# Rail Defect Classification with Deep Learning Method

**Shiyao Lu[1], Jingru Wang[2], Guoqing Jing[2], Weile Qiang[3], Majid Movahedi Rad[4*]**

[1]Business School, Hohai University, #8 West Focheng Road, Jiangning District, Nanjing, China, 211100

[2]School of Civil Engineering, Beijing Jiaotong University, No. 3 Shangyuancun, Haidian District, Beijing, China, 100044

[3]Infrastructure Inspection Research Institute, China Academy of Railway Sciences Corporation Limited, No. 2 Daliushu Road, Haidian District, Beijing, China, 100081

[4] Department of Structural and Geotechnical Engineering, Széchenyi István University, Egyetem tér 1, H-9026 Győr, Hungary

e-mail: 190413120010@hhu.edu.cn, 20121196@bjtu.edu.cn, gqjing@bjtu.edu.cn, qiangweile@rails.cn, majidmr@sze.hu

*Abstract: The good condition of railway rails is crucial to ensuring the safe operation of the railway network. At present, the rail flaw detectors are widely used in rail flaw detection, they are typically based on the principle of ultrasonic detection. However, the rail detection results analysis process involves huge manual work and the associated labor costs, with low levels of efficiency. In order to improve the efficiency, accuracy of results analysis and also reduce the labor costs, it is necessary to employ classification of ultrasonic flaw detection B-scan image, based on an artificial intelligence algorithm. Inspired by transformer models, with excellent performance in the field of natural language processing (NLP), some deep learning models differ from traditional convolutional neural networks (CNN), gradually emerge in the field of computer image processing. In order to explore the practicality of this model in the field of computer image processing (vision), in the paper, the Vision Transformer (ViT) is employed to train with rail defect B-scan images data and produce a rail defect classification. The model accuracy is more than 90% with the highest accuracy reaching 98.92%.*

*Keywords: railway; rail defect; artificial intelligence; vision transformer*

# 1 Introduction

Rail is the most important component of the track, and its function is to provide a continuous, smooth and minimum resistance rolling surface for the wheels, guide the locomotives and vehicles forward, directly bear the stresses of the wheel, etc. [1]. With the development of high-speed and heavy-load railway, rail is subjected to higher pressure and frequency of use, resulting in more rail defects, Figure 1 shows some common rail defects, such as rail transverse crack, spalling defect, screw hole crack, etc., which threaten the safety of railway operation and state of the railway track, thus need more maintenance costs. For example, until 2020 nearly 11% of railway tracks were renewed and rehabilitated in Hungary, which means millions of Euros [2]. In order to ensure the safe operation of railway, prolong the service life of rail and reduce investment in railway maintenance, it is very important to detect rail defect regularly. Traditional manual inspection requires the staff with strong professional ability, inefficient and risky. Nondestructive testing (NDT) technology can be conducted efficiently and accurately without damaging the rail, common nondestructive testing methods mainly include ultrasonic detecting, visual detecting, magnetic flux leakage detecting and eddy current detecting. Ultrasonic detection mainly detects internal rail defect, while other detection methods mainly detect rail surface and near surface defect [3]. Rail flaw detection cars and rail flaw detectors are mostly based on ultrasonic detecting principle, cooperate with manual detect rail internal defect regularly.

a) rail transverse crack b) spalling defect c) screw hole crack

Figure 1
Common rail defects

Ultrasonic rail flaw detection mainly uses some physical characteristics of ultrasonic wave in object propagation to find defects of rail. Currently contact pulse reflection method is widely used in the detection of rail flaw detection car and rail flaw detector. According to the different display contents, the images on the oscilloscope screen are divided into A-scan and B-scan. For the A-scan, the horizontal axis shows echo time, longitudinal axis shows echo height, the B-scan's horizontal axis shows mileage location and shares, longitudinal axis points directly to the defect shape. Compared with A-scan image, B-scan image can directly determine the location and shape of the defect, which is convenient for

identification [4]. At present, in the process of ultrasonic rail inspection, due to the lack of intelligence, it cannot fully realize the inspection and reporting at the same time. It still needs to rely heavily on manual second judgment, but the impact of human factors will appear inconsistent results and some other problems. Therefore, it is of great significance to optimize the existing rail defect detection process, save labor cost and improve detection efficiency and accuracy by combining artificial intelligence methods.

