Academic Journal of Science and Technology
ISSN: 2771-3032 | Vol. 9, No. 1, 2024

Inception meets Swin Transformer: A Novel Approach for Metal Defect Recognition

Donglin Tang*, Yunliang Zhao

School of Electrical and Mechanical Engineering, Southwest Petroleum University, Chengdu 610500, China
* Corresponding author

Abstract: The detection of metal defects with high precision and efficiency is a significant challenge in modern industry. Existing machine learning methods for recognizing common metal surface defects heavily rely on expert knowledge for manual feature extraction. Conventional deep learning methods face challenges in capturing global feature information from defect images or defect detection signals.To address this issue, we proposed a metal defect recognition method based on an Inception-fused Swin Transformer model. The method combines the adaptive local feature extraction capability of the Inception structure with the advantage of the Swin Transformer in capturing global feature information from defect signals. Additionally, it utilizes the Channel-Coordinate Attention module (CoordAttention) to highlight important feature channels. Experimental results demonstrate the effectiveness of the proposed method on the Ultrasonic Defect Grayscale Image dataset (ULFSL-DET) and the publicly available Image-based Metal Defect dataset (NEU-CLS), achieving recognition accuracies of 98.1% and 99.8%, respectively. The method exhibits high effectiveness in recognizing metal defect signals in grayscale images, and it demonstrates strong generality for image-based metal defect recognition.

Keywords: Metal Defect, Inception, Swin Transformer, Signals in grayscale images.

1. Introduction

Surface defects of industrial products, such as those found in machinery, chemical, and petroleum fields, can significantly impact the quality, safety, and availability of products. Therefore, the detection and identification of metal surface defects have become a prominent research area in modern industry[1, 2]. Traditional non-destructive testing (NDT) methods, including ultrasonic, eddy current, and magnetic particle testing, have been widely employed in metal defect detection[3]. However, these methods are often limited by slow detection speeds and high dependence on expert knowledge, which no longer meet the

requirements of modern industry. With the advancement of pattern recognition technology (PRT), the combination of NDT and PRT has led to significant improvements in defect detection efficiency, enabling automatic defect detection and evaluation.

Current methods for metal defect detection and classification based on pattern recognition technology (PRT) can be divided into two categories: machine learning (ML) and deep learning (DL) methods. ML methods are effective in identifying and classifying complex data, but they require expert knowledge to manually design and extract input features. Additionally, they are only useful for a single class of objects. For example, Mensah et al. used artificial neural networks (ANN) and nonlinear regression models to predict pipeline failure pressure by evaluating corrosion clusters in pipelines[4]. Lin et al. achieved real-time defect detection for metal additive manufacturing parts using laser- induced breakdown spectroscopy (LIBS) combined with plain Bayesian, K-nearest neighbor, decision trees, and random forests[5]. Gaja et al. reduced the risk of deposited material failure by identifying laser metal deposition defects using logistic regression models and ANN[6]. However, these methods all require manual feature extraction, which is time-consuming, labor-intensive, and lacks accuracy and versatility.

Convolutional neural network (CNN) has been widely used in deep learning due to its ability for automatic feature extraction and end-to-end learning. However, CNN methods such as VGG16[7], GoogLeNet[8] and ResNet34[9] have significant drawbacks, including large and complex parameters, high computational complexity, high training cost, and inability to capture global feature information. For instance, Balcioglu et al. [10] utilized a deep convolutional neural network (DCNN) to detect and recognize surface defects in metal gears, Meng Tian et al. [11] combined ResNet with an eddy current testing technique to evaluate the depth of metal surface defects effectively, and He et al. [12] employed a multi-scale convolutional neural network to classify hot rolled steel defects. These studies chose deeper convolutional neural networks to accomplish the defect recognition task. However, the networks require too many samples for model parameter training due to the excessive number of parameters and computation. Moreover, the complexity of these methods causes overfitting phenomena.

The Transformer algorithm has emerged as a powerful tool in natural language processing (NLP) and has recently garnered significant attention in image recognition[13]. Unlike CNN, which can struggle to capture global feature information, the Transformer algorithm excels at this task. For example, Dosovitskiy et al. applied Vision Transformer to image recognition tasks and outperformed CNN in accuracy on massive datasets[14]. Zhou[15] and Touvron et al. continued to improve the accuracy of the Vision Transformer model in image classification by deepening the network depth[16]. To reduce the number of parameters in the Vision Transformer model, Wang et al. reduced computational complexity by improving the original pyramidal vision transformer[17]. Liu et al. proposed Swin Transformer based on the Vision Transformer for self-attentive computation with moving windows. The Swin Transformer achieves global self-attentive computation by computing local windows and cross-window connections, while making the model

complexity linear to the image size[18]. Despite its advantages, the Transformer model still suffers from several limitations, such as vast and complex parameters, training difficulties, an excessive amount of data required, and an inability to capture local information.

In addressing the challenges of heavy dependence on expert feature extraction, low intelligence in defect detection, low accuracy, and the lack of global attention capabilities in traditional convolutional neural networks for image-based metal defect recognition, this paper proposes a metal defect recognition method based on an Inception-fused Swin Transformer model. Focusing on defects on steel surfaces in industrial components, we employ wavelet transform to denoise ultrasonic echo signals from steel surface defects. After transforming the signals into grayscale images, the Inception model automatically captures the necessary feature information for defect recognition. The Swin Transformer, with lower model complexity compared to Transformer models, captures global feature information. Simultaneously, the introduced Channel-Coordinate Attention module (CoordAttention) emphasizes crucial feature channels[19], achieving effective identification and classification of metal defects.

The remainder of this paper is organized as follows. In the next section, we introduce the theoretical knowledge related to the proposed method, including Convolutional Neural Network (CNN), Swin Transformer, and channel attention mechanism. The third section presents the structure, parameters, and innovations of the proposed model. In the fourth section, we present the datasets, data preprocessing process, experimental parameters, and platform configuration. The fifth section exhibits the experimental evaluation and analysis of the proposed method on two datasets. Finally, we summarize the analysis and experiments in the last section.

