International Journal for Housing Science and Its Applications
Publish August 6, 2025. Volume 46, Issue 3 Pages 8107-8119

https,//doi.org/10.70517/ijhsa463693

Research on trackside equipment identification and automatic detection based on multi-scale convolutional neural network

Duanyang Cai[1], Huafeng Zhuge[1], Ru Wang[1,*], Cong Wu[1], Lu Shen[1] and Guo Zhang[2]
[1] Zhejiang Haining Rail Transit Operation Management Co., Ltd., Haining, Zhejiang, 314400, China
[2] Chengdu Tangyuan Electric Co., Ltd., Chengdu, Sichuan, 610000, China
Corresponding authors: (e-mail: 17853291705@163.com).

Abstract The operating mileage of high-speed railroad is growing rapidly, and the detection of electric trackside equipment still mainly relies on manual visual inspection, which has the problems of shortage of personnel, low efficiency, low accuracy, and great influence by the environment. This study proposes a high-speed railroad trackside equipment identification and automatic detection method based on multi-scale convolutional neural network, which aims to solve the problems of low efficiency and poor accuracy of traditional manual inspection. The study adopts a modular-designed high-speed industrial camera for data acquisition, and constructs a dataset containing a total of 3274 pictures of five types of equipment, namely, choke transformer voltage box, cable diverter box, cable terminal box, transformer box and signaling machine. Based on the Faster-RCNN framework, ResNet101 is selected as the backbone network, and the trackside equipment detection model is designed by feature pyramid network, RoI pooling and improved loss function. The experimental results show that the model achieves an average accuracy of 97.65% in the detection of five types of trackside equipment, and the processing speed is 21.42 frames/second. Compared with other detection algorithms, this model improves the recognition accuracy, and the introduction of the feature pyramid network improves the average accuracy of the model by 4.17%. In addition, the detection accuracy is significantly improved by increasing the candidate region size to {128,256,512}. The proposed multi-scale convolutional neural network method provides an effective solution for the automated detection of trackside equipment in high-speed railroads, and provides technical support to ensure the safety of railroad operation.

Index Terms Multiscale convolutional neural network, Trackside equipment, Automatic detection, Faster-RCNN, Feature pyramid network, ResNet101

I. Introduction

The purpose of daily inspection operation of high-speed railroad is to timely investigate and solve the hidden safety problems in the process of train operation, and then ensure the safety of train operation [1], [2]. With the rapid development of high-speed railroad, railroad construction intelligence has become an inevitable trend of signal professional construction, the integration of emerging technologies into the traditional construction technology, to strengthen the digitalization of high-speed railroad construction, optimize the construction equipment intelligence, can promote the development of high-speed railroad at a high level [3]-[5].

High-speed railroad signal trackside equipment mainly includes signaling machine, transponder, LEU, choke transformer, rail transformer box, double body box, cable terminal box and so on [6]. In the construction process, the positioning and detection of trackside equipment still need to rely on manual work, and there are problems such as low construction efficiency, poor safety, and low accuracy [7]. Therefore, there is an urgent need to develop smarter trackside equipment fixing and measuring devices to realize the automatic positioning and re-testing of trackside equipment installation position in different scenarios, and at the same time, complete the measurement of trackside equipment limit data, to ensure that the trackside equipment is installed correctly and reliably [8]-[10]. Among them, high speed rail trackside equipment number is generally composed of English letters, numbers and some special characters [11]. High-speed rail trackside equipment number arrangement is more regular, and the font size is basically consistent, belongs to the more regular text scene, can use the target detection model to realize the automatic detection of equipment number area image [12]-[14].

Railroad transportation as a national important infrastructure, its safe and stable operation is related to economic development and people's travel safety. In the high-speed railroad system, trackside electrical equipment bears the signal transmission, power supply control and other key functions, and its working condition directly affects the safety and efficiency of train operation. At present, the detection and maintenance of trackside equipment of the electrical department mainly rely on the staff to enter the track area at night to carry out manual inspection, this way is not only a harsh working environment, high danger, and low detection efficiency, high leakage rate, strong subjectivity, it is difficult to meet the growing operational density and safety requirements of high-speed railroads. In recent years, artificial intelligence and machine vision technology have made significant progress in the field of industrial inspection, providing a new technical path for the automated inspection of trackside equipment. Deep learning algorithms, especially convolutional neural networks, have demonstrated superior performance in the field of target detection, and their powerful feature extraction and classification capabilities make them ideal for solving the problem of equipment identification in complex scenes. However, high-speed railroad trackside equipment has a wide variety of types, large differences in size, complex and changing installation environments, and unstable lighting conditions, all of which bring challenges to automated detection. How to construct a robust detection model to realize accurate identification and state assessment of trackside equipment is a key issue that needs to be solved urgently.

Based on the above analysis, this study proposes a trackside equipment identification and automatic detection method based on multi-scale convolutional neural network. Firstly, a data acquisition system suitable for high-speed railroad scenarios is designed, and a large-scale dataset covering five types of key trackside equipment is constructed. Then, Faster-RCNN is used as the basic framework, ResNet101 is selected as the backbone network, feature pyramid network is introduced to enhance the detection ability of targets of different sizes, and an improved loss function is designed to improve the model training effect. Finally, the effectiveness of each component of the model is verified through multi-group comparison experiments and ablation experiments, and the performance under different parameter settings is systematically evaluated, which provides a parameter optimization basis for the practical application of the model.

II. Introduction to data acquisition and algorithms

With the gradual construction and improvement of the railroad network, high-speed rail operation speed gradually “up to speed”, the daily operation of the railroad electrical department and the sky window construction requirements are increasingly high. At present, the operation and maintenance mode of the electric power department is relatively single, and the detection methods and maintenance modes of some trackside equipment and facilities are still stuck in "fault repair" and "planned repair", and the inspection and maintenance of signal and communication equipment and facilities such as stations, sections, bridges, and tunnels are completed through large-scale and long-distance "scattering" working methods. Low degree of mechanization, high leakage rate, low efficiency of operation and maintenance, personal safety can not be effectively guaranteed, can not meet the increasingly busy operational needs of the railroad and the need for high security, but also does not meet the requirements of the development of safety, green and efficient.

