An overview of deeply optimized convolutional neural networks and research in surface defect classification of workpieces

Quanyang Li1, Zhongqiang Luo1,2, Hongbo Chen3, Chengjie Li4
1School of Automation and Information Engineering, Sichuan University of Science and Engineering, Yibin 644000, China
2Artificial Intelligence Key Laboratory of Sichuan Province, Sichuan University of Science and Engineering, Yibin 644000, China
3Sichuan Shuneng Electric Power Co. LTD, Chengdu 610000, China
4School of Computer Science and Technology, Southwest Minzu University, Chengdu 610041, China

Corresponding author: Zhongqiang Luo (e-mail: zhongqiangluo@gmail.com) and Chengjie Li (lcj@swun.edu.cn).

This work was supported in part by the National Natural Science Foundation of China under Grant 61801319, in part by Sichuan Science and Technology Program under Grant 2020JDJQ0061, 2021YFG0099, in part by the Sichuan University of Science and Engineering Talent Introduction Project under Grant 2020RC33, in part by Innovation Fund of Chinese Universities under Grant 2020HYA04001, in part by Artificial Intelligence Key Laboratory of Sichuan Province Project under Grant 2021RZJ03.

ABSTRACT Currently, the industrial development is becoming increasingly rapid. Technicalization, informatization and industrialization gives the fundamental impetus for industrial development and progress. Nevertheless, there are numerous problems that hindering industrial progress and human security in the industrial field. The surface defects of the workpieces are one of the primary problems. Moreover, defects of multi-type, mixed and unapparent characteristics presented by workpieces make the detection and classification of workpiece more difficult. Deep convolutional neural networks (DCNN) show strong ability of feature extraction and mines deeper essential features of data because of its features of unique receptive field structure and weights of shared. It can represent original data information well and obtain results more accurately than the traditional methods. But there also remains a problem that conventional DCNN has a huge number of parameters and computation, which brings great pressure to the equipment in terms of computing power, memory, speed and so on. Based on this situation, the optimization methods of CNNs model in the aspects of data, structure, algorithm are summarized. Related lightweight structures and networks are also summarized in this paper. The purpose of these work is to reduce the number of parameters and computation and improve the training performance. At the same time, the research on defect classification of workpieces based on traditional machine learning and deep learning model is reviewed, and the research on defect classification of workpieces based on deeply optimized CNNs is referred and prospected.

INDEX TERMS Deeply optimized CNN, workpieces detection, lightweight, surface defect classification

I. INTRODUCTION

With the diversification of people's life and the development of industries, the requirement of industrial products becomes increasingly strict. Qualified industrial products are the basic condition in guaranteeing people a healthy life and work. Therefore, the task of workpieces defect classification occupies an extremely important position in the industry. The detection of industrial equipment and related workpieces is the most important aspects to reduce economic loss and safety risk. For different workpiece products have different surface defects. Magnetic tile exists "crack", "blowhole", "break" and other defects, steel plate surface exists "pitting", "burr", "scratch", "crack" and other defects. These defects not only affect the product quality, but also have certain safety risks. Therefore, the research of surface defect classification for industrial workpieces is a hot research field. Defects of workpieces can come from many fields, such as electrical work, manufacturing, building industry and so on. Some common examples of surface defect images are shown in Fig. 1.
Deep learning is a learning model developed based on machine learning. As traditional machine learning methods can no longer adapt to the information characteristics in the era of big data. There is an urgent need for a learning model that can adaptively complete the learning task of information characteristics, so deep learning arises at the historic moment. Deep learning imitates the mechanism of human brain to analyze data, such as languages, images, sounds and so on [1]. At present, deep learning has been widely used in image processing, text understanding, speech recognition, defect classification, fault diagnosis and other fields. This technology is now introduced in the field of workpieces surface defect classification and combined with some optimization methods to improve the identification accuracy. At the same time, China's "Made in China 2025" white paper proposed the requests that "The use of advanced molding and processing methods, online classification devices, intelligent production, logistics systems and classification equipment should be promoted. The stability, quality reliability, environmental adaptability, service life and other indicators of key physical product performance can reach to the international advanced level of similar products"[2]. Therefore, the surface defect classification method based on deep learning not only has important academic research value, but also has very broad application prospect.

In the field of workpieces surface defect detection, detection methods can be divided into supervised [3], unsupervised [4] and semi-supervised [5] methods. Under supervised learning, CNNs is commonly used for research. This paper only explores the optimization and lightweight methods of CNNs under supervised learning, so the defect detection task under unsupervised learning and semi-supervised learning is not described.

In view of the problems of the complexity, the poor efficiency in conventional DCNN model under supervised learning. Paper summarizes the optimization and lightweight methods of conventional DCNN and describes their principle. And the research of other scholars applied DCNN which were optimized to workpiece defect detection and classification are discussed. The main contribution of this paper is that research on optimization and lightweight DCNN methods has certain reference significance in the future.

The structure of this paper is as follows. In chapter 1, the methods of workpieces surface defect detection and related studies respectively based on traditional machine learning and deep learning are briefly introduced. The chapter 2 introduces the basic structure, principle and improved network of CNN. In the third chapter, the optimization methods of conventional DCNN are related for the complexity and poor performance in conventional DCNN. Chapter 4 summarizes the current lightweight networks and their optimization methods. In chapter 5, the current studies of deeply optimized CNN in defect detection and classification of workpieces is reviewed. The final chapter summarizes the above five chapters and prospects the deeply optimized CNN in the task of workpieces surface defect detection and classification.

II. RESEARCH STATUS OF WORKPIECES SURFACE DEFECTS CLASSIFICATION
### TABLE1 CLASSIFICATION METHODS FOR SURFACE DEFECTS OF WORKPIECES

| Methods          | Concrete model or algorithm | Advantages                                                                 | disadvantages                                      |
|------------------|-----------------------------|-----------------------------------------------------------------------------|---------------------------------------------------|
| Traditional Mechanism | ECT [6,7,14]                | Low surface defect requirements; Fast speed; low cost.                      | Low degree of automation; Limited applicable objects; Few kinds of defects can be judged. |
|                   | MFL [11,12]                 |                                                                            |                                                   |
|                   | ACFM [9]                    |                                                                            |                                                   |
|                   | Ultrasound [8,10]           |                                                                            |                                                   |
| Machine Learning  | KNN [15,19]                 | The classification effect is good for those with few data sets and obvious defects. | Features need to be extracted manually; The robustness of the system is poor; Most of the system cannot achieve end-to-end structure. |
|                   | SVM [7,16-18,21]            |                                                                            |                                                   |
|                   | RF [13,20]                  |                                                                            |                                                   |
| Deep Learning     | CNN [23-25, 28-30]          | Deep extraction of effective feature information; Good robustness; High efficiency. | Conventional deep neural network requires a large amount of computation and parameters and high memory usage. |
|                   | AE [23,27]                  |                                                                            |                                                   |
|                   | DBN [26]                    |                                                                            |                                                   |

### A. Traditional mechanism methods

Since the various kinds of high sensitivity detection devices have developed rapidly. In the field of workpieces surface defects, many detection and classification methods have been studied such as EddyCurrent Testing (ECT), Alternating Current Field Measurement (ACFM), MagneticFlux Leakage (MFL) and Ultrasonic. These detection methods are mainly based on high sensitivity mechanical and electrical technology or optical technology. They can realize the detection and classification of workpieces surface defects through the electrical or magnetic signal processing. [6] introduced a differential excitation type sensor based on ECT for detecting metal defects. Aiming to detect and analyze the cracks of different depth and width, an experiment is conducted. The test results showed that the differential excitation type ECT device can be used for detecting defects on metal with high sensitivity. [7] presented a method to detect sub-surface defects in non-ferromagnetic material with motion based on ECT. [8] proposed a method of completely non-contact laser ultrasonic imaging to detect austenitic stainless steel with an artificial defect. In view of the rail surface crack, [9] studied the influence of probe scanning speed, crack length and depth on the detection signal based on ACFM. Combined with the optical moving rail device and ACFM detector, a rail dynamic scanning platform is built. However, these methods have many shortcomings such as the limited applicable objects and few kinds of defects that can be judged. Thus, they fail to meet the requirements of current workpieces surface defect detection and classification.

### B. Traditional machine learning methods

At present, methods of workpieces surface defect classification can be divided into machine learning-based and deep learning-based. Traditional defect classification based on machine learning methods have Random Forest (RF) [13,20], K-Nearest Neighbor (KNN) [15,19], Support Vector Machine (SVM) [7,16,17,18,21] and so on. However, there exists many weaknesses such as relying on manual extraction of features, big workload, poor robustness, complex steps. More importantly, many of them can not realize the end-to-end structure. The appearance of deep learning has effectively changed the surface defect detection method. The deep learning model can automatically and efficiently extract defect features, with relatively high robustness, and can better complete the task of workpieces surface defect detection and classification.