2. Background Theory

2.1. GoogLeNet

CNNs have seen significant development in recent years, with variants such as LeNet, AlexNet, VGG, and other classical neural networks achieving high accuracy in image classification tasks. These networks deepen the network depth to improve non-linear expression ability and better fit features. However, as the number of network layers increases, the model parameters become large and may lead to overfitting. To address this issue, the Inception structure is used in GoogLeNet to fuse feature information of different scales. A 1x1 convolution kernel is applied to reduce the dimensionality. These approaches significantly reduce the model parameters and make the network deeper and wider. The Inception model comprises three convolutional kernels of different sizes and a pooling operation, and the results of the four parts are merged into a channel. Multiple Inception models are cascaded in series to form GoogLeNet.

2.2. Swin Transformer model

With the success of the Vision Transformer (Vit) model in vision tasks that can compete with CNN, the Transformer has become a promising research direction in the field of image processing. Among various Transformer-based models, Swin Transformer has drawn considerable attention by keeping the advantages of Vit while reducing model complexity. This reduction is mainly achieved by utilizing the Swin Transformer Block, which consists of two subunits.

Firstly, layer normalization (LN) is applied to input feature layers to coordinate the feature data distribution. Then, it passes to the self-attentive mechanism module. Finally, LN and two RELU nonlinear mapping (MLP) operations are performed. The two subunit self-attentive mechanisms use Window MSA (W-MSA) and Shifted Window MSA (SW-MSA) modules, respectively.

- Window MSA

The Window Multi-Head Self-Attention mechanism (W-MSA) used in Swin Transformer divides the input feature layers into windows of fixed size to reduce the computational complexity while achieving the same effect as the original Multi-Head Self-Attention mechanism in Vision Transformer. The attention calculation is performed within each window separately, and the results are merged to obtain the final output. The attention computation process could be described as follows.

$$Q = XW^q \quad K = XW^k \quad V = XW^v \tag{1}$$

$$Self\ \ Attention(Q,K,V) = Softmax(\frac{QK^T}{\sqrt{d}})V \tag{2}$$

Where is the input feature with the fixed window size, Q, K , V are the query matrix, matching matrix, and value matrix obtained by multiplying the trainable weight matrices W^q , W^k and W^v with the input X , respectively, $Softmax$ is the normalized exponential activation function, and d value is the query matrix dimension.

The workflow of the multi-head attention mechanism is shown in Figure 1(a). Firstly, the input image X is mapped to groups of Q, K ,and V by a linear transformation of n different trainable weights W^q , W^k , W^v matrices. Then the n different Z matrices obtained after $Self\ \ Attention$ calculation for each group of Q, K ,and V matrices are concatenated together. Finally, the concatenated Z matrix is multiplied by the weight matrix W^O to get the final matrix Z containing all the attention heads so that the model can focus on different positions in different subspaces.

- Shifted Window MSA

Shifted Window MSA achieves global attention capability by cross-connecting the individual windows. The procedure is shown in Figure 2(b). Firstly, the window dividing box is moved to the lower right corner by the distance of [window size/2] to connect different windows. Then, the isolated patches (A, B, C) are moved to windows 1, 2, 3 with incomplete patches. Finally, the correlation between non-adjacent patches in the original feature map is eliminated by masking operations in windows 1, 2, 3. After completing the window association, a Window MSA calculation is performed.

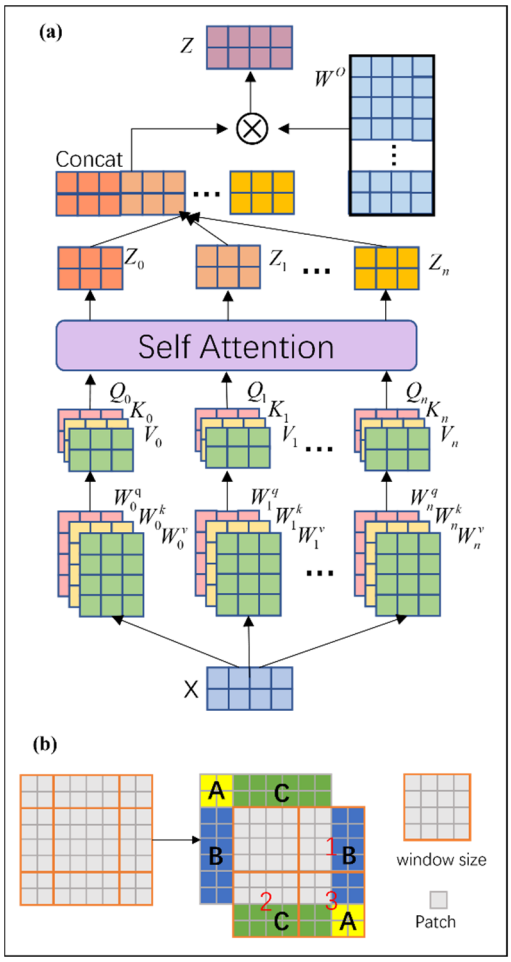

Figure 1. (a) Multi head Self Attention Mechanism. (b)Shifted Window Processing Method.

2.3. CoordAttention model

The channel attention mechanism is used to optimize the compressed channel weight values with feature learning to highlight important feature channels and suppress redundant channels. In this paper, the CoordAttention module is used, which retains the critical channel feature location information. Unlike other commonly used channel attention mechanisms such as SE[20] and CBAM[21], the CoordAttention module performs the following steps:

Firstly, the global pooling kernel is decomposed into two individual kernels: (H, 1) and (1, W). Pooling operation is carried out for every feature channel horizontally and vertically. Therefore, feature information compression and location encoding are achieved.

$$z_c^h(h) = \frac{1}{W} \sum_{0 \le i < \mathrm{w}} x_c(h, \ i) \tag{3}$$

$$z_c^w(w) = \frac{1}{H} \sum_{0 \le j < h} x_c(j, \ \mathrm{w}) \tag{4}$$

Where $z_c^h(h)$, $z_c^w(w)$, W , and H denote the output, channel width, and height of the c[th] channel at height h and width w, respectively.

Then, the joint channel feature map formed by joining $z_c^h(h)$, $z_c^w(w)$ is subjected to a 1x1 convolution, and a nonlinear activation operation.

$$f = \delta(F_1([z^h, \ z^{\mathrm{w}}])) \tag{5}$$

Where $[*,*]$ denotes the spatial feature layer stacking operation, F_1 is the convolution operation, δ is the nonlinear activation operation, $f \in R^{C/r \times (H+W)}$ is the compressed stacked feature channel in the horizontal and vertical directions, and r is the 1x1 feature convolution layer change ratio.

