Front. Agr. Sci. Eng. 2025, 12(2): 245–260

RESEARCH ARTICLE

Improved method for a pedestrian detection model based on YOLO

Yanfei LI (✉), Chengyi DONG

School of Mechanical and Electrical Engineering, Hunan Agricultural University, Changsha 410128, China.

KEYWORDS

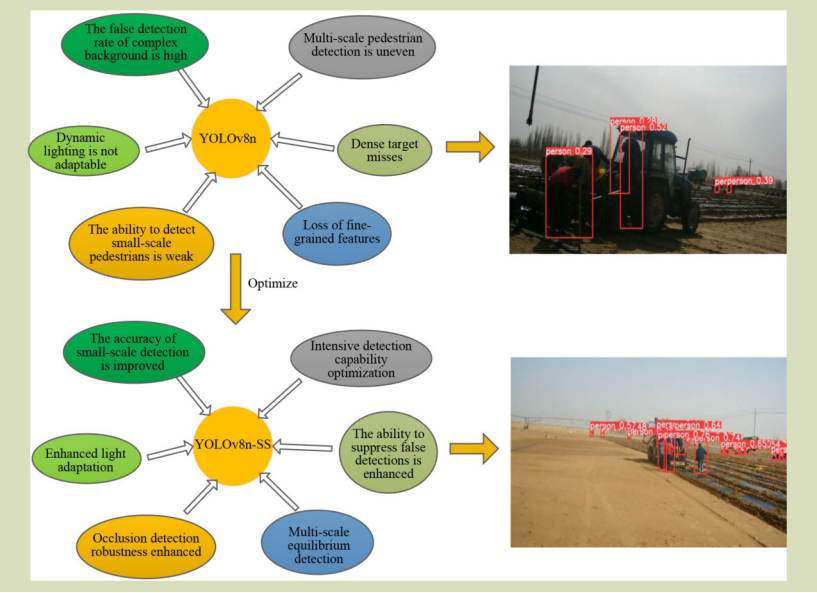
YOLOv8n, dense pedestrian detection, SPD-Conv, SK attention mechanism, adaptive extraction

HIGHLIGHTS

- An enhanced YOLOv8n-SS pedestrian detection algorithm is developed.
- Integrate spatial pyramid dilated (SPD) with convolution layers for SPD-Conv modules.
- Incorporate selective kernel attention mechanisms to enable context-aware feature selection and adaptive feature extraction.
- An efficient and robust intelligent monitoring solution for agricultural mechanization.

Received November 13, 2024;
Accepted March 24, 2025.

GRAPHICAL ABSTRACT



ABSTRACT

To address the dual challenges of excessive energy consumption and operational inefficiency inherent in the reliance of current agricultural machinery on direct supervision, this study developed an enhanced YOLOv8n-SS pedestrian detection algorithm through architectural modifications to the baseline YOLOv8n framework. The proposed method had superior performance in dense agricultural contexts while improving detection capabilities for pedestrian distribution patterns under complex farmland conditions, including variable lighting and mechanical occlusions. The main innovations were: (1) integration of spatial pyramid dilated (SPD) operations with conventional convolution layers to construct SPD-Conv modules, which effectively mitigated feature information loss while enhancing small-target detection accuracy; (2) incorporation of selective kernel attention mechanisms to enable context-aware feature selection and adaptive feature extraction. Experimental validation revealed significant performance improvements over the original YOLOv8n model. This enhanced architecture achieved 7.2% and 9.2% increases in mAP0.5 and mAP0.5:0.95 metrics respectively for dense pedestrian detection, with corresponding improvements of 7.6% and 8.7% observed in actual farmland working environments. The proposed method

ultimately provides a computationally efficient and robust intelligent monitoring solution for agricultural mechanization, facilitating the transition from conventional agricultural practices toward sustainable, low-carbon production paradigms through algorithmic optimization.

1 Introduction

Smart agricultural equipment serves as a critical enabler for the transition of modern agriculture toward ecological sustainability, intelligent advancement and long-term viability. With the increasing prevalence of advanced agricultural technologies, such as automated farming machinery and AI-driven tractors, in contemporary agricultural practices, operational safety has emerged as a paramount concern. These intelligent systems typically use sophisticated sensing technologies, including computer vision systems and LiDAR sensors, to comprehensively perceive their operational environments. Through real-time identification and continuous monitoring of personnel, pedestrians and agricultural contexts, they effectively mitigate potential hazards associated with farm machinery operations while maintaining agricultural productivity. The accurate detection of dynamic human presence within agricultural settings constitutes not only the fundamental requirement for implementing machine-human collision avoidance systems but also establishes the essential foundation for developing more sophisticated human-machine collaborative frameworks in precision agriculture.

An effective pedestrian detection system can identify, track or avoid pedestrians in a timely manner, greatly improving the efficiency and intelligence level of production safety management. The recognition of objects in agriculture can involve both sparse objects and crowded objects, which requires the establishment of a complete, universal and accurate recognition system. Current pedestrian detection algorithms often fail to meet the requirements of high accuracy and real-time performance, especially in crowded and complex environments[1]. Therefore, how to achieve efficient and accurate detection in dense pedestrian scenes is an important question for contemporary research to address[2]. This not only helps to improve the efficiency of urban safety management, but also promotes the process of agricultural modernization and provides a possible solution for achieving higher quality agricultural production.

Optimize workflows, improve work efficiency, and even use drones or robots to assist in monitoring and managing farmland through intelligent monitoring of the agricultural operating environment, to compensate for the lack of labor to a certain extent. When using certain equipment, such as self-driving tractors and harvesters, the pedestrian detection model can identify and judge pedestrians in the surrounding environment in real time, ensuring that the equipment operation does not conflict with people. This is essential to improve the safety of farm operations, especially in fields and orchards, where workers or farmers may enter the work area to follow harvesters to pick up missed produce. By installing cameras or sensors, pedestrian detection models can identify these pedestrians entering the area, warn them in time and even automatically stop the machine to avoid collisions and accidental injuries.

For this detection task, there have been multidirectional studies at home and abroad, such as YOLOv5 network improvement for traffic sign detection, and the efficiency and accuracy of YOLOv5 have been highly recognized[3]; The small-world Hopfield neural network of fuzzy synapses is used for digital recognition, which focuses on digital recognition tasks, where the targets are usually static and structured, but lacks the ability to adapt to complex scenes in pedestrian detection tasks[4]. The vehicle-mounted adaptive traffic sign detector focuses on the traffic sign detection task and is more optimized for the special needs of the vehicle-mounted environment, such as dynamic background, lighting change and motion blur, but it lacks versatility in the pedestrian detection task and is difficult to directly migrate to the pedestrian the detection task[5]. Compared with the above methods, the main advantages of the improved YOLOv8 in the pedestrian detection task are high precision, real-time, multiscale processing, strong adaptability and improved performance in dynamic target detection.

The main challenge of pedestrian detection technology is that the increase in pedestrian density leads to an increase in occlusion and overlap, which makes many existing detection algorithms face greater difficulties. Although deep learning technology has made significant progress in recent years[6], target detection models such as YOLO[7,8] (you only look once), SSD[9,10] (single shot multiBox detector) have performed well in many standard data sets. have excellent performance,

but they still have certain limitations when dealing with dense scenes. For example, the accuracy of the YOLO series models may be affected by the mismatch between the detection frame and the actual shape of the pedestrian. Especially in the case of small and dense targets, missed detections or false detections often occur.

To solve these problems, research in recent years has gradually tilted toward more complex network architecture and technical means. New researchers have introduced techniques such as multiscale feature fusion, attention mechanism, and adaptive anchor frame[11–13] to improve the performance of the model in complex scenarios. These methods aim to enhance the ability of the model to capture different small targets and dense objects, to improve the accuracy and robustness of detection. In addition, the wide application of data augmentation technology also provides more diverse samples for training models, so that the models can more effectively adapt to complex environments in practical applications. In recent years, Xie et al.[14] proposed the PSC-Net (pointwise spatial convolution network) method, which includes a dedicated module to explicitly capture inter- and intra-part co-occurrence information of different body parts of pedestrians through graphical convolutional networks. Chen et al.[15] introduced the spatial attention module on the basis of YOLO and added it to the backend of the backbone network Darknet-53 to realize the weight amplification of important feature information in the spatial dimension. For dense scenes, Li et al.[16] used the aggregation-distribution mechanism to reconstruct the neck structure of YOLOv8 on the basis of YOLOv8, so that multilevel information could be fused, more efficient information exchange was achieved, and the detection ability of the model was enhanced. Wang et al.[17] proposed to use dense blocks instead of residual blocks to reduce the number of network structure parameters and avoid unnecessary calculations. Gong et al.[18] proposed a multi-sensor module that uses multi-amplification convolution to sample feature images, which avoids information loss caused by repeated sampling, to improve the feature extraction and object detection performance of the algorithm. In 2023, the Ultralytics group[19] proposed the YOLOv8 algorithm, which has five models: n, s, m, l and x, and the main differences between them are the size of the model, the number of parameters, the computing resource requirements and the performance. Although YOLOv8n is slightly lacking in accuracy, it has significant advantages in speed and resource efficiency, and is suitable for targeted improvements in pedestrian detection.

