frontiers | Frontiers in Energy Research

TYPE Original Research
PUBLISHED 15 May 2024
DOI 10.3389/fenrg.2024.1345574

OPEN ACCESS

EDITED BY
Fuqi Ma,
Xi'an University of Technology, China

REVIEWED BY
Longchao Yao,
Zhejiang University, China
Wei Zhang,
Civil Aviation University of China, China

*CORRESPONDENCE
Guo Zhijun,
guozhijun5270000@126.com

RECEIVED 28 November 2023
ACCEPTED 18 April 2024
PUBLISHED 15 May 2024

CITATION
Zhijun G, Weiming L, Qiujie C and Hongbo Z (2024), Terminal strip detection and recognition based on improved YOLOv7-tiny and MAH-CRNN+CTC models .
Front. Energy Res. 12:1345574.
doi: 10.3389/fenrg.2024.1345574

Terminal strip detection and recognition based on improved YOLOv7-tiny and MAH-CRNN+CTC models

Guo Zhijun[1]*, Luo Weiming[1], Chen Qiujie[1] and Zou Hongbo[2]

[1]Dongguan Power Supply Bureau of Guangdong Power Grid Co., Ltd., Dongguan, Guangdong, China, [2]College of Electric Engineering and Renewable Energy, China Three Gorges University, Yichang, China

For substation secondary circuit terminal strip wiring, low efficiency, less easy fault detection and inspection, and a variety of other issues, this study proposes a text detection and identification model based on improved YOLOv7-tiny and MAH-CRNN+CTC terminal lines. First, the YOLOv7-tiny target detection model is improved by the introduction of the spatially invariant multi-attention mechanism (SimAM) and the weighted bidirectional feature pyramid network (BiFPN). This also improves the feature enhancements and feature fusion ability of the model, balances various scales of characteristic information, and increases the positioning accuracy of the text test box. Then, a multi-head attention hybrid (MAH) mechanism is implemented to optimize the convolutional recurrent neural network with connectionist temporal classification (CRNN+CTC) so that the model could learn data features with larger weights and increase the recognition accuracy of the model. The findings indicate that the enhanced YOLOv7-tiny model achieves 97.39%, 98.62%, and 95.07% of precision, recall, and mean average precision (mAP), respectively, on the detection dataset. The improved MAH-CRNN+CTC model achieves 91.2% character recognition accuracy in the recognition dataset.

KEYWORDS

terminal strip, improved YOLOv7-tiny model, convolutional recurrent neural network with connectionist temporal classification, spatially invariant multi-attention mechanism, weighted bidirectional feature pyramid network, multi-head hybrid attention mechanism

1 Introduction

A more significant piece of insulating equipment (Huang et al., 2023) in the secondary equipment (Zhong et al., 2023) of a substation is the secondary circuit terminal strip. It serves as a line transmission component, connects the equipment inside and outside the screen, and carries numerous groups of mutually insulated terminal components. The ability of the protection device to connect to the main equipment via the terminal strip is crucial, and the ability of the protection device to operate normally is directly correlated with proper wiring. Normalizing the terminal block can significantly lower the likelihood of accidents resulting from the secondary circuit and the frequency of wiring errors. Current worker point-to-point inspections are not only slow but also prone to incorrect and inadequate inspections (Liu et al., 2023). The rapid advancement of deep learning (Wang et al., 2018) has led to a surge in the use of image detection and recognition in power-related fields, including live detection, robot inspection of substations, and unmanned aerial vehicle

inspection of transmission lines. These applications benefit from the high stability and accuracy of the recognition features of the technology.

Currently, the advancement of deep learning in the field of artificial intelligence technology has progressively established the mainstream. In order to avoid the hidden hazard of substation operation, Zhou et al. (2018) integrated the efficient and accurate scene text (EAST) algorithm into the line end identification of the screen cabinet to identify the text information. This algorithm was then combined with manual experience judgment. By employing combined character placement and recognition, Wang et al. (2020) increased the text character recognition accuracy and expedited the recognition process. The accuracy of each module cannot be optimized by this training strategy; it can only improve the model overall performance. Wang and Yi (2019) trained YOLOv5 by incorporating the concept of structural clipping into the model and then pruned the model based on the training outcomes. Training precision was lost even if the model scale was shrunk and training speed increased. The maximum pooling layer was applied to the convolutional neural network (CNN) model by Masci et al. (2012) in order to increase the model recognition accuracy. Cui et al. (2013) used a pattern of template matching to ascertain the direction and location of the terminal row; nevertheless, it is limited to determining these two factors, and manual labor is still required for text recognition. Wang et al. (2019) suggested fresh segmentation results, processed the algorithm, and employed a progressive way to segment texts of varying scales. Although the detection speed is lost, the detection rate is increased. Yang et al. (2022) described a text identification model that combines support vector machines with quad-pronged splitting having strong positioning effects and good accuracy, but it is too complex to extract design elements. Although the detection speed was poor, Xiaoxuan et al. (2021) built a set of intelligent recognition systems based on the YOLOv3 network and paired it with deep transfer learning approach. The multi-dimensional long short-term memory recurrent neural network with connectionist temporal classification (MDLSTM-RNN+CTC) model was proposed by Messina and Louradour (2015) and applied to text line character identification. This approach incorporated feature information from four dimensions thoroughly; however, its recognition accuracy was not very excellent.

In response to the shortcomings of the existing detection and recognition models, this paper presents an innovative approach for terminal text detection and recognition that combines an enhanced YOLOv7-tiny model with a MAH-CRNN+CTC architecture. Initially, the proposed improved YOLOv7-tiny object detection model integrates the spatially invariant multi-attention mechanism (SimAM), which plays a pivotal role in enhancing the model capacity to discern and focus on essential features of the target while filtering out noise, thereby boosting the overall detection performance. Subsequently, the model adopts a weighted bidirectional feature pyramid network (BiFPN), which efficiently consolidates feature maps from varying scales. This strategy enables the bidirectional exchange of feature information and dynamic allocation of weights according to feature significance, further refining the model precision in detecting targets. The MAH-CRNN+CTC recognition model introduces a multi-head attention hybrid (MAH) mechanism. This component facilitates the comprehensive consideration of the entire sequence information, effectively addressing the issue of long-range dependencies. As a result, it accelerates the model training process, enhances feature extraction efficiency, and significantly boosts the model recognition accuracy.

The bidirectional long and short-term memory (Bi-LSTM) module offers a potent temporal modeling tool that empowers the model to decipher and leverage intricate contextual cues within the input sequences, thereby bolstering both the precision and resilience of the recognition process. Conversely, the CNN module concentrates on achieving end-to-end text recognition through multi-level analysis and abstraction of images, transforming intricate image data into sequential features compatible with subsequent processing by the bidirectional Bi-LSTM component.

This paper is structured as follows: Section 1 introduces the text detection module, elaborating on the YOLOv7-tiny network model and detailing the improvements made to it; Section 2 encompasses an introduction to the text recognition module, focusing on the enhanced methods employed for improving the text recognition model; Section 3 presents the analysis of the experimental outcomes for both the text detection and recognition processes; and lastly, Section 4 presents a summary of the paper.