II. A. Optical Configuration and Data Acquisition

II. A. 1) Acquisition hardware selection

The overall design of high-speed railroad trackside equipment safety inspection device adopts modular design, convenient to use when assembled, imaging system industrial grade high-frequency high-speed camera, and supplemented by light compensation, automatic control and other technologies to realize the automatic trigger control of the camera triggering, shooting image acquisition and storage, and through the human-computer interface to realize the management personnel to inspect the device and the shooting of the situation of control.

In order to realize the comprehensive integration of the various professional needs, the device selects the detection imaging range for the contact network pillar root to the field side extension of 3000mm, in the tunnel to cover the tunnel wall above ground 2400mm, the imaging range basically covers the professional trackside inspection range. At the same time, if the imaging angle adjustment device is set up, the angle can be adjusted according to personalized needs, in order to clearly reflect the differentiation of high-speed railroad trackside professional signs and markers, trench covers and auxiliary driving class, safety protection and other equipment abnormal state. Including deformation, missing, loose, parts of the rotation, displacement and cracks and other defects, the device selected Baumer high-speed industrial camera for imaging, the camera resolution of 5000 × 4000px, full frame shooting frame rate of 15fps, the sensor type of 36mm CMOS. If the assumption that the lens selected for the 60mm, then after calculating the inspection process, the minimum distance between the object 5.5m, the field of view of the camera is 5.5m, the camera's field of view is 5.5m, the camera's field of view is 5m, the camera's field of view is 5m. m, the field of view of the camera is 3500mm×2500mm, if the camera is set up in the existing high-speed rail contact network inspection car, then the length ratio of the camera imaging photo and high-speed rail line per unit of time is about 1.95, which fully meets the requirements of trackside equipment safety inspection.

According to the minimum 5.5m object distance and 60mm focal length lens for trackside equipment inspection, the actual length of each pixel on the imaging photo of the camera is 0.52mm×0.34mm, and the smallest device of the trackside equipment is the M13 fixing bolt, whose outer circle diameter is about 23mm, which is equivalent to about 45 effective pixels for imaging, and it is able to recognize whether the fixing bolt nut on the trackside equipment is dislodged or not, and can predict in advance whether the trackside equipment is safe or not. It can fully recognize whether the fixing bolt nut on the rail-side equipment is falling off, and can prejudge the defects of the rail-side equipment in advance.

This device is designed to use external hard disk for high-speed data storage, through the detection of the actual shooting image, the size of each photo is about 0.2MB, according to the 500km single line detection range, it is about 600,000 photos need to be stored, the need for storage space of 60G, if you take into account the changes in the speed of the train and the detection of the number of kilometers increase, it is necessary to further expand the storage space, plus High-speed storage, so choose 1TB SSD high-speed hard disk as the storage medium. The control, parameter design and monitoring of the camera are carried out by software, which is developed based on VC and utilizes the Baumer camera's dual Gigabit Ethernet interface to realize the control of the camera, relevant parameter settings, real-time data acquisition and storage, and other functions.

II. A. 2) Data acquisition and processing

Based on the trackside equipment data acquisition device given above, the categories of trackside equipment dataset in this paper include choke transformer box (BEX), cable diverter box (HF), cable terminal box (HZ), transformer box (XB) and signaling machine (Light).

In this paper, the dataset is manually labeled using the annotation tool LabelImg, and the device numbered area in each picture is framed with a rectangle to get the real location of the inspection object. The sample composition of the data set produced in this paper is shown in Table 1. In addition, the data labeling tool can automatically export the Label.txt file, which records the labeling results of all the samples in the dataset. Among them, one line of records is the labeling of all samples in a picture, from left to right, the relative path of the picture file, the specific content of the device number, and the pixel coordinates of the four points on the top, left, right and bottom of the real frame.

Table 1: Distribution of sample data volume

Label	Total data volume	Training data volume	Verify data volume
BEX	1031	722	309
HF	1279	895	384
HZ	963	674	289
XB	1184	828	356
Light	1172	820	352

The dataset produced in this paper contains 3274 pictures of trackside equipment of high-speed railroads, labeled with 5629 data samples, covering five categories, namely, choke transformer box, cable diverter box, cable terminal box, transformer box and signaling machine, which can provide data samples for the subsequent research work. In addition, the dataset produced in this paper can also facilitate research work in other related fields.

II. B.Introduction to Related Algorithms

II. B. 1) Convolutional Neural Networks

Convolutional Neural Network (CNN) is a deep learning model specialized for processing two-dimensional data, and has shown excellent performance especially in the field of image processing. A convolutional neural network extracts local features of an image through a hierarchical structure and abstracts them layer by layer into higher-level features. Its key components include convolutional layer, pooling layer, fully connected layer and activation function [15].

(1) Convolutional Layer

In convolutional neural networks, the convolution operation is one of the core components. The basic principle is to extract specific patterns or features in an image by weighting and summing each local region of the input image with a sliding convolutional kernel (also known as a filter). The convolution kernel is usually a small matrix that slides over the image pixel by pixel, multiplying the localized pixel values with the weights of the convolution kernel and summing the results to produce a new feature map. This operation enables the extraction of features at different levels from low-level edge information to high-level semantic information by applying the convolution kernel layer by layer.

The mathematical expression for convolution is:

$$Y(i,j)=\sum_{m=1}^{M}\sum_{n=1}^{N}X(i+m,j+n)\cdot K(m,n) \tag{1}$$

where $Y(i,j)$ is the pixel value of the output feature map, $X(i+m,j+n)$ is the pixel value of the input image and $K(m,n)$ is the weight matrix of the convolution kernel. Here M and N denote the number of rows and columns of the convolution kernel, respectively.

(2) Pooling layer

Pooling layer is a dimensionality reduction technique that is mainly used to reduce the size of the feature matrix, thereby reducing the computational complexity while retaining the important features. Common types of pooling operations include maximum pooling and average pooling. Maximum pooling retains salient features by extracting the maximum value in a local region, while average pooling calculates the average value in a local region to retain the overall information.

