Journal of Engineering and Applied Science

RESEARCH Open Access

Detection research of insulating gloves wearing status based on improved YOLOv8s algorithm

Caixia Tao[1], Chaoting Wang[1] and Taiguo Li[1*]

*Correspondence: leetg@mail.lzjtu.cn

[1] School of Automation and Electrical Engineering, Lanzhou Jiaotong University, Lanzhou 730070, China

Abstract

The safety hazards may be caused by power grid operators not wearing insulating gloves according to regulations for live electrical working. Additionally, existing methods for detecting the wearing status of insulating gloves suffer from low recognition accuracy, slow detection speed, and large memory occupation by weight files. To address these issues, a Mixup-CA-Small-YOLOv8s (MCS-YOLOv8s) algorithm is proposed for detecting the wearing status of insulating gloves. Firstly, the mixup data augmentation technology using image mixing is introduced, increasing the data's diversity and improving the model's generalization ability. Secondly, the coordinate attention (CA) module is added to the original backbone network to strengthen the channel and positional information, suppressing the secondary feature information. Finally, a small target detection structure is designed by removing the last bottom feature detection layer in the original neck network and adding a shallow feature. The ability of small targets' feature extraction is enhanced without increasing too much computation. The experimental results indicate that the mean average precision of the MCS-YOLOv8s algorithm on the test set is 0.912, the detection speed is 87 FPS, and the model's weight memory occupies 15.7 MB. It is verified that the model has the advantages of high detection accuracy, fast speed, and small weight memory, which has great significance in ensuring the safe and stable operation of the power grid.

Keywords: Data augmentation, CA module, Small target detection structure, YOLOv8s, Insulating gloves, Wearing status detection

Introduction

The safety hazards associated with the power grid involve various factors, such as personnel behavior and equipment failure [1]. Following the regulations, power grid operators are required to wear insulating gloves when conducting charged operations. Failure to comply with this requirement is considered unsafe behavior. Currently, within the power grid infrastructure, staff rely on monitoring videos to manually assess whether grid personnel are wearing insulating gloves. Depending on manual supervision alone can leave the grid with significant deficiencies in terms of operating costs, detection objectivity, and reliability [2, 3]. The intelligent detection of power grid personnel insulating gloves whether wearing condition, based on deep learning algorithms, holds significant importance in enhancing the efficiency of monitoring and ensuring the safe and stable operation of the power grid. With the advancement of computer vision technology, accomplishing this task has become more feasible.

In prior studies on detecting the condition of power grid personnel wearing insulating gloves, literature [4] employs traditional computer vision methods to fuse global color features and SIFT features. They utilize the center of mass distance and intersection over union between insulating gloves and body parts, to identify the status of insulating glove wearing. However, their approach does not account for the situation where the wrong gloves are worn. Additionally, their dataset images for the target are clear and high resolution, and the model exhibits poor generalization ability and could not meet the requirements of practical engineering application. Reference [5] utilizes the RetinaNet network to recognize the state of wearing insulating gloves. While achieving high detection accuracy for correctly wearing insulating gloves, the recognition accuracy for incorrectly wearing insulating gloves is considerably low. This is mainly attributed to the small size of the hand target and the weak feature of the hand, causing the prediction frame to deviate from the positioning of the small target, leading to misdetection or even missed detection. Reference [6] introduces an algorithm for detecting the status of wearing insulating gloves using an enhanced version of YOLOv3. The average accuracy improves by 31% compared to the YOLOv3 base model. Nevertheless, the overall detection accuracy still requires improvement, and the model's weight memory is 346.9 MB. According to the literature [7], the improved YOLOv3 model exhibits higher detection accuracy and speed compared to Faster R-CNN. However, the overall mean average precision is only 79.25%, and the detection speed is merely 19.4 FPS. Reference [8] proposes YOLOv4-based detection for insulating gloves wearing status, achieving high average accuracy. Nonetheless, targets not wearing insulating gloves correctly are not recognized. Furthermore, the detection speed and model's weight file memory are not mentioned. Therefore, it is difficult to ensure that engineering deployment is achieved. Both literature [9] and literature [10] utilize YOLOv5 for detecting insulating gloves. Despite their methods achieving high mean average precision, neither can detect the condition of incorrectly wearing insulating gloves.

The current model for detecting the wearing condition of insulating gloves, based on traditional computer vision and deep learning methods, suffers from low recognition accuracy, poor real-time performance, and a large memory occupation of weight files. The enhancement and attention to the recognition of small hand targets that are not properly wearing insulating gloves is urgently needed. The analysis indicates that the deep learning method surpasses the traditional computer vision method in all aspects of performance for recognizing insulating gloves. Furthermore, owing to the ongoing development of the deep learning method, the model's recognition performance for the insulating gloves wearing condition continues to improve consistently.

Deep learning methods have become increasingly prevalent in tasks related to target detection [11]. In the 2014 IEEE International Conference on Computer Vision and Pattern Recognition, the R-CNN for target detection was proposed by Girshick et al. [12]. In the following year, Fast R-CNN was proposed by Girshick et al. [13]. At the 2015 Conference and Workshop on Neural Information Processing Systems, Kaiming He et al.

introduced Faster R-CNN, which significantly improves average accuracy [14]. However, its detection speed is only 5 FPS. In 2016, Redmon et al. introduced the YOLO one-stage target detection network, achieving a detection speed of 45 FPS [15]. However, its recognition accuracy is lower compared to the two-stage target detection network. In 2017 and 2018, Redmon et al. proposed the YOLOv2 [16] and YOLOv3 [17] algorithms, respectively. These algorithms significantly improve recognition accuracy and detection speed. However, the detection of small targets remains unsatisfactory. In 2020, the YOLOv4 algorithm was proposed by Bochkovskiy et al. [18], while YOLOv5 was proposed by Glenn et al. [19] in the same year. Both algorithms employ a deeper network structure and the path aggregation network (PANet) to enhance the accuracy of small target detection. Following this, YOLOv6 [20], YOLOv7 [21], and YOLOv8 [22] were introduced. YOLOv8 demonstrates higher mean average precision and faster detection rates. After considering all performance metrics of the above target detection algorithms, it is concluded that YOLOv8 is the most suitable base model for real-time wearing status detection of the insulating gloves of power grid personnel.

To achieve higher recognition accuracy, faster detection speed, and a smaller weight file memory for detecting the wearing condition of insulating gloves, this paper proposes the MCS-YOLOv8s algorithm, an enhancement of the YOLOv8s base model. Firstly, the Mixup data augmentation method is employed to enhance the model's generalization. Secondly, the backbone network incorporates the CA mechanism to enhance the extraction of crucial features. Finally, a small target detection structure for detecting small targets is designed to enhance the identification of incorrectly wearing insulating gloves in complex backgrounds. This is crucial for detecting the wearing status of insulating gloves for power grid personnel.