2 Text detection

2.1 Improving the YOLOv7-tiny network model

YOlOv7 consists of three components: the neck, which fuses features, the head, which makes predictions, and the backbone, which extracts features (Wu et al., 2019). The major components of the feature extraction network of the YOLOv7 network are the MPConv, spatial pyramid pooling, cross-stage partial channel (SPPCSPC), E-ELAN, and Columbia Broadcasting System (CBS) modules. The E-ELAN module uses expand, shuffle, and merge cardinality to improve network learning while maintaining the original gradient path based on the original ELAN. After convolution of the feature map 3 times and 5 × 5, 9 × 9, and 13 × 13 maximum pooling, the SPPCSPC module uses the concept of spatial pyramid pooling to obtain image features under various receptive fields. This solves the issue of repetitive feature extraction from the image by the convolutional neural network. Subsequently, the characteristics of distinct receiving domains are combined, and following double convolution, they are ultimately split with the feature map. The MPConv module uses a 2 × 2 maximum pooling operation to increase the receptive field of the current feature layer. It then uses 1 × 1 convolution to adjust the number of channels. Finally, it fuses the feature information that has been processed with the feature information obtained by normal convolution to improve the feature extraction capability of the network. As the YOLOv7 feature fusion network, the path aggregation network (PANet) is utilized to fuse the deep semantic and shallow location characteristics of the image and produce feature maps of various sizes. The RepConv structure modifies the number of channels for characteristics with varying scales on the prediction side.

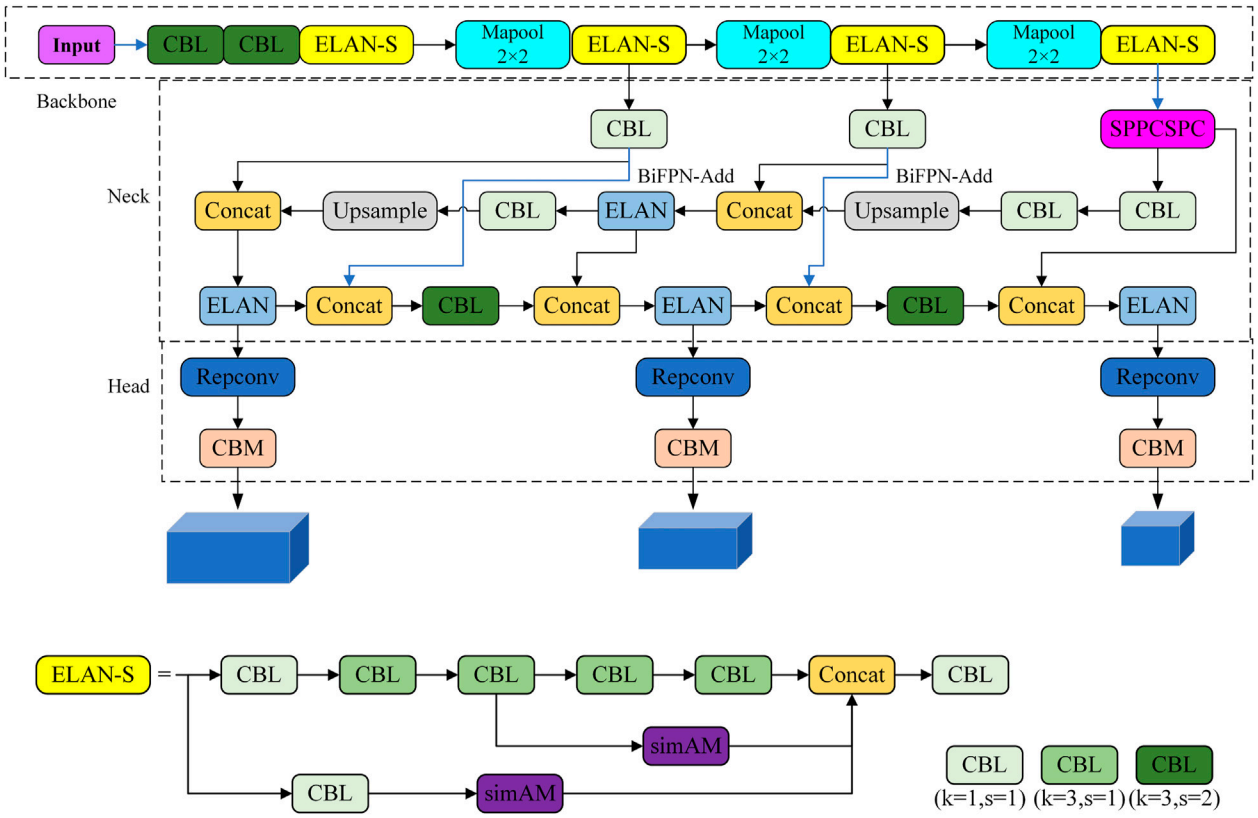


FIGURE 1
Improved YOLOv7-tiny network structure.

YOLOv7-tiny is an improvement over YOLOv7, as shown in Figure 1. ELAN-S is utilized in place of E-ELAN in the backbone section, and the SimAM module is added to the ELAN-S structure to improve the feature expression capabilities of the network. The Max pooling operation is exclusively used for down sampling, and the convolution process in MPConv is canceled. The BiFPN module is incorporated into the SPPCSPC structure for feature fusion in the neck section. The RepConv structure is still used in the head section to modify the number of channels for features with varying scales.

2.2 SimAM module

The module for attention mechanisms different from the channel attention mechanism and spatial attention that have been previously proposed, SimAM (also known as the SimAM module) (Yang et al., 2021) is a lightweight attention module that is both simple and incredibly effective. The SimAM module will not add further complexity to the network because it does not include any extra parameters. It is a feature map-derived 3D attention method. This module uses the energy function to optimize it in accordance with neuroscience theory and quickly arrive at an analytical solution; in other words, it uses the energy function to determine the attention mechanism weight. The energy function $e_t(*)$ is defined as follows:

$$e_t(w_t, b_t, y, x_i) = (y_i - \hat{t})^2 + \frac{1}{M-1}\sum_{i=1}^{M-1}(y_0 - \hat{x}_i)^2, \tag{1}$$

$$\begin{cases} \hat{t} = w_t t + b_t \\ \hat{x}_i = w_t x_i + b_t \end{cases}, \tag{2}$$

where t and x_i are the target neuron and other neurons of the input feature tensor X, respectively, and $X \in R^{C\times H\times W}$. C, H, and W are the pass number, height, and width of the feature tensor, respectively. i is the neuron index on a certain number of channels. M is the number of neurons on the channel, $M = H \times W$. w_t and b_t target neuron transform weights and bias, respectively. y, y_t, and y_0 are scalar quantities, of which y_t and y_0 are for different values; this paper introduced the binary label instead, with $y_t = 1$ and $y_0 = -1$.