Based on the research on the YOLO model, this study used the YOLOv8n model, and the detection accuracy was improved through optimization measures, which included two aspects. (1) For the occlusion problem, because the network will miss the key features of the labeled pedestrians in the learning process, this study introduced the SK attention mechanism module into the C2f (cross stage partial two fusion) module of YOLOv8n, and selects and fuses the feature information of different receptive fields through network learning[20]. In a weakly supervised way, the network is adaptively guided to give more attention to the visible parts of occluded pedestrians, that is, different channels in the feature map, and the passages of occluded pedestrians are given higher weights, to help the backbone network focus on key features and suppress non-critical features, to improve the accuracy of the model in detecting occluded objects. (2) For small and low-resolution targets, because convolution will lead to the loss of some fine-grained information in the process of downsampling and pooling, and the small target will not be able to obtain enough effective feature learning in the training process, this study replaced the partial convolutional neural network (CNN) Conv with the SPD-Conv (space-to-depth and pointwise convolution) module in the network structure, to replace the downsampling and pooling process, so that the model can reduce the loss of fine-grained information and improve the segmentation accuracy of small objects in the image[21].

2 Related work

2.1 YOLOv8

Object detection occupies an important position in the field of computer vision, and the YOLO series models have received considerable attention because of their real-time and high efficiency. Since the advent of YOLO v1 in 2016, the YOLO series has continued to evolve, with each release of the algorithm, which is composed of three main components: the backbone network, the neck network and the detection head[22].

With the continuous evolution of the YOLO series, the research requirement for object detection is also increasing[23–26], and compared with the previous version of YOLO, YOLOv8 has made the following optimizations.

(1) Optimization of network structure. Compared with the v7 version, YOLOv8 uses the C2f module to replace the commonly used C3 (cross-stage-partial bottleneck with three convolutions) module, which further reduced the amount of computation and improves the calculation speed. In addition,

YOLOv8 also introduced the SPPF (spatial pyramid pooling fast) module, which can extract features on different receptive fields, so that the network can more effectively capture the spatial information of the target, so that the model can effectively process targets of different sizes, thereby enhancing the object detection ability.

(2) Improvement of the loss function. While the standard YOLO models usually use GIoU (generalized intersection over union) as the loss function of the bounding box, YOLOv8 introduces a new type of comprehensive loss function, which effectively integrates classification loss, regression loss and confidence loss. Categorical loss (usually using cross-entropy loss), confidence loss (usually based on binary cross-entropy), regression loss DIoU (distance intersection over union)[27] and CIoU (complete intersection over union)[28] are substitutes. This integration method can effectively balance the influence of various losses in training, not only considering the overlapping area of the target bounding box, but also introducing the difference in the distance between the center points of the bounding box and the aspect ratio, to make the model more accurate when locating the target and improve the overall performance of the model.

(3) Enhancement of multiscale detection. In the YOLO series models, multiscale detection is one of the key technologies to improve the detection accuracy. YOLOv8 further improved the expression of multiscale features by using a combination of FPN (feature pyramid network)[29] and PANet (path aggregation network)[30]. FPN can propagate high-level semantic information through a top-down path, while PANet aggregates low-level detail information through a bottom-up path, so that features of different scales can be more effectively fused, thereby improving the accuracy of detecting small and large targets.

(4) Improvement of training strategies. YOLOv8 adopts mixed precision training and multi-task learning. The former is achieved by dynamically adjusting the calculation accuracy during the training process to reduce the memory footprint and speed up the training speed. The latter improves the overall performance of the model by optimizing both object detection and classification tasks. In addition, compared to the v5 version, YOLOv8 also uses an improved anchor mechanism, moving from the anchor-based approach of YOLOv5 to no anchor points, that is, by automatically selecting the best anchor point, making the model more flexible and efficient when dealing with targets of different shapes and sizes.

Figure 1 shows the YOLOv8 model structure incorporating the improvements made in this study. The Conv and C2f modules are briefly described as follows. Conv consists of a Conv2d (i.e., 2 dimensional) layer, a BatchNorm2d layer and a Conv2d layer for convolution operations and the results are summed to generate a feature map, and the BatchNorm2d layer is batch normalization to improve the stability and convergence speed of the model.

C2f (Fig. 2) is one of the main modules of YOLOv8, which works by first passing the input through the first convolutional layer, and then dividing the output into two parts. One part is passed directly to the output, and the other part is processed by multiple bottleneck modules. Finally, the results of the two parts are stitched together in the channel dimension and passed through the second convolutional layer to obtain the final output.

2.2 Improved YOLOv8n model

In this study, the Conv module in the YOLOv8n model was improved with the SK (selective kernel) attention mechanism was added to the neck network part to solve the problem of low accuracy in complex scenes and small targets.

2.2.1 SPD-Conv

The standard CNN does not perform well when dealing with low-resolution images or small objects because there are some flaws in the existing CNN architecture, that is, the use of step convolution and pooling layers, which will lead to the loss of fine-grained information and the degradation of feature representation capabilities. For this study, we adopted a new convolution method, SPD-Conv, to extract image features in two steps through the spatial depth (SPD) module and non-stepping convolution, to reduce information loss and ultimately optimize the problem of low-resolution and small-task objectives.

SPD-Conv (Fig. 3) is composed of SPD layer and non-step Conv layer, and the SPD layer is vital for the downsampling of the feature map within CNN and across feature maps, which can be rearranged through the downsampling operation to transform the spatial information into depth information. The SPD module extracts a sequence of sub-feature graphs $f_{x,y}$ by scale factor sampling for the feature map $X_{i,j}$ with size $S \times S \times C_1$, and the number of sub-features in this sequence is $scale^2$. Given that the sub-feature map $f_{x,y}$ is sliced by the feature map $X_{i,j}$, $i+x$ and $y+j$ can satisfy the divisibility scale, and each sub-feature map is the size $(\frac{s}{scale}, \frac{s}{scale}, scale^2 C_1)$, $x+y=scale^2$. The spatial dimension is reduced by scale and

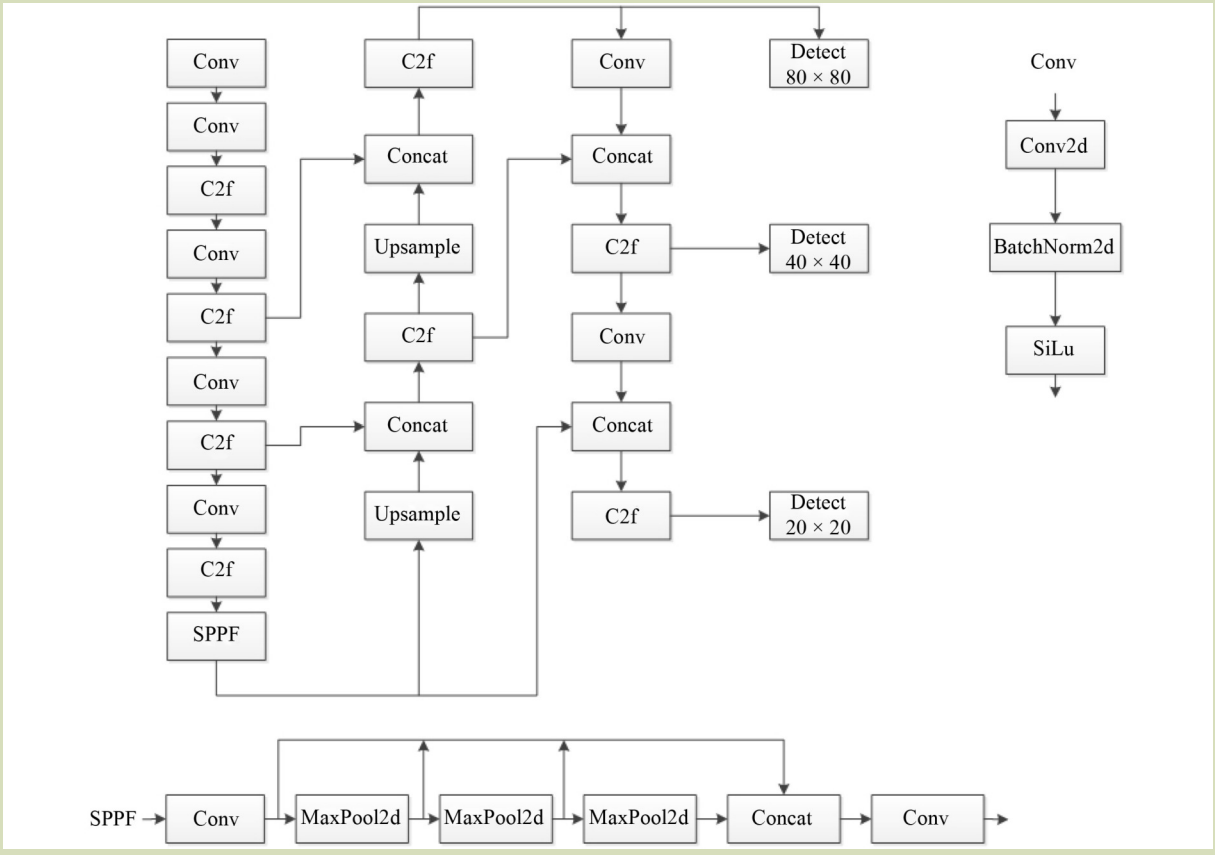


Fig. 1 YOLOv8 network architecture.