(3) Fully Connected Layer

The fully connected layer is responsible for integrating the local features extracted by the convolutional layer into global features and is used for final classification. The main task of the convolutional layer is to extract the local features of the image, while the fully connected layer generates a global understanding of the entire image by linearly combining these features. Usually, the fully connected layer is located in the last layers of the convolutional neural network and acts as a classifier, mapping the extracted features to specific output categories [16].

The fully connected layer works by processing the input feature vectors through a linear transformation, multiplying the input with a weight matrix and adding a bias to output the results of a classification or other task. Its computational formula is:

$$y_j=f\left(\sum_{i=1}^{n}W_{ji}x_i+b_j\right) \tag{2}$$

where W_{ji} denotes the weight parameter between the input node i and the output node j, which constitutes the weight matrix $W\in R^{m\times n}$, b_j is the bias term associated with the output node j, $f(\cdot)$ is the nonlinear activation function, e.g., ReLU, Tanh, etc., x_i is the input vector of the i th component from the feature output of the previous layer, and y_j is the j th component of the output vector y, which represents the final result obtained from the computation.

II. B. 2) ResNet network

As the network deepens, the optimization becomes worse and the accuracy of the test and training data decreases due to the problem of gradient explosion and gradient vanishing caused by the deepening of the network.The ResNet network constructs the residual module by “shortcutting”, where the residual module consists of a convolutional layer fitted to 1 and a shortcut constant mapping 2. The residual module consists of a convolutional layer fitted with $F(x)$ and a Shortcut constant mapping x, which defines the features learned by the residual module $H(x)$ as:

$$H(x)=F(x)+x \tag{3}$$

The ResNet family of models is a deep residual structure model, where the residual module can be understood as a subnetwork, and a very deep network model can be built by stacking subnetworks.ResNet is proposed to alleviate the problems of gradient vanishing, gradient explosion, and network degradation to a certain extent.There are two types of ResNet blocks, a two-layer structure applied to the ResNet18 model and the There are two types of ResNet blocks, a two-layer structure used in the ResNet18 and ResNet34 models, where the main branch at the core of the residual structure is built on top of two 3×3 convolutional layers, and the connecting lines located on the right side of the residual structure form the Shortcut branch, which is also known as the shortcut branch. Another three-layer structure is applied to the ResNet50 and ResNet101 models, which uses three convolutional layers for the main branch of the residual structure [17]. The first is a 1×1 convolutional layer used to compress the dimensionality of the channel so that the second is a 3×3 convolution unaffected by the input of the previous layer and the output does not affect the next layer, and the third is a 1×1 convolutional layer used to reduce the dimensionality of the channel. The first and second convolutional layers on the main branch use the same number of convolutional kernels, the third layer is 4 times the first layer, also on the shortcut branch there is a 1×1 convolutional layer which has the same number of convolutional kernels as the third convolutional layer on the main branch. The middle 3 × 3 convolutional layer is first reduced under a reduced 1 × 1 convolutional layer and then the reduction is done under another 1 × 1 convolutional layer. Both maintain the model accuracy and reduce the network parameters and computation, saving computation time.

III. Trackside equipment identification and detection model

The rapid growth of high-speed railroad mileage, compared with the lack of professional maintenance equipment and maintenance personnel is increasingly prominent. For the inspection of mechanical characteristics, it still mainly relies on the original manual visual inspection. Artificial visual inspection is mainly through the manual way on the road, assisted by portable lighting equipment in the sunroof time along the line walking inspection of electrical trackside equipment. Artificial visual inspection has problems such as shortage of personnel, low inspection efficiency, low inspection accuracy, and great influence by low temperature. In order to ensure the healthy, orderly and efficient development of high-speed railroad, for the electric trackside equipment which directly affects the safety of traveling, the inspection equipment should be developed vigorously, and the advanced machine vision technology should be used to improve the safety of the electric trackside equipment, reduce the manpower cost, improve the inspection efficiency and accuracy, and reduce the influence on the personnel due to the low temperature and other environmental factors.

III. A. Trackside equipment identification and detection model design

III. A. 1) Multi-scale feature extraction

A multiscale hierarchical feature is a scene-level feature that is invariant and consistent in scale space, allowing larger image environments (which can be as large as the entire scene) to be applied to local recognition decisions, containing appropriately centered and scaled targets and hierarchical natural attributes that provide a good basis for predicting potential target classes.

In visual analysis, pixels form edge segments, edge segments form patterns, patterns form objects, and objects form scenes. Convolutional neural networks provide a simple framework to learn such visual hierarchical features. In this study, in order to realize the category prediction of each pixel target in an image, this paper combines CNN with a multi-scale approach, i.e., the CNN network weights are replicated to multiple scales in the scale space.

Given an input image I, a fast local Laplace filter is applied to it to enhance the details of the target image. Construct a multi-scale pyramid X_s of this image, $s \in \{1,2,\cdots,N\}$, where X_1 is in the same scale as the original image I and has the same dimensions. The multi-scale pyramid constructed in this paper is a Gaussian pyramid, which is computed by the method proposed by Burt et al. The resulting image area of each layer is $1/5$ of the upper image, and the scaling down of the image resolution is 1.

After normalization, the local neighborhood of each image in the multi-scale pyramid is made to have 0 mean and unit standard deviation. Then given the CNN model f_s, and set its internal parameters as θ_s, the convolutional neural network is composed of a combination of CNN models corresponding to each scale image, and all model parameters are shared across the scales, i.e:

$$\theta_s = \theta_0, s \in \{1,2,\cdots,N\} \tag{4}$$

where θ_0 is the initial parameter of the model.

Under scale s, for a multi-scale convolutional neural network f_s with L stages exists:

$$f_s\left(X_s;\theta_s\right) = W_L H_{L-1} \tag{5}$$

where W_L is the weight matrix of the L th stage, H_{L-1} is the output of the $L-1$th stage, and there is $H_0 = X_s$. The output of each intermediate hidden stage L can be expressed as:

$$H_l = pool\left(\tanh\left(W_l H_{l-1} + b_l\right)\right), l \in \{1,2,\cdots,L-1\} \tag{6}$$

where pool function denotes the pooling operation, the maximum pooling method is used, $\tanh$ is the activation function, and W_l and b_l are the weight matrix and bias parameter vectors of the stage, respectively. Together, W_l and b_l constitute the trainable parameters θ_s of the CNN model.