Neurons inside the same channel can be trained to have their linear separability minimized by minimizing Eq. 1. By incorporating a regular term and employing binary labels, the energy function can be changed to

$$e_t(w_t, b_t, y, x_i) = \frac{1}{M-1}\sum_{i=1}^{M-1}[-1 - (w_i x_i + b_t)]^2 + [1 - (w_t t + b_t)]^2 + [1 - (w_t t + b_t)]^2 + \lambda w_i^2, \tag{3}$$

where λ is the regularization coefficient and w_i is the weight of the transformation of the i neuron.

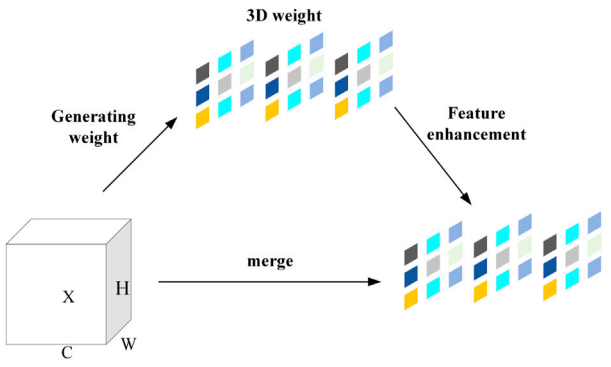


FIGURE 2
Spatially invariant multi-attention mechanism (SimAM) module structure.

Eq. 3 to Eq. 4.

$$\begin{cases} w_t = -\dfrac{2(t-u_t)}{(t-u_t)^2 + 2\sigma_t^2 + 2\lambda} \\ b_t = -\dfrac{1}{2}(t+u_t)w_t \end{cases}, \tag{4}$$

where u_t and σ_t^2 are both intermediate variables. Eq. 5,

$$\begin{cases} u_t = \dfrac{1}{M-1}\sum_{i=1}^{M-1} x_i \\ \sigma_t^2 = \dfrac{1}{M-1}(x_i - u_t)^2 \end{cases} \tag{5}$$

By substituting w_t and b_t into Eq. 2, the minimum energy e_t^* can be obtained, that is Eq. 6,

$$e_t^* = \frac{4\left(\hat{\sigma}^2 + \lambda\right)}{(t-\hat{u})^2 + 2\hat{\sigma}^2 + 2\lambda}. \tag{6}$$

u_t and σ_t^2 are replaced with the mean $\hat{u} = \frac{1}{M}\sum_{i=1}^{M} x_i$ and variance $\hat{\sigma}^2 = \frac{1}{M}\sum_{i=1}^{M}(x_i - \hat{u})^2$, respectively.

The lower the energy, the greater and more important the difference between the target neuron and the peripheral neuron t. The importance of neurons can be obtained by obtaining $\frac{1}{e_t^*}$, and then the enhanced feature tensor $\tilde{X}$ can be obtained by using Eq. 7:

$$\tilde{X} = sigmoid\left(\frac{1}{E}\right) \odot X, \tag{7}$$

where X is the input characteristic tensor. E is the sum of e_t^* in all channels and spatial dimensions. $\odot$ is the Hadamard product.

In Eq. 7, the sigmoid function is added to limit the excessive value of E, and the sigmoid function does not affect the relative importance of each neuron.

Figure 2 shows the SimAM module chart. It can be seen as a cell aimed at increasing the convolution characteristic expression ability of the neural network; any intermediate feature tensor can be taken as the input and the transformation output with the same size and have the feature of enhancing the characterization of the tensor, where X is the input feature tensor in the figure. The biggest advantage of this module is based on the defined energy function to choose from.

2.3 Weighted bidirectional feature pyramid network

As shown in Figure 3A top-down pathway of the feature pyramid network (FPN) allows for feature fusion. A certain amount of detection accuracy can be increased by the fused high-level semantic information. Prior feature fusion techniques frequently treated the feature information of various scales identically. Although it is impossible to determine the relative relevance of many input features, each contributes differently to the output features. This implies that the characteristics of some scales might be more significant and have a bigger influence on the outcome. Consequently, the weighted BiFPN is proposed in this research (Tan et al., 2020). As shown in Figure 3B, additional weights are applied for each input, utilizing a distinct blend of several input properties.

First, the nodes with a single input edge and little contribution are eliminated to simplify the network and decrease the amount of parameters. This effectively lowers the network complexity. Second, based on the properties of three distinct scales, the jump connection mechanism was established, increasing a feature fusion path in the quantity under the assumption of somewhat larger. Diagrams will be used to better integrate low-level and high-level semantic information, and weights can be used to focus network model studies on the most important informational properties, thereby enhancing network performance and characterization. The calculation of wighted feature fusion in BiFPN is represented by Eq. 8:

$$Out = \sum_{i=0} \frac{\omega_i * I_i}{\varepsilon + \sum_{j=0} w_j}, \tag{8}$$

where ω represents the learnable weight, I_i represents the input feature, and $\varepsilon = 0.0001$.

3 Text recognition methods

3.1 Improved CRNN+CTC algorithm

Text recognition is all that is needed to identify secondary device terminals. The convolutional recurrent neural network (CRNN) model not only performs well for more complicated texts, handwritten letters, and symbols but it also does not require segmenting the target to precisely mark the characters. It also has no restrictions on the length of the text sequence. There are not many model parameters, and training proceeds quickly. The model structure is thereby enhanced and optimized by making a reference to the network architecture of the traditional text recognition model or CRNN. Meanwhile, to better mine the long-distance data features of correlated time series, the MAH mechanism is introduced to the Bi-LSTM in the recurrent neural network module to accommodate the secondary equipment terminal strip identification of the substation. The network architecture of the substation secondary

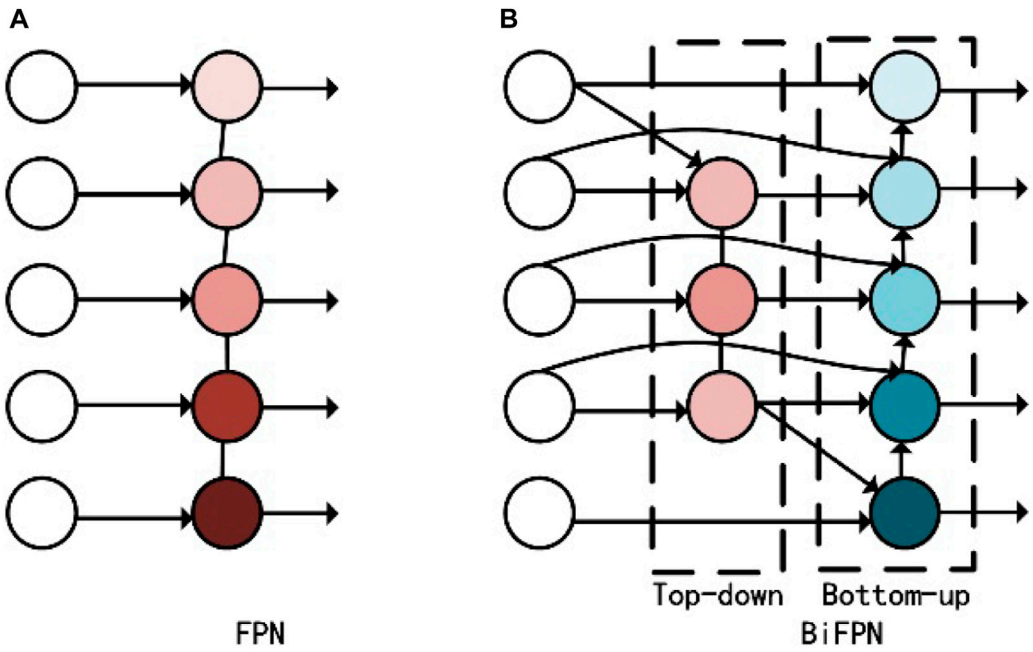


FIGURE 3
Improved feature structure pyramid.