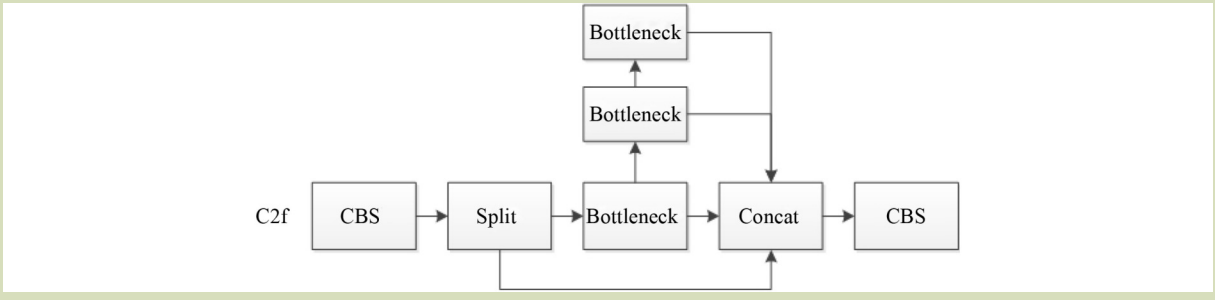


Fig. 2 C2F structure diagram.

the channel dimension is increased by scale. The calculation process of sub-feature mapping $f_{x,y}$ is described as follows:

$$\begin{aligned} &f_{0,0} = X[0:S:scale, 0:S:scale], \\ &f_{1,0} = X[1:S:scale, 0:S:scale], \ldots, \\ &f_{scale-1,0} = X[scale-1:S:scale, 0:S:scale]; \end{aligned} \tag{1}$$

$$\begin{aligned} &f_{0,1} = X[0:S:scale, 1:S:scale], f_{1,1}, \ldots, \\ &f_{scale-1,1} = X[scale-1:S:scale, 1:S:scale]; \end{aligned} \tag{2}$$

$$\begin{aligned} &f_{0,scale-1} = X[0:S:scale, scale-1:S:scale], f_{1,scale-1}, \ldots, \\ &f_{scale-1,scale-1} = X[scale-1:S:scale, scale-1:S:scale]. \end{aligned} \tag{3}$$

After the SPD layer operation, a non-stepping convolutional layer is added in step sizes of 1, which has C_2 filters, where $C_2 < scale^2 C_1$.

In the experiment, in order to improve the detection of small

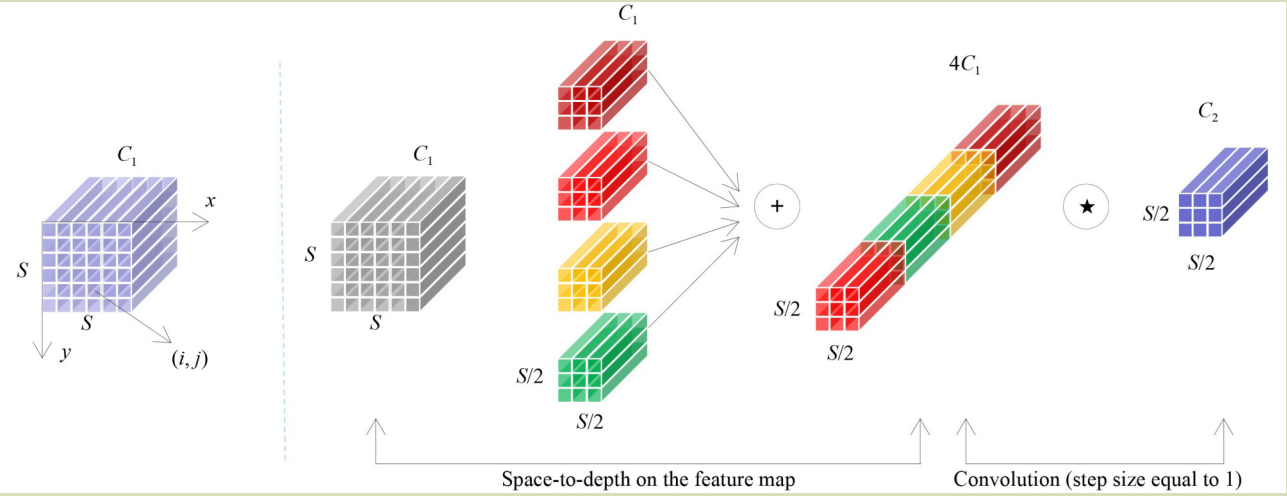


Fig. 3 The structure of SPD-Conv module.

targets and low-resolution targets, the Conv module was improved in the network model (Fig. 4), the SPD-Conv module used the SPD module to collect feature maps and then the first convolutional layer in the C2f module in YOLOv8 was converted into two parts, which are merged after different convolutional layers and then pass through the second convolutional layer. Finally, the SPD-Conv operation was completed by non-compiled convolution, so that the model can fully learn the feature information of the selected region.

2.2.2 SK attention mechanism

SKNet (selective kernel network) captures multiscale features at different levels by selectively applying convolutional kernels at different scales. To achieve this, a selection module was introduced that adaptively decides which scales of convolution kernels to use on each channel. This selective multiscale convolutional kernel helps to improve the ability of feature representation, making the network more adaptable and generalizable. In the SK attention mechanism, there are three key operations on the input feature map: split, fuse and select (Fig. 4).

The segmentation operation performs multi-branch separation convolution of the feature map, and input 3 × 3 and 5 × 5 size convolution kernels to extract the features of the input feature map, assuming that the branch is n, the dimension of the feature map is transformed from (c, h, w) to (n, c, h, w). The fusion operation is to add a plurality of feature extraction results, that is, U_1 and U_2 are summed to obtain U, the feature dimension is changed from (n, c, h, w) to (c, h, w), and then through the global average pooling layer F_{gp}, the feature dimension (c, h, w) is transformed into (c, 1, 1), F_{fc} is a two-layer fully connected layer, the process is to first reduce the dimension, then increase the dimension and finally the feature map dimension is n (c, 1, 1), and the output is the matrices a

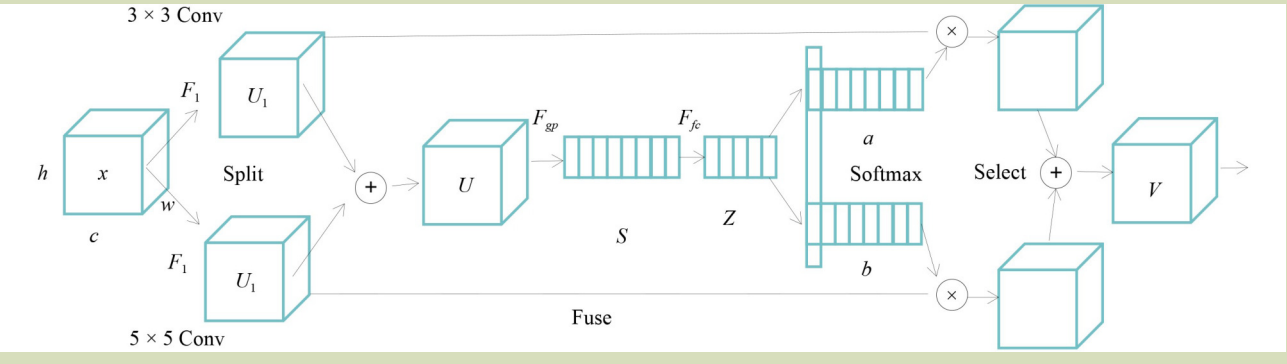


Fig. 4 The working principle of the SK attention mechanism.

and b, where the matrix b is a redundant matrix; The selection operation is to multiply the multiple feature vectors processed by the softmax function with the feature map extracted by multiple branches in the first step, that is, U_1 and U_2 are weighted by using two weight matrices of a and b, and finally the n feature maps are summed to obtain vector V.

The c-element s is calculated by the spatial dimensions $H \times W$ shrinkage U:

$$s_c = F_{gp}(U_u) = \frac{1}{H \times W} \sum_{i=1}^{H} \sum_{j=1}^{W} U_c(i, j) \tag{4}$$

Z is a compact eigenvector that compresses the dimension for greater efficiency:

$$\boldsymbol{Z} = \boldsymbol{F}_{fc}(\boldsymbol{s}) = \boldsymbol{\delta}(\boldsymbol{B}(\boldsymbol{W}\boldsymbol{s})) \tag{5}$$

where, δ is the activation function of ReLU and B is batch normalization.