Traditional machine learning methods for classifying surface defects of workpieces have achieved better results before the advent of deep learning. Many scholars have also adopted these traditional methods and made certain improvements to them for the task of classifying defects in industrial products. [14] proposed an integrated model of multi-level SVM and decision tree to detect surface defects of aluminum plates. It combined with ECT probe, an aluminum plate surface defect detection and classification system are built. A variety of defect cracks have been studied at surface and sub-surface locations. [15] detected several surface defects in the gun barrel. It first extracting multiple texture features using histograms and gray co-occurrence matrices, then Bayes, KNN, ANN and SVM are used to test the classifier, and the results show that SVM has the best effect. [16] used SVM to classify solder joint defects of PCB originals, which focused on classifying solder joints with multiple features using the optimal kernel function of SVM. The mis-detected solder joints were classified again by using...
the SVM multi-classification algorithm with HOG features, with a final accuracy of 98.46%; [17] studied defect classification of cylindrical-shaped metal workpieces using SVM. It improved the classification accuracy by classifying scratches and liquid spill defects. An algorithm of several typical cold-rolled strip steel surface defects identification based on principal component analysis (PCA) and SVM methods are studied by [18], which achieved a high discrimination accuracy. [19] proposed a defect classification technique based on the integration of SVM, KNN and Bayesian classifier for defects such as texture, color and shape of tiles, achieved high classification accuracy. [20] proposed to apply the RF model to the classification of terminal defects. Firstly, using the decision tree algorithm to establish a single base model. And then through the majority voting criterion of multiple base models to obtain the RF model. Finally, the RF model is applied to the classification of terminal defects. The result showed that the accuracy of the model in the classification of terminal defects is 94%. [21] proposed a classification algorithm based on SVM to the classification of tool surface defects. It extracts features such as shape and texture of the defective area on the tool surface to train the SVM classifier. Finally, the proposed defect classification algorithm was experimented. The result showed that the defect detection rate of the algorithm reached to 97.2% and the classification accuracy reached to 94.3%. The above literatures studied the task of surface defect classification through using traditional machine learning. Although the recognition accuracy of some experiments was good, features needed to be extracted manually, and multiple algorithms were fused and integrated, which was relatively complicated.

However, with the development of technicalization, informatization and industrialization, the trend of multi-type, mixed and inconspicuous features presented by the workpieces defects has become increasingly significant. Traditional defect classification methods have gradually failed to adapt to this status. Moreover, most of them cannot realize the end-to-end structure and traditional defect classification methods need to be improved continuously to achieve a good effect. But this consumes a lot of time, manpower and energy. So, the use of more accurate, fast and systematic deep learning methods for the classification of workpieces surface defect is a hot research topic nowadays.

C. Deep Learning Methods

Deep learning has been applied more maturely after more than a decade of studies by countless predecessors. It are widely used by scholars and scientists in defect classification. Deep learning currently has four mature learning models as follows: convolutional neural network (CNN), recurrent neural network (RNN), deep belief network (DBN), and auto-encoder (AE) [22]. The use of deep learning models for surface defect classification is a current research craze.

[23] comprehensively compared the performance of CNN, SAE, MP and SVM-RBF based on the data set of wafer surface defects, and concluded that CNN was superior to other models, and further proposed an automatic defect classification system based on CNN. As traditional defect detection methods rely on manual extraction and have limited application occasions, [24] uses deep learning Faster RCNN model and VGG16 as feature extractor to detect mirror-like material defects with 94% accuracy. [25] proposed a classification and recognition method for surface defects of sheet metal parts with small samples based on CNN model, built a classic CNN model and fine-tuned parameters, and used data augmentation technology to expand the number of samples. The accuracy and real-time performance of this method can meet the requirements of actual industrial production. [26] used DBN for solar cell surface defects to pre-train the initial parameters of the network and fine-tune the network parameters by BP algorithm. Then it can obtain the mapping relationship between the training image and the defect-free image. Finally, the comparison between the reconstructed image and the defective image is used to achieve the defect classification of the test image. [27] used convolutional denoising auto-encoder (CDAE) networks with different gaussian pyramid levels to reconstruct image patches. Then it used the reconstructed residuals of the training patches as indicators for the defect prediction of direct pixel orientation, combining the reconstructed residual maps generated by each channel to generate the final classification results. The results have made good results on cloth silk fabric classification. In view of the classification and detection problem of small samples of steel plate surface defects, a steel plate surface defect detection system with markers based on CNN was proposed by [28]. 1300 defect samples were firstly collected, and then the collected images were denoised and edge detected. Finally, the CNN structure was adopted to detect and classify these defects. [29] processed medical images, extracted edge features by using a multiscale edge classification algorithm. It has also fused the underlying information using ResNet to achieve image segmentation. Its feature extraction accuracy of more than 80%. Based on the analysis of common defects on steel plate surface and existing deep learning surface defect classification algorithms, [30] selected the improved Faster R-CNN deep learning model to classify steel plate surface defects. It significantly improved the classification accuracy of small defects.

In the field of workpieces image defect classification, there are a few applications of AE and DBN for workpieces defect classification. There are almost no applications of RNN because it is not suitable for direct image feature extraction. The reason for this is that CNN has the receptive field structure based on machine vision, which is much more capable of image feature extraction than other models. In addition, CNN is suitable for workpieces surface defect image data set which are labeled. It belongs to supervised
learning. This model has high accuracy and good adaptability. Supervised learning is one of the primary methods in deep learning methods, which is widely used by researchers in many fields. Models of DBN, AE are more widely used in unlabeled workpieces data, belonging to unsupervised learning.

III. DCNNs

A. The basic structure of CNN

The history of CNN models can be traced back to the LeNet-5 model proposed by Yann LeCun in 1998 [31]. LeNet-5 model defined the basic structure of CNN for the first time, and it is considered as the first work of CNN model in the field. The main components of CNN are convolutional layer, pooling layer and fully-connected layer, in which convolution and pooling are done for all data of each layer. The deep features of the input data are extracted, and the fully-connected layer performs weighting calculation to obtain the result of classification probability. The structure of CNN is shown in Fig.2.

The input layer is the start of the neural network, and its main function is to receive the original training and test data. Then it will pre-processing this data to unify and optimize the data specifications to improve the training speed and effect of the network. In the convolutional layer, each convolutional layer contains a large number of convolutional kernels. The feature extraction of an image is to recalculate the pixel values of each part of the original image by using convolutional computation. Usually, we need to do ‘zero-paddings’ in the boundary of the image before the convolution layer in order to retain the edge feature information. But the retention also makes the image size bigger and bigger, which affects the training effect. So the pooling layer is added after the convolution layer, whose main function is to downscale the features, remove the redundant information and reduce the number of parameters. The fully-connected layer and the output layer are located in the last layers of the CNN. The fully-connected layer calculates the weight and bias of the extracted features, and the output layer normalizes the results. They can be combined to output the classification probability of the samples visually. The above is the basic structural composition of CNN. Each part has its own role and interacts with each other to complete the classification task of image feature extraction. But the conventional CNN is still difficult to classify large data and high-quality image data, the fundamental reason is that the conventional CNN is simply a stack of convolutional layers, pooling layers to a certain depth. Feature information will continue to be lost after heavy feature extraction and screening. Finally, the recognition accuracy will significantly decline. Therefore, it is necessary to continuously improve and optimize the CNN in order to achieve the requirements of recognition accuracy.

B. The development of CNN and its variants

Since Yann LeCun proposed the LeNet-5 model in 1998 and successfully applied it to handwritten digit recognition, CNN have gained a high status in the recognition task of simple image data set. However, as the image size becomes higher and higher, simple CNN can no longer meet the recognition task and will have a large error and a significant decrease in accuracy rate. People urgently need a new "guiding light".

Deep learning can mine the data feature information deeply to achieve the requirement of accurate recognition, which is favored by many scholars. So many researchers have used this concept to continuously improve the CNN, and finally multiplied AlexNet [32], VGG [33], GoogLeNet [34], ResNet [35], DenseNet [36] and other classical networks. Among them, AlexNet introduces the ReLU nonlinear activation function for the first time. Even if the number of layers is increased, the nonlinear relationship between output and input can be well characterized. In addition, the computational simplicity makes the output of some neurons become 0, which makes the network output more sparse and more in line with the principle of biological neural networks. The pooling layer in VGG becomes the maximum pooling layer, which can better extract significant features. In addition, the proposed inception structure in GoogLeNet can extract image features at multiple scales, enrich the extracted feature information and improve the recognition accuracy.
The above networks obtained excellent results in several ILSVRC competitions. However, later researchers found that the loss value will suddenly increase and the accuracy rate will decrease after reaching stability when the number of layers of the above network are deepening. Based on this, ResNet creatively introduced the shortcut connection, with the aim of ensuring that the accuracy rate would not decrease when the number of layers of the network are deepening.