At present, the research on ultrasonic B-scan image recognition and classification can be divided into two main schools: the traditional machine learning method of "manual feature extraction plus classifier" and the deep learning method based on convolutional neural network. Huang et al. [5] used BP neural network as the framework to realize the classification and recognition of rail defect. During the training process, the characteristics of B-scan images were extracted manually, and the experimental results showed that the recognition rate could reach more than 95% as long as the appropriate threshold was set. Huang et al. [6] proposed a rail defect classification algorithm based on image processing. The algorithm combining Tamura texture features and local binary mode was used to extract image features, and the extracted feature vectors were sent into support vector machine for training. Finally, the image classification results were output in visual form. The results show that the total classification accuracy can reach 99% by using the combination of the two feature extraction algorithms, Wu et al. [7], extracted the color features of each type of defect, the regional features which represent the location of defect distribution, the contour features which represent the defect contour and the numerical constraint digital features related to the defect type. He proposed a parameter threshold adjustment method based on the law of data distribution, established a perceptron classification learning model, and transformed the defect detection problem into a contour classification problem. The recall rate reached 98.62%, and the accuracy rate reached 100%. According to the characteristics of B-scan of rail defects, he proposed an intelligent detection method based on generalized feature clustering, and developed k-means clustering algorithm for defect clustering analysis based on dimensionality reduction feature and strong constraint generalized feature, with an accuracy of 97.55% [8]. In recent years, deep learning has been widely used in image recognition and classification tasks. In terms of rail defect classification, deep learning is mainly used for rail surface defects identification and classification [9] [10], it is seldom used for B-scan images of rail defects. But the implementation process and methods can be migrated. Sun et al. [11] designed a deep learning neural network framework based on AlexNet convolutional neural network to transform the object detection problem into a classification problem. Rail defect identification is carried out by using the existing system, the designed deep convolutional neural network model and manual identification respectively. The results show that the accuracy of manual identification and deep convolutional neural network model is higher than the existing system, and the manual identification is better when there are more types of defect. Since features of B-scan are not obvious and prone to clutter interference,

Hu et al. [12] used residual neural network (ResNet-50) based on transfer learning to realize automatic identification and classification of rail defect. The network was evaluated using three dimensions and compared with the performance of BP neural network, support vector machine and Bayesian classifier. The results showed that, the residual neural network based on transfer learning has higher accuracy and efficiency. Luo et al. [13] combined the advantages of traditional machine learning and deep learning algorithms. Deep learning was used to locate objects, construct multi-dimensional features and establish image classification algorithm based on support vector machine. Compared with ResNet model and MobileNetV1 model, the classification accuracy, false positive rate and single frame running time of the three methods are similar, but the single frame running time of the model is shorter and the detection efficiency is higher. On the basis of YOLOV3, Chen et al. [14] adjusted the network structure to expand the receptive field, and used k-means clustering algorithm to obtain prior frames for B-scan image data to facilitate subsequent model training. The results showed that this model was considerable.

At present, the amount of rail defect data is limited, so traditional machine learning methods may be superior to deep learning methods somehow. However, the trend of expanding data sets and need of the ability to transfer learning show that deep learning is more promising than traditional machine learning. Convolution neural network has become the dominant model of computer vision tasks, the same for B-scan images classification. However, with the emergence of more efficient structure, computer vision (CV) and natural language processing (NLP) are integrating gradually. Transformer, which has excellent performance in the field of natural language processing, has become a research hotspot in the computer vision filed. Compared with convolutional neural networks and recursive networks, transformer models can show better performance in various computer vision tasks, with advantages, such as, a simplified structure and a strong transfer learning ability [15].

In this paper, Vision Transformer, the application model of Transformer in the image classification task, is used to study the classification of rail defect. First, the ultrasonic B-scan images of rail defect are divided into four types, which are divided into training sets and test sets, according to a ratio of 2.5:1. Then, the Vision Transformed system will be trained with the training set data, and use the data of the test set to verify the classification accuracy of the model. The results show that the model has good performance in rail defect classification.

# 2 Ultrasonic Flaw Detection B-Scan

The B-scan image on the display screen of ultrasonic flaw detection equipment is referred to as ultrasonic B-scan images. The horizontal axis contains mileage position, left and right track information, and the vertical axis is the corresponding position of the defect on the height of the rail. Compared with traditional images,

B-scan image has the following characteristics: It is not affected by external light and brightness. The noise is affected by various factors in the process of ultrasonic transmission and the regularity is not strong. There is no ready-made data set, so we need to build our own data set according to the specific road section, and the resolution is fixed, etc. It can be seen that in the process of ultrasonic rail flaw detection, analysis of B-scan images is more intuitive and efficient, but the application of traditional image processing methods may not be applicable, so it is necessary to select an appropriate image processing method based on the specific characteristics of B-scan image to optimize the process of flaw detection intelligently.

The background of B-scan images is usually black or light yellow, and the channel color of each probe can be set by itself. Combined with the rail structure, as shown in Figure 2, B-scan images can be roughly divided into three areas, as shown in Figure 3, which is convenient to determine the location of defect.