In this study, the most information-rich feature map of the original model and the detection head are fused. The feature map processed by the attention module is fused between the multiscale feature network and the detection head, and the target weight is reserved for the feature map with the most abundant information, which improves the detection accuracy.

2.2.3 YOLOv8n-SS

By adding the SK attention mechanism block and the SPD-Conv module, this study proposes the YOLOv8n-SS model (Fig. 5), which is used to solve the problems of complex environments and small detection objects. In this study, the SPD module is added before the backbone network C2f module in the YOLOv8n model to solve the influence of low-resolution images on the detection results. It progresses the subsequent Conv modules to form the SPD-Conv module, but the intensive sampling of the SPD-Conv module may make it

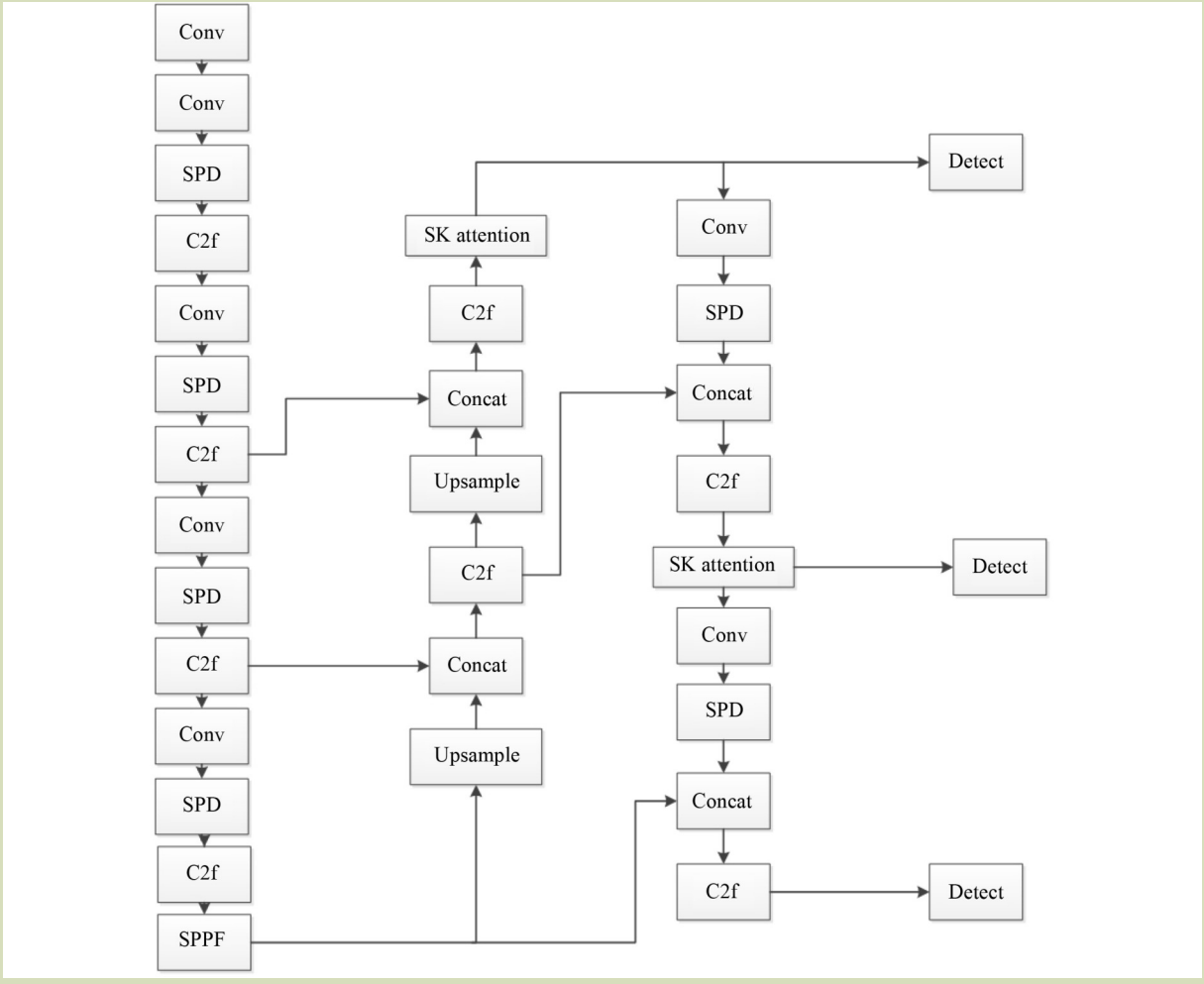


Fig. 5 YOLOv8n-SS network architecture.

difficult for the YOLOv8n model to locate the target region. Therefore, the SK attention mechanism module was added to the network structure of the YOLOv8n model to solve this problem. Finally, a YOLOv8n-SS model was obtained, which can deal with the above problems well (detection of pedestrians in environments with small targets and complex presence).

Figure 6 shows the configuration of the YOLOv8n-SS model, with the backbone network on the left and the head network on the right. From left to right, the source, the number of repetitions, the module and the parameters are displayed.

The backbone is responsible for feature extraction. The modifications included the SPD (space-to-depth) module having, after certain convolutional layers, additional downsample feature maps while preserving spatial information. The C2f Module, a custom module used for feature extraction, was repeated multiple times at different stages and the SPPF was used at the end of the backbone to aggregate features at different scales.

The head network is responsible for detecting objects using the features extracted by the backbone. The modifications included: upsampling used to increase the resolution of feature maps; Concat used to combine feature maps from different stages of the backbone; SK attention: function added to enhance feature representation by dynamically selecting kernel sizes based on attention mechanisms; and SPD modules were used in the head to downsample feature maps.

3 Experiment and analysis

3.1 Data sets

3.1.1 CrowdHuman data set

The CrowdHuman data set[31] is a public data set released by Queungshi for the public to do pedestrian detection research, with a huge amount of data, most of the pictures contain dense crowds, diverse crowd environments and complex human postures. At present, it has 15,000 training sets, 5000 test sets and 4370 verification sets, totaling more than 470,000 pedestrians, which are mainly used for human detection and human pose estimation tasks and are suitable for various application scenarios. For this study, 15,000 images of the training set were selected for experimentation, and the training set and the verification set were divided into 8:2, with the first 12,000 images as the training set and the last 3000 images as the verification set.

3.1.2 Farmland pedestrian data set

The farmland pedestrian data set was obtained from the video interception of agricultural machinery and equipment operating environment in six farmland contexts, and a total of 4800 images were obtained. A total of more than 26,000 pedestrians were imaged, and the pedestrians in the images were labeled using the LabelImg tool, of which 60% of the pedestrians were small targets and occluded pedestrians in agricultural production. For this study, the data set was divided into a 5:1 training set and a validation set.

```
backbone:

  - [-1, 1, Conv, [64, 3, 1]]  # 0-P1/2
  - [-1, 1, Conv, [128, 3, 1]]  # 1-P2/4
  - [-1, 1, SPD, [1]]  # 2-P2/4
  - [-1, 3, C2f, [128, True]]
  - [-1, 1, Conv, [256, 3, 1]]  # 4-P3/8
  - [-1, 1, SPD, [1]]
  - [-1, 6, C2f, [256, True]]
  - [-1, 1, Conv, [512, 3, 1]]  # 7-P4/16
  - [-1, 1, SPD, [1]]
  - [-1, 6, C2f, [512, True]]
  - [-1, 1, Conv, [1024, 3, 1]]  # 10-P5/32
  - [-1, 1, SPD, [1]]
  - [-1, 3, C2f, [1024, True]]
  - [-1, 1, SPPF, [1024, 5]]  # 13
```

```
head:
  - [-1, 1, nn.Upsample, [None, 2, 'nearest']]
  - [[-1, 8], 1, Concat, [1]]
  - [-1, 3, C2f, [512]]  # 16

  - [-1, 1, nn.Upsample, [None, 2, 'nearest']]
  - [[-1, 5], 1, Concat, [1]]
  - [-1, 3, C2f, [256]]
  - [-1, 1, SKAttention, [256]]

  - [-1, 1, Conv, [256, 3, 1]]
  - [-1, 1, SPD, [1]]
  - [[-1, 16], 1, Concat, [1]]
  - [-1, 3, C2f, [512]]
  - [-1, 1, SKAttention, [512]]

  - [-1, 1, Conv, [512, 3, 1]]
  - [-1, 1, SPD, [1]]
  - [[-1, 13], 1, Concat, [1]]
  - [-1, 3, C2f, [1024]]

  - [[20, 25, 29], 1, Detect, [nc]]
```

Fig. 6 YOLOv8n-SS model structure configuration.