Finally, f is divided into two separate feature layers $f^h \in R^{C/r \times H}$ and $f^w \in R^{C/r \times W}$, then f^h and f^w are transformed into the original input feature channel layers using two 1x1 convolutions.

$$g^h = \sigma(F_h(f^h)) \tag{6}$$

$$g^w = \sigma(F_w(f^w)) \tag{7}$$

Where σ denotes the sigmoid activation function, g^h and g^w are the attention weights of the feature channels in the horizontal and vertical directions, respectively. Finally, the output of the CoordAttention module can be expressed as:

$$y_c(i, j) = x_c(i, j) \times g_c^h(i) \times g_c^w(j) \tag{8}$$

3. Methods

3.1. Overview of ICST

The framework of ICST is shown in Figure 2, and its specific steps are divided into 3 stages.

Stage 1: the input image features are learned at different scales using Inception1, and Swin Transformer is applied to capture the global feature correlation of the input image, and then the global image information is obtained.

Stage 2: Firstly, deeper feature extraction of the upper layer feature map is carried out by Inception2. The task of enriching the number of feature layers and downsampling is completed. Then, the CoordAttention module is used to highlight essential feature channels. At the same time, the output feature layer of the CoordAttention module is processed by batch normalization (BN) for batch normalization to ensure better delivery of network parameter gradient optimization. Finally, Swin Transformer is used to proceed learning global information.

Stage 3: After stage 2 is repeated three times, defect detection and recognition are achieved by applying a fully connected layer after an average pooling layer.

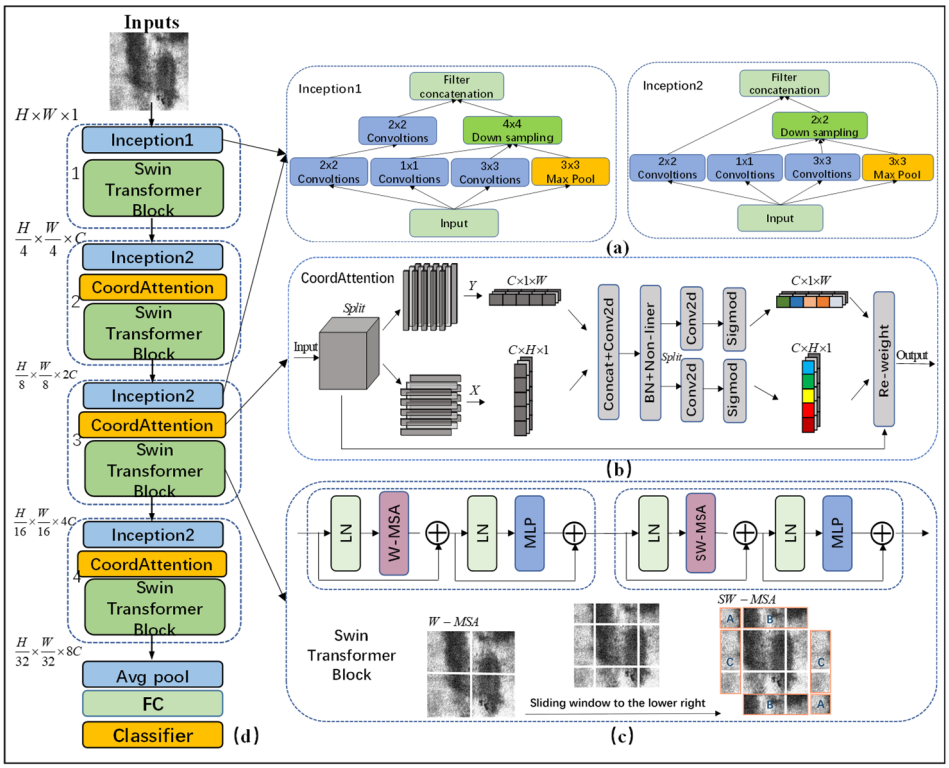

Figure 2. ICST structure diagram. (a)Inception. (b)CoordAttention. (c)Swim Transformer Block

In the ICST, four Inception modules are distributed in three stages to change the feature channel size and number, while other operations do not change the channel size and number. The specific output size of each layer and Inception module parameters are shown in Table 1. In the table, H, W and C indicate the image's height, width and number of channels, respectively. $((1\times1)+(4\times4),3C/8)$ in the column of Inception module parameters indicates that the 1x1 size kernel is used for convolution first by step 1, and then the 4x4 size kernel is used for downsampling mapping to avoid the loss of image information by pooling. $3C/8$ is the number of convolutional kernels, and the value of C in this paper is taken as 24.

Table 1. Specific Parameters of Network Layers

Steps	Output Size	Inception module parameters
Input	$H\times W\times 1$	—
stage1	$\frac{H}{4}\times\frac{W}{4}\times C$	$\begin{bmatrix}(1\times1)+(4\times4),\frac{3}{8}C\\(2\times2)+(2\times2),\frac{2}{8}C\\(3\times3)+(4\times4),\frac{2}{8}C\\(3\times3\max pool)+(4\times4),\frac{1}{8}C\end{bmatrix}$
Stage2	$\frac{H}{8}\times\frac{W}{8}\times 2C$	$\begin{bmatrix}(1\times1)+(2\times2),\frac{3}{8}\times2C\\(2\times2),\frac{2}{8}\times2C\\(3\times3)+(2\times2),\frac{2}{8}\times2C\\(3\times3\max pool)+(2\times2),\frac{1}{8}\times2C\end{bmatrix}$
Stage3	$\frac{H}{16}\times\frac{W}{16}\times 4C$	$\begin{bmatrix}(1\times1)+(2\times2),\frac{3}{8}\times4C\\(2\times2),\frac{2}{8}\times4C\\(3\times3)+(2\times2),\frac{2}{8}\times4C\\(3\times3\max pool)+(2\times2),\frac{1}{8}\times4C\end{bmatrix}$
Stage4	$\frac{H}{32}\times\frac{W}{32}\times 8C$	$\begin{bmatrix}(1\times1)+(2\times2),\frac{3}{8}\times8C\\(2\times2),\frac{2}{8}\times8C\\(3\times3)+(2\times2),\frac{2}{8}\times8C\\(3\times3\max pool)+(2\times2),\frac{1}{8}\times8C\end{bmatrix}$
Avg pool	$1\times 8C$	—
FC	number classes	—

3.2. Network Innovation

Firstly, The proposed ICST model adopts a multi-scale convolutional kernel approach to learn local features of images, which allows the network to autonomously select the required kernel size and complete the task of down-sampling during layer-by-layer propagation, while also reducing computational complexity.

