EasyChair Preprint

№ 8950

SFYOLO: a Lightweight and Effective Network Based on Space-Friendly Aggregation Perception for Pear Detection

Yipu Li, Yuan Rao, Xiu Jin, Zhaohui Jiang, Lu Liu and Yuwei Wang

October 3, 2022

SFYOLO: a lightweight and effective network based on space-friendly aggregation perception for pear detection

Yipu Li, Yuan Rao *, Xiu jin, Zhaohui Jiang, Lu Liu, and Yuwei Wang

[1] College of Information and Computer Science, Anhui Agricultural University, Hefei 230036, China
[2] Key Laboratory of Agricultural Sensors, Ministry of Agriculture and Rural Affairs, Hefei 230036, China
[3] Anhui Provincial Key Laboratory of Smart Agricultural Technology and Equipment, Hefei 230036, China

Abstract. It is always challenging for efficiently conducting accurate detection of small and occluded pears in modern orchards. In the past few years, the aforementioned detection tasks remained unsolved though lots of researchers attempted to optimize the adaption of background noise and viewpoints, particularly compliant models suitable for simultaneously detecting small and occluded pears with low computational cost and memory usage. In this paper, we proposed a lightweight and effective object detection network called as SFYOLO based on space-friendly aggregation perception. Specifically,a novel space-friendly attention mechanism was proposed for implementing the aggregate perception of spatial domain and channel domain. Afterwards, an improved space-friendly transformer encoder was put forward for enhancing the ability of information exchange between channels. Finally, the decoupled anchor-free detectors were used as the head to improve the adaptability of the network. The mean Average Precision (mAP) for in-field pears was 93.12% in SFYOLO, which was increased by 2.03% compared with original YOLOv5s. Additional experiments and comparison were carried out considering newly proposed YOLOv6 and YOLOv7 that aimed at optimizing the detection accuracy and speed. Results verified that small and occluded pears could be detected fast and accurately by the competitive SFYOLO network under various viewpoints for further orchard yield estimation and development of pear picking system.

Keywords: YOLOv5s · Object detection · Visual attention mechanism · Transformer encoder · Aggregate perception

1 Introduction

To meet the consumption requirements of the world's growing population, horticulture has been trying to find new ways to increase orchard productivity [1, 2].

* Corrsponding author:Yuan Rao, Anhui Agricultural University, Hefei 230036, China
Email: raoyuan@ahau.edu.cn (Yuan Rao)

The development of artificial intelligence and robot technology provides a feasible scheme for the improvement of production efficiency and efficiency [3, 4]. However, in the practical application scenario of agriculture, there are still great challenges in the application of the above technologies. Although labor-intensive agronomic management based on manual labor is difficult to meet the needs of agricultural development under the background of rising agricultural costs and shortage of skilled labor, however, the technology-intensive agronomic management dominated by computer science and technology is still lack of practical experience. If the transition from labor-intensive orchard to technology-intensive orchard can be realized, the development of automatic agronomic management, such as pear growth monitoring, yield estimation and automatic picking of fruits, will help in reducing economic and environmental costs.

In the latest development of intelligent agriculture and independent production, deep learning technology has been widely used practically in agricultural management to improve and estimate agricultural production. Traditionally, it is common for machine vision methods based on convolutional neural networks (CNN) to implement fruit detection. However, due to various sizes from big to small, severe occlusion caused by dense distribution, different viewpoints captured by different device, and several other factors result in obstacles and restrictions of object detection with satisfactory accuracy. Taking the detection of small pear as the main task, how to accurately detect small fruits and severely occluded fruits in complex orchard environments is the key to realize fruit growth monitoring and intelligent yield estimation. For the purpose of improving detection accuracy, current detection networks generally tend to increase the depth of the neural network and use dense connections [5,6], but this may lead to feature loss for small fruits. Therefore, an appropriate detection method is still required for implementing agronomic management in an effective and efficient way.

In order to overcome the above shortcomings, based on the existing YOLO series networks [7–11], we strive to balance the detection accuracy and computational cost, and design a Space-Friendly YOLO (SFYOLO) network. Firstly, a novel SFA (Space-Friendly Attention) mechanism was designed, which enabled the aggregation of spatial and channel attention at a low computational cost. Subsequently, in order to improve the overall perception ability of the SFYOLO, we introduced transformer encoder as the global feature extraction and modeling module, and the SFA mechanism was embed into it to build SF-TE (Space-Friendly Transformer Encoder). Afterwards, SF-TE was used as the feature extraction unit of the neck part of the network to re-extract multi-level features. Finally, the decoupled anchor-free detectors were used as the head of the network to improve the adaptability and the accuracy of the detection network, especially for the dense regions of small and occluded pears.

2 Related Works

Self-Attention in Computer Vision Visual attention mechanism is derived from studies of the human vision system. Different parts of the human retina

have different levels for information processing. In order to make rational use of limited visual information processing resources, humans usually focus on specific parts of the visual area. Generally speaking, the attention mechanism determines which part of the input needs to be paid more attention to. The ability and efficiency of feature extraction will be improved if the network focuses only on task-related regions, rather than useless regions. Self-Attention mechanisms, such as SE [12], CBAM [13] and CA [14], assign different weights to the region that need attention and eliminate irrelevant information, thus improving the quality of the extracted features.