Methods

YOLOv8s algorithm

The YOLOv8 algorithm comprises five different network structures with varying widths and depths, namely n, s, m, l, and x. When evaluating the performance of different YOLOv8 algorithm structures for the same task, such as detection accuracy and speed, it is found that the YOLOv8s algorithm is better suited for detecting the condition of insulating gloves wearing. Figure 1 displays the structure of the YOLOv8s network, which comprises three parts: backbone network (backbone), neck network (neck), and output (output). The backbone network includes the convolution module (Conv), bottleneck layer (C2f), and spatial pyramid pooling-fast layer (SPPF). The C2f module enhances the model's ability to extract gradient flow information while maintaining a lightweight design. It enhances the feature extraction of input images in the backbone network. The neck network utilizes a PANet structure that fuses strong semantic and localization feature information through top-down and bottom-up path aggregation. Finally, at the output, multiple bounding boxes undergo non-maximum suppression (NMS) for filtering. The prediction category with the highest output confidence value is then selected, and the coordinates of the bounding boxes at the target location are returned.

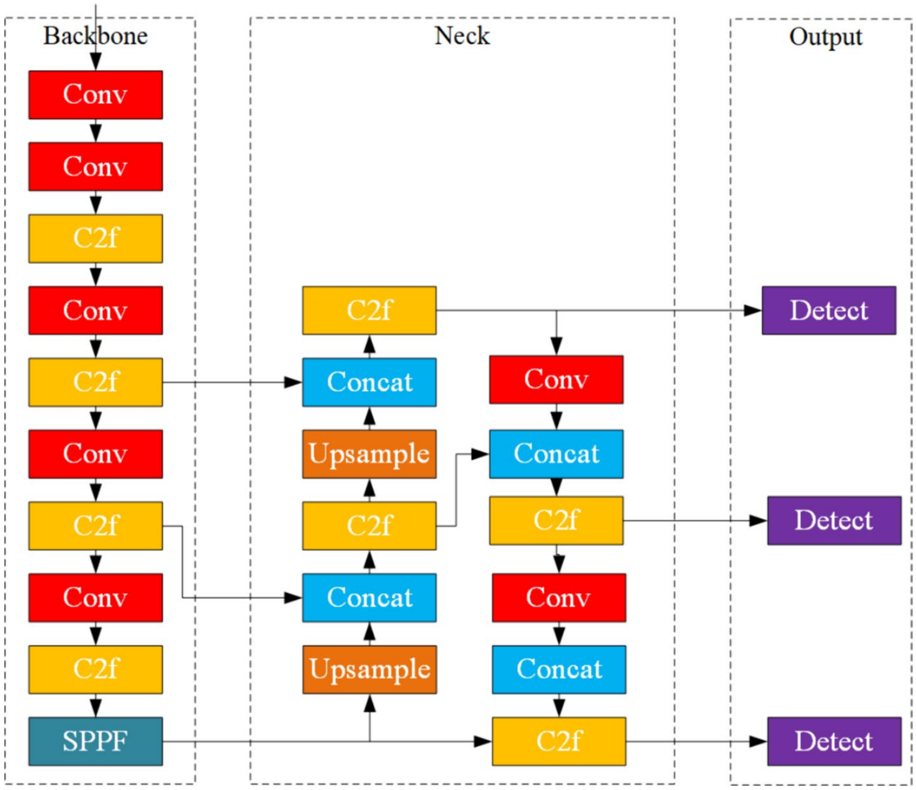


Fig. 1 YOLOv8s network structure

MCS-YOLOv8s algorithm improvement methods

Figure 2 illustrates the network structure of the MCS-YOLOv8s algorithm, which is enhanced on the YOLOv8s model. The following outlines the overall improvement ideas:

1) Firstly, the data enhancement strategy has been enhanced by incorporating the Mixup data augmentation technique. This approach elevates the training images' background complexity and enriches the training set's diversity. Improve the model's detection performance and robustness while retaining the current network structure and computational efficiency.
2) Next, the CA module is introduced in the backbone of YOLOv8s between the Conv module and C2f module and following the SPPF module. The intermediate feature map can efficiently integrate spatial positional information and accurately localize the position of small targets through the CA module, which strengthens the attention to channel and positional information by taking into account their positional relationship. This ensures a more objective and precise evaluation of the target's position.
3) Finally, the PANet structure in the neck network is enhanced by removing the bottom feature detection layer and introducing a shallow feature, thereby designing a new small target detection structure. This approach can preserve more effective information about small targets. The utilization of shallow feature maps with small receptive fields can significantly enhance the detection of small targets. The small

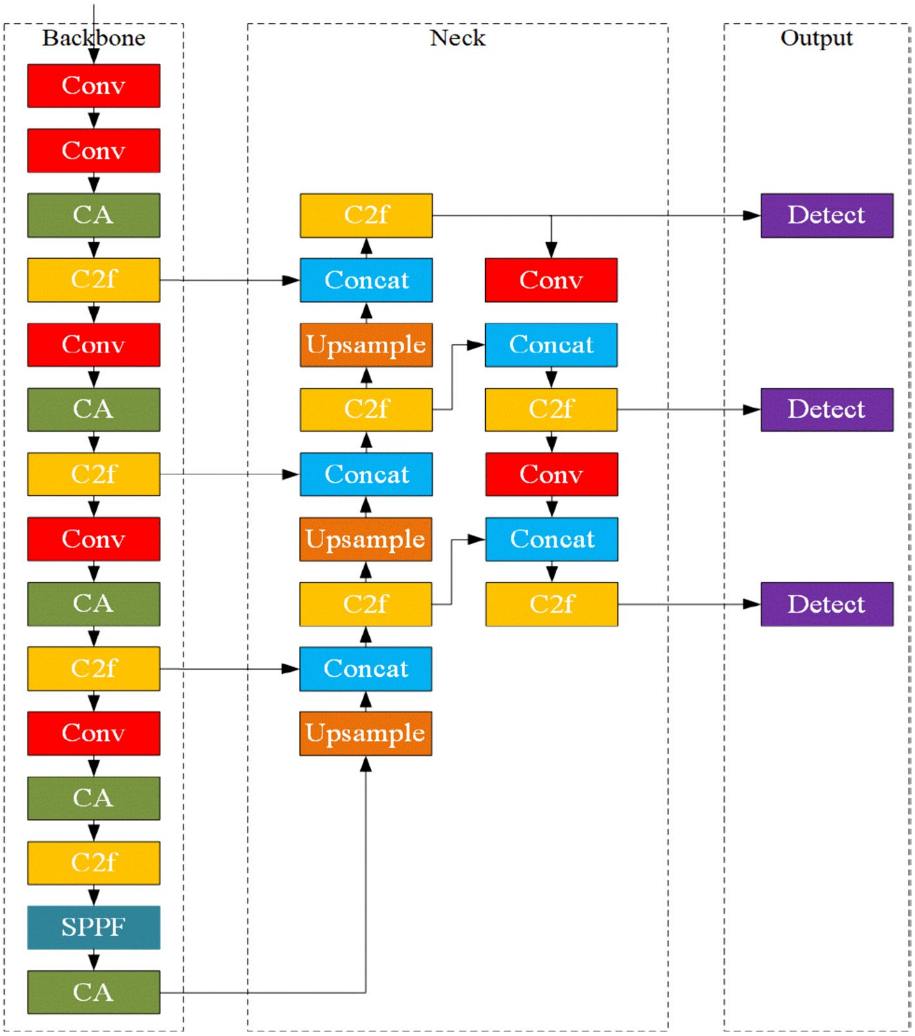


Fig. 2 MCS-YOLOv8s algorithm network structure

target detection structure can more accurately detect hard-to-recognize small targets without significantly increasing computational effort.