In addition, to improve recognition accuracy for time series signal recognition, the Swin Transformer Block is introduced to establish global feature relationships and expand the image perception field.

Finally, to address the redundant feature layers introduced by the Inception module, the CoordAttention module is used to highlight important feature channels while retaining precise location information to preserve the temporal order of time series defect sample information.

These design choices enable the model to incorporate a priori knowledge of feature scale, translation invariance, and feature localization, and reduce the training difficulty and overfitting caused by a large number of global information parameters, resulting in improved model performance.

4. Datasets Description and Experimental Configuration

4.1. Experimental dataset

In order to verify the effectiveness of ICST for the recognition of metal defect ultrasonic signal grayscale images, a self-built ultrasonic detection defect dataset (ULFSL-DET) was constructed. At the same time, the public defect dataset (NEU-CLS) on steel surface was applied to verify the generality of the recognition of metal defects.

ULFSL-DET contains 2mm, 5mm, 8mm of rectangular, circular, elliptical and irregular shape artificial steel plate ultrasonic echo surface defect signals. There are 200 samples in each type, forming a total of 600-sample dataset.

In the ultrasonic detection experiment, a self-receiving straight probe with a center frequency of 2.5MHz and 20mm diameter was employed. The experiment was conducted using a digital ultrasonic flaw detector to collect the echo signals reflected by the defects. Three artificial defect plates were coated with ultrasonic coupling agent, and the probe was moved on them to locate the signal with the largest echo peak, which was observed on the oscilloscope. Subsequently, the echo signal data was saved in the ultrasonic detector as a CSV file. The experimental platform and defective samples are illustrated in Figure 3(a).

To improve the accuracy of signal recognition, pre-processing of the echo signal obtained from the ultrasonic experiment is required due to the presence of noise and redundant information. In this paper, the pre-processing processes are as follows.

Firstly, wavelet packet decomposition is applied to decompose the signal with Coif2 as the decomposition wavelet basis function. The resulting decomposed signal is then denoised using the soft threshold criterion.

Secondly, to meet the input requirements of ICST, the filtered ultrasonic defect detection signal needs to be converted into a two-dimensional grayscale image of 128x128 pixels. The conversion is processed by normalizing the voltage value of the ultrasonic echo signal. The normalized formula is as follows. Due to the different peaks of the three depth defect echo signals, $x_{\max}$ and $x_{\min}$ are taken as fixed values 13 and -2.

$$x^* = \frac{x - x_{\min}}{x_{\max} - x_{\min}} \tag{9}$$

Finally, after the voltage value is multiplied by 255, we convert it into a matrix with a width of 128. Part of the experimental echo signal and the pre-processing process are shown in Figure 3(b).

The NEU-CLS dataset[22], developed by Northeastern University, comprises of 1800 images of typical surface defects on hot rolled steel strips, with 300 images for each of the six defect types: cracking (Cr), inclusions (In), plaque (Pa), pitting (Ps), oxide (Rs), and scratches (Sc). Sample images of these defects are shown in Figure 3(c).

4.2. Experimental data configuration and parameter settings

In this paper, the experimental platform is Pycharm. The experimental code is developed and implemented based on the deep learning framework Pytorch. The hardware devices are AMD EPYC 7R13 48-Core CPU@2.65GHz, NVIDIA GeForce RTX 3060 (12GB) GPU.

As the two datasets contain limited data, the cross-validation method is employed in five independent experiments to obtain more training samples. The final experimental results are averaged across these five experiments. In the ULFSL-DET dataset, 80% of the total samples are used as training data, with the remaining 20% used as test data. Similarly, in the NEU-CLS dataset, 70% of the total samples are used as training data, while 30% of the total samples are used as test data. The specific parameters of the dataset division and the model training parameters are listed in Table 2.

Table 2. Dataset partition parameter and Model Parameter Setting table

Category	ULFSL-DET				NEU-CLS						
	2mm	5mm	8mm	Total	Rs	Pa	Cr	Ps	In	Sc	Total
Training set	160	160	160	480	210	210	210	210	210	210	1260
Test Sets	40	40	40	120	90	90	90	90	90	90	540
Image Size	1×128×128				1×200×200						
Learning Rate	0.001				0.001						
Number of Iterations	150				100						
Rate Adjustment	Cosine Annealing (T = 130)				Cosine Annealing (T = 80)						
Optimization	AdamW				AdamW						
Loss Function	Cross-entropy				Cross-entropy						