Eventually, the output feature maps of all N CNN models are upsampled, uniformly sized and combined together in order to generate a 3-dimensional feature matrix F. At this point, F can be regarded as a multi-scale scene-level hierarchical image descriptor, represented as follows:

$$F=\left[f_1,u(f_2),\cdots,u(f_N)\right] \tag{7}$$

where u is the upsampling function.

III. A. 2) Backbone network selection

The backbone network of this paper is ResNet, which is one of the most popular and effective models, and its structure is simple and clear, and its residual structure has provided inspiration and ideas for many subsequent researches. 1×1 convolution kernel can greatly reduce the number of parameters, which can be used for dimension upgrading and downgrading. In this paper, we use ResNet101 as the backbone network. For ResNet101, the first layer of CONV_2 is a dashed residual structure because the output result of maximum pooling downsampling is [64,64,56].The input feature matrix of CONV_3 is [64,56,64] and the output is [32,32,128], and the first layer is a dashed residual structure.The first layers of CONV_4 and CONV_5 are also both dashed residual structure.

III. A. 3) Overall framework of the detection model

In this paper, target detection of trackside equipment images is based on Faster-RCNN, the core idea is to extract a number of regions containing higher probability of target from the inspection image to be detected through the region suggestion strategy, and then detect the target trackside equipment.The framework of the Faster-RCNN recognition and detection model is shown in Fig. 1.

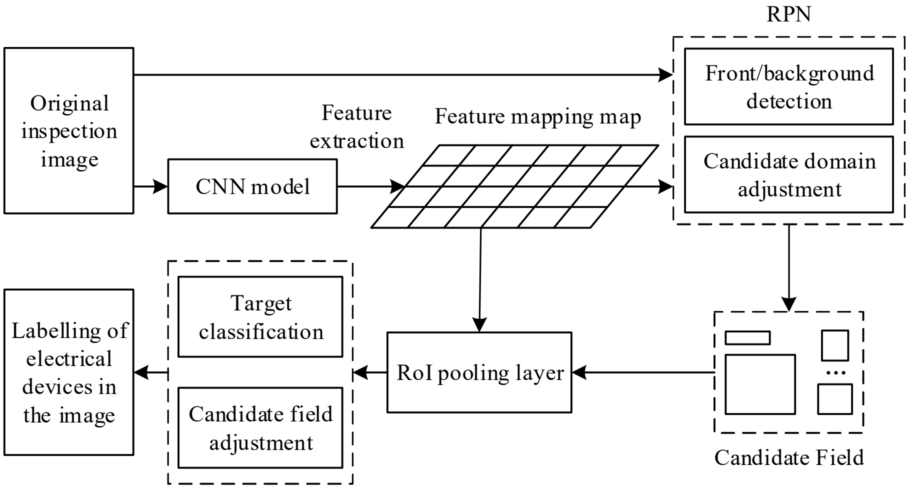
RPN
Original inspection image
Feature extraction
Feature mapping map
Front/background detection
Candidate domain adjustment
CNN model
Target classification
Candidate field adjustment
Labelling of electrical devices in the image
RoI pooling layer
Candidate Field

Figure 1: Faster-RCNN recognition and detection model

For any 1 image/frame of trackside equipment, firstly, feature extraction is performed by CNN to get the feature mapping map, and the region proposal policy network (RPN) uses SoftMax activation function to calculate the probability that each feature point in the mapping map belongs to the foreground (target), and meanwhile, a number of candidate domains with different dimensions are generated in the corresponding positions of the original image. The region of interest RoI pooling layer maps the candidate domains into feature vectors with fixed dimensions, and then SoftMax is used again to determine the object categories in the candidate domains. Among them, the feature mapping map serves as a location index, and when candidate domains of the same device are merged, the feature mapping map can assist in adjusting the size of the candidate domains and marking the trackside devices in the image.

III. B. Recognizing the Detection Model Component Modules

III. B. 1) Characteristic Pyramid Networks

RPN network is the main improvement of Faster RCNN over its predecessor algorithm, which serves to output high-quality candidate regions of different sizes and aspect ratios using the input image extracted by a convolutional neural network as input [18].

The RPN structure utilizes a sliding window based mechanism to determine candidate regions instead of generating candidate regions such as selective search. The convolutional feature layer is processed by the sliding window to obtain a 512-dimensional feature vector, when it is slid on the convolutional feature map, k anchor frames are obtained at each position with its center point as the anchor point, and then 2k fractional values and 4k coordinates are obtained by 2 full joins, where the fractional values are used to evaluate whether the candidate region contains the detection target or not, and the coordinates contain the candidate region length and width values w, h, and the center position coordinates x, y. The number, size and aspect value of anchor frames are related to the output effect of the whole network, in the original algorithm, the k value is taken as 9, i.e., 9 anchor frames are determined for each anchor point, with 3 scales of 128×128, 256×256 and 512×512, and the 3 aspect ratios of 1:1, 1:2 and 2:1, respectively.