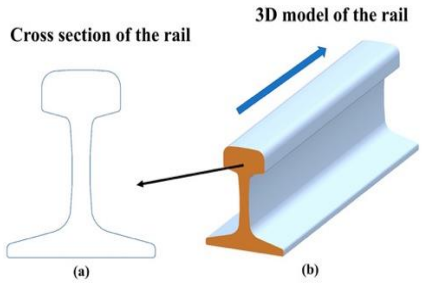



Figure 2
Rail cross section [16]

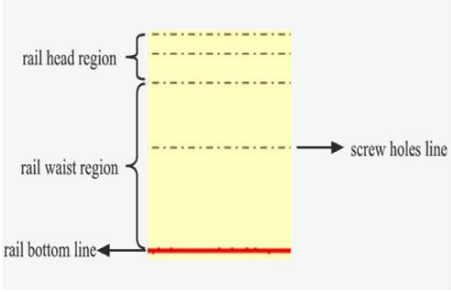



Figure 3
B-scan partition

Common defects are generally divided into: rail head transverse crack, rail joint damage, rail horizontal, vertical, oblique crack, rail bottom crack, rail weld defects. Rail transverse crack mostly occurs in rail head, which is one of the most harmful rail damage in all kinds of defects. The main defects of rail joint are screw hole crack, rail head jaw crack, joint drop block and saddle wear. The screw hole crack and rail head jaw crack can be detected by ultrasonic inspection. The other three kinds of rail defects can be detected by ultrasonic inspection. However, due to the limitation of the current detection capacity of flaw detection equipment and the professional ability of flaw detection workers, the detection rate of all kinds of damage has not reached 100%, and there are problems such as missed detection and false positives.

Based on the study of B-scan images data of rail damage in Hohhot Railway Bureau in 2020, it is found that different types of B-scan images of rail defect have their own characteristics, which are summarized as follows. First, it is determined that the background of B-scan images in this study is light yellow, and the state setting of each probe is shown in Figure 4.

| | | A | B | C | D | |
|---|---|---|---|---|---|---|
| State of the probe | Channel | A: anterolateral | B: posteromedial | C: anteromedial | D: posterolateral | |
| | Gain value | 47.5 | 47.5 | 46.0 | 47.5 | |
| | Integral value | 61 | 69 | 105 | 58 | |
| | Grass shape wave% | 10 | 10 | 15 | 5 | |
| | | E | F | G | H | I |
| | Channel | E: forward | F: back | G: forward | H: back | I: 0 |
| | Gain value | 44.5 | 44.0 | 52.5 | 52.5 | 57.5 |
| | Integral value | 96 | 93 | 35 | 44 | 110 |
| | Grass shape wave% | 15 | 15 | <5 | 5 | 15 |

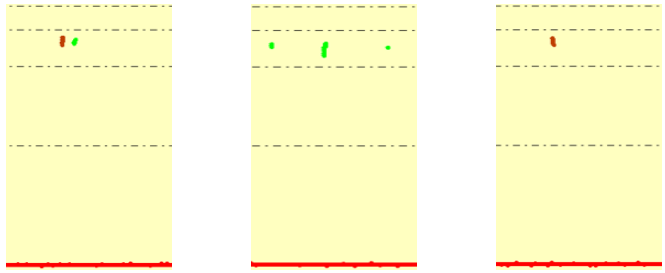
Figure 4

Probe state setting parameters

If there is abnormal echo in the rail head area in B-scan images, the existence of rail head transverse cracks can be confirmed, and the rail head transverse cracks can be divided into inner rail head transverse crack, central rail head transverse crack and lateral rail head transverse crack. The characteristics of B-scan images of inner rail head transverse crack are summarized as follows:

1. The anteromedial and posteromedial 70° probes are used to evaluate the flaw
2. Three wave conditions:
   - anteromedial 70° probe/posteromedial 70° probe appear echo same time
   - anteromedial 70° probe appears echo
   - posteromedial of 70° probe appears echo
3. When the anteromedial 70° probe appears echo alone, the echo distribution is sporadic
4. When the posteromedial 70° probe appears echo alone, the echo distribution is dense.

Figure 4 shows the B-scan images of the inner rail head transverse crack. Similar to the inner rail head transverse crack, if the anterolateral or posterolateral 70° probe echoes in the region of the rail head in B-scan images, the lateral rail head transverse crack may occur.