3.2 Introduction to the experimental environment

The operating system of the experimental platform was Windows11, the CPU is Intel Core i7-12650H @2.3GHz, the memory was 16 GB, the graphics card is Nvidia GeForce RTX 4060, and the video memory is 8GB. The CUDA version used by Torchvision was 11.6, the deep learning framework was Pytorch-GPU 1.13.1 + cu117, the deep neural network GPU acceleration library was cuDNN8.4.0, and the image size used in the data set was 640 × 640. All models were trained using the default optimizer SGD, with a training batch (batch-size) of eight, a training count of 200 and an initial learning rate of 0.01.

3.3 Evaluation indicators

The accuracy (*P*), recall rate (*R*) and average accuracy (*mAP*) were the main evaluation indexes of the model. For this study, *mAP* was used as a reference, and the higher its value, the improve the ability to detect the effect of the model. The calculation formula is as follows:

$$Precision = \frac{TP}{TP+FP} \tag{6}$$

where, *TP* (true positive) is the number of pedestrians predicted by the classifier and the actual result, that is, the number of samples correctly identified, *FP* (false positive) is the predicted result of the classifier is pedestrians, but the actual result is not pedestrians, that is, the number of incorrectly identified samples, and *TP* + *FP* is the sum of the number of pedestrians samples.

$$Recall = \frac{TP}{TP+FN} \tag{7}$$

where, *Recall* is the recall rate, *FN* (false negative) is the classifier predicts that the result is not a pedestrian, the actual result is a pedestrian, that is, the number of missing recognition samples, *TP* + *FN* is the actual result is the sum of the number of pedestrian samples, and the larger the Recall, the less it is missed.

$$AP = \int_0^1 P(r)\mathrm{d}r \tag{8}$$

where, *AP* is the average accuracy and *P(r)* is the accuracy of the recall rate *r*.

$$mAP = \frac{\sum_{n=1}^{N} AP_n}{N} \tag{9}$$

where, an *A* higher than *mAP* indicates that the model performs more effectively overall across the entire data set.

3.4 Comparison with other algorithms

In this study, the improved model was compared with other detection models on the same data set, including YOLOv5n, YOLOv6n, YOLOv8n, YOLOv8n-small (with a small object detection layer) and YOLOv8n-CBAM. All models were trained for 200 rounds and none of them used pre-training weights. Table 1 and Table 2 give the training results of each model.

In Table 1, compared with other models, the YOLOv8n-small model was superior compared with YOLOv8n, with P, mAP@0.5 and mAP@0.5:0.95 reaching 81.4%, 77.4% and 50.8%, respectively. Although the accuracy was not improved, only 0.9% lower than that of the YOLOv8n model, the model recall and average accuracy numerically higher but not significantly. It is worth noting that the YOLOv8n-SS model gave the best performance in accuracy, recall, mAP@0.5 and mAP@0.5:0.95, with very significant improvements, reaching 83.1%, 73.8%, 83.6% and 58.7%, respectively. Compared to YOLOv8n, the improved algorithm improved mAP@0.5 by 7.2% and mAP@0.5:0.95 by 9.2%, but with an increased computational load.

In Table 2, compared with other models, the YOLOv8n-SS model achieved the best performance in accuracy, recall,

Table 1 Detection accuracy of each model in CrowdHuman data set

Model	P (%)	R (%)	mAP@0.5 (%)	mAP@0.5:0.95 (%)	Parameter (10^6)	GFLOPs
YOLOv5n	81.5	63.8	75.2	48.1	2.5	7.1
YOLOv6n	81.9	64.7	75.6	48.8	4.2	11.8
YOLOv8n	82.3	65.1	76.4	49.5	3.0	8.9
YOLOv8n-small	81.4	66.2	77.4	50.8	3.0	12.5
YOLOv8n-CBAM	82.4	64.9	76.2	49.4	3.1	8.2
YOLOv8n-SS	83.1	73.8	83.6	58.7	5.2	72.6

Table 2 Detection accuracy of each model in farmland pedestrian data set

Model	P (%)	R (%)	mAP@0.5 (%)	mAP@0.5:0.95 (%)	Parameter (10^6)	GFLOPs
YOLOv5n	78.6	61.0	71.1	41.0	2.5	7.1
YOLOv6n	78.3	61.6	70.7	41.4	4.2	11.8
YOLOv8n	79.1	61.6	71.4	41.6	3.0	8.9
YOLOv8n-SS	80.1	70.5	79.0	50.3	5.2	72.6

mAP@0.5 and mAP@0.5:0.95, with significant improvements of 80.1%, 70.5%, 79.0% and 50.1%, respectively. Compared with YOLOv8n, the improved algorithm improved by 7.6% on mAP@0.5 and 8.7% on mAP@0.5:0.95.

According to these results (Fig. 7 and Fig. 8), YOLOv8n-SS gave a significant improvement compared with other models in Recall, mAP@0.5 and mAP@0.5:0.95, with a faster change trend in the first seven rounds, obvious differences by round 60, and gradual stabilization from round 150.

To evaluate the effects of each augmentation module proposed in the YOLOv8n-SS and to assess the extent of these effects, an ablation study with consistent parameters and training procedures was performed on the same data set. The results are summarized in Table 3 and Table 4, where the tick indicates the use of the corresponding module.

A series of ablation experiments were performed on the benchmark model YOLOv8n to evaluate the effect of adding different modules on the performance of pedestrian detection (Table 3). The accuracy (P) of the initial benchmark model was 82.3%, (R) 65.1%, (mAP@0.5) 76.4% and (mAP@0.5:0.95) 49.5%, respectively. We added SK and SPD-Conv, respectively. The results showed that P was 81.8%, R was 66.4%, mAP@0.5 was 77.0%, and mAP@0.5:0.95 was 50.3%, respectively, and the

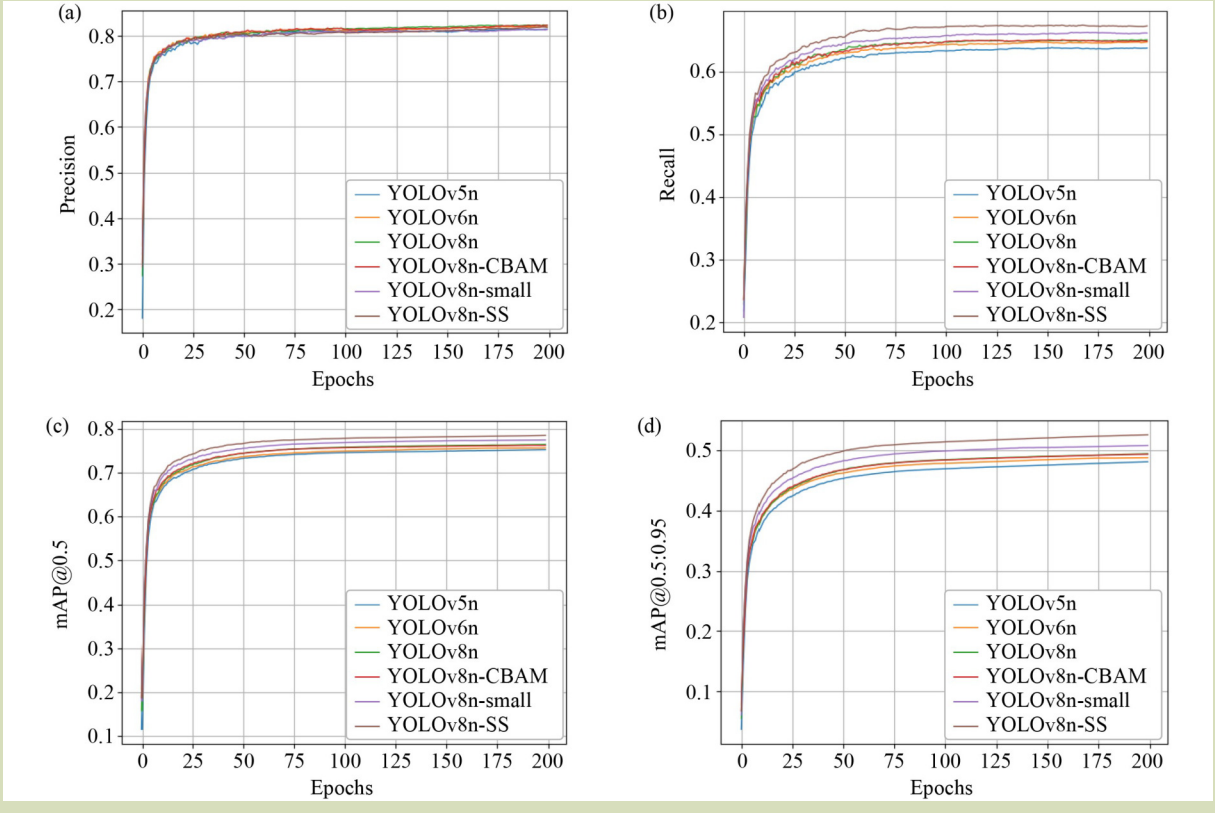


Fig. 7 Comparison of detection performance of each models on Crowdhuman: (a) precision, (b) recall, (c) mAP@0.5, (d) mAP@0.5:0.95.