Vision Transformer Due to the great success of Transformer in the field of NLP (Natural Language Processing), lots of research has been done trying to transfer it to the field of computer vision. Vision Transformer [15] firstly proves that the images can be directly applied in Transformer, it makes one image into sequence of patches and reaches SOTA performance in large datasets than traditional CNN-based networks. DETR [16] is the first network based on Transformer for object detection task. It simplifies the process of target detection by treating the detection problem as ensemble prediction, offering an effective way of combining CNN with Transformer. However, the limitations of two networks above are that they both require large-scaled datasets and take too much time for training. Some networks such as LeViT [17], CvT [18], and Visformer [19] alleviate these problems by means of multi-scale feature fusion. The aforementioned work has demonstrated that the proper combination of CNN and Transformer can help in reducing the inference time and the network size, enabling it to be applied in the real-time object detection task.

Object Detection Models With the development of deep learning, various object detection networks and methods have been proposed. The existing object detection methods can be divided into two categories: 1) One-stage detectors, such as YOLO series, FCOS [20], SSD [21]. 2) Two-stage detectors, such as VFNet(17), Faster RCNN(18). In recent research, several novel one-stage object detection networks, e.g. YOLOv6 [22] and YOLOv7 [23], were proposed to further improve the object detection performance in general scenarios. However, the computational cost increased significantly due to the use of high-capacity feature extraction backbone networks and feature extraction modules, which made it difficult to be directly applied to real-time detection tasks in agricultural scenarios with limited resource of computing hardware. At the same time, the overall performance of YOLOv6 and YOLOv7 in detecting small and occluded pears remained to be verified.

3 Image acquisition and processing

3.1 Image dataset acquisition

The pear dataset was collected in September 2021 from the experimental pear orchard located in Suzhou City, Anhui Province, China. We divided the data

collection work into two parts: collecting ground photography dataset and aerial photography dataset. The CCD camera mounted on the tripod and the CCD camera mounted on the Unmanned Aerial Vehicle (UAV) were used to capture pear images from the ground and the air, respectively. A total of 1840 pear images were collected at a shooting distance of about 2 metres to form a ground dataset. The aerial data set was collected by UAV at the height of 2 metres above the pear tree canopy, and 985 images were obtained to form the aerial dataset. The collected images had a dense distribution of target pears of different sizes and contain a large number of target pears with different distances and shapes, which brought a great challenge to accurate detection.

3.2 Annotation procedure and data augmentation

We annotated the ground 1840 images and the annotation information was saved in COCO format [24]. During data annotation, the percentage of small targets was especially increased, and each image contained up to 100 annotation boxes. The images were resized from 1280 × 720 pixels to 640 × 384 pixels with some pears covered by only a few pixels, which further increased the detection difficulty. The training of deep learning usually required a large amount of data. As we know, the limited data collected in real scenarios was often insufficient for network training. Therefore, we expanded the training samples by data augmentation for improving the generalization ability and robustness of the network. In this paper, we used random left-right flip, random up-down flip, HSV space transformation, random blur, Mosaic enhancement [9], Mixup enhancement [25], and other image preprocessing approaches provided in YOLOv5s for performing online enhancement of the data during training to expand the training set.

4 SFYOLO network design

4.1 Overall structure of SFYOLO

Although YOLOv5 has been widely used in various fields, there are still some problems remained to be solved. First of all, although YOLOv5s had a lighter network architecture and faster detection speed than majority of networks, YOLOv5s had a certain sacrifice in accuracy. Thus, how to make up for the lack of detection accuracy without significantly improving the detection speed is an urgent problem to be solved. Secondly, the probe head of YOLOv5s adopted no anchor frame structure, although it performed well in the general scene, but in the scene where a large number of small and occluded pears need to be detected, YOLOv5s was prone to suffering from missed and false detection. Developing appropriate methods that could improve the detection ability of these hard-to-difficult pears would definitely promote the transformation of the network from theory to practical application. In order to facilitate the deployment on the intelligent picking platform in the agricultural scene and reduce the computing power and storage space requirements of the network as much as possible, we followed the general architecture of YOLOv5 to meet the real-time requirements.