Related work

Coordinate attention (CA) mechanism module

The CA module enables the network to gain information about a broader area without incurring significant computational overhead by embedding positional information into channel attention. The CA module considers both channel and direction-dependent positional information. Additionally, it is flexible and lightweight enough to be easily integrated into the network architecture [23]. Figure 3 shows the CA module, which consists of the residual network module, X Avg Pool, and Y Avg Pool.

In Fig. 3, *C*, *H*, and *W* respectively represent the channel, height, and width of the input feature map. X Avg Pool and Y Avg Pool respectively refer to one-dimensional average pooling along the horizontal and vertical directions of the input feature map. Re-weight refers to the weighted aggregation of spatial features from input feature maps.

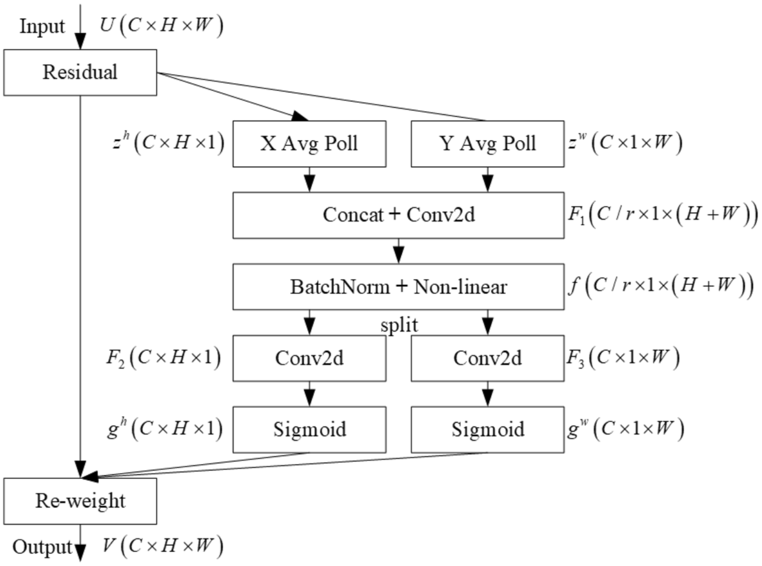


Fig. 3 Schematic diagram of the CA module

The CA module comprises two parts: positional information generation and position attention aggregation. Initially, the positional information of the feature map is generated. Pooling kernels of dimensions (*H*, 1) and (1, *W*) are employed to respectively encode features along the horizontal and vertical directions of a given input feature map *U*. The outputs from encoding the *c*-th channel on the input feature map *U*, with a height of *h* and a width of *w*, can be expressed as follows.

$$z_c^h(h) = \frac{1}{W}\sum_{0\leq i<W} u_c(h,i) \tag{1}$$

$$z_c^w(w) = \frac{1}{H}\sum_{0\leq j<H} u_c\left(j,w\right) \tag{2}$$

where $u_{\mathrm{c}}(h, i)$ represents the eigenvalue of the *c*-th channel on the input feature map *U* with height *h* and width *i*. $u_{\mathrm{c}}(j, w)$ represents the eigenvalue of the *c*-th channel on the input feature map *U* with height *j* and width *w*.

Position attention aggregation is initiated using the horizontal and vertical positional information generated by the above encoding. The feature maps, which have encoded both spatial directions, undergo a channel splicing operation. The spliced feature maps are then transformed using the 1×1 convolutional transform function F_1. This process can be expressed as follows.

$$f = \delta\left(F_1\left(\left[z^h, z^w\right]\right)\right) \tag{3}$$

where δ represents the nonlinear activation function. [·, ·] denotes the channel splicing operation along the spatial dimensions. The intermediate feature mapping of the positional information is denoted by $f \in R^{C/r\times(H+W)}$, to minimize computation and model complexity, with *r* being the reduction rate set to 16.

For the intermediate feature mapping f, it is decomposed into tensors $f^h \in R^{C/r \times h}$ and $f^w \in R^{C/r \times w}$ in horizontal and vertical directions, which are along the spatial dimension. The 1×1 convolutional transform functions F_2 and F_3 are then used to transform f^h and f^w into tensors with the same number of channels C. The computational process can be expressed as follows.

$$g^h = \sigma\left(F_2\left(f^h\right)\right) \tag{4}$$

$$g^w = \sigma\left(F_3\left(f^w\right)\right) \tag{5}$$

where g^h and g^w are transformed tensors output. σ is a sigmoid activation function.

Finally, the feature information of each position on the input feature map U is weighted and aggregated using g^h and g^w as the attention weights in the horizontal and vertical directions, respectively. The final feature map V after feature aggregation is then output. The procedure for weighted aggregation in computation can be expressed as follows.

$$v_c\left(j, i\right) = u_c\left(j, i\right) \times g_c^h\left(j\right) \times g_c^w(i) \tag{6}$$

where $v_c(j, i)$ represents the eigenvalue of the c-th channel on the output feature map V, with position coordinates (j, i). $u_c(j, i)$ represents the eigenvalue of the c-th channel on the input feature map U, with position coordinates (j, i). $g_c^h\left(j\right)$ and $g_c^w(i)$ are the horizontal attentional weights for the height j and the vertical attentional weights for the width i of the c-th channel on the input feature map U, respectively.

The backbone network feature extraction process in YOLOv8s involves the convolutional layer, which calculates the feature information of neighboring positions for each feature map. It is effective in extracting local features but struggles to capture global features. The convolutional layer neglects the inter-mapping between the information of each channel because each channel in the feature map contains distinct feature information [23]. Power grid personnel frequently wear insulating gloves when working with electricity outdoors. However, this practice can give rise to issues such as difficulty in detecting the target due to the distance between the hand target and detection equipment, as well as low pixel value. Therefore, the addition of the CA module enhances the learning of feature relationships between channels and captures long-range dependencies within a channel, enabling the retention of precise positional information for small targets. Literature [24] has demonstrated that the CA module can optimize the learning of various types of target feature information in feed-forward neural networks and efficiently integrate spatial coordinate information. This enhancement improves the network's ability to accurately locate the target position and effectively enhances target detection performance. Therefore, this paper proposes adding the CA module between each Conv module and C2f module in the backbone network, as well as at the last layer of the backbone. The introduction of the CA module is illustrated in Fig. 4. The enhanced backbone network can extract more precise information about small targets by enhancing attention to channel and