The RPN loss function consists of the edge classification loss function L_{cls} and the edge regression loss function L_{reg}, and the three formulas are calculated as follows:

$$L_{cls}(p_i, p_i^*) = -\log\left[(p_i, p_i^*) + (1-p_i)(1-p_i^*)\right] \tag{8}$$

$$L_{reg}(t_i, t_i^*) = \sum_i smooth_{L1}(t_i, t_i^*) \tag{9}$$

$$smooth_{L1}(x) = \begin{cases} 0.5x^2 & |x| < 1 \\ |x| - 0.5 & |x| \geq 1 \end{cases} \tag{10}$$

$$L(p_i, t_i) = \frac{1}{N_{cls}} \sum_i L_{cls}(p_i, p_i^*) + \frac{1}{N_{reg}} \sum_i t_i^* L_{reg}(t_i, t_i^*) \tag{11}$$

where N_{cls} is the number of minimum batch images in the input network, N_{reg} is the total number of anchor coordinates, both are normalized weight parameters, $smooth_{L1}$ is the loss function, p_i is the probability that the i th anchor frame is a target, and p_i^* is the sample labels; t_i is the predicted border coordinates obtained, and t_i^* is the border coordinate of the real target; the calculation formula is as follows:

$$\begin{cases} t_x = \dfrac{(x - x_r)}{w_r}, t_y = \dfrac{(y - y_r)}{h_r} \\ t_w = \log\left(\dfrac{w}{w_r}\right), t_h = \log\left(\dfrac{h}{h_r}\right) \\ t_x^* = \dfrac{(x^* - x_r)}{w_r}, t_y^* = \dfrac{(y^* - y_r)}{h_r} \\ t_w^* = \log\left(\dfrac{w^*}{w_r}\right), t_h^* = \log\left(\dfrac{h^*}{h_r}\right) \end{cases} \tag{12}$$

where (x, y) are the predicted border coordinates, (x_r, y_r) are the candidate box coordinates, and (x^*, y^*) are the real target border coordinates, with which w and h are the box width and height.

III. B. 2) Loss function design

In this paper, there are five types of detection object types, namely, choke transformer box, cable diverter box, cable terminal box, transformer box and signal machine.The loss function of Faster RCNN consists of classification loss and regression loss. For the classification part, the cross-entropy loss is directly utilized, and for the regression loss of the edge position, Smooth_L1Loss is used.For each area candidate box its loss function is:

$$L\left(\{p_i\}, \{t_i\}\right) = \frac{1}{N_{cls}} \sum_i L_{cls}(p_i, p_i^*) + \lambda \frac{1}{N_{reg}} \sum_i p_i^* L_{reg}(t_i, t_i^*) \tag{13}$$

where i is the index of a candidate box in a small batch of data, and p_i is the predicted probability of candidate box i as a target, with the true label $p_i^* = 1$ if the candidate box is a positive sample, and $p_i^* = 0$ if the candidate box is a negative sample. λ is the balance weight normalization value, by default, set $\lambda = 10$. $t_i = \left\{t_x, t_y, t_w, t_h\right\}$ denotes the vector of the 4 parameterized coordinates of the predicted bounding box, and t_i^* is the vector of the true bounding box associated with the positive sample. N_{cls} is the number of mini-batches during training. N_{reg}

is the number of candidate frames. L_{reg} is the regression loss function. L_{cls} is the logarithmic loss over the two categories as:

$$L_{cls}(p_i, p_i^*) = -\ln\left[p_i p_i^* + (1-p_i^*)(1-p_i)\right] \tag{14}$$

The regression loss function L_{reg} is:

$$L_{reg}(t_i, t_i^*) = \sum_{i\in|x,y,w,h|} smooth(t_i, t_i^*) \tag{15}$$

where the smooth function is defined as follows:

$$smooth_{L1}(x) = \begin{cases} 0.5x^2 \dfrac{1}{\sigma^2} & |x| \le \dfrac{1}{\sigma^2} \\ |x| - 0.5 & Other \end{cases} \tag{16}$$

where x is the error in edge prediction and the parameter σ is used to control the region of smoothing.

III. B. 3) Model Optimization Strategy

In order to better improve the detection effect of the Faster-RCNN model on trackside device recognition, this paper uses the Stepwise Inductive Batch Normalization (DBN) algorithm for model optimization training. During neural network training, DBN updates the mean and variance used in the normalization process in a moving average, and the BN layer adjusts the output based on the mini-batch dataset that has been used by all networks in the past, not just the mini-batch dataset currently input into the network.

For a neural network, the objective function of neural network optimization after using DBN is:

$$\min \ell(\theta, \lambda) = \sum_{i=1}^{N} \ell_t\left(x_i : \theta, \mu, \sigma\right) \tag{17}$$

where μ, σ are the sliding mean and sliding mean standard deviation used by DBN, and $\theta = \left(W, \gamma, \beta\right)$ is the set of neural network training parameters.

The DBN algorithm selects one mini-batch dataset at a time on the training dataset $\left\{x_i \in R^K\right\}_{i=1}^{n}$ to train the network, and assumes that there are a total of J mini-batch datasets selected, and the j th mini-batch set is $X^j = \left\{x_i^j \in R^K\right\}_{i=1}^{m_j}$, $j \in \{1, \cdots, J\}$. The random variable corresponding to x_i^j is X_i^j, and the realization of the k th random component of the random vector X_i^j in the mini-batch dataset is $x_{i,k}^j$, $k \in \{1, \cdots, K\}$. DBN finds the mean and variance for each dimensional component of x_i^j, the mean vector on the j th mini-batch dataset $\mu^j = \left(\mu_{\cdot,1}^j, \cdots, \mu_{\cdot,K}^j\right)$, the variance vector $(\sigma^j)^2 = \left((\sigma_{\cdot,1}^j)^2, \cdots, (\sigma_{\cdot,K}^j)^2\right)$, and each dimension of the mean vector component is:

$$\mu_{\cdot,k}^j = E\left[X_{i,k}^j\right] = \frac{1}{m_j}\sum_{i=1}^{m_j} x_{i,k}^j \tag{18}$$

Each dimensional component of the variance vector is:

$$(\sigma_{\cdot,k}^j)^2 = Var\left[X_{i,k}^j\right] = \frac{1}{m_j}\sum_{i=1}^{m_j}\left(x_{i,k}^j - \mu_{B,k}^j\right)^2 \tag{19}$$

where μ, σ are the sliding mean and sliding mean standard deviation, which are computed from the mean and variance of all the mini-batch datasets input to the neural network in the previous $j-1$ iterations using the sliding average method. At the current j th iteration, the neural network inputs the j th mini-batch dataset and updates the DBN sliding average using the mean μ^j and σ^j as:

$$\mu = \alpha^j \mu^j + \left(1 - \alpha^j\right) \mu \tag{20}$$

and the sliding average standard deviation, i.e:

$$\sigma = \alpha^j \sigma^j + \left(1 - \alpha^j\right) \sigma \tag{21}$$

where α^j is a human-set hyperparameter.DBN normalizes the input x_i^j using the sliding mean and sliding mean standard deviation, and performs a translation and scaling transformation using γ^j and β^j to obtain an output vector as:

$$\tilde{x}_i^{j+1} = \gamma^j \times \frac{x_i^j - \mu}{\sigma} + \beta^j \tag{22}$$

where x is the input to the DBN layer, which can be the input vector of the neural network or the input vector of one of the hidden layers of the neural network. The training parameters θ^j (containing W^j, γ^j, β^j) of the neural network are then updated using the BP algorithm to obtain the new training parameters θ^{j+1}, and then the neural network undergoes the next training cycle with the input of the $j+1$th mini-batch data set. The use of DBN in the network is the same as when using BN, where the normalization process is performed in all layers of the network and is performed after the pre-activation and before the activation function.