IV. DEEPLY OPTIMIZED CNN

After more than two decades of development, CNNs have gradually expanded its network depth. From LeNet-5 with 5 layers, to AlexNet with 7 layers, to VGG with 16 and 19 layers, to GoogLeNet with 22 layers, and then ResNet gradually increased to 34, 50, 101 and 152 layers. Even ResNet and DenseNet network structure can reach thousands of layers. Although the shortcut connection is innovatively introduced into the ResNet structure so that the accuracy of the network does not decrease, the number of parameters to be trained and the computational effort of the network increases exponentially. It puts great pressure on the computing power, speed and memory of current devices. Therefore, it has great value of practicability and research to optimize the CNN in the training, structure, algorithm and other ideas such as transfer learning and lightweight. Finally, the improved networks can reduce the number of parameters and computation. The common optimization methods of CNN are shown in the Fig.3.

![Diagram](image-url)  
**FIGURE 3.** Schematic diagram of CNN optimization method.

A. Training Optimization

(a) Data optimization

1) Mean-variance normalization

Mean-variance normalization is a standardization method that transforms data into a standard normal distribution. In
regression models, independent and dependent variables that obey normal distributions often correspond to better regression predictions. And in the task of image recognition, image normalization is to centralize the data by de-meaning. According to the knowledge of convex optimization theory and data probability distribution, data centralization conforms to the data distribution law, which can improve the training speed and increase the generalization ability after training. The image normalization formula is as follows.

\[ N_{img\_std} = \frac{X - \mu}{\max(\sigma, \frac{1}{\sqrt{N}})} \]  

where \( X \) denotes the image pixel matrix, \( \mu \) is the mean image pixel value, \( \sigma \) denotes the standard deviation of all pixel values, and \( N \) denotes the number of image pixels.

2) Data augmentation

Defect data set of workpieces are not easily collected, while DCNN requires a large amount of data as input to obtain better training process and predictive classification. Based on this problem, The method of single-sample data augmentation can expand data set by changing the shape (clipping, twisting, flipping, etc.) and the color effect of the photo (brightness, contrast, saturation, etc.). The method of multiple-sample data augmentation can expand data set through inter-sample fusion, inter-sample combination and other methods. Common methods include Mixup, Cutmix, Mosaic and so on. This step can increase the data set by several or tens of times. [37] used data augmentation to expand their data set by a factor of 10. And [38] used data augmentation to increase their data set by removing some facial features, changing image saturation, and image rectangular segmentation of face samples.

(b) Gradient Descent Optimization

Training process of CNN contains two stages, namely forward propagation and backward propagation. The forward propagation can get the results which is achieved by convolution, pooling of the input data. Then comparing with the actual results, calculating the cross-entropy loss error. At this moment, each layer hyperparameters are in a stationary state.

The main work of CNN back propagation stage is to update and optimize the parameters of each layer based on the error value calculated by forward propagation so as to minimize the loss function and make the predicted value close to the correct value. In general, using the gradient descent method to make the parameters reach to the optimal convergence value, at the same time, the neural network also reaches to the optimal state.

Forward propagation of the calculation formula is as follows.

\[ Z^{(l)} = W^T \ast X^{(l)} + b \]  

Nonlinear activation part:

\[ A^{(l)} = g(Z^{(l)}) \]  

Loss function part:

\[ J(W,b) = \frac{1}{m} \sum J(W,b;x,y) + \frac{\lambda}{2} \sum W^2 \]  

Where \( W, b \) are the weights and biases of this layer, \( X \) is the input vector of this layer, \( Z \) is the result of linear operation of this layer, \( g \) is the activation function, \( A \) is the result of nonlinear operation and the input vector of the next layer at the same time. The first term of the loss function part is the mean square deviation term. The second term is the regularization term. It will make the regularization for \( W \) but not for \( b \), this is because it will not have much effect without processing for \( b \). \( \lambda \) is the weight decay coefficient.

There must be an error between the sample classification result and the actual result. The error is calculated by the cross-entropy loss function. And the gradient of the weight and bias is then calculated according to the error. Finally, the parameters are back-propagated by layer to achieve the optimization of the network and minimize the error. The back propagation stage is calculated as follows.

\[ \Delta W = \frac{\Delta J}{\Delta W}, \Delta b = \frac{\Delta J}{\Delta b} \]  

\[ W^{(l)} = W^{(l)} - \alpha \Delta W, b^{(l)} = b^{(l)} - \alpha \Delta b \]  

Where \( J \) is the cross-entropy loss function, \( \Delta W, \Delta b \) are the results of the partial derivatives of \( W, b \) according to \( J \) respectively. And \( \alpha \) is the learning rate.

The basic principle of traditional gradient descent method and stochastic gradient descent (SGD) method [39] both is to update the parameters in the negative gradient direction according to the loss function. However, these two methods have a slow convergence rate and the learning rate needs to be selected manually. It is very likely to fall into local minima when facing complex and nonlinear problems. [40] introduced the concept of momentum based on the basis of SGD by calculating the exponentially weighted average of the gradient. This can avoid falling into local minima.

Although the momentum gradient descent method has very excellent performance compared to the first two methods, the learning rates of these algorithms are the same value. Moreover, in deep neural networks, updating the parameters with the same learning rate may lead to very slow convergence or even divergence. So Adagrad [41], Adadelta [42], RMSprop [43], Adam [44], and other adaptive learning rate optimization algorithms to allow the learning rate to be updated and changed as well. Among them, Adam algorithm combines RMSprop and momentum gradient descent method with the following equation.
\[ v_t = \beta_1 v_{t-1} + (1 - \beta_1) g_t \]  
\[ s_t = \beta_2 s_{t-1} + (1 - \beta_2) g_t^2 \]  
\[ \theta_t = \theta_{t-1} - \frac{\alpha v_t^*}{\sqrt{s_t^* + \epsilon}} \]  
\[ v_t^* = \frac{v_t}{(1 - \beta_1)} \]  
\[ s_t^* = \frac{s_t}{(1 - \beta_2)} \]  
\[ \alpha = \frac{1}{(1 + t)^{\gamma}} \]

Where \( v_t \) is the average of the historical squared gradient of exponential decay, \( s_t \) is the exponentially decaying mean of the historical gradient. \( \beta_1 \) and \( \beta_2 \) are both momentum coefficients. \( g_t \) represents the differential variable. \( \epsilon \) represents a very small number to prevent the denominator from being zero. Equation (7) draws on the idea of momentum. Equation (8) draws on the idea of second-order momentum and exponential moving average in RMSprop. Equation (9) and Equation (10) use bias correction to avoid the values of \( v_t \) and \( s_t \) converging to zero at the beginning of training so as to obtain better statistics. Equation (11) considers both the gradient mean and the squared gradient to adjust adaptively the updated magnitude of the parameters.

The Adam's algorithm for optimal parameter work is a common method that people use. It has achieved good results in parameter convergence. In order to verify the performance of different optimization algorithm, ResNet model, image recognition data set named CIFAR10 are selected. The default settings of learning rate and momentum are used, other hyperparameters are unchanged. And the loss value and accuracy are used as dependent variables, the number of iterations is the independent variable. The results are shown in Fig. 4. Our experiments prove that the Adam algorithm has better performance compared to other optimization algorithms. In addition, the performance of the SGD with momentum is similar to that of the Adam algorithm. Because the result of SGD with momentum is generally the best under good tuning of the parameters.

![Graphical representation of the image classification performance of five optimization algorithms. (a) Accuracy curves of optimization algorithm, (b) Loss value curve of optimization algorithm.](image)

### Hyperparameter optimization

The hyperparameters in the neural network contains the number of layers, learning rate, momentum, mini-batch size, and so on. They are directly affecting the training effect. So, optimizing the hyperparameters and selecting a suitable set of hyperparameters is a prerequisite to ensure good training effect, robustness, and strong generalization ability of the neural network. In addition, optimizing the hyperparameters is a combinatorial optimization problem. It is very difficult to optimize hyperparameters. And the time cost of evaluating a set of hyperparameter configurations is very high. Commonly hyperparameter optimization methods have Grid search, Random search, Bayesian optimization, Genetic algorithm, Particle swarm optimization et al.

1) **Grid search.**

Grid search as its name implies, the upper and lower boundaries of the hyperparameters are set. And then putting the combination of different hyperparameters in the upper and lower boundaries into a grid. Multiple groups of hyperparameters are combined to form a large grid and traversed one by one, which is actually the exhaustive method. Finally, select the best performance of a group as the network hyperparameters. If a large search range and a small step size are used, the global optimum can be found with a high probability. However, this search scheme is very computationally intensive and time consuming especially when there are more hyperparameters to be tuned. Therefore, in practice, the grid search method generally uses a wide search range and a large step size to find the possible locations of the global optimum, then it will gradually reduce the search range and step size to find a more accurate optimum. This operation reduces the time and computational effort required. However, the objective function is generally non-convex, the global optimum is likely to be missed.