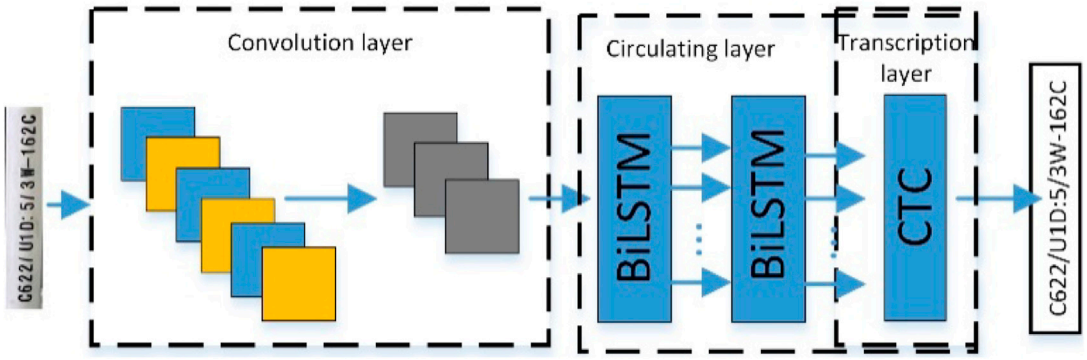


FIGURE. 4
Network architecture of the identification model of the secondary equipment terminal strip in the substation.

equipment terminal strip recognition model is shown in Figure 4. The three primary components of the terminal strip recognition model are the connection temporal classification (CTC), Bi-LSTM neural network, and convolutional neural network (CNN). These include the CNN for picture feature extraction, the Bi-LSTM for character sequence extraction, and CTC for character mismatch resolution.

3.2 Feature extraction network

The third and fourth max pooling kernel scales in the CNN module are set to 1 × 2 pixels, making it simple to use the CNN features that have been extracted as the recurrent neural network (RNN) input. To expedite the network training process, batch normalization layers are incorporated after the fifth and sixth layers of convolution. The original image height will be decreased to a fixed value of 32 pixels before it is entered into the CNN. The width of each feature vector in the feature sequence is set to 1 pixel, and they are all generated in the same direction as the feature map sequence, that is, from left to right. The first feature vector is linked to the first feature map.

In order to increase the network training speed, the BN layer is added to the CNN module in this research. The variable body of the ReLu function, known as the Leaky ReLu function, is adopted by the activation function. To address the issue of neurons not learning after the negative interval of the ReLu function, a leak value is added to the negative interval of the ReLu function, causing the output to slope slightly toward the negative input. As shown in Table 1, the CNN

TABLE 1 CNN network structure.

Network layer	Input size
Convolution layer	64 × 32 × 160
Maximum pooling layer	64 × 16 × 80
Convolution layer	128 × 16 × 80
Maximum pooling layer	128 × 8 × 40
Convolution layer	256 × 8 × 40
Maximum pooling layer	256 × 4 × 40
Convolution layer	512 × 4 × 40
Maximum pooling layer	512 × 2 × 40
Convolution layer	512 × 1 × 40

module gains 4 maximum pooling layers in this study, with the final 2 pooling layers having convolution kernel sizes of 1 × 2 pixels. The remaining convolution kernel sizes are 3 × 3 pixels and padding = 1, with the exception of the final convolution layer, which has a convolution kernel size of 2 × 2 pixels and padding = 0. The input image is processed in this article to create a 32 × 160-pixel image. After the CNN, the resulting feature map size is 512 × 1 × 40 pixels, meaning that there are 512 feature maps in total, each with a height of 1 pixel and a width of 40 pixels.

3.3 Sequence prediction network

The sequence properties in the sequence label distribution of each frame are predicted using the model prediction module. The RNN is highly proficient at capturing contextual connections in the realm of sequential text recognition. However, while processing lengthy texts, the standard RNN loses its ability to connect distant information and becomes vulnerable to the gradient disappearance issue, which makes the network difficult to converge and results in low training accuracy. An exceptional variety of the RNN that excels at acquiring long-term dependent data is the long short-term memory (LSTM). It can selectively recall the information that must be retained for a long time and forget the irrelevant information. It can also regulate the information transferred through the gate empty state, as shown in Figure 5.

Through its forgetting gate, input gate, output gate, and other gating structures, the LSTM cell structure may efficiently save and regulate the cell state update; the update rules are shown in Eq. 9. Eqs 9–14 illustrate how the gating unit is realized.

$$C_t = f_t * C_{t-1} + i_t * \tilde{C}_t, \tag{9}$$

$$f_t = \sigma\left(\boldsymbol{W}_f [\boldsymbol{H}_{t-1}, x_t] + \boldsymbol{b}_f\right), \tag{10}$$

$$i_t = \sigma\left(\boldsymbol{W}_i [\boldsymbol{H}_{t-1}, x_t] + \boldsymbol{b}_i\right), \tag{11}$$

$$\tilde{C}_t = \tanh\left(\boldsymbol{W}_c [\boldsymbol{H}_{t-1}, x_t] + \boldsymbol{b}_c\right), \tag{12}$$

$$O_t = \sigma\left(\boldsymbol{W}_O [\boldsymbol{H}_{t-1}, x_t] + \boldsymbol{b}_O\right), \tag{13}$$

$$\boldsymbol{H}_t = O_t * \tanh\left(C_t\right), \tag{14}$$

where $\boldsymbol{H}_{t-1}$ is the output of the hidden layer of the previous unit. x_t is the input of the current cell. f_t, i_t, and O_t represent the output of structures such as forgetting, input, and output in the gating structure, respectively. C_t, C_{t-1}, and $\tilde{C}_t$ represent the cell state of the current moment, the cell state of the previous moment, and the cell state of the output layer, respectively. $\boldsymbol{W}_i$, $\boldsymbol{W}_c$, and $\boldsymbol{W}_O$ are the corresponding weight parameters of the gate, respectively. $\boldsymbol{b}_f$, $\boldsymbol{b}_i$, $\boldsymbol{b}_c$, and $\boldsymbol{b}_O$ are the bias parameters corresponding to the gate, respectively. [,] is the vector connection symbol.