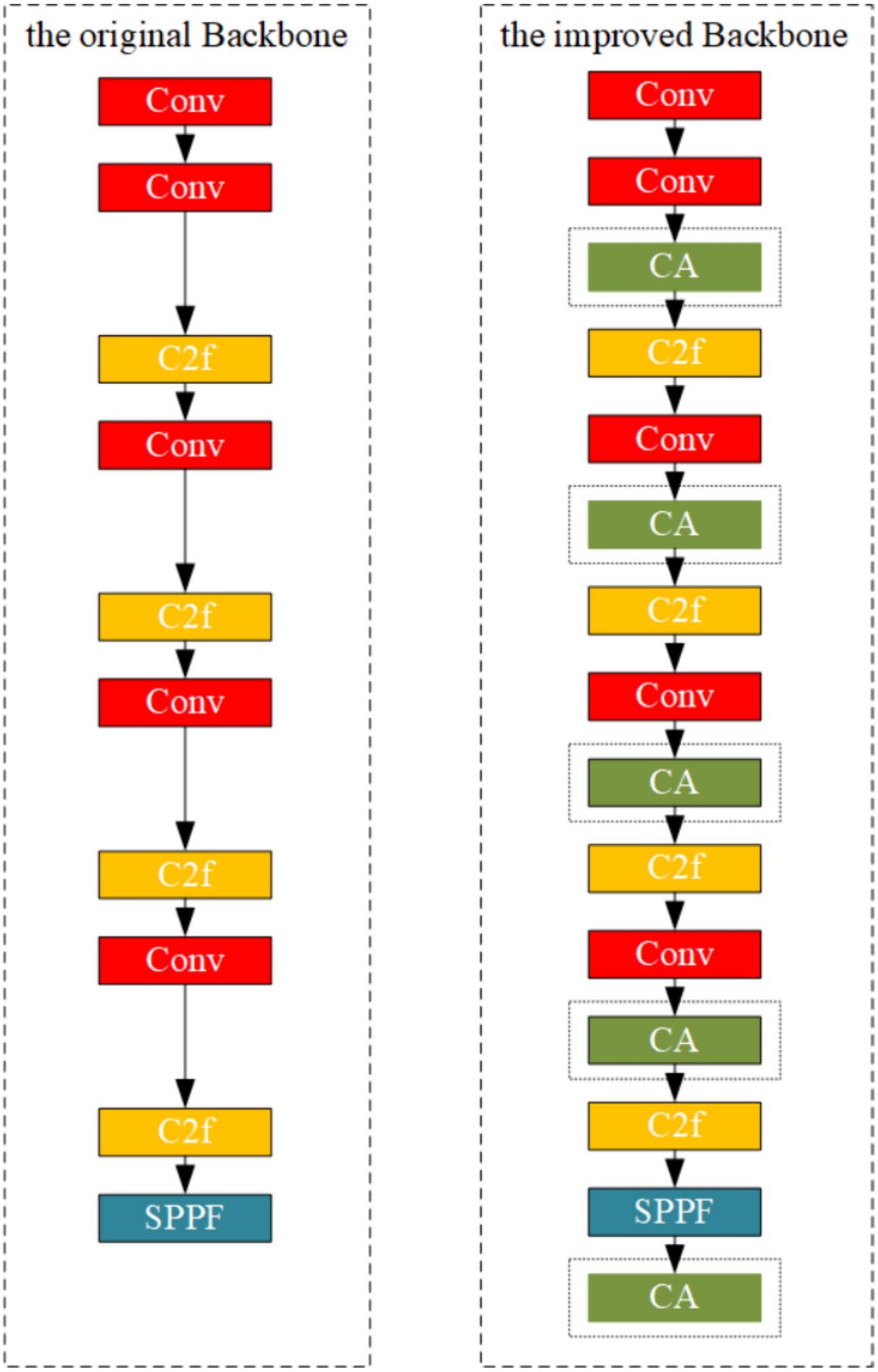


Fig. 4 Introduction of the CA module

positional information through the CA mechanism module while reducing attention to secondary information. The network model has been improved to enhance the detection of power grid personnel wearing insulating gloves status, making it more robust.

Small target detection structure

The YOLOv8s network requires the input image to undergo feature extraction in the backbone network, followed by processing in the neck network, before finally outputting to the detection layer. Figure 5 shows the backbone feature extraction and neck structure of YOLOv8s. The input feature map is denoted by C_i ($i=0, 1, 2, 3, 4, 5$), where i is the feature extraction level. The output feature map is denoted by P_n ($n=3, 4, 5$), where n is the feature output level. The neck network of YOLOv8s adopts the PANet structure, a feature fusion of strong semantic information from the top layer and robust localization information from the bottom layer, better retaining features and information related to small targets.

Addressing challenges such as the weak target features and the small size of insulating gloves wearing detection for power grid personnel, this paper enhances the PANet structure and proposes a small target detection structure specifically designed for small target detection. In the backbone network, the feature information of small targets is consistently lost as the network deepens due to multiple downsampling operations. Therefore, a sizing mechanism is introduced at the level of larger feature maps, serving as shallow features for small target detection. Shallow feature maps with smaller spatial receptive fields, which contain more edge information and have a stronger ability to represent geometric details, can significantly enhance the detection of small targets [25]. To reduce network complexity and computation while obtaining more effective information about small targets, this structure initiates channel fusion at the top feature extraction layer and removes the last bottom feature detection layer. The structure of the MCS-YOLOv8's backbone network for feature extraction and small target detection is illustrated in Fig. 6.

The MCS-YOLOv8s network modifies the original YOLOv8s network by removing the P_5 feature detection layer. It initiates feature layer channel fusion starting from the C_2 feature extraction layer, and the specific fusion operation can be expressed as follows.

$$P_2 = C_2 + [C_3 + (C_4 + C_5 \uparrow_{2\times}) \uparrow_{2\times}] \uparrow_{2\times} \tag{7}$$