2) **Random search.**

Using grid search will make unnecessary attempts on unimportant parameters. Therefore, an improved method is to perform a random combination of hyperparameters, instead...
of testing all values between the upper and lower bounds. Sample points are randomly selected in the search range and then a configuration of hyperparameters with the best performance is chosen. The rationale is that if the set of sample points is large enough, then the global optimum or its approximation can also be found with high probability by random sampling. Random search will generally be faster than grid search, but like the fast version of grid search, its results are not guaranteed.

3) Bayesian optimization.

Bayesian optimization algorithm uses a completely different method from grid search and random search in finding the optimal value parameters. Bayesian optimization method is an adaptive hyperparameter optimization method. The hyperparameter combinations based on existing experiments are used to predict the next combination that may bring the greatest benefit. Grid search and Random search will ignore the information of the previous point when testing a new point. Bayesian optimization algorithms take full advantage of the previous information. The Bayesian optimization algorithm learns the shape of the objective function to find the parameters that raise the objective function toward the global optimum. Specifically, at the first, assuming a collection function based on the prior distribution. Then, each time the objective function is tested using new sampling points, this information is used to update the prior distribution of the objective function. Finally, the algorithm tests the points at the most likely location of the global optimum given by the posterior distribution. For Bayesian optimization algorithm, there is one thing to note, once a local optimum is found, it will keep sampling in that region. So it is easy to fall into the local optimum. To compensate for this, the Bayesian optimization algorithm finds a balance between exploration and exploitation, where "exploration" is to obtain sampling points in the unsampled region, and "exploitation" is to sample in the region where the global optimum most likely to appear based on the posterior distribution.

4) Genetic algorithms.

Genetic algorithms are computational models of biological evolution that simulate genetic selection and elimination in nature. Suppose there are N learning models with some predefined hyperparameters. Then the accuracy of each model is calculated and it will keep only half of the models (the better performing model). After that, offspring with hyperparameters similar to the best model can be generated in order to obtain again a population of N models. The accuracy of each model is again calculated and the cycle is repeated in the defined generations. In this way, only the best model will survive at the end of the process.

5) Particle swarm optimization.

Particle swarm optimization (PSO) algorithm is inspired by the phenomenon of foraging behavior of bird flocks. Unlike the genetic algorithm (GA), the PSO has a memory, and the optimal position and direction of the previous iterations are retained and used in the update of the particle swarm. In PSO, using a particle to simulate individual birds. All particles have a fitness value determined by the function being optimized, and particles have only two properties: velocity and position. Velocity represents the speed of movement and position represents the direction of movement. PSO is initialized as a group of random particles (random solutions) and then iterates to find the optimal solution. In each iteration, the particles update themselves by tracking two extremums. The first one is the optimal solution found by the particle itself, which is called the individual extremum; the other extremum is the optimal solution found by the whole population, called the global extremum. Alternatively, instead of using the whole population, only a part of them can be used as a neighbor of the particle, and then the extremum among all the neighbors is the local minima.

(d)Weight initialization

In deep neural networks, as the number of layers increases, it is extremely prone to gradient disappearance or gradient explosion in the process of gradient descent. Therefore, the initialization of the weights w is crucial. A good initialization of the weights cannot completely solve the problem of gradient disappearance and gradient explosion, but it is very helpful to deal with these two problems. And it is very beneficial to the performance and convergence of model. The current initialization methods are gaussian distribution initialization, uniform distribution initialization, xavier initialization, he initialization. The first two initialization methods are based on the fixed variance method, which controls the variance of the initialization value within the gaussian distribution or uniform distribution. However, the output value of the activation function will gradually converge to zero as the number of network layers deepens. To solve this problem, [45] proposed Xavier initialization, but this method only performs well for deep learning models with tanh activation functions. And the problem that activation values converging to zero as the number of network layers deepens will reappears in the face of ReLu, Leaky ReLu activation functions. Thus, [46] proposed an initialization method for the ReLu function named he initialization and proved its effectiveness.

(e)Transfer learning

The idea of transfer learning is applied to a new model by using the parameters of previous model which is well trained. Moreover, these two models are both to solve similar problems. If a DCNN model is retrained at the beginning of task. It is very likely to encounter some problems such as overfitting, parameters do not converge and so on. Transfer learning can solve these problems [47].

There are three approaches for transfer learning as follows. The first one is the full transfer approach, which uses all trained weight coefficients from similar data set as initialization weight parameters in the new model and completes the next training; The second is the partial transfer approach, which freezes a part of the pre-trained layers and
only trains the convolutional and fully-connected layers close to the output layer; The third is the top transfer approach, which uses all weight parameters before the prediction layer (top layer) to train the top layer and complete the next training. [48] used VGG16 as a model and successfully applied it to a high-resolution palmprint image recognition task with the help of transfer learning ideas. The three deep transfer learning approaches are schematically shown as follows.

![Deep transfer learning approach](image)

**Figure 5.** Deep transfer learning approach. (a) Complete transfer approach, (b) Partial transfer approach, (c) Top-level transfer approach.

### B. Structure Optimization

The number of parameters and computation of networks are mainly reflected in the convolution layer and the fully-connected layer. For the convolution layer, the convolution mode is optimized. The conventional convolution calculation is shown in Fig. 6. \( H \) and \( W \) denote the height and width of the input features. \( K \times K \) denotes the size of the convolution kernel, \( N \) is the number of input feature channels. \( C \) denotes the number of convolution kernels in the layer, and also is the number of channels of the output feature map by convolution, the computational volume of the conventional convolution \( F_{\text{normal}} \) can be obtained as follows:

\[
F_{\text{normal}} = H \times W \times K^2 \times N \times C
\]  

(a)Group convolution

Group convolution is a variant of standard convolution, which first originated from AlexNet. Since the GPU GTX 580 memory used in AlexNet training is too small to train the whole model, Alex used group convolution to divide the whole network into two groups and put them into one GPU card for training. In other words, the input features are convolved in groups by channel, and then the results obtained from group convolution are concatenated by channel to obtain the final output features. It has the idea of...
If divided into \( G \) groups, the required computation \( F_{\text{group}} \) is:

\[
F_{\text{group}} = \frac{(H \times W \times K^2 \times N \times C)}{G} \quad (13)
\]

**FIGURE 7.** Schematic diagram of the group convolution. (b) Depthwise separable convolution

The idea of depthwise separable convolution can be traced back to inception module, which is the essence of lightweight networks such as Xception and MobileNet. And the depthwise separable convolution can decompose the traditional convolution into a depthwise convolution and a pointwise convolution, as shown in Fig. 8, where (a)(b) represents the depthwise convolution and the pointwise convolution respectively.

**FIGURE 8.** Schematic diagram of the Depthwise separable convolution method. (a) Left: Depthwise convolution, (b) Right: Pointwise convolution.

In (a), the input feature map of \( N \times H \times W \) is divided into \( N \) groups, and then each group does \( K \times K \) convolution, which is computed as:

\[
F_{\text{depthwise}} = H \times W \times K^2 \times N \quad (14)
\]

It is \( 1/C \) of the computational effort of conventional convolution, and significantly reduces the computational effort by ignoring the convolution of channel dimensions. Depthwise is equivalent to collecting the spatial features of each channel individually.

In (b), \( C \) ordinary \( 1 \times 1 \) convolutions are done on the output feature map after depthwise convolution, which is mainly used to change the output channel feature dimension. The computational volume is:

\[
F_{\text{pointwise}} = H \times W \times C \times N \quad (15)
\]

The total calculation volume \( F_{\text{dsc}} \) is:

\[
F_{\text{dsc}} = H \times W \times N(K^2 + C) \quad (16)
\]

Depthwise separable convolution operates by convolving the input features separately by different channels to obtain the first layer of output features. Then using the first layer of output features as the input features for point convolution and using a \( 1 \times 1 \) convolution kernel for convolution. Next, feature information of different channels can be obtained. Finally, information feature fusion and further classification can be performed.

In the MixNet paper proposed by Google team in 2019, mixed depthwise separable convolution emerged [49]. The input feature map is grouped by channel, and two channel grouping strategies are proposed in the paper. They are average grouping and exponential decay grouping, in which each channel is convolved according to different convolution kernel sizes. Finally, they are concatenated. The schematic diagram of exponential decay grouping mixed depthwise separable convolution is as follows.
The feature map between convolution layers can be regarded as a mixture of high-frequency information and low-frequency information. In conventional convolution, both high-frequency and low-frequency information are stored in the same way, which causes storage redundancy and increases the computational cost for low-frequency information [50]. Octave convolution (OctConv) is proposed to address this problem by converting conventional feature storage into a lightweight storage method based on both low and high frequencies.