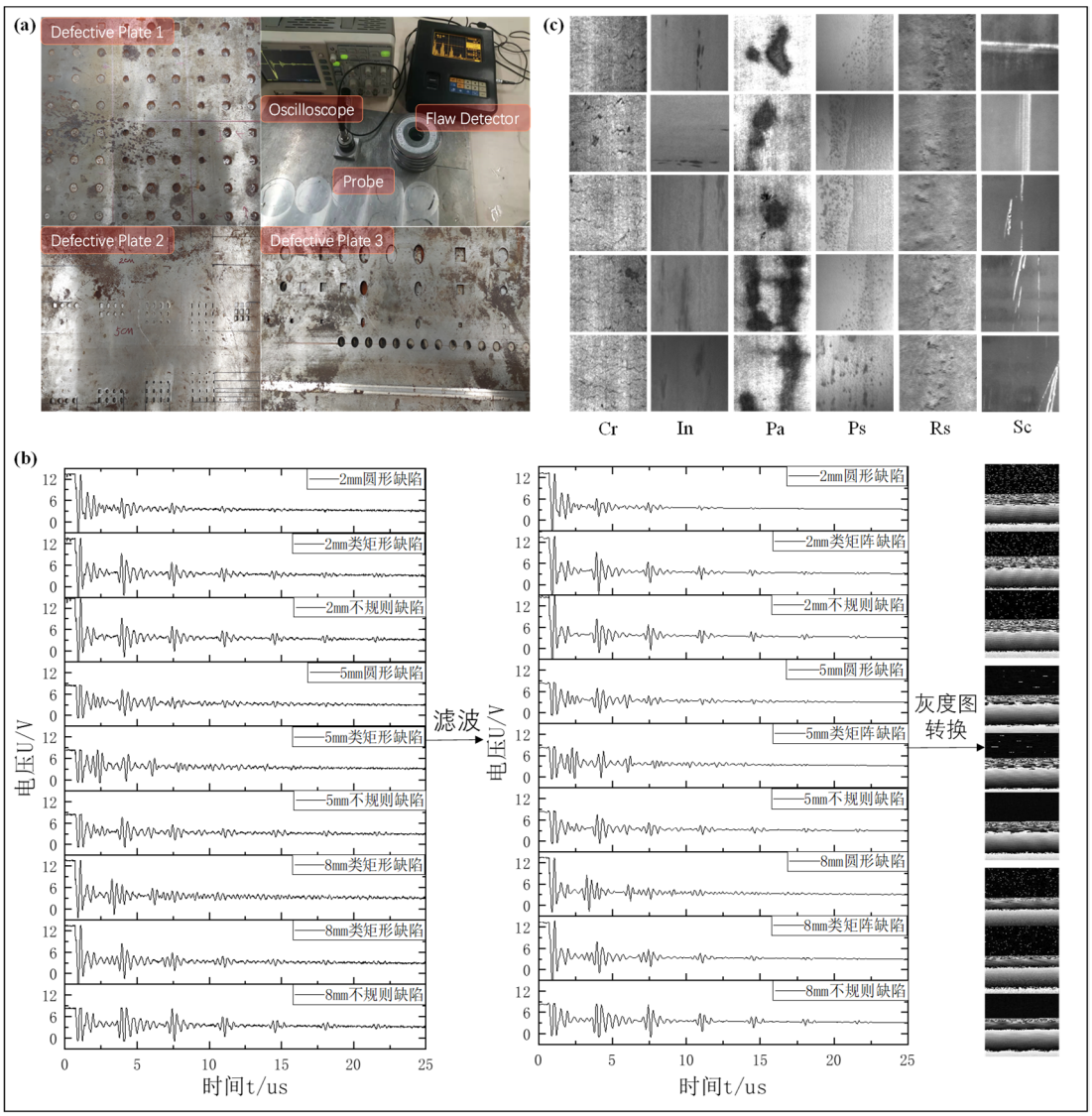

Figure 3. (a)Experimental Diagram of Ultrasonic Testing of Steel Plate Defects. (b) Partial ultrasonic echo signal and the preprocessing process. (c) Partial defect samples of the NEU-CLS dataset

5. Experimental Results

In this study, the accuracy of the model was evaluated using two different sets of data: the training set and the test set. The training set accuracy measures the consistency between the predicted labels and the actual labels of the samples used for parameter learning. On the other hand, the test set accuracy measures the consistency between the predicted labels and the actual labels of the samples that were not used for model training or parameter learning. The test set is crucial for evaluating the generalization ability of the model and is used as the primary basis for reporting the experimental results.

5.1. Defect classification experiment

To evaluate the recognition performance of ICST for metal defect ultrasonic signal grayscale images, experiments were carried out using the ULFSL-DET dataset, which was pre-processed by wavelet transform. The test set accuracy and loss value obtained from the experimental results are presented in Figure 4(a). Since the ULFSL-DET dataset has a smaller number of samples and a larger learning rate, the accuracy rate fluctuates significantly at the beginning of the iteration, but it shows an upward trend in general. Under the AdamW gradient optimization strategy, the accuracy rate steadily increases and gradually reaches the convergence state after 100 iterations. The experimental results show that the proposed ICST method achieves a high recognition accuracy of 98.1% in the ULFSL-DET dataset. The accuracy rate is calculated as the average of the accuracy of the five test sets.

To further investigate the recognition errors on the ULFSL-DET dataset, a confusion matrix was used to show the correspondence between the predicted values and the actual labels. As shown in Figure 4(b), the true labels are in the horizontal coordinates and the predicted values are in the vertical coordinates. In five independent experiments on the ULFSL-DET dataset, errors were observed in identifying 5mm defects as 2mm and 8mm defects as 5mm. These errors were mainly due to the presence of irregularly shaped defects in the dataset, which affected the depth judgment, and their echo signals had similar characteristics. These results suggest that the accuracy of defect recognition can be further improved by enhancing the ability of the model to distinguish between defects with similar characteristics.

In order to assess the ICST's ability to recognize metal surface defects in image form, experiments were conducted on the public NEU-CLS dataset. The experimental results, as shown in Figure 4(c), demonstrate the effectiveness of the model in recognizing surface defects in images. The larger sample size of the NEU-CLS dataset allows for more efficient optimization of model parameters, resulting in a quick and

accurate convergence of the test set accuracy. This suggests that the model possesses excellent generalization capabilities in the domain of image metal surface defect recognition. The average test set accuracy of five independent experiments was taken as the experimental result, and the ICST achieved a recognition accuracy of 99.8% on the NEU-CLS dataset.

The NEU-CLS dataset has demonstrated high recognition accuracy in general, but the results also revealed some errors in the recognition of oxidized skin (Rs) and plaque (Pa) defects. This may be attributed to the fact that these defects share significant similarities in their image shape and characteristics. Further research may be necessary to explore how to distinguish and classify these types of defects more accurately.