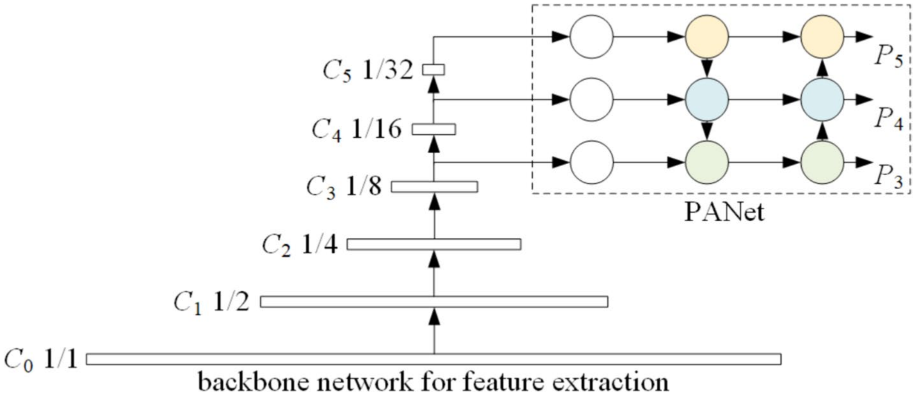


Fig. 5 The backbone and neck structure of YOLOv8s

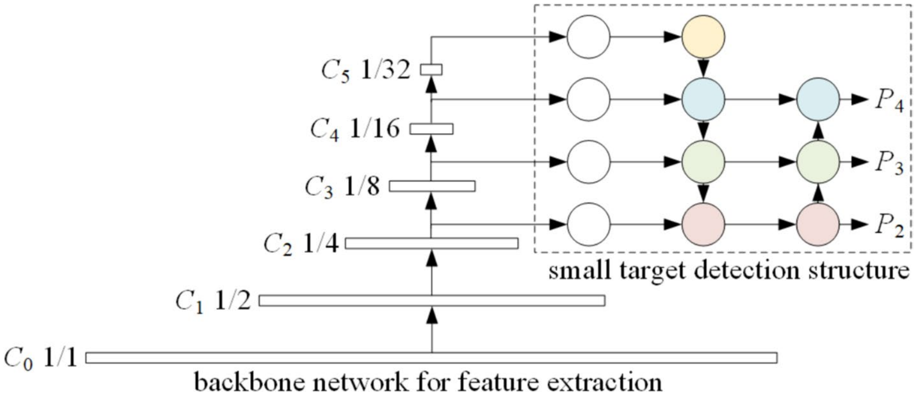

C_5 1/32
C_4 1/16
C_3 1/8
C_2 1/4
C_1 1/2
C_0 1/1
P_4
P_3
P_2
small target detection structure
backbone network for feature extraction

Fig. 6 The backbone and neck structure of MCS-YOLOv8s

$$P_3 = P_2 \downarrow_{2\times} + C_3 + (C_4 + C_5 \uparrow_{2\times}) \uparrow_{2\times} \tag{8}$$

$$P_4 = P_3 \downarrow_{2\times} + C_4 + C_5 \uparrow_{2\times} \tag{9}$$

where + denotes the superposition of feature map channels with the same length and width dimensions. $\uparrow_{2\times}$ represents the double upsampling operation, and $\downarrow_{2\times}$ represents the double downsampling operation. P_2, P_3, and P_4 are the output feature maps obtained from the enhanced PANet network. The algorithm presented in this paper conducts wearing state recognition of the insulating gloves based on the P_2, P_3, and P_4 feature maps.

Differing from the PANet structure used in the neck network of the YOLOv8s algorithm, the MCS-YOLOv8s algorithm removes the last bottom small-scale feature map detection layer (P_5), introduces a shallow large-scale feature map detection layer (P_2), and conducts target detection on three scales of feature maps (P_2, P_3, and P_4). P_2, P_3, and P_4 feature maps are obtained by aggregating spatial features from the C_2, C_3, C_4, and C_5 feature maps. Combined multichannel detection information is utilized to recognize the wearing condition of insulating gloves. Figure 7 illustrates the improvements made to the neck network.

Data augmentation strategy

The image samples of grid personnel wearing insulating gloves are limited. To address this limitation and enhance the diversity of input images, data augmentation techniques can be employed. This helps avoid the overfitting phenomenon in the convolutional network, ensuring that the trained model exhibits improved robustness and generalization ability [26, 27]. The original YOLOv8s algorithm employs seven data augmentation methods: random hue, saturation, value, translation, rotation, scale, and mosaic data enhancement. Mosaic data augmentation is a technique that entails randomly selecting four images from a training batch and then applying random cropping, scaling, and rotating operations to them. The resulting images are subsequently stitched together into a single image, thereby expanding the dataset and increasing the number of small targets in the sample. The six remaining data augmentation methods are traditional data augmentation techniques that apply

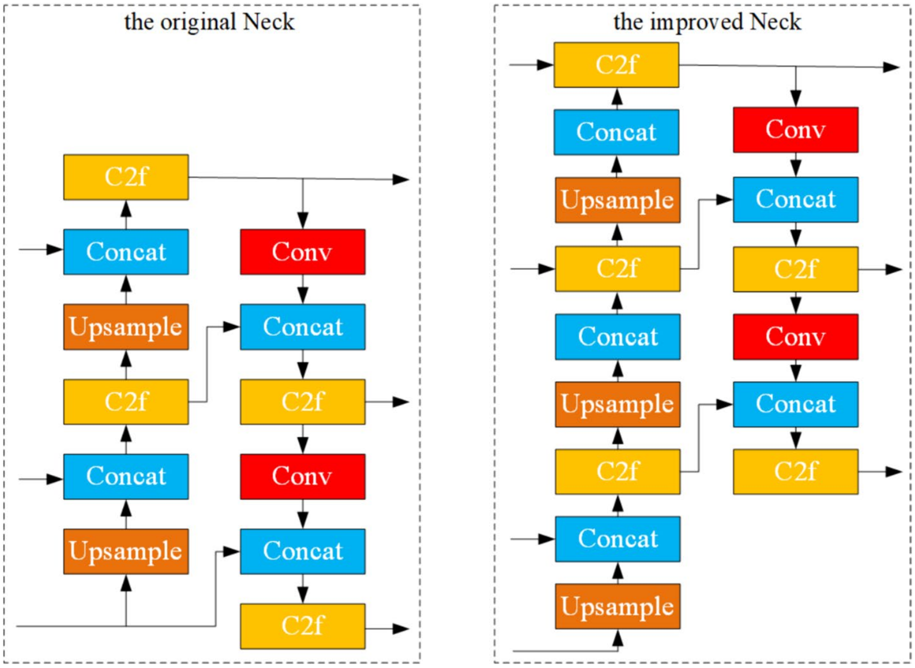


Fig. 7 Comparison of neck network improvement

geometric and color transformations to the images. These methods exhibit limited effectiveness in enhancing the performance of detecting the wearing status of insulating gloves for power grid personnel. In this paper, Mixup is employed, a data augmentation technique based on image mixing, to enhance the generalization ability of the network. This is achieved by augmenting the background complexity of the image. Mixup data enhancement is a regularization technique that randomly mixes two training image pixels to create a single image with two input labels.

The Mixup data enhancement method is randomly selecting two images from a training batch and obtaining their corresponding time series data, denoted as $\left(x_i, y_i\right)$ and $\left(x_j, y_j\right)$, where $i \neq j$. Here, x_i and x_j represent the two images, while y_i and y_j represent their respective label information. Finally, the mixed image is obtained through a calculation process. The resulting image and label are represented as X and Y, respectively, and the calculation process is described in (10) and (11).

$$X = \lambda x_i + (1 - \lambda)x_j \tag{10}$$

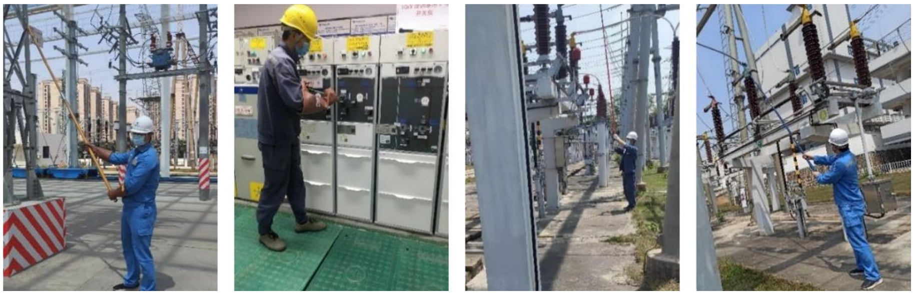
$$Y = \lambda y_i + (1 - \lambda)y_j \tag{11}$$

where λ falls within the range of [0, 1] and follows the beta (α, α) distribution. α is the hyper-parameter utilized to control the interpolation strength between feature targets. As α → 0, the Mixup data enhancement effect is close to failing.