While the feature sequence recognition of the secondary device terminal number considers both the past and future context information to be beneficial, one-way LSTM only employs the past context information. Consequently, this article employs the Bi-LSTM network module, which uses the future information backward and the past information forward, as shown in Figure 6.

In this paper, a two-layer Bi-LSTM is set up. The output of the CNN is a feature map of size m×T, where T is the output sequence length of the feature module and m is the number of channels. After

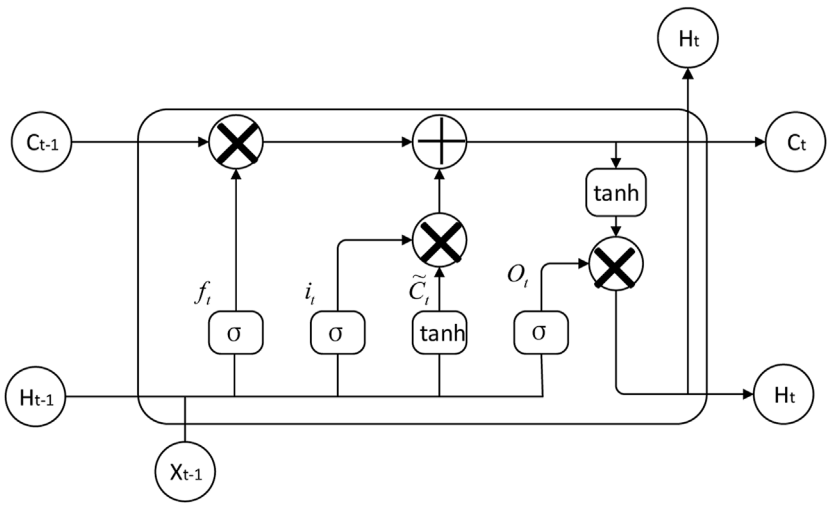


FIGURE 5
Long short-term memory (LSTM) cell structure.

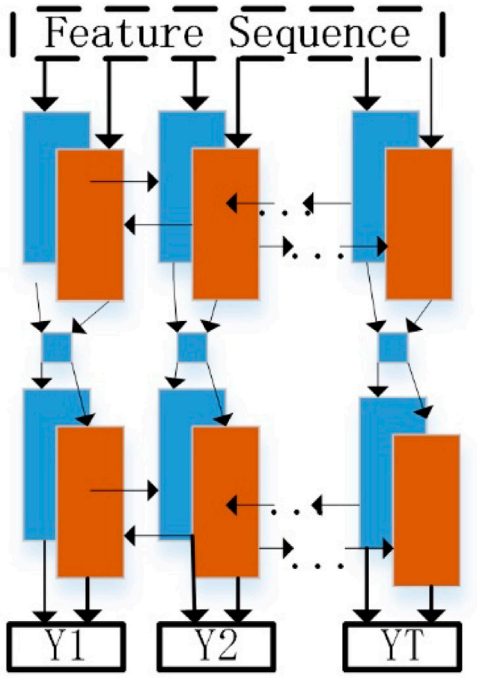


FIGURE 6
Structural diagram of the bidirectional (Bi)-LSTM.

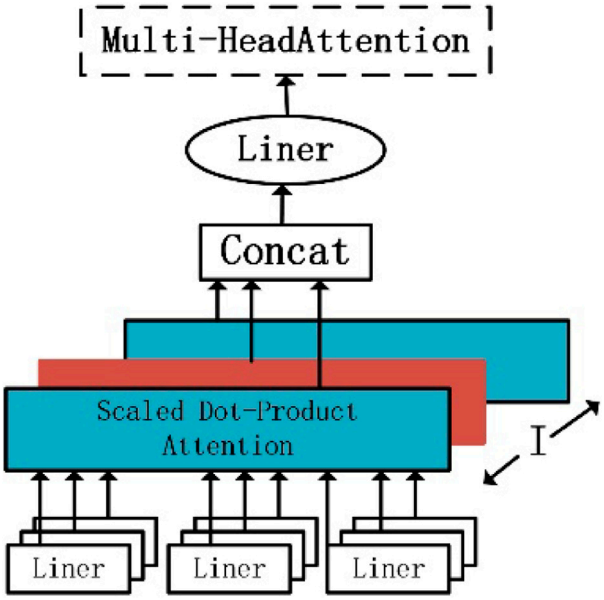


FIGURE 7
Structural diagram of the multi-head attention hybrid mechanism.

transforming $x = (x_1, x_2, \cdots, x_T)$ of each column through "map to sequence," it is input into the Bi-LSTM module and the output vector $T \times n_{class}$ of length $y = (y_1, y_2, \cdots, y_T)$, where n_{class} is the number of sub-row character categories of the secondary device terminal.

3.4 Multi-head attention hybrid mechanism

This article presents a model in a lengthy attention mechanism that supplements the Bi-LSTM module. This helps the Bi-LSTM module better address the correlation characteristic of long time-series data mining as the problem of long sequences making it easier to lose information arises during the training process. The output vector is transformed into three input matrices of dimension d_k by three different mapping operations, Q (Query), K (Key), and V (Value), and the attention output matrix is given by Eq. 15:

$$Attention(Q, K, V) = soft\,max\left(\frac{QK^T}{\sqrt{d_k}}\right)V, \tag{15}$$

where d_k is the feature dimension of each key, which is used for weight scaling, and softmax is normalized to the interval [0,1].

The multi-head attention hybrid mechanism divides the time series into an I subspace, and each head performs self-attention calculation on the subspace to enhance its expressive power. Then, the results of head I are spliced and integrated to obtain multiple heads, and each head is splice to obtain the final through linear trans formation, that is Eq. 16, 17.

$$heads_i = Attention(QW_i^Q, KW_i^K, VW_i^V), \tag{16}$$

where W_i^Q, W_i^K, and W_i^V represent the weight matrix of Q, K, and V, respectively.

$$MultHead(Q, K, V) = Concat(head_1, \cdots, head_I)W^o, \tag{17}$$

where W^o represents the weight of the linear transformation; $head_i$ represents head i in the bull attention module; and $Concat$ represents the splicing operation. $MultHead(Q, K, V)$ is the final output result, which can learn more feature information from different spaces, and its model structure is shown in Figure 7:

3.5 Transcription layer

The problem of difficult-to-align input and output is a common occurrence in the text recognition sector. Thus, in this article, the recurrent neural network is decoded using CTC, and the Bi-LSTM output is transformed into a sequence format.