IV. Experimental results and analysis

Train safety in high-speed railroad operation in addition to the most basic track devices, trackside equipment is also critical, high-speed railroad signals are transmitted by the trackside signal box, so the signal box and other trackside equipment for regular testing is also particularly important. With the current detection method investigation and analysis, the main way is still rely on workers to enter the train track side at night for one by one inspection and inspection, in the detection of the efficiency and artificial layer of the book have insufficient. This paper proposes a trackside equipment identification and detection model based on multi-scale neural network, which aims to improve the accurate detection of trackside equipment and provide help for the safe operation of trackside equipment.

IV. A. Analysis of identification and detection performance

IV. A. 1) Effect of different parameters

Faster-RCNN involves some parameters such as Dropout ratio, maximum number of iterations, batch size, and number of regions retained before and after NMS (non-maximum suppression), which have a large impact on the mean average precision (mAP). Tables 2 and 3 show the experimental results of mAP for Faster-RCNN model recognition detection when varying Dropout ratio and different number of NMS, respectively. In the table, MNI, BDB, and BIS denote the maximum number of iterations, the batch size of the region proposal stage, and the batch size of the detection stage, respectively.

In Table 2, when the percentage of dropout increases from 0.15 to 0.85, the mAP is generally decreasing, but there is a maximum value (0.835) at 0.65. There is no relevant theory to explain the effect of Dropout on mAP, and empirical values are usually taken. Let Dropout take the value of 0.64, change the number of candidate regions before and after NMS, and test its effect on mAP. According to Table 3, it can be seen that as the number of NMS decreases, the mAP also decreases gradually, which is because after NMS, the retained candidate region also decreases, leading to a decrease in the accuracy of the detection results. Therefore higher NMS can get better detection results. In addition, the number of candidate regions before and after NMS was taken as 2500 and 500, respectively, and the batch size was changed to test its effect on mAP. It was found that different batch sizes yielded different mAP showing an increasing trend. That is, as the batch size becomes smaller gradually, the mAP increases gradually. Related studies show that the optimization is fastest when the Batch_size is 2.

IV. A. 2) Identifying the Detection Confusion Matrix

In this paper, a total of five different types of devices are included in the collected trackside device dataset, i.e., choke transformer box (BEX), cable diverter box (HF), cable termination box (HZ), transformer box (XB) and signaling machine (Light). And the number of different device types in the dataset varies, after training the Faster-RCNN model using the training set, the validation set is used to verify the model recognition detection effect. In this paper, the confusion matrix is chosen as the evaluation index, and Figure 2 shows the confusion matrix for trackside device recognition detection of the Faster-RCNN model.

Table 2: Different dropout rate effects on mAP

Dropout	MNI	BDB	BIS	Before NMS	After NMS	mAP
0.15	6000	512	256	2500	500	0.842
0.25	6000	512	256	2500	500	0.825
0.35	6000	512	256	2500	500	0.831
0.45	6000	512	256	2500	500	0.784
0.65	6000	512	256	2500	500	0.835
0.75	6000	512	256	2500	500	0.796
0.85	6000	512	256	2500	500	0.779

Table 3: Different numbers of NMS effects on mAP

MNI	BDB	BIS	Before NMS	After NMS	mAP
6000	512	256	2500	500	0.831
6000	512	256	2100	400	0.824
6000	512	256	1700	300	0.815
6000	512	256	1300	200	0.812
6000	512	256	900	100	0.808
6000	512	256	500	50	0.803

The detection model is tested using data from the validator, and the average accuracy of the detection of the five types of trackside equipment in the statistical test set is between 97.5% and 99.5%. The number of correct identifications reached 302, 376, 282, 354, and 350 for the choke transformer box (BEX), cable diverter box (HF), cable termination box (HZ), transformer box (XB), and signaling machine (Light), respectively. This fully demonstrates that the trackside equipment recognition detection designed in this paper has a high accuracy rate and can provide reliable data support for the operation and maintenance of trackside equipment.

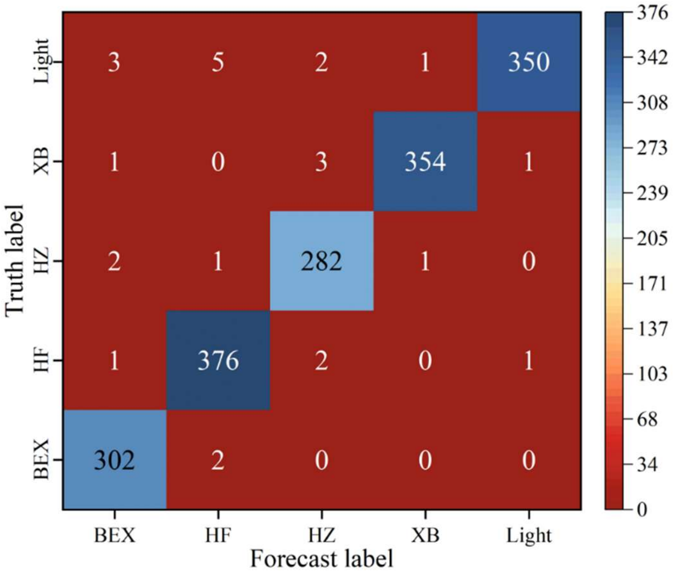

Figure 2: Identify and detect the confusion matrix

IV. A. 3) Impact of different backbone networks

In this paper, the algorithm is based on the ResNet101 network as the backbone network, and the Faster-RCNN model is established for the identification and detection of trackside devices. In order to verify the effectiveness of ResNet101 network, this paper chooses four basic networks, VGG16, DenseNet, ResNet18, and ResNet50, as comparisons, and compares the recognition and detection effects of the models under different backbone networks under multiple IoU thresholds. Table 4 shows the comparison results of the recognition detection effect of different models under the same IoU threshold.