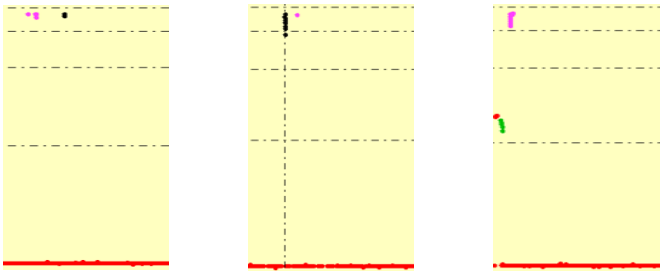


a) B and C channel echoes b) B channel echo c) C channel echo

Figure 5

Inner rail head transverse crack

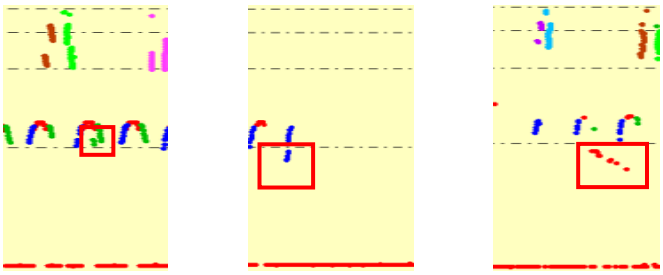
The characteristics of B-scan images of central rail head transverse crack are summarized as follows: 1) The forward and backward 70° straight probe echo is mainly used to judge the defect; 2) Three kinds of wave output conditions: the forward 70° straight probe and the back 70° straight probe appear at the same time, the forward 70° straight probe echo, the back 70° straight probe echo. Figure 6 shows the B-scan images of the specific central rail head transverse crack.



a) E and F channel echoes b) E channel echo c) F channel echo

Figure 6
Central rail head transverse crack

Screw hole cracks are mainly concerned with the damage of rail joints. The screw hole cracks can be divided into oblique upper screw hole cracks, oblique lower screw hole cracks and horizontal screw hole cracks. Characteristics of B-scan image of screw hole crack are summarized as follows: 1) The forward and backward 37° straight probe echo is mainly used to judge the defect; 2) If the echo exceeds the scale line of the lower boundary of the screw hole, it is determined as the oblique lower crack of the screw hole; 3) If echo overlaps, discontinuities or exceeds the upper boundary of the screw hole, it is determined as the oblique upper crack of the screw hole. 4) If there is abnormal 0° probe echo near the screw hole line, it is judged as horizontal screw hole crack. Figure 7 shows the B-scan images of the specific screw hole crack.



a) G channel echo b) H channel echo c) I channel echo

Figure 7
Screw hole crack

The rail inspection data of Hohhot, Jining and other places in 2020 were obtained by using the rail inspection vehicle RT-18, and nearly a thousand ultrasonic B-scan images were extracted from the playback software. Combining the classification principle in TB/T1778-2010 Rail Damage Classification and rail defects that could be detected by ultrasonic inspection, the data were labeled as four types: rail head transverse crack, screw hole crack, other cracks and normal.

# 3 Defect Classification Based on Transformer Model

## 3.1 Self-Attention

The self-attention mechanism is an integral component of Transformers, which explicitly models the interactions between all entities of a sequence for structured prediction tasks. Each word has three different vectors, Query (Q), Key (K), and Value (V), all 64 in length. These three vectors are obtained by multiplying the embedding vector $X$ by three different weight matrices $W^Q$、$W^K$、$W^V$, each of which has dimensions of $512 \times 64$. Figure 8 shows an example and the calculation process is as follows:

Step1: Transform the input word into the embedding vector $X$

Step2: Multiply the embedding vector with three weight matrices $W^Q$、$W^K$、$W^V$ to obtain Q, K, V.

Step3: Calculate the dot product between Q and K and divide it by $\sqrt{d_k}$. $d_k$ is the dimension of Q and K.

Step4: Use softmax to normalize the results, and then multiply by the matrix V to obtain the summation representation of the weights, $Attention(Q、K、V) = softmax\left(\frac{QK^T}{\sqrt{d_k}}\right)V$ [17]

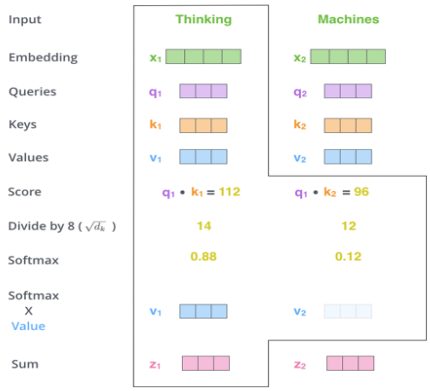



Figure 8
Self-Attention calculation process [17]

The multi-attentional mechanism used in Transformer is equivalent to the integration of $h$ different self-attentional mechanisms. Assuming $h = 8$, the output of multi-attentional mechanism is divided into three steps:

Step1: Vector $X$ is input into 8 self-attention, and 8 weighted eigenmatrices $Z_i, i \in \{1,2, \ldots ,8\}$

Step2: Assemble 8 $Z_i$ into a large eigenmatrix

Step3: Output $Z$ is obtained after the eigenmatrix passes through a fully connected layer. [17]

Multi-headed self-attention mechanism uses multiple query vectors $Q$ to compute in parallel to obtain multiple information from the input information. Each attention focuses on different parts of the information, and finally splicing the information, which makes the model pay more attention to the global information.