π is defined as the text sequence path composed of the Bi-LSTM output. For the Bi-LSTM module, the probability of output x given input l is calculated by the following Eq. 18:

$$p(l|x) = \sum_{\pi \in \beta^{-1}(l)} p(\pi \mid x), \tag{18}$$

where β is a multi-to-one mapping function, the purpose of which is to remove duplicate labels and blank labels. $\pi \in \beta^{-1}(l)$ represents all l paths that are π after transformation, and any path, as shown in Eq. 19.

$$p(\pi \mid x) = \prod_{t=1}^{T} y_{\pi_t}^t, \forall \pi \in L'^T, \tag{19}$$

where T represents the length of the input sequence and l is the label of the output. π_t represents the output character corresponding to path π at time t, which corresponds to the probability of obtaining the character at time t.

TABLE 2 Comparison of experimental results for text detection.

Model	P/%	R/%	Mean average precision (mAP)@0.5/%	Model size/MB	FPS (*f*/s)
YOLOXs (Yin et al., 2023)	95.58	79.14	87.21	16.4	86.90
YOLOv4-tiny (Zhao et al., 2023)	83.57	73.06	74.30	22.5	77.41
YOLOv5s (Han et al., 2024)	91.67	79.35	85.66	14.5	83.95
YOLOv7-tiny	94.91	84.82	92.15	12.2	103.42
Improved YOLOv7-tiny	**97.39**	**89.62**	**95.07**	**12.08**	**95.87**

The bold values represents the improved experimental results of this paper.

TABLE 3 Comparison of methods.

Model	P/%	R/%	mAP@0.5/%
YOLOv7-tiny	94.91	84.82	92.15
YOLOv7-tiny + BiFPN	94.89	87.42	93.16
YOLOv7-tiny + SimAM	96.16	87.01	93.36
YOLOv7-tiny + BiFPN + SimAM	**97.39**	**89.62**	**95.07**

The bold values represents the improved experimental results of this paper.

In Eq. 20, the training process of CTC is to adjust the parameter $\frac{\partial p(l|x)}{\partial \omega}$ of Bi-LSTM through the gradient ω so that $\pi \in \beta^{-1}(l)$ is maximized when the input sample $p(l|x)$ is obtained as

$$h(x) = arg\max_{l \epsilon L \leq T} p(l|x). \tag{20}$$

4 Analysis of experimental results

4.1 Experimental environment

In this paper, the operating system used for model training is Windows 10 with a 64-bit processor. Intel(R) Core(TM) i5-10200H CPU @ 2.40 GHz 2.40 GHz is used as hardware. Running memory was 12 GB. PyTorch is chosen as the deep learning framework. The programming language is Python 3.6. The CUDA version is 11.6.

4.2 Evaluation index

In this paper, precision (P), recall (R), and mean average precision (mAP) are used as evaluation indicators for text detection, as shown in Eqs 21–24.

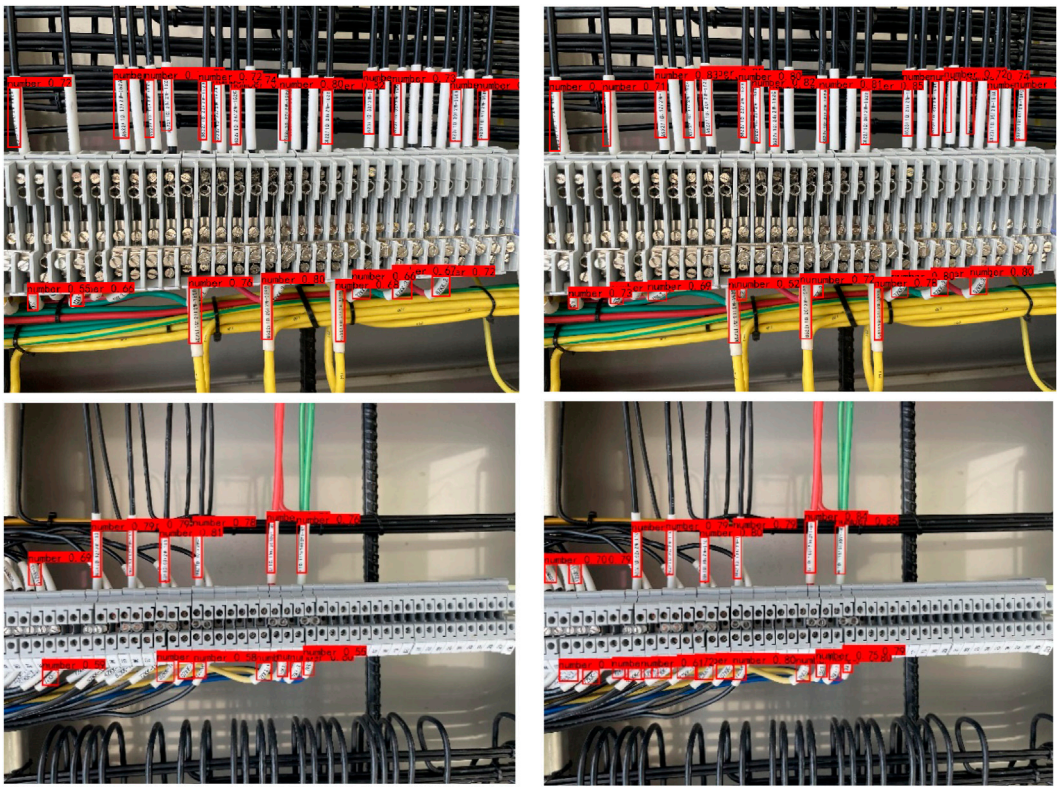


FIGURE 8
Comparison of the terminal strip text detection effect.

TABLE 4 Comparison with other methods.

Model	ACC (%)	Loss
CRNN+CTC	85.4	0.0981
MAH-CRNN+CTC (Guo et al., 2022)	87.59	—
Improved MAH-CRNN+CTC	**91.2**	**0.0329**

The bold values represents the improved experimental results of this paper.

$$P = \frac{TP}{TP + FP} \times 100\%, \tag{21}$$

$$R = \frac{TP}{TP + FN} \times 100\%, \tag{22}$$

$$AP = \int_0^1 P(R)dr, \tag{23}$$

$$mAP = \frac{\sum_{j=1}^{s} AP_j}{S}, \tag{24}$$

where *TP* stands for true positive, indicating the samples that the network detects following detection and classification match samples that have been labeled. False negatives or labeled samples that the network did not detect or classify—also known as missed detection—are represented by *FN* and *FP*, respectively. False positives are incorrectly classified detection samples that are not included in the labeled samples or false detection. The average precision (*AP*) of a single class is the area measured between the P(R) curve and the axis. Averaging the APs of all categories yields the mAP, where *S* is the total number of categories.

Average loss (Loss) and character recognition accuracy (Acc) are often used evaluation metrics for text recognition. Eq. 25, which illustrates the condition of incorrect recognition and multiple recognitions, shows that Acc is the ratio of the number of characters identified by model A to the total number of characters identified by model B. The average loss of character recognition is shown in loss. The better the model, the larger the Acc, and the smaller the loss.

$$A_{cc} = \frac{A}{B}. \tag{25}$$

4.3 Text detection experiment and result analysis

The dataset employed in the text detection module within this paper is sourced from a collection of 1,000 high-resolution images (1,024 × 1,024), depicting terminal rows of secondary equipment in substations. The division of data in this set follows a 8:2 ratio for training and testing subsets, respectively; moreover, the training subset itself is further stratified into a training set and a validation set according to a 9:1 allocation principle.