OctConv is represented as a single, generic, plug-and-play convolution unit that can directly replace the regular convolution without any adjustments to the network architecture. By simply replacing the regular convolution with OctConv, the accuracy of image recognition tasks can be continuously improved while reducing memory and computational costs [51]. The OctConv operation requires information transfer and exchange of low and high frequency information. Otherwise, the network performance is poor. As Fig.10 shows, \(X_L\) and \(X_H\) denotes the low and high frequency information of the input feature map respectively. And \(W_{L,L}\), \(W_{L,H}\), \(W_{H,L}\), \(W_{H,H}\) represents the weight matrix of feature extraction between low frequency and high frequency respectively. \(Y_L\) and \(Y_H\) are the low frequency and high frequency information of the output feature map respectively. The information transfer and exchange are shown in Fig.11.

Dilated convolution avoids the use of down-sampling and provides a larger receptive field with the same amount of computation. In addition, dilated convolution has an additional hyperparameter dilation rate, we will get different receptive fields and multi-scale information by setting different dilation rates. The effect of dilated convolution on the receptive field is shown in the Fig. 12.
FIGURE 12. The effect of dilated convolution on the receptive field. (a) Left: conventional convolutional receptive field, (b) Middle: dilation rate=2 receptive field, (c) Right: dilation rate=3 receptive field.

The gray part indicates the convolved receptive field. (a) is the regular convolution, equivalent to dilation rate=1, and the convolved receptive field size is 3. (b) is the dilated convolution, and the convolved receptive field size is 5, (c) is the dilated convolution, and the convolved receptive field size is 7.

(e) Inception module

FIGURE 13. Inception module structure schematic. (a) Top: InceptionV1, (b) Bottom: InceptionV2.

Fig. 13 shows the structure of Inception V1 and V2. Inception V3 follows the network structure of Inception V2, with improvements in optimization algorithms, regularization, etc. The basic Inception module in inception V4 still follows the structure of inception v2/v3, but the structure looks more concise and uniform. It uses more Inception modules, the experimental effect is also better. In addition, inception V4 combines the idea of residuals in ResNet to develop two versions of Inception-ResNetV1, V2, which improves the performance of the network to a certain extent.

(f) Global pooling layer

The number of parameters in the fully-connected layer of the neural network is very large. When the number of classification tasks is big, the number of parameters will be larger, and the computational cost will increase. The principle of global pooling layer is to pool the feature map of the last layer to form feature points, and these feature points will form the final feature vector, which will then be sent to the softmax layer for computing. The global pooling layer has no parameters, so it will not be involved in the computation in the neural network training [52]. ResNet introduced it to reduce the computational effort.
pooling layer can be divided into a global maximal pooling layer and a global average pooling layer. Both of them reduce the number of parameters to avoid overfitting on the one hand. On the other hand, this is more in line with the working structure of CNNs, where each feature information is associated with the output instead of the sub-information of the feature information being directly associated with the output.

V. LIGHTWEIGHT NETWORKS

Lightweight networks are networks with a smaller number of parameters and computational effort compared to ordinary CNNs by manual design or neural network structure search based on the idea of lightweight (mainly for the optimization of convolutional approach).

A. SqueezeNet

The core module of SqueezeNet [53] is the fire module, which borrows the idea of inception structure from GoogLeNet. It is divided into squeeze layer and expand layer. Squeeze layer is a convolutional layer with 1*1 convolutional kernel, which convolves the feature map of the previous layer to reduce the dimension of the feature map. The feature maps obtained from 1*1 and 3*3 are concatenated in the expand layer. SqueezeNet achieves the same correct rate as AlexNet on ImageNet with only 1/50th of the parameters of AlexNet. Furthermore, the deep compression technique is used, which consists of three means: cropping, quantization, and encoding. It can compress SqueezeNet to 0.5MB, which is 1/510th of AlexNet.

B. MobileNet

MobileNet [54] is an artificially designed lightweight network containing 3 versions, which is lightweight because it used depthwise separable convolution to replace conventional convolution. MobileNetV1 mainly uses the depthwise separable convolution method to reduce the computational effort and the accuracy is similar to the traditional convolution method. MobileNetV2 also follows the depthwise separable convolution method and introduces the inverse residual structure and linear bottleneck on the basis of V1. The schematic diagram of inverse residual structure is as follows.

![Schematic diagram of MobileNet](image)

FIGURE 14. Left: Residual structure ; Right: Inverse residual structure.

The input of the inverse residual structure first undergoes 1*1 convolution for channel expansion (increasing the number of convolution kernels to expand the number of channels). Then uses 3*3 depthwise convolution, and finally uses 1*1 pointwise convolution to compress the number of channels back. The whole process can be described as "dilation-convolution-compression" so as to obtain more channels for feature extraction. In addition, doing ReLu operations in low dimensions can easily cause loss of information. If ReLu is performed in high dimensions, the loss of information will be minimal. Another explanation is that when the high-dimensional information is transformed back to low-dimensional information, it is equivalent to do a feature compression, which will lose some information. And the lost information will be increased after sending it to ReLu again. Therefore, V2 replaces ReLu with a linear activation function can reduce the information loss.

MobileNetV3 follows the network and modules of V2 and introduces the SE module. SE module is an attention mechanism module, which is produced by SENet which won the ImageNet 2017 classification competition. SENet mainly learns the correlation between channels and filters the attention for channels. This can make the effective features weighted heavily and the weight of invalid features is small. Although this increases the number of parameters and computational effort, the performance is significantly improved.

![Schematic diagram of SE module](image)

FIGURE 15. Schematic diagram of SE module.

C. ShuffleNet

The main idea of ShuffleNet [55] is to improve the ResNet by using group convolution kernel and channel scattering. Group convolution can reduce the computation of traditional
convolution. Channel scattering can reorganize the feature map after group convolution and ensure that the information can be communicated between different groups. ShuffleNet is divided into two versions V1 and V2.

ShuffleNetV1 improved ResNet by replacing the dense 1*1 convolution with a 1*1 group convolution. And adding a channel scattering operation after the first 1*1 convolution. In addition, It not using the ReLu activation function after the 3*3 depthwise convolution. There is also a down-sampling version that uses stride=2 for the original input with 3*3 average pooling. It takes stride=2 at the 3*3 depthwise convolution to ensure that the two channels are of the same size. Then the obtained feature map is concatenated with the output instead of summing. The calculation amount and parameter size are reduced effectively. A comparison of three versions is shown below, the left figure shows the basic ResNet lightweight structure, the middle and right figures show two improvements for ResNet.

**FIGURE 16.** ShuffleNet Basic Module.

ShuffleNetV1 has 4 problems: 1) Extensive use of 1*1 group convolution operations in the basic unit; 2) Different number of channels for input and output features in the bottleneck layer of the residual network; 3) Excessive use of group convolution; 4) Excessive use of elementwise operations in shortcut. ShuffleNetV2 is an improvement of ShuffleNetV1 based on the above four problems. It introduces the channel separation. The following diagram shows the structure of ShuffleNetV2. As shown in the Fig. 17, ShuffleNetV2 replaces the 1*1 group convolution with regular convolution to balance the channel size of input and output. Channel split is to divide the features into two groups by channel. Finally, these two branches are concatenated to reduce the number of elementwise operations.

**FIGURE 17.** Two types of ShuffleNetV2 structure diagram.

D. Xception

Xception [56] network is an improved model proposed by Google based on inceptionV3. Its idea is to replace the operation of multi-size convolutional kernel feature response in the original inceptionV3 with depthwise separable convolution. The depthwise separable convolution applied in Xception network is the opposite of the ordinary depthwise separable convolution operation. The ordinary depthwise separable convolution is a 3*3 operation followed by a 1*1 operation, while in the Xception model, 1*1 operation is followed by a 3*3 operation. In addition, there is no relu layer is added in the middle of the network to ensure that the data is not corrupted. This operation can be reflected in the Fig 18.

**FIGURE 18.** Schematic diagram of Xception module operation.

It has been proven that Xception is comparable to inceptionV3 in terms of computation but better performance than inceptionV3 and better parameter utilization.

E. NasNet, MnasNet

NasNet [57] is a lightweight network based on the
Method of neural network architecture search. The main purpose of neural network architecture search is to use reinforcement learning to search for the most suitable hyperparameters of the basic units in the search space. Then stack these basic units searched to get a lightweight network. NASNet architecture features two AutoML designed layers named normal layer and reduction layer. It is no longer necessary to build the convolutional network architecture with human knowledge, and the hyperparameters are directly calculated by RNN. So AI automatic learning can be realized.