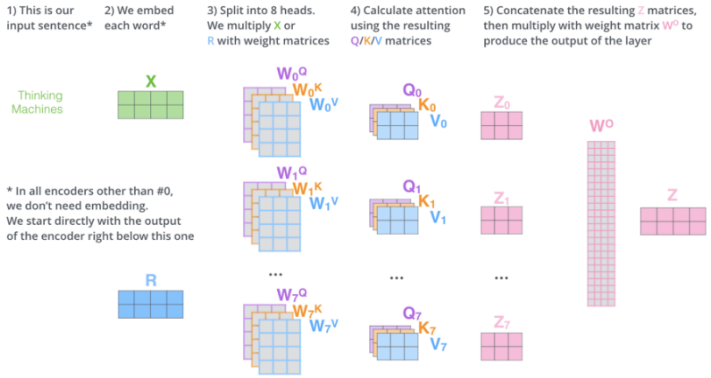



Figure 9
Multi-Head Attention [18]

## 3.2 Transformer

Transformer was originally applied in natural language processing tasks and was first proposed by Vaswani et al in 2017. It performs better than RNN and CNN in machine translation tasks and can achieve good results only using the encoding and decoding structure and attention mechanism. The biggest advantage is found in the efficient parallelization [19]. After that, BETR and GPT-3-based Transformer gradually emerged, and these models showed strong performance and contributed to the breakthroughs in the field of natural language processing [20] [21].

Transformer is a deep neural network completely based on self-attention mechanism, eliminating the circulation layer and convolution layer. Its structure is shown in Figure 10. The model is divided into Encoder and Decoder, with Encoder block on the left and Decoder block on the right. The part in the red circle is multi-head Attention, which is composed of multiple self-attention. It can be seen that the Encoder block contains one multi-head Attention. The Decoder Block contains two multi-head Attention (one of which uses Masked). There is also an Add & Norm Layer on top of multi-head Attention. Add means Residual Connection to prevent network degradation and Norm means Layer Normalization. Used to normalize the activation values for each layer. The Transformer workflow is as follows:

Step1: Get the representation vector of each word in the input sentence. The representation vector is obtained by adding the embedding of the word and the embedding of the word position.

Step2: Input the obtained word representation vector matrix into Encoder, and get the coding information matrix of all words in sentences after 6 Encoder blocks.

Step3: Transfer the coding information matrix output by Encoder to Decoder, which will translate the next word $i+1$ in turn according to the currently translated word $1{\sim}i$. In the process of use, when the word $i+1$ is translated, the word after $i+1$ needs to be covered by Mask operation. [22]

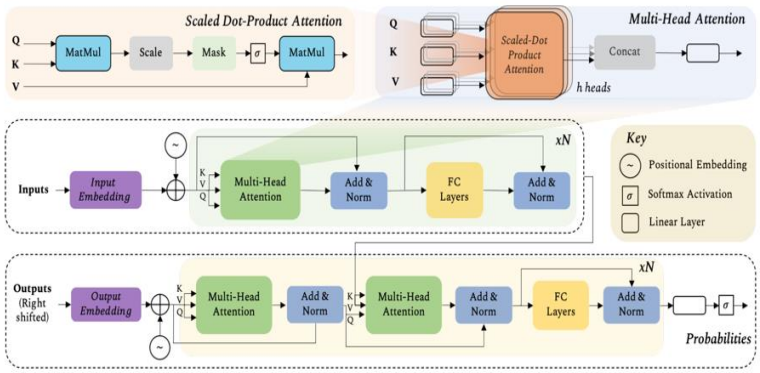



Figure 10
The Transformer architecture [22]

## 3.3 Vision Transformer (ViT)

Inspired by the successful application of Transformer in the NLP field, Parmar et al. generalized Transformer model to an image generation sequence modeling formula with easy to handle likelihood. He encoded each pixel with a value of [0,255] into a d-dimensional vector as the input of the encoder. The special feature of this model is the decoder, where each output pixel is calculated by the input pixel and the Attention between the generated pixels. This was the first image generation work using complete Transformer [23]. Dosovitskiy et al. proposed Vision Transformer (ViT) to deal with the problem of image classification. Specifically, THE ViT algorithm firstly cuts the image into image blocks and forms linear serialized data into Transformer to perform the image classification task. Then, supervised learning is used for image classification training. The overall structure of the ViT algorithm is shown in Figure 11. The results show that after pre-training with large-scale data sets, the method of transfer learning can be applied to other small and medium-sized data sets to reach or even exceed the current level of SOTA [24].