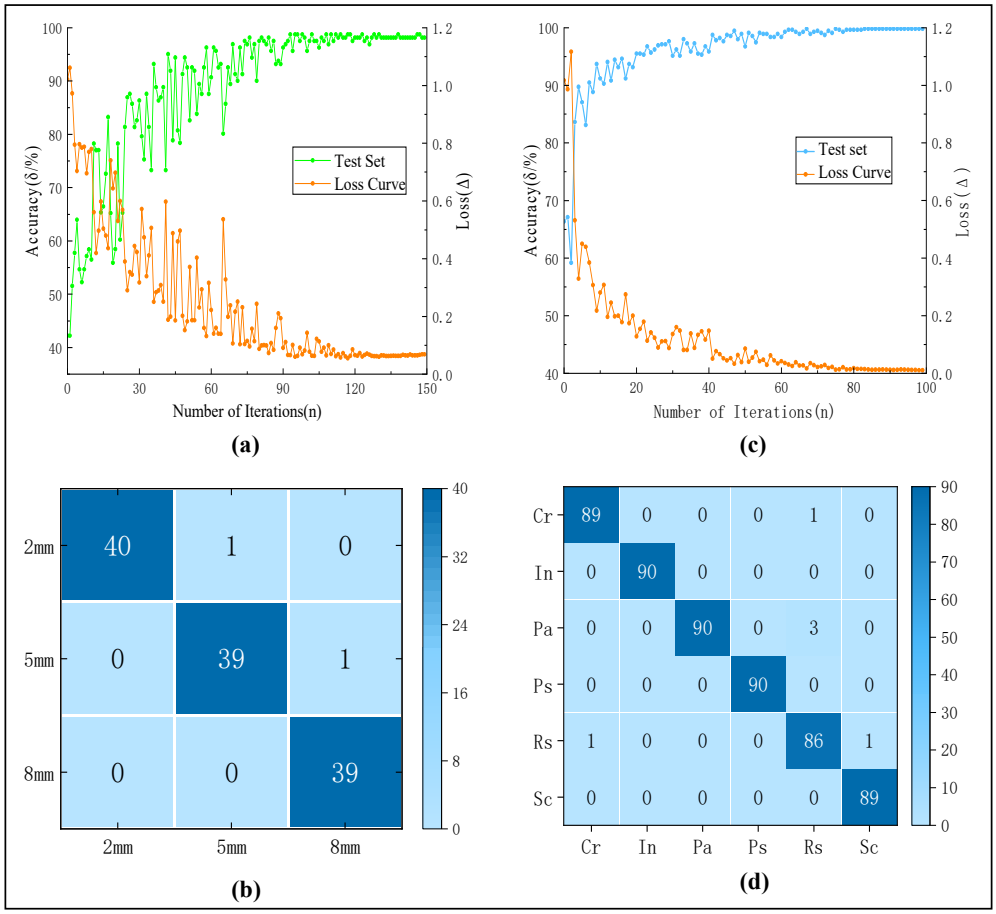

Figure 4. Defect Identification Performance Curve and Identification Confusion. (a)(b) ULFSL-DET; (c)(d) NEU-CLS.

5.2. Ablation experiments

In order to investigate the effects of the Swin Transformer Block and CoordAttention modules, we compared the performance of different networks including the Inception network, Inception_CoordAttention network, Inception_Swin Transformer Block network, and ICST. The evaluation metrics included accuracy (Acc), floating point computations (FLOPs), and the number of parameters (Params). The results are presented in Table 3.

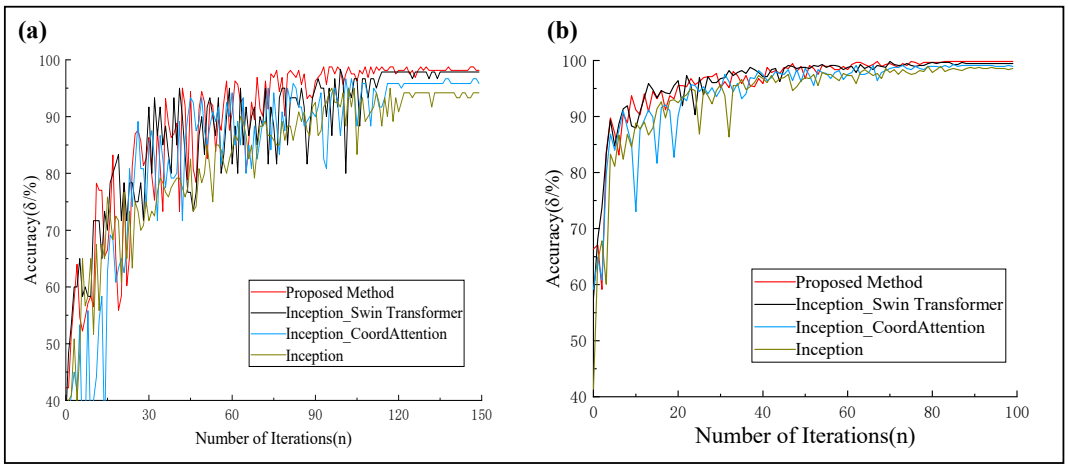

Figure 5. Defect identification performance curve of ablation experiment. (a) ULFSL-DET; (b) NEU-CLS.

The results of the experiments on the ULFSL-DET and NEU-CLS datasets are presented in Figure 5, respectively, which demonstrate the performance enhancement achieved by the Swin Transformer Block and CoordAttention modules. The Swin Transformer Block has a more significant impact on the recognition performance of time series metal defect signal recognition than that of the NEU-CLS dataset, indicating its effectiveness in improving the model performance in such tasks. On the other hand, the CoordAttention module enhances the model's ability to focus on essential feature channels and suppress redundant feature channel weights, leading to overall performance improvement. Furthermore, the CoordAttention module achieves performance improvement with a smaller number of parameters and computation cost than the Swin Transformer Block. Therefore, the CoordAttention module is more suitable when there is a need to reduce the number of model parameters and computations while sacrificing less accuracy.

Table 3. Ablation experiment

Inception	CoordAtt	Swin Transformer	ULFSL-DET			NEU-CLS		
			Acc(%)	FLOPs(G)	Params(M)	Acc(%)	FLOPs (G)	Params(M)
√			94.1	0.0172	0.140	98.5	0.0453	0.140
√	√		95.8	0.0173	0.149	98.9	0.0455	0.149
√		√	97.8	0.0980	1.340	99.4	0.2039	1.340
√	√	√	**98.1**	0.0981	1.350	**99.8**	0.2042	1.350

5.3. Comparison experiments of different classifiers

To comprehensively evaluate the performance of ICST, we compared it with traditional machine learning methods, including decision trees and SVMs, as well as classical convolutional neural network models such as VGG16[7], GoogLeNet[8], ResNet34[9], and Transformer model ViT. We employed accuracy (ACC), floating point computations (FLOPs), and the number of parameters (Params) as evaluation metrics.

In the ULFSL-DET dataset, conventional machine learning methods require manual extraction of ultrasound detection echo signal features. To extract data features with higher accuracy and less noise, we applied the Variational mode decomposition[23] (VMD) method to perform the eigenmode decomposition of the ultrasound signals after the filtering operation. Figure 6 illustrates that each ultrasound signal is decomposed to obtain seven eigenmode components IMF and one residual RES. We extracted nine defective features from each eigenmode component under the time domain indicator parameter, totaling 63 features for one sample signal. Each eigenmode component was manually extracted and divided into two categories: dimensionless parameters and dimensioned parameters. The dimensionless parameters included skewness, kurtosis, and peak, while the dimensioned parameters included variance, mean, maximum, minimum, amplitude, and standard deviation.