For the case that only complex units are searched in NasNet and the same units are stacked repeatedly, MnasNet [58] uses decomposed hierarchical search space, which focuses on the diversity of network layers and considers the model latency to get a model that is balanced between accuracy and latency. The search steps of MnasNet are: 1) Generating a network with RNN as the controller; 2) Training and validating with the target datasets to obtain the model accuracy; 3) Run the CNN model on the mobile side with the real inference latency. Then update the controller with the Proximal Policy Optimization (PPO) algorithm and repeat it until all parameters converge.

F. MixNet
MixNet [49] is mainly aimed at exploring a collection of convolutional kernels of different sizes. Because small convolutional kernels have a small receptive field and few parameters, but the accuracy is not high. Large convolutional kernels have a large receptive field, and the accuracy is slightly higher but the parameters are also relatively increased a lot. Thus, a new mixed depth separable convolution (MDConv) is proposed. The common depthwise separable convolution is to divide the input channels into M groups of size M, and then perform the convolution calculation with the same convolution kernel size for each group. However, MDConv in MixNet divides the input channels into C groups of size M, and then convolves each group with a different kernel size.

G. GhostNet
Redundancy in feature mapping is an important feature of those successful CNNs, but it has rarely been studied in neural architecture design. [59] paid attention to the redundant nature of feature maps and tried to replace the original convolution operation with some cheap operations to generate more feature maps. Ghost Module is a method of model compression, that is, while ensuring the accuracy of the network, reduce network parameters and computation, so as to improve the calculation speed and reduce latency. Ghost module can replace every convolutional layer in the existing CNNs.

![Regular convolutional layer with Ghost module](image)

**FIGURE 19.** Regular convolutional layer with Ghost module. (a) Top: the convolutional layer, (b) Bottom: the Ghost module.

H. Related properties of lightweight networks
The above lightweight networks are all optimized by convolution, and the following table lists the above lightweight networks and introduces their structures and other features.
TABLE 2. RELATED PROPERTIES OF LIGHTWEIGHT NETWORKS

| Lightweight Network | Type                | Convolution mode | Features                                      |
|---------------------|---------------------|------------------|-----------------------------------------------|
| SqueezeNet          | Artificial design   | Inception Module | Based on inception module, Model compression  |
| MobileNetV1         | Artificial design   | DWConv           | Small size, Low computational effort, Deployable on mobile devices |
| MobileNetV2         | Artificial design   | DWConv           | Follow V1 structure, Shortcut, Inverse residual module |
| MobileNetV3         | Artificial design   | DWConv           | SE module, Channel attention mechanism        |
| ShuffleNetV1        | Artificial design   | DWConv, GConv    | Change of channel order, low computational effort |
| ShuffleNetV2        | Artificial design   | DWConv, Conv     | Channel Split,1*1 regular convolution, Channel transformation |
| Xception            | Artificial design   | DWConv           | Following inceptionV3, Reversed convolution operation |
| GhostNet            | Model Compression   | Ghost Module     | Accepts feature redundancy, Linear operations |
| NasNet              | Network architecture search | DWConv, Dilated Conv | Global optimal search based on a priori knowledge, attempting feasible architectures |
| MnasNet             | Network architecture search | DWConv, Conv | More 5*5 deep convolution, layer hierarchical search space |
| MixNet              | Network architecture search | MDConv, GConv | Different convolution kernel sizes, Channel grouping |

VI. STUDY OF DEEPLY OPTIMIZED CNN ON WORKPIECES SURFACE DEFECT CLASSIFICATION

ECT MFL ACFM Ultrasonic SVM RF Bayes KNN CNN DBN SAE Training Structure Lightweight

![Diagram](image)

FIGURE 20. Development history of surface defect classification methods.

Conventional DCNNs has a huge amount of computation and parameters as its complicated structure. Then, a large amount of memory will be occupied in the training process, the training speed is also slow, and even the problem of non-convergence may be encountered. So, the performance of conventional DCNN model for defect detection and classification is poor. In addition, in the actual task of most workpieces surface defect classification, data collection usually is limited by acquisition conditions, human factors, environmental factors and other factors. This may result in a small number of data sets and low quality, and the training effect may not be satisfactory. The training of deep learning model is also limited by equipment conditions. Due to the huge amount of calculation and parameters, the training time is too long and the problem of non-convergence may be encountered. Moreover, at present, many industrial enterprises have certain requirements on the performance and mobile deployment of neural network, but the current high precision neural network often has a deep depth, requires a large memory footprint. This is not suitable for deployment in the mobile terminal. In view of the above problems, three solutions described in this paper including training optimization, structure optimization and lightweight can be adopted respectively. Therefore, using the above approaches of optimization to deeply optimize CNN has some practical significance for the defect classification task of workpieces.

Currently, deeply optimized CNN based on workpieces detection and classification tasks are gradually gaining popularity. Many researchers have also adopted relevant optimization methods for their own problems.
In view of the limited data available for training and verification of photovoltaic solar panels, [60] adopts the transfer learning method in training optimization in this paper, and further uses DCNN to identify and classify surface defects. [61] applied DCNN to the wafer surface defect classification task, and adopted data standardization, data augmentation technique and transfer learning in the training optimization presented in this paper to study their impact on calculation cost and test accuracy. [62] studied the optimal configuration of CNN based on the classification of track surface defects. In the aspects of data set, CNN structure and training, such as gradient descent optimization algorithm, hyperparameters, collocation between different numbers of convolution layers, pooling layers and fully connected layers, he conducted differential experiments and obtained the optimal configuration results of CNN. In [63], ResNet and transfer learning idea was utilized to classify image defects for insufficient data volume. [64] classified the defects of the core components of printing equipment. Firstly, an improved HOG eigenvalue algorithm was proposed, and then proposed a lightweight ResNet model to classify it, which has a certain accuracy improvement compared to the traditional SVM classification. In addition, the improved network reduces the number of parameters and computation of the model while ensuring the accuracy. It is practical. [65] proposed a very efficient CNN-based Mobile-Unet model to achieve end-to-end fabric defect segmentation, introducing a depthwise separable convolutional approach with high segmentation accuracy and classification speed. [66] designed a CNN-based defect classification model for silicon wafers named WDD-Net, which used the optimization of depthwise separable convolution and global average pooling to reduce parameters and computations. Multiple 1*1 standard convolution are used to increase the network depth. Finally, the author made an experimental comparison analysis with the models VGG-16 and MobileNet-v2, which converged fastly and achieved 99% recognition accuracy. [67] proposed a lightweight fully CNN model named LFCSSD by using lightweight ideas, mainly using the inceptionV3 module to reduce the computational effort. The accuracy on DAGM 2007 and KolektorSDD is 99.72% and 98.74% respectively. This model has a high efficiency and fast speed. [68] used depthwise&pointwise convolution layer, strided depthwise convolution layer, and upsample depthwise convolution layer to replace the standard convolution layer, pooling layer, and deconvolution layer, respectively. It has improved the performance of Fully Convolutional Networks and got the validation on the dataset of DAGM 2007. To solve the problem that the sample set of workpiece surface defects does not contain border label information and the sample annotation cost is high, [69] implemented a weakly supervised workpieces surface defect detection algorithm based on the improved NasNet. In addition, it replaced the backbone network of SSD with MobileNetv2 to improve the accuracy of small target detection. [70] proposed an improved YOLOv3 network detection method to the task of surface defect classification. Concretely, it used a MobileNet to replace Darknet-53 in YOLOv3's original network. Dilated convolution is also added to improve the network's ability to detect small target defects.

The above study of deeply optimized CNN on defect classification gives an inspiration to us. In the defect detection and classification task, it is advisable to use the above optimization method to deeply optimize CNN model when the adopted model encounters certain bottlenecks or difficulties. It may improve the detection speed and classification accuracy.

| Optimization methods          | References    | Effect                                      |
|-------------------------------|---------------|---------------------------------------------|
| Training optimization         | [60-63,72]    | Improved the speed; Network convergence is accelerated; |
| Structure optimization        | [65,66,68,70,71] | The number of computation and parameters is significantly reduced. |
| lightweight                    | [64,67,69,70,71] | Optimized memory space; Increases the possibility of mobile deployment. |

VIII. SUMMARY AND PROSPECTS

The workpieces is one form of presentation for the stable development of industrialization. The quality of workpieces directly affects the industrial progress and shows the degree and strength of a country's industrialization. Therefore, the task of classifying the defects of workpieces is of the utmost importance.