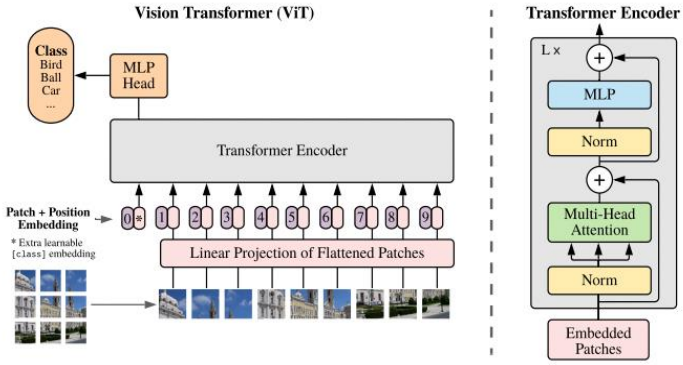



Figure 11
The Vision Transformer architecture [24]

The general workflow of ViT is as follows:

Step1: Convert 3D image into 2D input. In Transformer, the input is a two-dimensional matrix$(N, D)$ , where $N$ is the length of the sequence and $D$ is the dimension of each vector in the sequence. Therefore, the three-dimensional image of $H \times W \times C$ needs to be converted into the two-dimensional input of $(N, D)$first. This is done by changing the $H \times W \times C$ image into a sequence of $N \times (P^2 * C)$, which can be seen as a series of flattened image blocks, the number of image blocks is $N = {HW}/{P^2}$, each of which has a dimension of $(P^2 * C)$.Where $P$ is the size of image blocks and $C$ is the number of channels. Then the image block is added and linear transformation is performed on each $(P^2 * C)$ image block to compress the dimension into $D$.

Step2: Input the serialized image into Transformer for feature extraction. Transformer Encoder is the base module of ViT model, which consists of multi-head Attention, Norm, MLP and DropPath modules. One of the most important is multi-head Attention, which is implemented using DropPath (Stochastic Depth) instead of the traditional Dropout structure. DropPath can be understood as a special Dropout structure. The result is that you get rid of a subset of layers randomly during training, and you normally use the full Graph in prediction.

Step3: Use MLP Head to classify the output of the basic module. MLP Head is composed of LayerNorm and two fully connected layers, and RELU activation function is adopted. [25] [26]

It can be seen that the realization process of Vision Transformer has no convolution, pooling and other processes of convolutional neural network. The image is divided into image blocks (patches), and the linear embedding sequence of these image blocks is used as the input of Transformer, but only use the encoder block, and the image classification model is trained in a supervised manner. The application of Vision Transformer in the field of computer vision is still in the preliminary stage of exploration and trial. In order to verify the practicability of this model, more attempts need to be made. In this study, the model is used to deal with the rail defect classification, which not only verifies the effectiveness of the model, but also provides an intelligent solution for the task.

## 3.4 Model Test and Verification

With CPU as the computing core, the classification of rail defect B-scan images by ViT model is realized by Tensorflow, a deep learning framework developed by Google. Firstly, it is necessary to preprocess the B-scan images, set playback software parameters, filter out part of the noise, and intercept the B-scan images with obvious echoes. According to the knowledge of ultrasonic testing and rail defect, the type of rail defect is judged and the corresponding echo is circled. Then the rail defect in the test were divided into four types: rail head transverse crack, screw hole crack, other cracks and normal, and the training set and test set were divided according to the ratio of about 2.5:1 to form the data set, the number of images in the training set is 203, and the test set is 80. The data in the training set and the test set classified rail defects into the above four types. The B-scan images in the training set were marked, which is used to train the model. The B-scan images in the test set were unmarked, and this part of data was used to test the classification accuracy of the model. Table 1 lists the main parameters of the model.

Table 1
Parameters setting

| | | | |
|---|---|---|---|
| Learning rate | 0.001 | Epoch | 100 |
| Weight rate | 0.0001 | Heads number | 4 |
| Batch size | 256 | Transformer layer | 8 |
| Classes | 4 | Train : Test | 2.5:1 |

After installing the library required for model training and setting some basic parameters, the data needs to be resized to the size of $550 \times 100$, which can make the height range cover all areas from the head the bottom of the rail. The width range can include echoes where the defect occurred. Before input into the model, use the data augmentation method to preprocess. Then according to the workflow of Vision Transformer, the picture should be divided into pixel blocks and converted into patches. In the test, the patch size is $6 \times 6$, per image has 144 patches and per patch has 108 elements. For example, the normal B-scan image without echo segmentation is shown in Figure 12. Then input the serialized image into Transformer Encoder for feature extraction and use the MLP to achieve the classification.

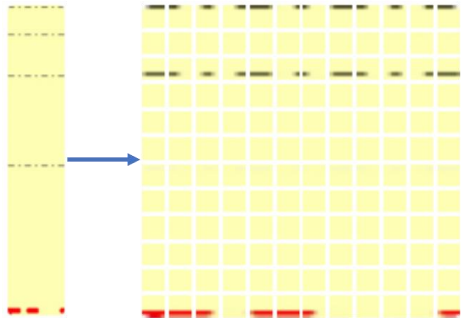


Figure 12
The B-scan image segmentation sample

Test results: The test results are shown in Figure 13 and Figure 14. It can be seen that after 100 epochs, the accuracy of test set reached 93.69%. The highest accuracy was 98.92%, which occurred many times during the epoch process. The results show that the model is suitable for the classification task of B-scan images of rail defect and it performs well.