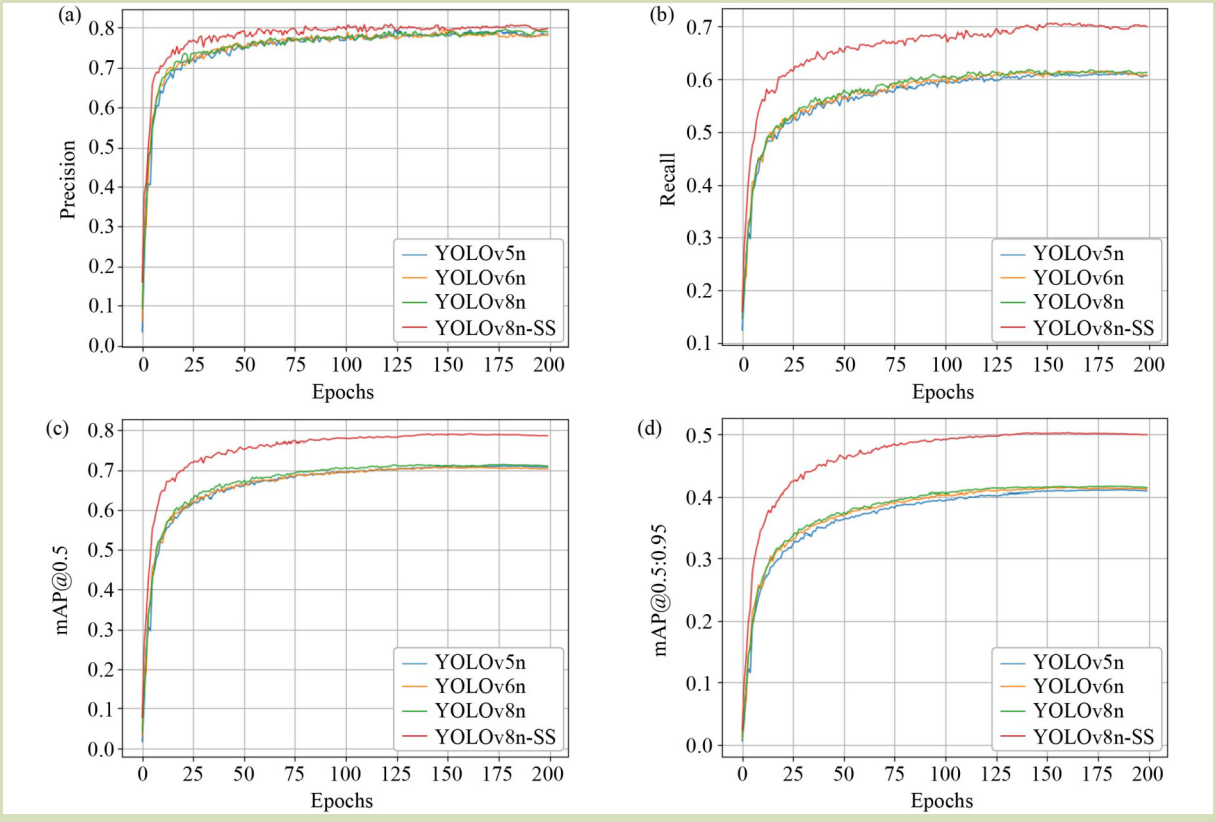


Fig. 8 Comparison of detection performance of each models on farmland pedestrian: (a) precision, (b) recall, (c) mAP@0.5, (d) mAP@0.5:0.95.

Table 3 Module ablation experiment on Crowdhuman

Model	SK	SPD-Conv	P (%)	R (%)	mAP@0.5 (%)	mAP@0.5:0.95 (%)	Parameters (10^6)	GFLOPs
YOLOv8n			82.3	65.1	76.4	49.5	3.0	8.9
	√		81.8	66.4	77.0	50.3	4.8	16.9
		√	82.9	73.0	83.0	57.9	3.5	50.9
	√	√	83.1	73.8	83.6	58.7	5.2	72.6

Table 4 Module ablation experiment on farmland pedestrian

Model	SK	SPD-Conv	P (%)	R (%)	mAP@0.5 (%)	mAP@0.5:0.95 (%)	Parameters (10^6)	GFLOPs
YOLOv8n			79.1	61.6	71.4	41.6	3.0	8.9
	√		79.4	62.3	72.6	42.9	4.8	16.9
		√	79.9	68.6	77.8	49.5	3.5	50.9
	√	√	80.1	70.5	79.0	50.3	5.2	72.6

accuracy of the model was slightly reduced, but the recall rate and average accuracy improved. When only SPD-Conv was added, the various items of the model increased significantly, P reached by 82.9%, R reached 73.8%, and mAP@0.5 and mAP@0.5:0.95 increased significantly, reaching 83.0% and 57.9%, respectively. After further combinatorial experiments,

we observed that the model was superior in P, R, mAP@0.5 and mAP@0.5:0.95 compared to SPD-Conv alone.

As given in Table 4, the accuracy (P) of the initial benchmark model was 79.1%, (R) 61.6%, (mAP@0.5) 71.4%, and (mAP@0.5:0.95) 41.6%, respectively. We added SK and SPD-Conv, respectively. The results showed that P was 79.4%, R was 62.3%, mAP@0.5 was 72.6% and mAP@0.5:0.95 was 42.9%, and the accuracy, recall and average accuracy of the model were improved. When SPD-Conv was added alone, all items of the model increased significantly, P reached 79.9%, R reached 68.6%, and mAP@0.5 and mAP@0.5:0.95 increased significantly, reaching 77.8% and 49.5%, respectively. After further combinatorial experiments, we observed that P, R, mAP@0.5 and mAP@0.5:0.95 were significantly improved by combination of modules compared to SPD-Conv alone.

3.5 Detect performance

To intuitively prove the superiority of the algorithm in recognizing pedestrians in complex environments, we use YOLOv8n and YOLOv8n-SS pairs to detect the images of dense data sets, which are roughly divided into three scenarios, namely, dark light, dense crowds and serious occlusion. The detection performance is reflected from these three aspects, and the detection performance is shown in Figs. 9–11. Figures 9–11 were modified based on photos, which are open source in a benchmark data set, CrowdHuman by Shao et al.[31].

With dim light (Fig. 9), the original image has six standing people and the detection results of YOLOv8n are inaccurate; it detects only five pedestrians and the pedestrians partially blocked on the left side are not detected. YOLOv8n-SS detects all six pedestrians and the ability to detect targets in dark places is also greatly improved, and the confidence of the detection frame is also high. Withs crowds (Fig. 10), the original image has 43 people and the detection results of YOLOv8n include some false detections. It detected 49 pedestrians, and mistakenly detects the lower left luggage and distant objects as pedestrians. YOLOv8n-SS detected 45 pedestrians, which effectively matches the detection of pedestrians compared with the original model. With severe occlusion (Fig. 11), the original image has 33 standing people but YOLOv8n did not detect the people in the back row with only heads visible and mistakenly fails to detect the leftmost person. Although the number of people detected was 33, there were false and missed detections. YOLOv8n-SS detected 32 people and accurately detect the

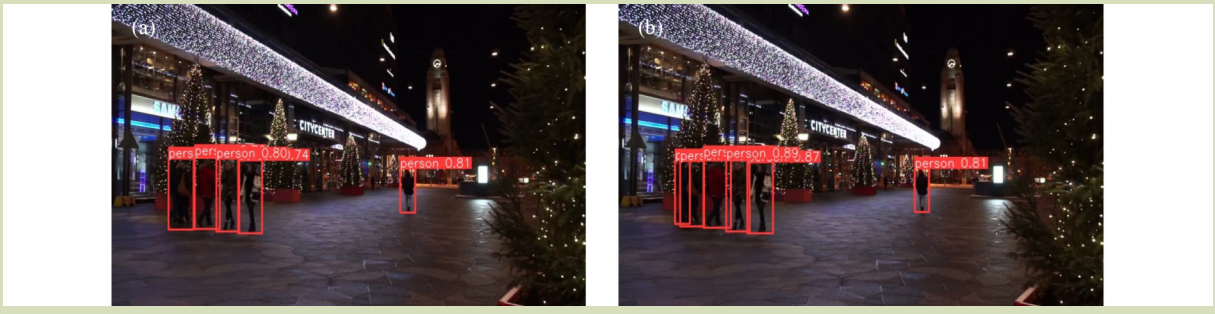


Fig. 9 In dim light: (a) YOLOv8 recognition, (b) YOLOv8n-SS recognition. Data from CrowdHuman by Shao et al.[31] under open access.