Currently, the use of DCNN on image recognition, defect classification, fault diagnosis and other fields is a hot research point. Numerous researchers have achieved excellent results in this field. However, many researchers have encountered certain bottlenecks such as the data acquisition conditions, the number of parameters and computation, economic capability, and the current development of intelligent devices et al. If we want to meet
the industrial requirements of high precision classification of workpiece defects without any degradation of the deep network performance, some optimization of DCNN is needed. Therefore, this paper reviews the optimization methods of DCNN in terms of data, algorithm and structure. These optimization methods can significantly reduce the number of parameters and computation, enhance the speed of operation and training, save memory space and reduce the equipment requirements. For instance, at the electrical operation site, surface defect detection of electric power tools can effectively prevent accidents and ensure the safety of workers. Therefore, the use of deeply optimized CNN to complete the task of surface defect classification of workpiece is a research hotspot and trend. It can ensure the development and progress of industry with stable steps.

REFERENCES

[1] Y. Wang, M. Liu and Z. Bao, "Deep learning neural network for power system fault diagnosis," 2016 35th Chinese Control Conference (CCC), 2016, pp. 6678-6683, doi: 10.1109/ChiCC.2016.7554408.

[2] X. Tao, W. Hou and D. Xu, “A survey of surface defect detection methods based on deep learning," Acta Automatica Sinica, May. 2021, vol. 47, no. 5, 1017–1034, DOI: 10.16383/j.aas.c190811.

[3] M. Haselmann and D. Gruber, “Supervised Machine Learning Based Surface Inspection by Synthesizing Artificial Defects," 2017 16th IEEE International Conference on Machine Learning and Applications (ICMLA), 2017, pp. 390-395, doi: 10.1109/ICMLA.2017.0-130.

[4] W. Liu, Z. Liu, H. Wang and Z. Han, "An Automated Defect Detection Approach for Catenary Rod-Insulator Textured Surfaces Using Unsupervised Learning," in IEEE Transactions on Instrumentation and Measurement, vol. 69, no. 10, pp. 8411-8423, Oct. 2020, doi: 10.1109/TIM.2020.2987503.

[5] M. Mayuravaani and S. Manivannan, “A Semi-Supervised Deep Learning Approach for the Classification of Steel Surface Defects," 2021 10th International Conference on Information and Automation for Sustainability (ICIAIS), 2021, pp. 179-184, doi: 10.1109/ICIAIS52020.2021.9606143.

[6] M. Cao, W. Zhang, W. Zeng and C. Chen, “Research on the device of differential excitation type eddy current testing for metal defect detection," 2013 Far East Forum on Nondestructive Evaluation/Testing: New Technology and Application, 2013, pp. 155-158, doi: 10.1109/FENDT.2013.663554.

[7] T. Rocha, H. Ramos, A. L. Ribeiro and D. Pasadas, "Sub-surface defect detection with motion induced eddy currents in aluminum," 2015 IEEE International Instrumentation and Measurement Technology Conference (I2MTC) Proceedings, 2015, pp. 930-934, doi: 10.1109/I2MTC.2015.7151339.

[8] W. Zeng, H. -t. Wang, G. -y. Tian, G. -x. Hu, X. -m. Yang and M. Wan, "Application of a non-contact laser ultrasonic imaging technique in surface defect detection," 2014 IEEE Far East Forum on Nondestructive Evaluation/Testing, 2014, pp. 30-33, doi: 10.1109/FENDT.2014.6928227.

[9] J. -P. Peng, J. -J. Fu, J. -M. Zhao, X. Zhang and H. Yin, "Dynamic Detection of Rail Surface Crack Based on ACFM," 2020 IEEE Far East NDT New Technology & Application Forum (FENDT), 2020, pp. 111-114, doi: 10.1109/FENDT50467.2020.9337521.

[10] D. Cerniglia, B. B. Djordjevic and V. Nigrelli, "Quantitative subsurface defect detection in composite materials using a non-contact ultrasonic system," 2001 IEEE Ultrasonics Symposium. Proceedings, An International Symposium (Cat. No.01CH37263), 2001, pp. 751-754 vol.1, doi: 10.1109/ULTSYM.2001.991833.

[11] M. Pan, D. Zhou, X. Chang, "Analysis of surface defect detection characteristics of new pulsed MagneticFlux Leakage detection method," Sensors and Microsystems, 2017, vol. 36, no. 12, 32-35, DOI: 10.13873/j.1000-9787(2017)12-0032-04.

[12] Z. Kang, F. Luo and M. Pan, et al. "Alternating Magnetic-Flux Leakage detection system for steel pipe surface defect detection," Nondestructive testing, 2006, vol. 4, 189-191.

[13] B. -. g. Kwon and D. -. j. Kang, "Fast defect detection algorithm on the variety surface with random forest using GPUs," 2011 11th International Conference on Control, Automation and Systems, 2011, pp. 1135-1136.

[14] D. J. Pasadas, P. Baskaran, C. G. Ramos and A. L. Ribeiro, "Detection and Classification of Defects Using ECT and Multi-Level SVM Model," in IEEE Sensors Journal, vol. 20, no. 5, pp. 2329-2338, 1 March 2020, doi: 10.1109/JSEN.2019.2951302.

[15] Shannugamani R. , Sadique M , Ramamoorthy B . Detection and classification of surface defects of gun barrels using computer vision and machine learning. Measurement, 2014, 60:222-230.

[16] S.H. Chen, S. Zhao, J. Ma, Y. Zhang, L. Guo. "Solder joint defect detection method of SVM multi-classification PCB based on multi-feature", Laser Journal, Jun. 2019, vol. 40, no. 6, pp. 21-26. DOI: 10.14016/j.cnki.gzz.2019.06.021.

[17] Morimoto M. - Heterogeneous Defect Detection of Metal Surface using SVM. " IEICE Tech. Rep., vol. 109, no. 56, SIP2009-17, pp. 87-92, May 2009.

[18] X. Chen, "The study of steel surface defects recognition based on PCA-SVM methods." Mechanical Design and Manufacturing Engineering, Jun, 2017, vol. 46, no. 6, pp. 100-103, DOI: 10.3969/j.issn.2095-509X.2017.06.023.

[19] Sharma M . , Kaur G . "Integrated approach for defect detection in ceramic tiles," International Journal of Computers & Technology, 2012, vol. 3, no. 2, 463-468.

[20] C. Chen , H.D. Zhang, Y.X. Wu. "A study on the classification of terminal defects based on random forest classification," Journal of Bengbu College, Mar.2021, vol. 10, no. 2, 97-102. DOI : 10.13900/j.cnki.jbc.2021.02.020.

[21] H. Liu J.N. Chen, L. Zhang. “Machine learning based approach for tool surface defect classification," Computer Measurement & Control, Jun. 2021, vol. 29, no. 6, pp. 64-68, DOI: 10.16526/j.cnki.11-4762/p.2021.06.014.

[22] C.L. Li, J.B. Xiong, N.Q. Su, et al., “Application Review of Deep Learning in Fault Diagnosis. " Machine Tool & Hydraulics,2020, vol. 48, no. 13,pp.174-184.DOI: 10.3969/j.issn.1001-3881.2020.13.039.

[23] S. Cheon, H. Lee, C. O. Kim and S. H. Lee, "Convolutional Neural Network for Wafer Surface Defect Classification and the Detection of Unknown Defect Class," in IEEE Transactions on Semiconductor Manufacturing, vol. 32, no. 2, pp. 163-170, May 2019, doi: 10.1109/TSM.2019.2902657.

[24] J. Karangwa, L. Kong, T. You and J. Zheng, “Automated Surface Defects Detection on Mirrorlike Materials by using Faster R-CNN," 2020 7th International Conference on Information Science and Control Engineering (ICISCE), 2020, pp. 2288-2294, doi: 10.1109/ICISCE50968.2020.00341.
[25] Z. Xie, et al, “Surface defect classification and recognition method of sheet metal parts based on convolutional neural network,” Computer Measurement & Control, 2020, vol. 28, no. 6, pp. 187-190, DOI:10.16552/j.cnki.11-4762/tp.2020.06.038.

[26] X.B. Wang, J. Li, M.H. Yao, et al., “Solar Cells Surface Defects Detection Based on Deep Learning,” Pattern Recognition and Artificial Intelligence, Jun. 2014, vol. 27, no. 6, pp.517-523. DOI:10.3969/j.issn.1003-6059.2014.06.006.

[27] Shuang M., Yudan W., Guojun W., “Automatic Fabric Defect Detection with a Multi-Scale Convolutional Denoising Autoencoder Network Model,” Sensors, Apr. 2018, vol. 18, no. 4,1064. DOI: 10.3390/s18041064.

[28] B. Tang, C. Dai, W. Huang, “Surface Defect Detection of Steel Plate with Labeling Based on Convolutional Neural network,” Manufacturing Automation, vol. 42, no. 9, Sep, 2020, 34-40. DOI:CNKI:SUN:Jxgy.0.2020-09-009.