The MCS-YOLOv8s algorithm employs eight data enhancement methods to enhance the mixing of contextual information, increase dataset diversity and complexity, and improve overall performance.

Table 1 The number of experimental datasets

Dataset name	Glove	Wrongglove	Total
Training set	1760	1763	3523
Validation set	225	226	451
Test set	486	487	973
Total	2471	2476	4947



(a) glove (b) glove (c) wrongglove (d) wronglove

Fig. 8 Typical images of various categories

Results and discussion

Experimental dataset

To verify the effectiveness of the proposed network model for insulating gloves detection, we conduct experiments using images taken at a power grid maintenance site. The dataset comprises 4947 images, randomly divided into training, validation, and test sets with a ratio of 8:1:2. Table 1 provides the number of images in each set.

The dataset comprises real-time images from a power grid maintenance site, captured under complex shooting conditions. The targets are small in scale and frequently occluded. This dataset is representative and useful for training models and enhancing detection performance. Labelimg is employed for labeling the images, with categories divided into "glove" for correctly wearing insulating gloves and "wrongglove" for wearing wrong gloves or not wearing gloves. Figure 8 displays typical images for each category. Figure 8a and b depicts grid personnel correctly wearing insulating gloves. In contrast, Fig. 8c shows grid personnel wearing non-insulating gloves, and Fig. 8d shows grid personnel without gloves.

Experimental environment

To ensure efficient training and testing of the improved YOLOv8s model, we utilize the hardware environment configuration presented in Table 2. A deep learning environment is established using PyCharm 2022, PyTorch 1.7, and Python 3.8 on a Windows 10 operating system. The batch size is set to 16, and the number of training epochs is 200, with both the initial and termination learning rates set to 0.01.

Table 2 Hardware environment

Hardware	Type
CPU	Intel Core i5-12400F
GPU	NVIDIA GeForce GTX 3060 (12G)
Mainboard	PRIME B660M-K
Memory	6G (DDR4 3200 MHz) × 2
Solid-state drives	Kingston M.2 250G

Table 3 F1 score of different models

Model	Glove	Wrongglove	F1 score
YOLOv8s	0.931	0.814	0.874
YOLOv8s + Mixup	0.935	0.814	0.876
YOLOv8s + CA	0.941	0.822	0.879
YOLOv8s + Small	0.936	0.850	0.894

Evaluation metrics

To compare the detection performance of the MCS-YOLOv8s algorithm with other models, this paper employs four evaluation metrics: F1 score, mean average precision (mAP), detection speed (V), and weight size (Ws). V represents the number of images processed per second, measured in frames per second (FPS).

The F1 score is calculated as the weighted harmonic mean of the precision (P) and the recall (R).

$$F1 - score = 2 \times \frac{P \times R}{P + R} \tag{12}$$

The value of mAP is typically calculated at an intersection over union (IoU) of 0.5 between the prediction frame and the true frame.

$$mAP = \frac{1}{m} \sum \int_0^1 P(R) dR \tag{13}$$

Ablation experiments results and discussion

The ablation experiments involve analyzing the performance of each model by comparing the original and improved models using the same dataset. Three improvements are made to the YOLOv8s model: Mixup data enhancement, the addition of a CA module to the backbone network, and the implementation of a designed small target detection structure for the neck network. YOLOv8s is the original model, and YOLOv8s + Small, YOLOv8s + CA, and YOLOv8s + Mixup are the three improved models. Performance comparison experiments are conducted among these models.

Tables 3 and 4 respectively display the F1 score and mAP performance of the model for each detection category. The table includes the following information: "glove" indicates the proper wearing of insulating gloves by grid workers, and "wrongglove"

Table 4 mAP of different models

Model	Glove	Wrongglove	mAP
YOLOv8s	0.952	0.816	0.884
YOLOv8s + Mixup	0.955	0.823	0.889
YOLOv8s + CA	0.960	0.831	0.895
YOLOv8s + Small	0.953	0.858	0.906

indicates improper wearing of insulating gloves or lack of use of gloves by grid workers. mAP represents the mean average precision of all categories, and the F1 score is computed based on the P and R values of all categories. Upon comparing the data in the two tables, it is evident that the improved models YOLOv8s + Small, YOLOv8s + CA, and YOLOv8s + Mixup exhibit higher F1 score and mAP than the original model YOLOv8s. Notably, the most significant performance improvement is in detecting the incorrect wearing of insulating gloves. The effectiveness of the three improved methods proposed for recognizing the wearing status of insulating gloves is verified.

In ablation experiments, the following three models YOLOv8s + Small + CA, YOLOv8s + Small + Mixup, and YOLOv8s + CA + Mixup are obtained by fusing the above three improved methods two by two. Tables 5 and 6 demonstrate that the same experiments are conducted using the same dataset to test the model's performance, with the YOLOv8s + Small + CA model exhibiting the best performance. The model's detection performance has significantly improved with the utilization of the small and CA modules. In addition, the other two models outperform the original model.

The MCS-YOLOv8s algorithm is derived from the YOLOv8s base model by integrating three improvement points: Mixup data enhancement, CA module, and small target detection structure. To evaluate the performance of the MCS-YOLOv8s algorithm, it is compared with the original YOLOv8s model. Table 7 shows the performance of the two models in terms of F1 score, mAP, Ws, and detection speed V.

The results display that the MCS-YOLOv8s model achieved a maximum mAP of 0.912, which is 2.8% higher than the original YOLOv8s model's mAP of 0.884. The F1 score also improved by 1.6%, from the original 0.874 to 0.89. At this point, the model

Table 5 F1 score of different models

Model	Glove	Wrongglove	F1 score
YOLOv8s	0.931	0.814	0.874
YOLOv8s + Mixup + CA	0.948	0.823	0.887
YOLOv8s + Mixup + Small	0.936	0.849	0.894
YOLOv8s + Small + CA	0.938	0.852	0.896

Table 6 mAP of different models

Model	Glove	Wrongglove	mAP
YOLOv8s	0.952	0.816	0.884
YOLOv8s + Mixup + CA	0.957	0.841	0.899
YOLOv8s + Mixup + Small	0.956	0.856	0.906
YOLOv8s + Small + CA	0.956	0.859	0.908

Table 7 Comprehensive performance of different models

Model	F1 score	mAP	Ws/MB	V/FPS
YOLOv8s	0.874	0.884	22.5	127
MCS-YOLOv8s	0.890	0.912	15.7	87

Table 8 Performance comparison with other algorithms

Model	mAP	Ws/MB	V/FPS
Mobilenet-SSD	0.709	91.1	49
YOLOv3-tiny	0.742	27.3	53
YOLOv4-tiny	0.753	24.2	57
YOLOv7-tiny	0.767	12.3	112
YOLOv5s	0.873	14.4	114
YOLOX-S	0.875	44.3	63
MCS-YOLOv8s	0.912	15.7	87

recognition performance is at its highest level, and the weight file size is reduced to 15.7 MB. Although the detection speed of the MCS-YOLOv8s model is reduced, achieving a speed of 87 FPS is still a high value for fast detection. It can be seen that the MCS-YOLOv8s model is better than the original YOLOv8s model in all other performances under the guarantee of real-time detection, which is more conducive to the realization of the deployment of the mobile terminal, and accurately detects the status of insulating gloves worn by power grid workers in the actual engineering to ensure the safe and stable operation of the power grid.