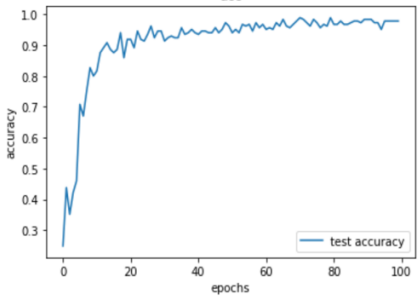



Figure 13
Classification accuracy

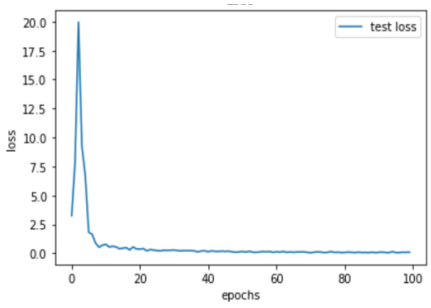



Figure 14
The loss function

## 3.5 Cost Analysis

By using artificial intelligence to improve rail flaw detection, labor costs and times can be dramatically decreased, during a year, as shown in Figure 14 and Figure 15. Traditional detection requires 8 workers and 6 senior technicians for every 10 kilometers of rail. As the use of new technology with Ultrasonic rail flaw detection becomes more widespread, the cost of monthly performed flaw detections can be reduced by 13,195,000 Euro in a year. Time savings may reach 600 hours a year. In that case, basic workforces will be released and professional inspectors will be needed to check the machine and improve the efficiency regularly, which encourages the workforce to improve professional knowledge.

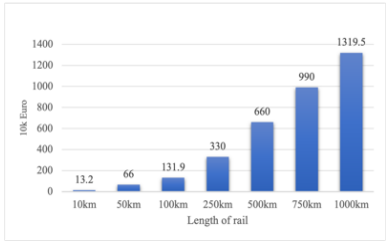



Figure 14
Reduction of labor cost

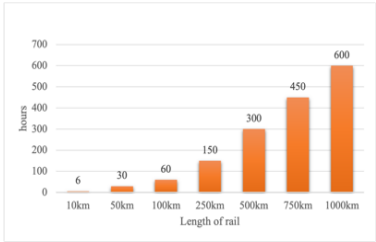



Figure 15
Reduction of time consumption

### Conclusions

To solve the problem of automatic classification of B-scan images of rail defects, Vision Transformer, a novel image classification model, based on Transformer, is examined in this paper, to verify and test the data set of B-scan images of rail defects. The following conclusions are drawn:

(1) Vision Transformer is completely different from convolutional neural network, which pays more attention to the global information of the image. The results show that the model can extract the B-scan images' features of

rail damage well, thus completing the rail damage classification task efficiently and accurately.

(2) Limited by time and other factors, the research in this paper is only based on the Vision Transformer basic model. With the expansion of data sets and the fine-tuning and improvement of the model, it is believed that the image classification model based on Transformer can perform better, which is also the focus of the following research.

(3) This study provides an intelligent solution for optimizing the flaw detection process, which not only improves the detection efficiency and accuracy but also saves costs and reduces time consumption. Plagued by slow population growth, more urgency is needed to accelerate the application of artificial intelligence. However, due to current concerns from society and corporations about the accuracy of existing artificial technology, application and successful cases should be promoted to more audiences. In addition, applying artificial intelligence to more scenarios will also revise and improve the technology itself. Based on the current situation of a gradually ageing population in global and the successful model proved in this article, artificial intelligence should be more preferred to help release workforces in this area.

**Acknowledgement**

This work was supported by Science and Technology Research and Development Program of China State Railway Group Co., Ltd. [K2020G006]：Research on condition assessment and evolvement law of Wuhan-Guangzhou high-speed railway track based on long-term service data.