From the table, it can be seen that under different intersection and concurrency ratio thresholds, under the same conditions, the detection accuracy of the model in this paper when ResNet101 is used as the backbone network are all performed optimally. Comparing the experimental results, it can be seen that the detection time of VGG16 is the shortest, but the detection accuracy is the worst. The detection accuracy of ResNet18~50 is improved while the detection time consumes longer. The use of DenseNet as a backbone not only improves the detection accuracy, but also reduces the detection time compared to ResNet because the transmission of parameters reduces the amount of computation. It can be concluded that the deep optimization of the basic network structure can improve the detection accuracy of the model, but the deepening of the network brings an increase in the number of parameters and sacrifices part of the detection time. The use of ResNet network as the backbone network of the identification detection model, which simplifies the connection of Dense block and achieves faster detection speed while guaranteeing the model detection accuracy, has obvious advantages in the identification and detection of trackside equipment.

Table 4: Effect comparison of different algorithms under the same IoU

Network	mAP/%			Detection time/ms
	IoU=0.4	IoU=0.6	IoU=0.8	
VGG16	53.02	35.48	14.25	100.21
DenseNet	63.91	48.96	22.38	164.85
ResNet18	63.42	42.15	21.47	152.36
ResNet50	65.16	43.47	20.59	178.59
ResNet101	68.57	49.81	24.62	142.17

IV. B. Model comparison and ablation experiments

IV. B. 1) Comparison experiment of different algorithms

In order to verify the reliability of the Faster RCNN model in performing trackside device identification and detection, it is compared with the first-order YOLO algorithm and several different types of models. The corresponding mAP values and detection speeds for each model are shown in Table 5. Comparing the experimental data, the detection speed of the YOLO algorithm is significantly better than the basic CNN and RCNN algorithms. Among them, the improved PPYOLO algorithm based on YOLOv3 is the fastest, reaching 24.51 frames/sec, and its accuracy is better than the YOLOv3 algorithm. However, further observation of the output image recognition results reveals that the YOLO model is prone to ignoring small-sized targets, resulting in target omission, which needs to be avoided in engineering applications.

For the Faster RCNN algorithm, the mAP is lowest (92.01) when VGG16 is the backbone network. It can be found that the more layers of the backbone network, the higher the accuracy, but the detection speed decreases with it. The mAP of the model using ResNet101 with more layers as the backbone network reaches 97.65%, which is 4.41% higher than that of the model using ResNet18, but it is not obvious relative to that of the model using ResNet50, and it can process 21.42 frames of images per second. The trackside equipment identification and detection model designed in this paper meets the requirements of trackside equipment identification and can provide reliable data support for the accurate detection and identification of trackside equipment.

Table 5: Results of different target detection models

	Accuracy/%	Recall/%	F1/%	mAP/%	FPS
CNN	84.75	80.34	20.62	88.16	14.42
RCNN	85.64	82.71	21.04	90.48	14.83
YOLOv3[ResNet30]	86.32	84.06	21.29	93.24	22.38
YOLOv3[DarkNet50]	86.79	85.93	21.59	94.56	22.79
PPYOLO[ResNet50]	87.85	86.72	21.82	95.03	24.51
Faster RCNN[VGG16]	80.17	85.84	20.73	92.01	18.16
Faster RCNN[ResNet18]	81.24	89.29	21.29	93.24	22.64
Faster RCNN[ResNet50]	82.95	90.36	21.62	96.79	21.85
Faster RCNN[Resnet101]	89.46	92.15	22.69	97.65	21.42

IV. B. 2) Analysis of model ablation experiments

In the Faster RCNN model established in this paper, it mainly contains four components: the ResNet101 backbone network, the feature pyramid network, RoI pooling, and the loss function. In order to study the effect of different modules on improving the model recognition detection performance, this paper designs the ablation experiment. Table 6 shows the results of the model ablation experiments.

As can be seen from the table, on the basis of ResNet101 backbone network, after continuously adding the feature pyramid network, RoI pooling and loss function, the average accuracy value of the model for trackside equipment shows a significant upward trend, and its value is improved from 90.21% in model A to 97.65% in model E, with an overall improvement of 7.44%. This illustrates the feasibility of the individual modules in the model, with the feature pyramid network showing the greatest improvement of 4.17% compared to the base ResNet101 backbone network. The feature pyramid network uses the input image extracted by the convolutional neural network as input and outputs high quality candidate regions of different sizes and aspect ratios. This also shows the feasibility and effectiveness of the feature pyramid network introduced in the Faster RCNN model in this paper. Although the speed of the model in performing trackside device recognition detection decreases from 22.35 frames/sec in Model A to 21.42 frames/sec in Model E, the decrease in detection speed is within an acceptable range relative to the increase in average recognition accuracy.

Table 6: Model ablation experiment analysis

Model	ResNet101	RBN	RoI	Loss	mAP/%	FPS
A	√	×	×	×	90.21	22.35
B	×	√	×	×	94.38	22.06
C	×	×	√	×	92.73	21.83
D	×	×	×	√	93.46	21.56
E	√	√	√	√	97.65	21.42

On this basis, this paper further analyzes the detection effect of the model on different types under different candidate zone sizes. The candidate zone size is divided into three different sizes, {32,64,128}, {64,128,256} and {128,256,512}, to recognize and detect the five types of trackside equipment, and its specific detection results are shown in Table 7.