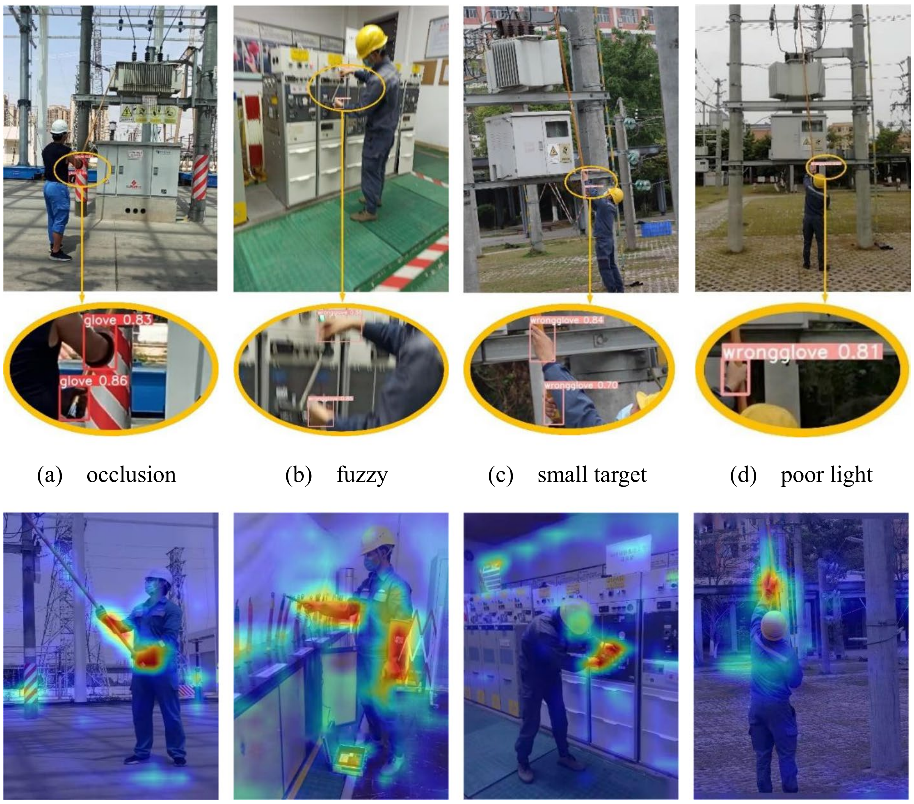
Comparison with other algorithms

To further verify the detection performance of the MCS-YOLOv8s model, we compare the MCS-YOLOv8s model with six mainstream target detection models using the same dataset, hardware, software, and experimental parameters. The results of the comparison experiments are shown in Table 8.

Table 8 shows that the Mobilenet-SSD algorithm has the lowest detection speed and mAP value, with the model weights file memory being 91.1 MB. The YOLOv3-tiny and YOLOv4-tiny models, lightweight models in the traditional YOLO series, respectively exhibit mAP values of 0.742 and 0.753. The latest lightweight model in the YOLO series is YOLOv7-tiny, boasting an impressive image processing speed of 112 FPS and a weight file memory of only 12.3 MB. However, its recognition accuracy remains relatively low. The s model of YOLOv5 has demonstrated a significant improvement in mAP value, reaching 0.873, compared to YOLO's tiny series of models. The YOLOX-S model achieves a mAP value of 0.875, slightly lower than the 0.884 achieved by the YOLOv8s algorithm in the YOLO series. However, unlike the tiny series and YOLOv5s, the weight file of the YOLOX-S model has a larger memory of 44.3 MB. The algorithm presented in this paper, MCS-YOLOv8s, improves recognition accuracy by 3.9% compared to YOLOv5s and 3.7% compared to YOLOX-S. Additionally, it has a weight memory of only 15.7 MB and processes images at a speed of 87 FPS, surpassing YOLOX-S.

Visualization of model detection results and discussion

The MCS-YOLOv8s model is employed to detect images in the test set, and the detection results are displayed in Fig. 9. Figure 9a displays an image with severe target occlusion, where incomplete target feature information in these images causes significant interference with recognition. Many images suffer from this issue, and resolving it is crucial to enhancing the effectiveness of model detection. The model accurately detects heavily occluded targets, as demonstrated by the results. Figure 9b displays an image of fuzzy targets with inadequate texture features, which may lead to missed detection. However, as shown in the actual detection results graph, this type of target can still be effectively detected. Figure 9c displays the image for detecting small targets. As the number of feature layers increases, the feature information for small targets is lost. To enhance the accuracy of small target detection, the neck network is improved, and the small target detection structure is utilized. Figure 9d shows an image taken on a cloudy day with poor outdoor lighting. In low-light environments, the quality of captured images decreases, making detection more difficult. However, the improved model with the added CA module can still accurately recognize such targets due to its powerful feature extraction network. Figure 9e and f displays class activation maps of power plant personnel wearing insulating gloves, while Fig. 9g and h shows class activation maps without wearing insulating gloves. From the four class activation maps, it can be seen



(a) occlusion (b) fuzzy (c) small target (d) poor light

(e) glove (f) glove (g) wrongglove (h) wronglove

Fig. 9 Visualization of detection effect

that the detection targets of the MCS-YOLOv8s model are more focused on the hand region, and the highly responsive regions are concentrated in the parts that are most helpful in making category judgments for correct classification. The experimental results show that the MCS-YOLOv8s model can effectively detect the insulating gloves wearing condition of power grid workers, which is of great significance in ensuring the safety of power grid workers and the safe and stable operation of the power grid.

Conclusion

The proposed MCS-YOLOv8s model is implemented to recognize the wearing status of insulating gloves for power grid personnel working with electricity in this paper. The Mixup data enhancement method enhances the dataset diversity without increasing computational effort, improving the model's detection performance and generalization. Introducing the CA module in backbone enhances the attention and extraction of effective feature information. The design of a new structure for detecting small targets in the neck improves the acquisition of small target feature information, resulting in better classification and localization of targets. The experimental results indicate that the MCS-YOLOv8s model achieves a mAP value of 91.2% on the test set, showcasing a 2.8% improvement in detection performance compared to the YOLOv8s base model.

The MCS-YOLOv8s model has a final weight file memory occupation of only 15.7MB and can process images at a speed of up to 87 FPS. Despite this decrease in speed, it still maintains a high value, making it suitable for real-time detection requirements. Additionally, the model has low hardware requirements, such as CPU, which renders it feasible for deployment on embedded devices.