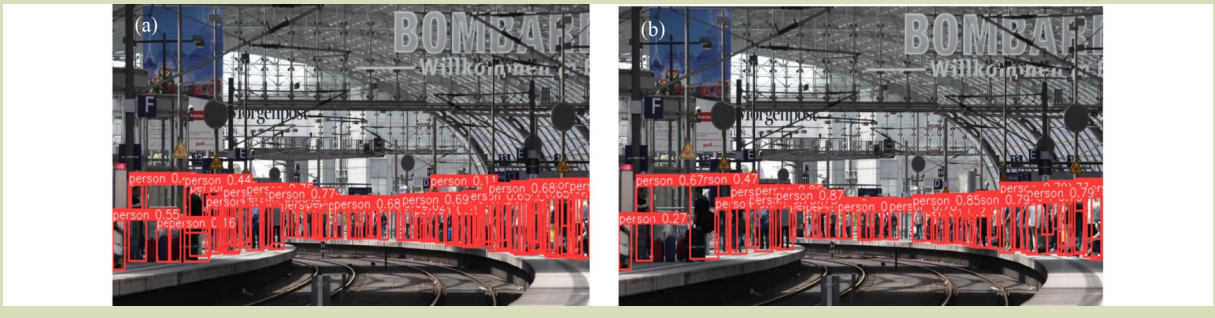


Fig. 10 Crowded: (a) YOLOv8 recognition, (b) YOLOv8n-SS recognition. Data from CrowdHuman by Shao et al.[31] under open access.

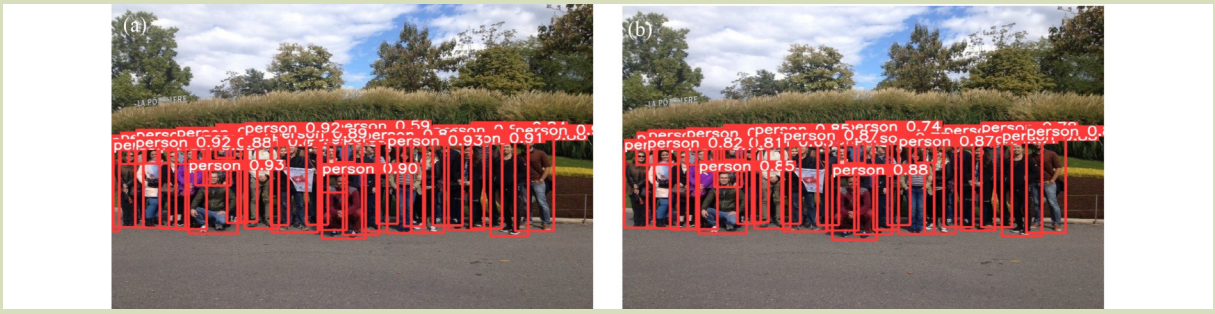


Fig. 11 Heavily occluded: (a) YOLOv8 recognition, (b) YOLOv8n-SS recognition. Data from CrowdHuman by Shao et al.[31] under open access.

occluded people. It greatly improved the detection of the people in the back row who are severely occluded and also maintains a high confidence level. YOLOv8n-SS has a clear advantage when dealing with complex scenes, such as dim light, dense crowds and heavy occlusion. From this experiment, in Fig. 12 we present a comparison of the loss curves of the YOLOv8n and YOLOv8n-SS bounding boxes, and it evident that the improved loss function converged faster.

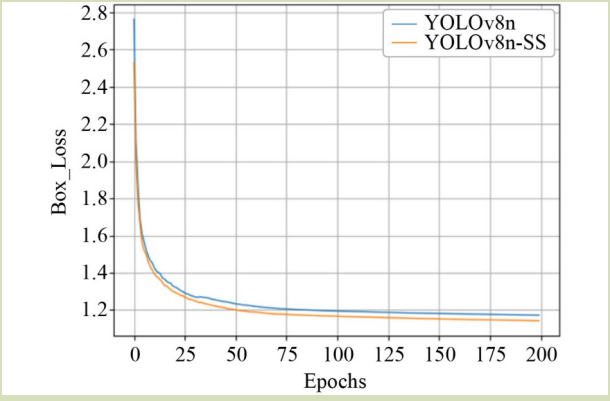


Fig. 12 Loss of the border frame.

Figure 13 and Fig. 14 provide a comparison of the pedestrian detection effect of the original YOLOv8n and the improved Crowd-YOLOv8 proposed for crowded scenes. In an agricultural scene, where pedestrians work in the farmland, the performance of the proposed model is again a significant improvement on the original model for individuals in dim light, occlusion by crowds or agricultural machinery and equipment, and small targets. The superiority of the proposed model in dense pedestrian detection was fully demonstrated. The loss curves (Fig. 15) for the YOLOv8n and YOLOv8n-SS bounding boxes converged faster in the improved model.

4 Conclusions

In this study, an improved pedestrian detection model was developed and tested, and is proposed, the main purpose of which was to solve some problems existing in the existing pedestrian detection object detection models in farmland operation scenarios, especially the problems of occlusion and small scale. From the perspective of low-resolution and small-target optimization, the proposed YOLOv8n-SS model first combines the Conv module with SPD to reduce information loss and finally optimizes the low-resolution and small-task

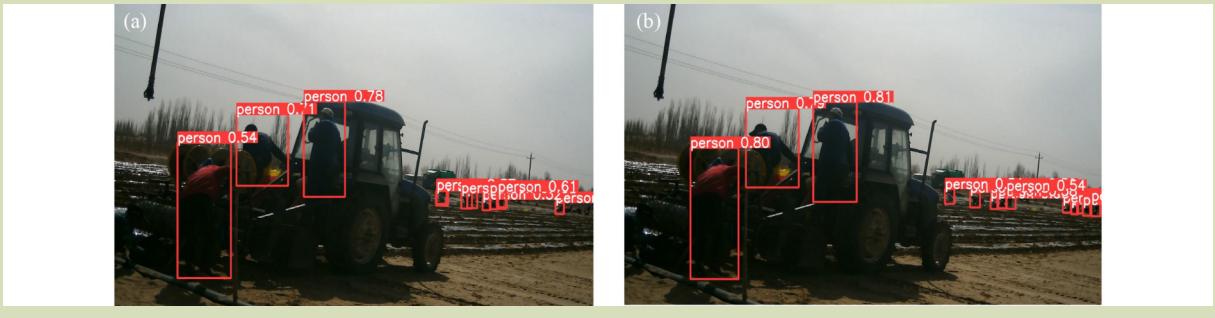


Fig. 13 Dim light and small targets: (a) YOLOv8 recognition, (b) YOLOv8n-SS recognition.

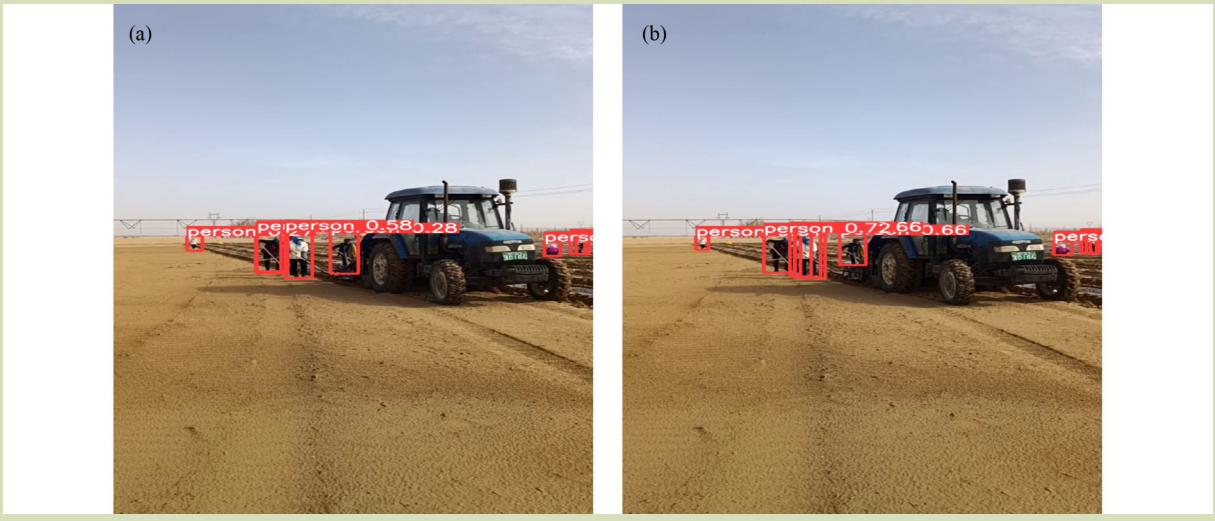


Fig. 14 Heavy occlusion and small targets: (a) YOLOv8 recognition, (b) YOLOv8n-SS recognition.