The test results of each model on the ULFSL-DET and NEU-CLS dataset are shown in Table 4. Compared with ICST, the advantage of machine learning method is that the model computation and the number of parameters are negligible. However, their recognition accuracy of metal defect ultrasonic signal grayscale images is much lower than the ICST. Still, the machine learning method is also highly dependent on expert opinions and a cumbersome work process in the feature extraction work. It shows that ICST can be primarily referred to other tasks that machine learning method is applied for time series defect signal recognition. Compared with the classical deep learning methods, the difference in accuracy is very small. However, in terms of model computation and the number of parameters, ICST is much smaller than the classical deep learning methods. Because the structure of ICST is simple and the number of parameters is small, ICST is more advantageous in small and medium-sized datasets, and the model parameters are easier to train.

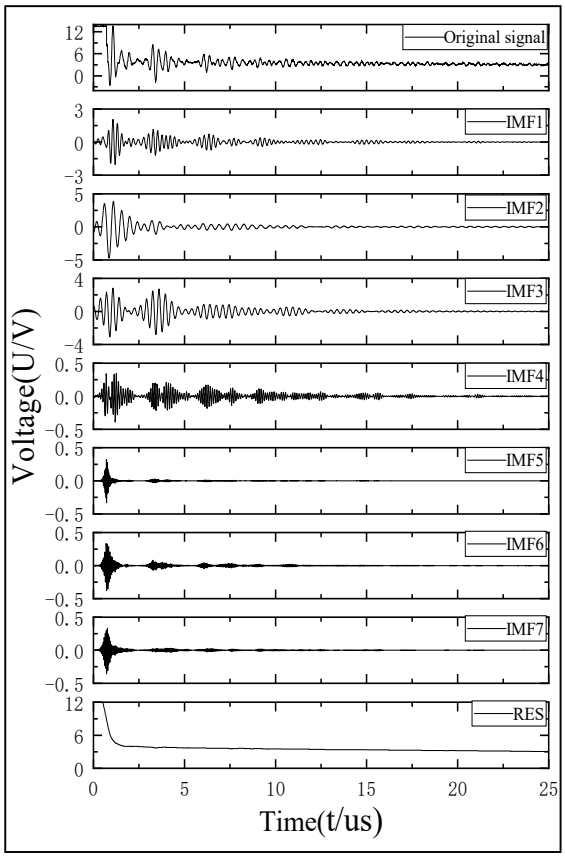

Figure 6. VMD Exploded View of ULFSL-DET Dataset Sample

Since NEU-CLS is an image-based mental defect dataset and does not require manual feature extraction, it is only selected for comparison experiments of ICST versus with the classical convolutional neural network model and Transformer model. Since the NEU-CLS dataset has a larger sample size than the ULFSL-DET dataset, each model gets better parameter optimization. So the accuracy of each model

on the NEU-CLS dataset is slightly improved. ResNet34, VGG16 and ICST are similar in recognition accuracy, but conventional methods' network computation and the number of parameters are too large, which is not conducive to small and medium-sized datasets. In contrast, ICST achieves the detection of metal defect ultrasonic signal grayscale images and image-based mental defects with lower computation cost, while obtaining optimal accuracy in small and medium-sized datasets.

Table 4. Test results of different models

Dataset	Indicators	Machine Learning		Deep Learning				Proposed Method
		DT	SVM	ResNet34	GoogLeNet	VGG16	Vit	ICST
ULFSL-DET	Acc	74.3%	75.8%	**98.2%**	96.5%	97.9%	93.8%	**98.1%**
	FLOPs	—	—	1.172G	0.494G	5.023G	5.530G	0.0981G
	Params	—	—	21.28M	5.97M	65.07M	85.22M	1.35M
	Extraction	Artificial	Artificial	Convolution	Convolution	Convolution	Convolution	Convolution
NEU-CLS	Acc	—	—	99.2%	97.8%	98.3%	97.5%	99.8%
	FLOPs	—	—	3.083G	1.160G	12.15G	12.34G	0.2042G
	Params	—	—	21.28M	5.97M	65.07M	85.22M	1.350M
	Extraction	—	—	Convolution	Convolution	Convolution	Convolution	Convolution

6. Conclusion

In conclusion, the proposed Inception fusion Swin Transformer-based method has shown promising results in the recognition of metal defect ultrasonic signal grayscale images. By combining the Inception structure and the Swin Transformer model, the method can effectively extract both local and global features of the defect echo signals, achieving high recognition accuracy with fewer parameters and computational cost compared to classical deep learning models. The experiments on the ULFSL-DET dataset demonstrate the effectiveness of the proposed method in the ultrasonic detection defect task, and the experiments on the NEU-CLS dataset show its generality in image-based metal defect recognition. These results indicate the great potential of the proposed method in real-time detection fields for metal defect ultrasonic signal grayscale images and image-based metal defects.

Acknowledgment

This work was supported by the Sichuan Provincial Science and Technology Support Plan Project (2017FZ0033), Sichuan Provincial Bureau of Market Supervision and Pipeline Science and Technology Plan Project (CSCJZ2022007) and Chengdu Technology Innovation R&D Project (1).

Author's Contributions

Donglin Tang and Yunliang Zhao completed the research and implemented the network. Yuanyuan He, Henghui Li and Simeng Yi contributed to the ULFSL-DET dataset and provided meaningful discussion.

Funding

This study was supported by the Sichuan Provincial Science and Technology Support Plan Project (2017FZ0033), Sichuan Provincial Bureau of Market Supervision and Pipeline Science and Technology Plan Project (CSCJZ2022007) and Chengdu Technology Innovation R&D Project (2018-YF05-00201-GX).

Availability of data and code The NEU-CLS dataset can be as a public dataset downloaded via the link http://faculty.neu.edu.cn/yunhyan/NEU_surface_defect_database.html. The ULFSL-DET Dataset cannot be shared at this time due to the need for other company project. The code can be shared by contacting email 1490608930@qq.com.

Declarations

Ethics approval The authors declare that this manuscript was not submitted to more than one journal for simultaneous consideration. The submitted work is original and not has been published elsewhere in any form or language Competing interests The authors declare that they have no conflict of interest.