As can be seen from the table, by adjusting the size of the candidate frame from {32,64,128} to {128,256,512}, the size distribution is better adapted to the feature map feeling wild while better covering the size range of the objects in the dataset, and the Faster-RCNN, which uses ResNet-101 as the feature network, is able to recognize the five types of trackside equipment. -RCNN's detection AP value on trackside devices shows that the larger the size of the candidate frame, the higher the recognition accuracy of the model on trackside devices. This proves that a larger predefined anchor box can effectively improve the performance of the Faster-RCNN model's detection results on trackside devices with the detection of border regression.

Table 7: Comparison of parameters between different candidate region sizes

-	{32,64,128}	{64,128,256}	{128,256,512}
BEX	90.15	90.54	91.72
HF	90.27	90.63	91.08
HZ	90.38	90.71	90.94
XB	90.21	90.59	91.26
Light	90.19	90.42	91.51
mAP/%	90.24	90.58	91.32

V. Conclusion

The automatic high-speed rail trackside equipment detection system adopts a multi-scale convolutional neural network framework, which effectively solves the problems faced by traditional manual detection. Experimental validation shows that the Faster-RCNN model with ResNet101 as the backbone network performs well in trackside equipment recognition, with a detection accuracy of 97.65%, and the processing speed is maintained at 21.42 frames/second. Comparison experiments show that the ResNet101 backbone network improves significantly over other network models, with a 4.41% improvement in accuracy compared to the ResNet18 model. The ablation experiment confirms that the introduction of the feature pyramid network contributes the most to the model performance, improving the recognition accuracy by 4.17%. Candidate region size experiments show that the model recognizes best when the size is set to {128,256,512}, with an average accuracy of 91.32%. The detection accuracies of the five types of trackside equipment are all higher than 90%, among which the signaling machine and choke box have the highest recognition rate, reaching 91.51% and 91.72%, respectively. The trackside equipment dataset of 3274 images and 5629 data samples constructed in the study provides valuable resources for related fields. The successful application of multi-scale convolutional neural network in trackside equipment detection provides technical support for intelligent railroad inspection, which will significantly improve the maintenance efficiency and safety of the electrical department.

References

[1] Mao, Q., Cui, H., Hu, Q., & Ren, X. (2018). A rigorous fastener inspection approach for high-speed railway from structured light sensors. ISPRS Journal of Photogrammetry and Remote Sensing, 143, 249-267.

[2] Jiang, X., & Wang, S. (2021). Railway panorama: A fast inspection method for high-speed railway infrastructure monitoring. IEEE Access, 9, 150889-150902.

[3] Rui, Y., Ruifang, M., & Feng, J. (2018). A High-Speed Train Operation Plan Inspection Simulation Model. Mathematical Problems in Engineering, 2018(1), 9202986.

[4] Zhang, J., & Zhang, J. (2021). Comprehensive Evaluation of Operating Speeds for High-Speed Railway: A Case Study of China High-Speed Railway. Mathematical Problems in Engineering, 2021(1), 8826193.

[5] Gong, W., Akbar, M. F., Jawad, G. N., Mohamed, M. F. P., & Wahab, M. N. A. (2022). Nondestructive testing technologies for rail inspection: A review. Coatings, 12(11), 1790.

[6] Wu, N. (2018, March). High-speed railway signal trackside equipment patrol inspection system. In Young Scientists Forum 2017 (Vol. 10710, pp. 623-630). SPIE.

[7] Guo, G., Cui, X., & Du, B. (2021). Random-forest machine learning approach for high-speed railway track slab deformation identification using track-side vibration monitoring. Applied Sciences, 11(11), 4756.

[8] Zhou, L., Liu, X. Z., & Ni, Y. Q. (2019). Contemporary Inspection and Monitoring for High-Speed Rail. High-speed rail, 27.

[9] Yin, M., Li, K., & Cheng, X. (2020). A review on artificial intelligence in high-speed rail. Transportation Safety and Environment, 2(4), 247-259.

[10] Bustos, A., Rubio, H., Castejón, C., & García-Prada, J. C. (2018). EMD-based methodology for the identification of a high-speed train running in a gear operating state. Sensors, 18(3), 793.

[11] Gao, L., Jiu, Y., Wei, X., Wang, Z., & Xing, W. (2020). Anomaly detection of trackside equipment based on GPS and image matching. IEEE Access, 8, 17346-17355.

[12] Li, W., Li, J., & Liu, Y. (2020, August). Classification of trackside equipment based on convolutional neural network. In 2020 Chinese Control And Decision Conference (CCDC) (pp. 4806-4811). IEEE.

[13] Peng, X., Zeng, J., Wang, Q., & Zhu, H. (2023). Research on an identification method for wheelset coaxial wheel diameter difference based on trackside wheelset lateral movement detection. Sensors, 23(13), 5803.

[14] Li, C., Xie, Z., Qin, Y., Jia, L., Guan, L., & Ma, X. (2021, October). Abnormal Perception of Railway Perimeter Based on Trackside Surveillance Video. In International Conference on Electrical and Information Technologies for Rail Transportation (pp. 149-157). Singapore: Springer Singapore.

[15] Fares Al Mohamad,Leonhard Donle,Felix Dorfner,Laura Romanescu,Kristin Drechsler,Mike P Wattjes... & Keno Kyrill Bressem. (2025). Open-source Large Language Models can Generate Labels from Radiology Reports for Training Convolutional Neural Networks. Academic radiology,32(5),2402-2410.

[16] Amelia E.H. Bridges,Eleanor Cross,Kyran P. Graves,Nils Piechaud,Antony Raymont & Kerry L. Howell. (2025). Practical application of artificial intelligence for ecological image analysis: Trialling different levels of taxonomic classification to promote convolutional neural network performance. Ecological Informatics,88,103146-103146.

[17] Gyubin Lee,Sangun Kim,Ji seon Kim & Jooyong Kim. (2025). Deep Learning-Based Mapping of Textile Stretch Sensors to Surface Electromyography Signals: Multilayer Perceptron, Convolutional Neural Network, and Residual Network Models. Processes,13(3),601-601.

[18] Justice Kwame Appati & Isaac Adu Yirenkyi. (2024). A cascading approach using se-resnext, resnet and feature pyramid network for kidney tumor segmentation. Heliyon,10(19),e38612-e38612.