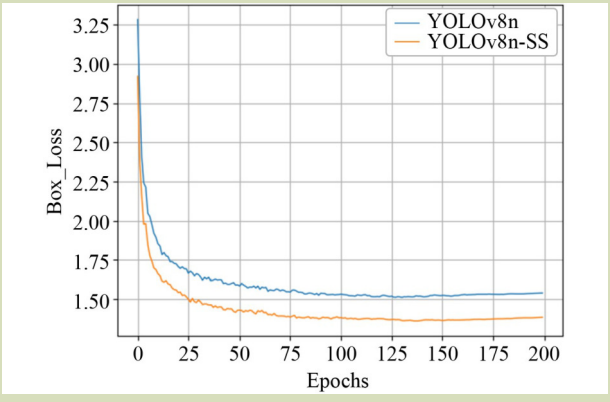


Fig. 15 Loss of the border frame.

target problems, which significantly improves the ability of the network to extract pedestrian features. Secondly, for the occlusion problem, the SKAttention function is introduced, the feature map and the detection head are fused to solve the influence of low-resolution images on the detection results, and the features of different scales and levels are adaptively extracted, and the most important features are selected. By combining these methods, the model more effectively adapted to pedestrian detection tasks in the farmland scenes, thereby improving the detection performance of the whole model.

To evaluate the effectiveness of the model, an experimental study was conducted using a widely recognized public data set CrowdHuman, and the experimental results showed that compared with YOLOv8n, the accuracy P increased by 0.8%, mAP@0.5 and mAP@0.5-0.95, respectively, by 7.2% and 9.2%, respectively, which not only showed excellent performance in identifying crowded pedestrians, but also significantly improved the recall rate. In addition, the practicality of the method was considered in this study. With the intensification of agricultural labor shortage and the development of intelligent agricultural machinery technology, the demand for pedestrian monitoring in complex farmland scenarios is becoming increasingly prominent, and the YOLOv8n-SS model has good portability and detection efficiency, which can meet the needs of pedestrian detection in complex scenarios, so the model has a wide application prospect in agricultural scenarios.

This research will continue introducing more advanced feature fusion methods and enabling the integration of object tracking and detection, further optimize the network structure, improve the inference speed, reduce the consumption of computing resources, deliver a lightweight model and combine the efficient detection ability of YOLOv8 with advanced object tracking technology to achieve long-term tracking and behavior analysis of pedestrians.

Acknowledgements

This research was supported by the General Program of the Natural Science Foundation of Hunan Province of China (2021JJ30359).

Compliance with ethics guidelines

Yanfei Li and Chengyi Dong declare that they have no conflicts of interest or financial conflicts to disclose. This article does not contain any studies with human or animal subjects performed by any of the authors.

REFERENCES

1. Chen N, Li M L, Yuan H, Li Y H, Yang D, Liu Z J. Review of pedestrian detection with occlusion. *Computer Engineering and Applications*, 2020, **56**(16): 13–20 (in Chinese)
2. Zhou H Z, Yu G. Research on pedestrian detection technology based on the SVM classifier trained by HOG and LTP features. *Future Generation Computer Systems*, 2021, **125**: 604–615
3. Zhang R, Zheng K, Shi P, Mei Y, Li H, Qiu T. Traffic sign detection based on the improved YOLOv5. *Applied Sciences*, 2023, **13**(17): 9748
4. Washizawa T. Application of Hopfield network to saccades. *IEEE Transactions on Neural Networks*, 1993, **4**(6): 995–997
5. Wang J F, Chen Y, Ji X Y, Dong Z K, Gao M Y, Lai C S. Vehicle-mounted adaptive traffic sign detector for small-sized signs in multiple working conditions. *IEEE Transactions on Intelligent Transportation Systems*, 2024, **25**(1): 710–724
6. Wei W, Cheng Y, He J F, Zhu X Y. A review of small object detection based on deep learning. *Neural Computing & Applications*, 2024, **36**(12): 6283–6303
7. Redmon J, Divvala S, Girshick R, Farhadi A. You Only Look Once: Unified, Real-time Object Detection. In: 2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Las Vegas, NV. *IEEE*, 2016, 779–788
8. Redmon J, Farhadi A. YOLO9000: Better, Faster, Stronger. In: 2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Honolulu, HI. *IEEE*, 2017, 6517–6525
9. Liu W, Anguelov D, Erhan D, Szegedy C, Reed S, Fu C Y, Berg A C. SSD: Single Shot MultiBox Detector. In: Leibe B, Matas J, Sebe N, Welling M, eds. Computer Vision–ECCV 2016. *Springer*, 2016, 21–37
10. Zhang H B, Qin L F, Li J, Guo Y C, Zhou Y, Zhang J W, Xu Z. Real-time detection method for small traffic signs based on Yolov3. *IEEE Access: Practical Innovations, Open Solutions*, 2020, **8**: 64145–64156
11. Liao H, Zhu W. YOLO-DRS: a bioinspired object detection algorithm for remote sensing images incorporating a Multi-Scale efficient lightweight attention mechanism. *Biomimetics*, 2023, **8**(6): 458
12. Ye H, Wang Y. Residual transformer YOLO for detecting Multi-Scale crowded pedestrian. *Applied Sciences*, 2023, **13**(21): 12032
13. Tian Y N, Wang S H, Li E, Yang G D, Liang Z Z, Tan M. Tan M. MD-YOLO: multi-scale dense YOLO for small target pest detection. *Computers and Electronics in Agriculture*, 2023, **213**: 108233
14. Xie J, Pang Y, Cholakkal H, Anwer R, Khan F, Shao L. PSC-Net: learning part spatial co-occurrence for occluded pedestrian detection. *Science China. Information Sciences*, 2021, **64**(2): 120103
15. Chen X W, Jia Y P, Tong X Q, Li Z R. Research on pedestrian detection and DeepSort tracking in front of intelligent vehicle based on deep learning. *Sustainability*, 2022, **14**(15): 9281
16. Li N F, Bai X L, Shen X F, Xin P Z, Tian J, Chai T F, Wang Z Y. Dense pedestrian detection based on GR-YOLO. *Sensors*, 2024, **24**(14): 4747
17. Wang Z Z, Xie K, Zhang X Y, Chen H Q, Wen C, He J B. Small-object detection based on YOLO and dense block via Image Super-Resolution. *IEEE Access: Practical Innovations, Open Solutions*, 2021, **9**: 56416–56429
18. Gong L X, Wang Y Y, Huang X, Liang J L, Fan Y M. An improved YOLO algorithm with multisensing for pedestrian detection. *Signal, Image and Video Processing*, 2024, **18**(8-9): 5893–5906
19. Ultralytics. YOLOv8. GitHub Repository, 2023. Available at GitHub website on April 8, 2025
20. Li X, Wang W, Hu X, Yang J. Selective Kernel Networks. In: 2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), Long Beach, CA. *IEEE*, 2019, 510–519
21. Sunkara R, Luo T. No more strided convolutions or pooling: a new CNN building block for low-resolution images and small objects. In: Amini M R, Canu S, Fischer A, Guns T, Kralj Novak P, Tsoumakas G, eds. Machine Learning and Knowledge Discovery in Databases. ECML PKDD 2022. *Springer*, 2023, 443–459
22. Jiao L, Abdullah M I. YOLO series algorithms in object detection of unmanned aerial vehicles: a survey. *Service Oriented Computing and Applications*, 2024, **18**(3): 269–298
23. Zhang Y, Shen Y, Jun Z. An improved tiny-yolov3 pedestrian detection algorithm. *Optik*, 2019, **183**: 17–23
24. Kamil R, Nowicki M R, Skrzypczynski P. Adopting the YOLOv4 architecture for low-latency multispectral pedestrian detection in autonomous driving. *Sensors*, 2022, **22**(3): 1082
25. Lv H H, Yan H B, Liu K Y, Zhou Z W, Jing J J. YOLOv5-AC: attention mechanism-based lightweight YOLOv5 for track pedestrian detection. *Sensors*, 2022, **22**(15): 5903
26. Kaya Ö, Çodur M Y, Mustafaraj E. Automatic detection of pedestrian crosswalk with faster R-CNN and YOLOv7. *Buildings*, 2023, **13**(4): 1070
27. Song Q, Wang H, Yang L, Xin X S, Liu C, Hu M J. Double parallel branches FCOS for human detection in a crowd. *Multimedia Tools and Applications*, 2022, **81**(11): 15707–15723

28. Zhang H X, Yang X F, Hu Z Y, Hao R X, Gao Z H, Wang J H. High-density pedestrian detection algorithm based on deep information fusion. *Applied Intelligence*, 2022, **52**(13): 15483–15495
29. Yin R H, Zhang R F, Zhao W, Jiang F. Jiang F. DA-Net: pedestrian detection using dense connected block and attention modules. *IEEE Access: Practical Innovations, Open Solutions*, 2020, **8**: 153929–153940
30. Shi J L, Zhang R, Guo L J, Gao L L, Ma H F, Wang J H. Discriminative feature network based on a hierarchical attention mechanism for semantic hippocampus segmentation. *IEEE Journal of Biomedical and Health Informatics*, 2021, **25**(2): 504–513
31. Shao S, Zhao Z J, Li B X, Xiao T T, Yu Gang, Zhang X Y, Sun J. CrowdHuman: a benchmark for detecting human in a crowd. *arXiv*, 2018 [Published Online] doi:10.48550/arXiv.1805.00123