[29] X.FLi, D. Li, Y.W. Wang, “Multi-scale edge detection algorithm for medical ultrasound images based on deep residual networks,” Journal of Jinlin University (Science Edition), 2021, vol. 59, no. 4,900-908. DOI: 10.13431/cnki.jsxbxb.2020169.

[30] Y. Wu, Y.Z. Yang, et al., “Surface defect detection method of steel plate based on Faster R-CNN,” Journal of Donghua University (Natural Science), 2021, vol. 47, no. 3, 84-89. DOI: 10.1088/j.cnki.dhuzb.20200159.

[31] Y. Lecun, L. Bottou, Y. Bengio and P. Haffner, “Gradient-based learning applied to document recognition,” in Proceedings of the IEEE, vol. 86, no. 11, pp.2278-2324, Nov. 1998, doi: 10.1109/5.726791.

[32] Alex Krizhevsky, I Sutskever, G Hinton, “ImageNet Classification with Deep Convolutional Neural Networks,” NIPS. Curran Associates Inc, 2012.

[33] Simonyan K., Zisserman A. “Very Deep Convolutional Networks for Large-Scale Image Recognition,” Computer Science, Sep. 2014, arXiv:1409.1556.

[34] C. Szegedy et al., “Going deeper with convolutions,” 2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2015, pp. 1-9. DOI: 10.1109/CVPR.2015.7298594.

[35] K. He, X. Zhang, S. Ren and J. Sun, “Deep Residual Learning for Image Recognition,” 2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2016, pp. 770-778. DOI: 10.1109/CVPR.2016.90.

[36] G. Huang, Z. Liu, L. Van Der Maaten and K. Q. Weinberger, "Densely Connected Convolutional Networks," 2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2017, pp. 2261-2269, DOI: 10.1109/CVPR.2017.243.

[37] W. Tang, Y. Zheng, et al. “Roller Bearing Fault Diagnosis Based on Stacked Auto-encoder with Dynamic Learning Rate,” Computer Engineering and Applications, 2020, vol. 56, no. 20, pp.264-269. DOI:10.3778/j.issn.1002-8331.1909-0243.

[38] Y. Bao, T. Lu, et al., “A deep forged video classification method based on i_ResNet34 model and data enhancement,” Computer Science, Jul. 2021, vol. 48, no. 7, 77-85.

[39] Bottou L. “Stochastic Gradient Descent Tricks,” Lecture Notes in Computer Science, 2012, vol. 7700 Springer,Berlin,Heidelberg. https://doi.org/10.1007/978-3-642-35289-8_25.

[40] N. Qian. “On the momentum term in gradient descent learning algorithms,” Neural Networks, vol. 12, no. 1, doi:10.1016/S0893-6080(98)00116-6.

[41] Duchi J., Hazan E., Singer Y., et al., “Adaptive Subgradient Methods for Online Learning and Stochastic Optimization,” Journal of Machine Learning Research, 2011, 12: 2121-2159.

[42] Zeiler D. “ADADELTA: An Adaptive Learning Rate Method,” arXiv:1212.5701, Dec. 2012.

[43] Dauphin Y N, Vries H, Bengio Y. “RMSProp and equilibrated adaptive learning rates for non-convex optimization,” arXiv:1502.04390v2, 2015.

[44] Diederik P. Kingma & Jimmy Ba. “Adam: A Method for Stochastic Optimization,” CoRR(2014).

[45] Glorot X, Bengio Y. “Understanding the difficulty of training deep feedforward neural networks,” Journal of Machine Learning Research, 2010, 9: 249-256.

[46] K. He, X. Zhang, S. Ren and J. Sun, “Delving Deep into Rectifiers: Surpassing Human-Level Performance on ImageNet Classification,” 2015 IEEE International Conference on Computer Vision (ICCV), 2015, pp. 1026-1034, DOI: 10.1109/ICCV.2015.123.

[47] W.Q. Weng, “Image recognition based on improved convolutional Neural Network,” Huazhong University of Science and Technology, May, 2019. DOI : 10.27157/d.cnki.gzhuzu.2019.000661.

[48] B. Wu, Y. Xing, et al., “High-resolution palm print image recognition based on VGG16 and transfer learning,” Intelligent Computer and Applications, 2021, vol. 11, no. 5, 37-42.

[49] M.X. Tan, Q. V. Le, “MixConv: Mixed Depthwise Convolutional Kernels,” arXiv :1907.09595, BMVC, Jul. 2019.

[50] J. Wang, S. Feng, Y. Cheng. “A Review of Research on Lightweight Neural Network Structures for Deep Learning,” Computer Engineering, Apr. 2021, vol. 47, no. 8, pp. 1-13. DOI:10.19678/j.issn.1000-3428.060931.

[51] Y. Chen et al., “Drop an Octave: Reducing Spatial Redundancy in Convolutional Neural Networks With Octave Convolution,” 2019 IEEE/CVF International Conference on Computer Vision (ICCV), 2019, pp. 3434-3443, doi: 10.1109/ICCV.2019.00353.

[52] M. Lin, Q. Chen, S. Yan, “Network In Network,” Computer Science, Dec. 2013, arXiv:1312.4400.

[53] F. N. Landola, S. Han, M. W. Moskewicz, et al. SqueezeNet: AlexNet-level accuracy with 50x fewer parameters and <0.5MB model size. 2016. arXiv:1602.07360.

[54] A. G. Howard, M. Zhu, B. Chen, et al. MobileNets: Efficient Convolutional Neural Networks for Mobile Vision Applications, 2017. arXiv:1704.04861.

[55] X. Zhang, X. Zhou, M. Lin, J. Sun. ShuffleNet: An Extremely Efficient Convolutional Neural Network for Mobile Devices.//IEEE/CVF Conference on Computer Vision and Pattern Recognition.2018.

[56] F. Chollet, "Xception: Deep Learning with Depthwise Separable Convolutions," 2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2017, pp. 1800-1807, doi: 10.1109/CVPR.2017.195.

[57] B. Zoph, V. Vasudevan, J. Shlens, Q. V. Le, “Learning Transferable Architectures for Scalable Image Recognition,” //IEEE/CVF Conference on Computer Vision and Pattern Recognition.2018.

[58] M. Tan, B. Chen, R. Pang, et al. “MnasNet: Platform-Aware Neural Architecture Search for Mobile," Jul, 2018. arXiv:1807.11626.

[59] K. Han, Y. Wang, Q. Tian, J. Guo, C. Xu and C. Xu, "GhostNet: More Features From Cheap Operations," 2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2020, pp. 1577-1586. DOI:10.1109/CVPR42600.2020.00165.

[60] I. Ziyout and A. Oatawne, "Detection of PV Solar Panel Surface Defects using Transfer Learning of the Deep Convolutional Neural Networks," 2020 Advances in Science and Engineering Technology International
His research interests include blind source separation, data mining and intelligent information processing.

Quanyang Li received the B.S. degrees in communication engineering from The Engineering&Technical College of Chengdu University of Technology, Leshan, China, in 2020. He was admitted to Sichuan University of Science and Engineering in 2020, where he is currently studying the M.S. degree. His main research direction is computer vision.

Zhongqiang Luo received the B.S. and M.S. degrees in communication engineering and pattern recognition and intelligent systems from Sichuan University of Science and Engineering, Zigong, China, in 2009 and 2012, respectively. He received the Ph.D. degree in communication and information systems from University of Electronic Science and Technology of China (UESTC), in 2016. Since 2017, he has been with the Sichuan University of Science and Engineering, where he is currently an associate professor. From December 2018-December 2019, he was a visiting scholar with Department of Computer Science and Electrical Engineering of University of Maryland Baltimore County (UMBC). His research interests include information fusion, blind source separation, signal processing for wireless communication system and intelligent signal processing.

Hongbo Chen received M.S. degrees in Electrical engineering and automation from Wuhan University, Wuhan, China, in 2012. As the main researcher, he participated in "Research and application of non-electric quantity testing technology in condition based maintenance of power transmission and transformation equipment", "Detection and evaluation of insulation defects and pollution ultraviolet rays of high voltage power transmission and transformation equipment", "Pollution investigation, measurement and analysis research along the Sichuan section of Xiangjiaba-Shanghai and Jinping-Sunan ± 800kV UHV DC transmission lines" and other key scientific and technological projects of State Grid Corporation of China.

Chengjie Li received the B.Sc. degree in Shandong normal University, Qufu (Confucius’s hometown), China, 2004, the M.Sc. degree in computer software and theory in Xihua University, Chengdu, China, 2009. He received the Ph.D. degree in Communication and Information System in University of Electronic Science and Technology of China (UESTC), Chengdu, China, in 2017. Since 2017, he has been with the Southwest Minzu University, where he is currently a lecturer. His research interests include blind source separation, data mining and intelligent information processing.