**References**

[1] C. Esveld, "Modern railway track," Zaltbommel: MRT-productions, Vol. 385, 2001

[2] B. Eller, S. Fischer, "Tutorial on the emergence of local substructure failures in the railway track structure and their renewal with existing and new methodologies, " *Acta Technica Jaurinensis*, Vol. 14, No. 1, pp. 80-103, 2021

[3] G. Tian, B. Gao, Y. Gao, P. Wang, H. Wang, Y. Shi, "Review of Railway Rail Defect Non-destructive Testing and Monitoring," *Chinese Journal of Scientific Instrument*, Vol. 37, No. 8, pp.1763-1780, 2016

[4] Z. Huang, F. Shi, "Rail flaw detection and fracture prevention knowledge," *Beijing: China Railway Publishing House Co., Ltd*, 2015

[5] X. Huang, Y. Shi, Y. Zhang, P. Li, L. Xiong, Y. Zhong, "BP Neural Network Based on Rail Flaw Classification of RFD Car's B-scan Data," *Chinese Railways*, Vol. 03, pp. 82-87, 2018

[6] M. Huang, J. Luo, W. Wang, J. Cao, "Research on Classification of Rail Defects Based on Image Processing Algorithm," *Electric Drive for Locomotives*, Vol. 04, pp. 41-46, 2020

[7] F. Wu, Q. Li, S. Li, T. Wu, "Train rail defect classification detection and its parameters learning method," *Measurement*, 2020, 151: 107246

[8] F. Wu, X. Xie, J. Guo, Q. Li, "Internal Defects Detection Method of the Railway Track Based on Generalization Features Cluster", 2021

[9] S. Faghih-Roohi, S. Hajizadeh, A. Núñez, R. Babuska and B. De Schutter, "Deep convolutional neural networks for detection of rail surface defects," 2016 International Joint Conference on Neural Networks (IJCNN), IEEE, pp. 2584-2589, 2016

[10] Hajizadeh, Siamak, Alfredo Núnez, and David MJ Tax, "Semi-supervised rail defect detection from imbalanced image data." IFAC-PapersOnLine, Vol. 49, No. 3, pp. 78-83, 2016

[11] C. Sun, J. Liu, Y. Qin, Y. Zhang, "Inteligent Detection Method for Rail Flaw Based on Dep Learning," *China Railway Science*, Vol. 39, No. 05, pp. 51-57, 2018

[12] W. Hu, S. Qiu, X. Xu, X. Wei, W. Wang, "Ultrasonic Detection and Classification for Internal Defect of Rail Based on Deep Learning," *Journal of the China Railway Society*, Vol. 43, No. 04, pp. 108-116, 2021

[13] J. Luo, X. Yu, J. Cao, W. Du, "Intelligent Rail Flaw Detection System Based on Deep Learning and Support Vector Machine," *Electric Drive for Locomotives*, Vol. 02, pp. 100-107, 2021

[14] Z. Chen, Q. Wang, K. Yang, T. Yu, J. Yao, Y. Liu, P. Wang, Q. He, "Deep Learning for the Detection and Recognition of Rail Defects in Ultrasound B-scan Images," 2021

[15] K. Han, Y. Wang, H. Chen, X. Chen, J. Guo, Z. Liu, Y. Tang, A. Xiao, C. Xu, Y. Xu, Z. Yang, Y. Zhang, D. Tao, "A survey on visual transformer," arXiv preprint arXiv:2012.12556, 2020

[16] Y. Guo, L. Huang, Y. Liu, J. Liu, G. Wang, "Establishment of the Complete Closed Mesh Model of Rail-Surface Scratch Data for Online Repair," *Sensors* 2020, 20, 4736, https://doi.org/10.3390/s20174736

[17] TB/T1778-2010 Rail Damage Classification, in Chinese

[18] The Illustrated Transformer Online: https://jalammar.github.io/illustrated-transformer/

[19] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, L. Kaiser, "Attention is all you need, " Advances in neural information processing systems, pp. 5998-6008, 2017

[20] J. Devlin, M. Chang, K. Lee, K. Toutanova, "Bert: Pre-training of deep bidirectional transformers for language understanding," arXiv preprint arXiv:1810.04805, 2018

[21] A. Bhandare, V. Sripathi, D. Karkada, V. Menon, S. Choi, K. Datta, V. Saletore, "Efficient 8-Bit Quantization of Transformer Neural Machine Language Translation Model," 2019

[22] S. Khan, M. Naseer, M. Hayat, S. Zamir, F. Khan, M. Shah, "Transformers in vision: A survey," arXiv preprint arXiv:2101.01169, 2021

[23] N. Parmar, A. Vaswani, J. Uszkoreit, L. Kaiser, N. Shazeer, A. Ku, D. Tran, "Image Transformer," 2018

[24] A. Dosovitskiy, L. Beyer, A. Kolesnikov, D. Weissenborn, X. Zhai, T. Unterthiner, M. Dehghani, M. Minderer, G. Heigold, S. Gelly, J. Uszkoreit, N. Uszkoreit, "An image is worth 16x16 words: Transformers for image recognition at scale," arXiv preprint arXiv:2010.11929, 2020

[25] H. Neil, W. Dirk, "Transformers for Image Recognition at Scale, "Online: https://ai.googleblog.com/2020/12/transformers-for-image-recognition-at.html, 2020

[26] ViT (Vision Transformer) Online: https://paddlepedia.readthedocs.io/en/latest/tutorials/computer_vision/classification/ViT.html