References

[1] Czimmermann T, Ciuti G, Milazzo M, Chiurazzi M, Roccella S, Oddo CM, Dario P (2020) Visual-Based Defect Detection and Classification Approaches for Industrial Applications-A SURVEY. Sensors 20:1459. https://doi.org/10.3390/s20051459

[2] Fang X, Luo Q, Zhou B, Li C, Tian L (2020) Research Progress of Automated Visual Surface Defect Detection for Industrial Metal Planar Materials. Sensors 20:5136. https://doi.org/10.3390/s20185136

[3] Ming-Jian S, Ting L, Xing-Zhen C, De-Ying C, Feng-Gang Y, Nai-Zhang F (2016) Nondestructive detecting method for metal material defects based on multimodal signals. Acta Phys Sin 65:167802. https://doi.org/10.7498/aps.65.167802

[4] Mensah A, Sriramula S (2022) Machine learning based integrity decision management of pipeline corrosion clusters. In: 2022 International Conference on Decision Aid Sciences and Applications (dasa). Ieee, New York, pp 795–799

[5] Lin J, Yang J, Huang Y, Lin X (2021) Defect identification of metal additive manufacturing parts based on laser-induced breakdown spectroscopy and machine learning. Appl Phys B-Lasers Opt 127:173. https://doi.org/10.1007/s00340-021-07725-3

[6] Gaja H, Liou F (2018) Defect classification of laser metal deposition using logistic regression and artificial neural networks for pattern recognition. Int J Adv Manuf Technol 94:315–326. https://doi.org/10.1007/s00170-017-0878-9

[7] Zhang Y, Chan W, Jaitly N (2017) Very Deep Convolutional Networks for End-to-End Speech Recognition. In: 2017 Ieee International Conference on Acoustics, Speech and Signal Processing (icassp). Ieee, New York, pp 4845–4849

[8] Szegedy C, Liu W, Jia Y, Sermanet P, Reed S, Anguelov D, Erhan D, Vanhoucke V, Rabinovich A (2015) Going Deeper with Convolutions. In: 2015 Ieee Conference on Computer Vision and Pattern Recognition (cvpr). Ieee, New York, pp 1–9

[9] He K, Zhang X, Ren S, Sun J (2016) Deep Residual Learning for Image Recognition. In: 2016 IEEE Conference on

Computer Vision and Pattern Recognition (CVPR). pp 770–778

[10] Balcioglu YS, Sezen B, Gok MS, Tunca S (2022) Image Processing with Deep Learning: Surface Defect Detection of Metal Gears Through Deep Learning. Mater Eval 80:44–53. https://doi.org/10.32548/2022.me-04230

[11] Meng T, Tao Y, Chen Z, Avila JRS, Ran Q, Shao Y, Huang R, Xie Y, Zhao Q, Zhang Z, Yin H, Peyton AJ, Yin W (2021) Depth Evaluation for Metal Surface Defects by Eddy Current Testing Using Deep Residual Convolutional Neural Networks. IEEE Trans Instrum Meas 70:2515413. https://doi.org/10.1109/TIM.2021.3117367

[12] He D, Xu K, Wang D (2019) Design of multi-scale receptive field convolutional neural network for surface inspection of hot rolled steels. Image Vis Comput 89:12–20. https://doi.org/10.1016/j.imavis.2019.06.008

[13] Vaswani A, Shazeer N, Parmar N, Uszkoreit J, Jones L, Gomez AN, Kaiser L, Polosukhin I (2017) Attention Is All You Need. In: Guyon I, Luxburg UV, Bengio S, Wallach H, Fergus R, Vishwanathan S, Garnett R (eds) Advances in Neural Information Processing Systems 30 (nips 2017). Neural Information Processing Systems (nips), La Jolla

[14] Dosovitskiy A, Beyer L, Kolesnikov A, Weissenborn D, Zhai X, Unterthiner T, Dehghani M, Minderer M, Heigold G, Gelly S, Uszkoreit J, Houlsby N (2021) An Image is Worth 16x16 Words: Transformers for Image Recognition at Scale

[15] Zhou D, Kang B, Jin X, Yang L, Lian X, Jiang Z, Hou Q, Feng J (2021) DeepViT: Towards Deeper Vision Transformer

[16] Touvron H, Cord M, Sablayrolles A, Synnaeve G, Jégou H (2021) Going Deeper With Image Transformers. pp 32–42

[17] Wang W, Xie E, Li X, Fan D-P, Song K, Liang D, Lu T, Luo P, Shao L (2022) PVT v2: Improved baselines with Pyramid Vision Transformer. Comput Vis Media 8:415–424. https://doi.org/10.1007/s41095-022-0274-8

[18] Liu Z, Lin Y, Cao Y, Hu H, Wei Y, Zhang Z, Lin S, Guo B (2021) Swin Transformer: Hierarchical Vision Transformer using Shifted Windows. In: 2021 Ieee/Cvf International Conference on Computer Vision (iccv 2021). Ieee, New York, pp 9992–10002

[19] Hou Q, Zhou D, Feng J (2021) Coordinate Attention for Efficient Mobile Network Design. In: 2021 Ieee/Cvf Conference on Computer Vision and Pattern Recognition, Cvpr 2021. Ieee Computer Soc, Los Alamitos, pp 13708–13717

[20] Hu J, Shen L, Sun G (2018) Squeeze-and-Excitation Networks. In: 2018 Ieee/Cvf Conference on Computer Vision and Pattern Recognition (cvpr). Ieee, New York, pp 7132–7141

[21] Woo S, Park J, Lee J-Y, Kweon IS (2018) CBAM: Convolutional Block Attention Module. In: Ferrari V, Hebert M, Sminchisescu C, Weiss Y (eds) Computer Vision - Eccv 2018, Pt Vii. Springer International Publishing Ag, Cham, pp 3–19

[22] He Y, Song K, Meng Q, Yan Y (2020) An End-to-End Steel Surface Defect Detection Approach via Fusing Multiple Hierarchical Features. IEEE Trans Instrum Meas 69:1493–1504. https://doi.org/10.1109/TIM.2019.2915404

[23] Dragomiretskiy K, Zosso D (2014) Variational Mode Decomposition. IEEE Trans Signal Process 62:531–544. https://doi.org/10.1109/TSP.2013.2288675