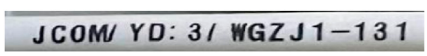
In the experiment, the epoch is set to 200, batch size is 8, the Adam optimizer is used to update the optimization gradient, and the cosine annealing algorithm is used to dynamically adjust the learning rate attenuation strategy. The initial learning rate of the model is 0.001, the weight attenuation parameter is 0.0005, and the learning rate momentum parameter is 0.937.

To confirm that the revised model presented in this work is superior, Table 2 compares the revised model with lightweight models like YOLOXs (Yin et al., 2023), YOLOv4-tiny (Zhao et al., 2023), YOLOv5s (Han et al., 2024), and YOLOv7-tiny based on the terminal strip wiring dataset of secondary devices. The enhanced model in this study has an average accuracy (mAP) of 95.07%, which is 7.86%, 20.77%, 9.41%, and 2.92% greater than that of YOLOXs, YOLOv4-tiny, YOLOv5s, and YOLOv7-tiny, respectively, based on the experimental findings shown in Table 2. With a memory occupation of only 12.08 MB, the upgraded model outperforms the YOLOv4-tiny model by 46.3%. In order to guarantee accuracy, the enhanced model outperforms the other models in terms of accuracy and recall rate, that is, by 97.39% and 89.62%, respectively. The average detection speed (FPS) of the enhanced model is 95.87 f/s, which is marginally slower than the quickest detection speed of YOLOv7-tiny; nevertheless, this model performs better in other detection algorithm performance tests. Therefore, the upgraded model in this research still exhibits significant improvements in the identification of speed and accuracy with respect to the total detection performance of the model.

An array of ablation experiments was created for comparison analysis in order to confirm the efficacy of the modified YOLOv7-tiny model suggested in this paper. The trials were carried out using the same training conditions to guarantee the accuracy of the experiments. The comparative findings are shown in Table 3 for the original model, each upgraded module, and the test set.

Table 3 shows how the precision rate, recall rate, and mAP increased by 1.25%, 2.91%, and 1.21%, respectively, when the SimAM was added to the original model. It demonstrates that compared to the original model, the SimAM module is more capable of feature extraction and expression. The precision, recall, and mAP of the model improved to 94.89%, 87.42%, and 93.16%, respectively, after the FPN module was swapped out for the BiFPN module in the neck network. This improvement was

TABLE 5 Comparison of recognition renderings of the model.

Picture	CRNN+CTC	Improved MAH-CRNN+CTC
	2-32KK1-**0**	2-3ZKK1-6
	J04/Y**0**:13/WGZJ1-181	J04/YD:13/WGZJ1-131
	JCOM/YD:3/WGZJ1-131	JC0M/YD:3/WGZJ1-131

The bold values represents the improved experimental results of this paper.

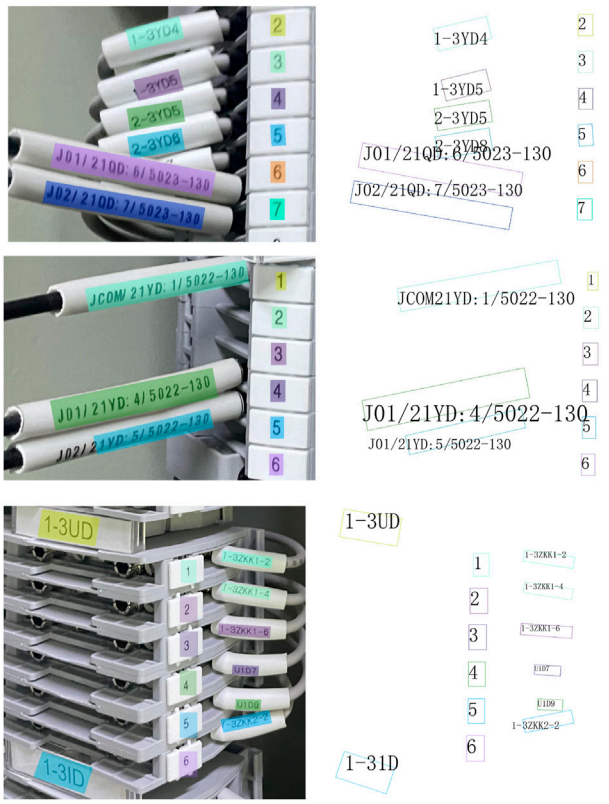


FIGURE 9
Detection and recognition performance on terminal blocks.

attributed to the superior ability of the BiFPN module to fuse multi-scale weighted feature information while maintaining lightweight.

Figure 8 compares the text detection effectiveness of the terminal strip. The detection effect of the previous model is on the left, and the detection effect of the improved algorithm is on the right. Figure 8 shows that the old model had low detection accuracy and more missed detections for text information in densely dispersed terminal strips and occluded terminal strips. The improved model has high detection accuracy, and only the significantly obstructed portion is undetected. The remaining portion does not show error detection or missed detection. As a result, compared to the original model, the improved YOLOv7 model has better detection accuracy.

4.4 Text recognition experiment and result analysis

Four-hundred images of the terminal strip of the secondary equipment in the substation make up the dataset utilized by the text recognition module in this work. The partition of the dataset into training and test data follows the 8:2 concept, while the training data are split into training and validation sets using the 9:1 approach.

In this experiment, we specified certain parameter configurations, where the epoch is set to 300 and the batch size is configured as 8. For gradient optimization, the Adam optimizer was employed, alongside a cosine annealing algorithm that dynamically tunes the learning rate decay strategy. The initial learning rate of the model is assigned a value of 0.0008, the weight decay parameter is set at 0.0001, and the momentum for the learning rate is 0.937.

This study presents the optimization of the CNN module using the original CRNN model, with a modified activation function (Leaky ReLu) and a single-layer BN at the network end. To avoid losing sequence information from taking too long, the sequence prediction module incorporates the MAH mechanism. In order to accommodate the secondary device terminal labeling dataset, the number of Bi-LSTM hidden layer cells in the RNN portion is set at 512. Table 4 shows that the improved MAH-CRNN+CTC model has a character recognition accuracy of 91.2%, which is 5.8% higher than that of the traditional model, and has a low average loss at the same time. The traditional CRNN+CTC model cuts character recognition accuracy to only 85.4%. Furthermore, the improved MAH-CRNN+CTC model in this paper still has higher identification precision that that presented by Guo et al. (2022).

Table 5 displays the recognition effect in real time. The classical paradigm has issues with character recognition loss and simple recognition mistakes of related characters, such as misrecognizing "Z" as "2," "3" as "8," and "D" as "0." In an effort to enrich the variety within the dataset, supplementary fuzzy images are incorporated. Experimental findings demonstrate that the model maintains strong recognition capabilities even when dealing with instances characterized by indistinct recognition features. This highlights the superior generalization performance of the improved model to its predecessor.

