of Dalian Polytechnic University. His research interests include machine vision and mechatronics design.

Author contributions Dr. **WEI Honglei** proposed the research direction and summarized the background and significance of the research. Mr. **YANG Yining** led the progress of the research and the improvement and implementation of the model and made a summary. The authors commented on the manuscript draft and approved the submission.

Competing interests The authors declare no competing interests.

(Production Editor: WANG Jing)

基于改进ResNet的晶圆缺陷模式识别研究

杨祎宁， 魏鸿磊

(大连工业大学机械工程与自动化学院,大连116034,中国)

摘要:晶圆缺陷检测是半导体制造的重要环节,通过对由缺陷形成的晶圆图进行缺陷模式的识别可以追溯生产过程中问题并进行专项改进,从而提高晶圆制造的良品率。因此,针对晶圆缺陷的模式识别问题,探究采用改进的ResNet网络对7种常见晶圆缺陷进行自动识别。在原ResNet的基础上,将SE注意力机制嵌入到网络中,提高网络的特征提取能力,发现关键特征,抑制无用特征。改进残差结构,加入深度可分离卷积代替传统卷积,降低网络的计算量和参数量使得网络轻量化,从而方便在工业环境中更好的进行部署。实验表明,改进后的ResNet模型准确率达到98.5%,参数量较原模型大幅减少,与常见的卷积神经网络相比具有较好的效果。综合来看,该方法能够很好地进行常见晶圆缺陷类型的模式识别,具有一定的应用价值。

关键词: ResNet;深度学习;机器视觉;晶圆缺陷模式识别