The MCS-YOLOv8s algorithm model is horizontally compared with mainstream target detection algorithms such as YOLOv8s, YOLOX-S, YOLOv7-tiny, and YOLOv5s. The experimental results indicate that the MCS-YOLOv8s model exhibits faster detection speed, higher detection accuracy, and a smaller memory occupation of weight files. These findings confirm the advancement of the MCS-YOLOv8s algorithm.

In future work, the model will be trained using a more diverse and extensive dataset. Furthermore, we will explore additional optimizations for the current model, enhancing its lightweight and detection performance.

Abbreviations

YOLO	You Only Look Once
MCS-YOLOv8s	Mixup-CA-Small-YOLOv8s
CA	Coordinate attention
SIFT	Scale Invariant Feature Transform
R-CNN	Region-based Convolutional Neural Network
Fast R-CNN	Fast Region-based Convolutional Neural Network
Faster R-CNN	Faster Region-based Convolutional Neural Network
PANet	Path Aggregation Network
Conv	The convolution module
C2f	Bottleneck layer
SPPF	Spatial pyramid pooling-fast layer
NMS	Non-maximum suppression
P	Precision
R	Recall
FPS	Frames per second
mAP	Mean average precision
V	Detection speed
Ws	Weight size
Mobilenet-SSD	Mobilenet-Single Shot MultiBox Detector

Acknowledgements

Not applicable.

Authors' contributions

CXT provided guidance and ideas about the article. CTW collected and analyzed the experimental data and edited the manuscript. TGL provided technical guidance about the algorithm and language modification for the manuscript. All authors approved the manuscript.

Funding

Funded by the science and technology program of Gansu province (Grant No. 23JRRA880) and funded by Gansu Provincial Department of Education: University Teacher Innovation Fund Project (Grant No. 2023A-039).

Availability of data and materials

The data used to support the findings of this study can be shared upon request.

Declarations

Ethics approval and consent to participate

The human photos used in the paper are from the publicly available image datasets of grid operator behavior (https://aistudio.baidu.com/datasetdetail/93633/0) that do not contain personally identifiable information or violate any privacy rights. Thus, ethics approval was not required for this research.

Consent for publication

All authors have agreed.

Competing interests

The authors declare that they have no competing interests.

Received: 22 April 2024 Accepted: 18 May 2024
Published online: 03 June 2024

References

1. Xue MH, Ai CM, Lv HJ et al (2022) Intelligent image processing technology for safety helmet wearing in power plant. Proc CSEE 42(09):3346–3354
2. Zheng X, Yao J, Xu X (2019) Violation monitoring system for power construction site, IOP Conference Series: Earth and Environmental Science. IOP Publishing 234(1):012062
3. Zhao Z (2023) Power safety management and control based on the risk fusion model of object detection and power operation. In: 2023 IEEE 6th International Electrical and Energy Conference (CIEEC). IEEE, p 1626–1631
4. Yu K, Liu H, Li T et al (2021) A protective equipment detection algorithm fused with apparel check in electricity construction. In: 2021 33rd Chinese Control and Decision Conference (CCDC). IEEE, p 3122–3127
5. Zhang WK, Pan LZ, Guo ZB et al (2022) Visual detection method of abnormal state of insulating gloves based on RetinaNet in power scenarios. J Hunan Univ Sci Tech (Natural Science Edition) 37(01):85–91
6. Zheng HY, Song CH, Wu TT et al (2023) Small target detection and matching algorithm for wearing condition detection of insulating gloves. J Chinese Comp Syst 44(09):1989–1995
7. Zhao B, Lan H, Niu Z et al (2021) Detection and location of safety protective wear in power substation operation using wear-enhanced YOLOv3 Algorithm. IEEE Access 9:125540–125549
8. Cheng XL, Gao C, Ye XQ et al (2021) Electric intelligent safety monitoring identification based on YOLOv4. In: 2021 IEEE International Conference on Electrical Engineering and Mechatronics Technology (ICEEMT). IEEE, p 688–691
9. Wang K, Liu JL, Hu QQ et al (2023) Research on automatic detection of PPE under multiple operation scenarios. J Saf Environ 23(11):3960–3967
10. Wang YS, Zhu JJ, Wang ZY et al (2023) Detection of insulation gloves worn by power plant personnel based on improved YOLOv5. Comp Measure Control 31(11):60–65
11. Ren N, Fu Y, Wu YX et al (2022) Review of research on imbalance problem in deep learning applied to object detection. J Front Comp Sc Tech 16(09):1933–1953
12. Girshick R, Donahue J, Darrell T et al (2014) Rich feature hierarchies for accurate object detection and semantic segmentation. In: Proceedings of the IEEE conference on computer vision and pattern recognition. p 580–587
13. Girshick R (2015) Fast R-CNN. In: Proceedings of the IEEE International Conference on Computer Vision. p 1440–1448
14. Ren S, He K, Girshick R et al (2016) Faster R-CNN: towards real-time object detection with region proposal networks. IEEE Trans Pattern Anal Mach Intell 39(6):1137–1149
15. Redmon J, Divvala S, Girshick R et al (2016) You only look once: unified, real-time object detection. In: Proceedings of the IEEE conference on computer vision and pattern recognition. p 779–788
16. Redmon J, Farhadi A (2017) YOLO9000: better, faster, stronger. In: Proceedings of the IEEE conference on computer vision and pattern recognition. p 7263–7271
17. Redmon J, Farhadi A (2018) YOLOv3: an incremental improvement. arXiv preprint arXiv 1804:02767
18. Bochkovskiy A, Wang CY, Liao HYM (2020) YOLOv4: optimal speed and accuracy of object detection. arXiv preprint arXiv 2004:10934
19. Glenn J, (2020) YOLOv5, https://github.com/ultralytics/yolov5.
20. Li C, Li L, Jiang H et al (2022) YOLOv6: a single-stage object detection framework for industrial applications. arXiv preprint arXiv 2209:02976

21. Wang C Y, Bochkovskiy A, Liao H Y M, (2023) YOLOv7: trainable bag-of-freebies sets new state-of-the-art for real-time object detectors, Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. 7464–7475
22. Glenn J, (2023) YOLOv8, https://github.com/ultralytics/ultralytics.
23. Li TG, Zhang YZ, Zhang TC et al (2023) Train driver gesture recognition based on improved YOLOv5s algorithm. J China Railway Soc 45(01):75–83
24. Hou Q, Zhou D, Feng J (2021) Coordinate attention for efficient mobile network design. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. p 13713–13722
25. Yan JH, Zhang K, Shi TJ et al (2022) Multi-level feature fusion based dim small ground target detection in remote sensing images. Chinese J Sci Instr 43(03):221–229
26. Wang BG, Yang CM, Xia P (2021) YOLOv3-based wooden beam column defects detection combined with data enhancement and lightweight model. Electr Mach Control 25(04):123–132
27. Gao W, Zhou C, Guo MF (2021) Insulator defect identification via improved YOLOv4 and SR-GAN algorithm. Electr Mach Control 25(11):93–104

Publisher's Note

Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.