The performance of the detection and recognition of terminal rows in practical applications, as presented in this paper, is shown in Figure 9. The experimental findings indicate that our proposed detection and recognition model consistently achieves strong detection capability and high accuracy across various real-world scene images.

5 Conclusion

This work proposes a terminal strip detection and recognition model based on the improved YOLOv7-tiny and MAH-CRNN+CTC models to address the issue of confined arrangement and varying terminal block lengths of secondary equipment in substations. First, the SimAM and BiFPN attention mechanism modules were added to improve the model capacity for feature extraction and information fusion, increase the model detection accuracy, and increase its accuracy rate, summon rate, and mAP to 97.39%, 89.62%, and 95.07%, respectively. Second, the MAH mechanism was introduced to address the low recognition accuracy of the CRNN. This improves the model capacity to predict and process character sequence information, minimizes the loss of character feature information, and increases the model recognition accuracy to 91.2%, which is 5.8% higher than that of the traditional model. The findings demonstrate the good detection and identification effects for terminal strips of the improved YOLOv7-tiny and MAH-CRNN+CTC approaches presented in this work.

Data availability statement

The raw data supporting the conclusion of this article will be made available by the authors, without undue reservation.

Author contributions

GZ: writing–review and editing. LW: writing–review and editing. CQ: writing–review and editing. ZH: writing–original draft.

Funding

The authors declare that financial support was received for the research, authorship, and/or publication of this article. This work was supported in part by the National Natural Science Foundation of China under Grant No. 52107108.

Conflict of interest

Authors GZ, LW, and CQ were employed by Dongguan Power Supply Bureau of Guangdong Power Grid Co., Ltd.

The remaining authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Publisher's note

All claims expressed in this article are solely those of the authors and do not necessarily represent those of their affiliated organizations, or those of the publisher, the editors, and the reviewers. Any product that may be evaluated in this article, or claim that may be made by its manufacturer, is not guaranteed or endorsed by the publisher.

References

Cui, Y. X., Li, Y., Wang, H. J., and Wang, X. L. (2013). Identification of location and orientation for terminal blocks based on template matching. *Key Eng. Mater. Zurich* 561, 515–520. doi:10.4028/www.scientific.net/kem.561.515

Guo, Ke, Bai, Y., Shao, X., Wang, X., and Ma, J. (2022). Terminal block text recognition based on multi-scale attention and convolutional recurrent neural networks. *J. Anhui Univ. Nat. Sci. Ed.* 46 (06), 49–56.

Han, C., Yang, ZHOU, Wang, L., Lei, H., Yao, D., and Liang, W. (2024). Research of CCTV drainage pipeline defect recognition method based on YOLOv5s. *Munic. Technol.* 42 (0 3), 230–236. doi:10.19922/j.1009-7767.2024.03.230

Huang, H., Wu, J., Xiao, H., Liang, Z., Wang, J., Tan, X., et al. (2023). Terminal text detection and recognition based on attention mechanism. *Mech. Electr. Eng.* 52 (06), 202–206.

Liu, W., Lin, G., Fu, D., and Wang, S. (2023). Substation secondary loop terminal row design text detection and recognition. *J. hubei Univ. Natl. Nat. Sci. Ed.* 9 (02), 198–205. doi:10.13501/j.carol.carroll.nki.42-1908/n.2023.06.009.06

Masci, J., Meier, U., Ciresan, D., Schmidhuber, J., and Fricout, G. (2012). Steel defect classification with max-pooling convolutional neural networks. *2012 Int. Jt. Conf. Neural Netw. IJCNN*, 1–6. doi:10.1109/ijcnn.2012.6252468

Messina, R., and Louradour, J. (2015). "Segmentation-free hand-written Chinese text recognition with LSTM-RNN," in Proceeding of the 13th IAPR International Conference on Document Analysis and Recognition, Nancy, France, August, 2015, 171–175.

Tan, M., Pang, R., and Le, Q. V. (2020). "Efficient det: scalable and efficient object detection," in Proceedings of the IEEE/CVF conference on computer vision and pattern recognition, Seattle, WA, USA, June, 2020, 10781–10790.

Wang, X., and Yi, Z. (2019). Research on Obstacle detection method of Mowing robot Working environment based on improved YOLOv5. *Chin. J. Agric. Mech.* 44 (3), 171–176.

Wang, J., Ma, Y., Zhang, L., Gao, R. X., and Wu, D. (2018). Deep learning for smartmanufacturing: methods and applications. *J. Manuf. Syst.* 48, 144–156. doi:10.1016/j.jmsy.2018.01.003

Wang, L., Huang, Li, Zhang, L., Long, Z., Li, Y., and Zhou, J. (2020). Fault diagnosis technology of transformer substation panel cabinet based on joint training method. *Electr. Power Supply* 37 (Suppl. 5), 85–90.

Wang, W., Xie, E., Li, X., Hou, W., Lu, T., Yu, G., et al. (2019). "Shape robust text detection with progressive scale expansion network," in Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Long Beach, CA, USA, June, 2019, 9336–9345.

Wu, X., He, Y., Zhou, H., Cheng, L., and Ding, M. (2019). Research on environmental personnel identification of monitored waters based on improved YOLOv7 algorithm. *J. Electron. Meas. Instrum.* 37 (5), 20–27.

Xiaoxuan, Hu, Xijin, Z., Qi, Z., and Wang, H. (2021). Intelligent Detection system of Marine Welding Surface defects based on deep transfer learning. *Shipbuild. Technol.* 49.

Yang, Z., Huang, H., He, L., Liu, Z., Li, X., and Liu, Z. (2022). Surface defect detection of circuit board based on color histogram. *Comput. Integr. Manuf. Syst.*,

Yang, L., Zhang, R. Y., Li, L., and Xie, X. (2021). "SimAM: a Simple,Parameter-free attention module for convolutional neural networks," in Proceedings of the 38th International Conference on Machine Learning, Virtual, July, 2021, 11863–11874.

Yin, Z., Qi, Y., and Wang, L. (2023). TERMINAL_PIN welding surface defect detection method based on YOLOXs. *Comput. Appl.* 43 (S2), 209–215.

Zhao, J. G., Han, Z. S., Fan, J. J., and Zhang, J. (2023). Safety helmet wearing detection algorithm based on improved YOLOv7-tiny. *J. Hebei Univ. Archit. Eng.* 41 (04), 240–245.

Zhong, M., Jun, T., Fu, A. M., and Yang, Y. (2023). Based on attention mechanism of secondary loop terminal text detection and recognition method. *Electr. power Sci. Technol. J.* 38 (03), 132–139. doi:10.19781/j.iSSN.1673-9140.2023.03.014

Zhou, J., Yan, Li, and Zhang, K. (2018). Intelligent identification System for line sleeve label of screen cabinet equipment in substation secondary system. *Mach. Electron.* 36 (11), 67–70.