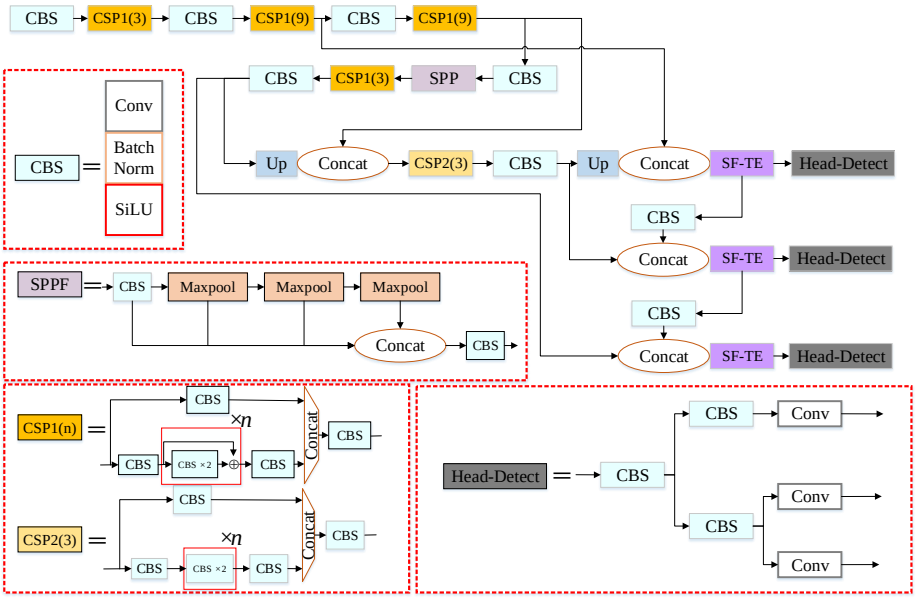

Fig. 1. The network architecture of SFYOLO (where SF-TE corresponds to space-friendly transformer encoder in Figure 3)

The overall structure of SFYOLO is shown in the Figure 1, and its framework can be summarized into three main parts, namely, the backbone, neck, and head. The backbone extraction network is CSPDarknet. After loading the pre-training weights on the COCO dataset, it can extract the necessary feature information from the original three-channel input image for subsequent detection and classification tasks. The neck PANet was used to reprocess the multi-scale feature images extracted by backbone at different stages. The basic feature extraction module was replaced with SF-TE, which could better perceive local and non-local aggregate features. The main part of head consists of three detectors. We chose anchor-free detectors instead of the original anchor-based detectors to improve the generalization ability of the network, with potential improvement in detection of small and occluded pears. At the same time, the detectors were decoupled and the classification process of frame and category was separated, which not only greatly improved the convergence speed, but also increased the classification and localization performance of the head.

4.2 Technical route of pear detection network

The flow of the SFYOLO-based pear detection network proposed in this study could be concluded as follow. Firstly, improvements were made in the neck and head of the original YOLOv5 network. For purpose of improving the neck, we proposed a Space-Friendly Attention (SFA) mechanism based on the channel-

aware and coordinate-aware aggregation, which aimed at enhancing the aggregation perception ability of the network between the spatial domain and the channel domain to make up for the lack of the original network's ability to detect small and occluded pears. Secondly, a construction method of Space-Friendly Transformer Encoder (SF-TE) was developed to establish spatial long-range dependencies. Through the joint perception of Transformer and SFA, the robustness of the network was improved, and the local perception and global perception were effectively combined. Then a novel neck was proposed by means of utilizing the aforementioned SF-TE to replace the original CSP. In the improved head, for each level of feature extracted from neck, we adopted a 1×1 convolution layer to reduce the feature channel to 256 and then added two parallel branches with two 3 × 3 convolution layers each for classification and regression tasks respectively. Then IoU branch was added on the regression branch. Finally, qualitative and quantitative experiments were carried out to verify the effectiveness of the improvements.

4.3 Improvements of pear detection network based on YOLOv5s

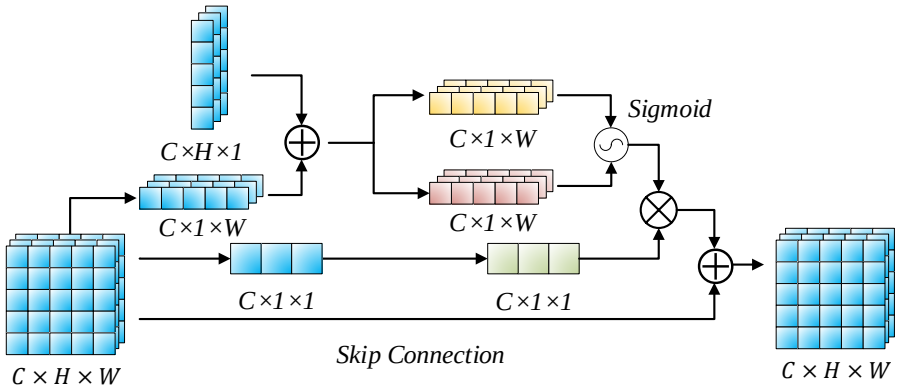

Fig. 2. The structure of Space-Friendly Attention mechanism (SFA)

Space-friendly attention mechanism (SFA) In machine vision tasks, spatial information was equally important as coordinate information. Achieving an aggregate perception of spatial and channel features was beneficial to the overall perception ability of the network. Self-attention mechanism within both spatial and channel domains of the feature maps contributed to capturing dimensionally richer information in local and global way. As illustrated in Figure 2, on the basis that the coordinate attention mechanism summarized the feature map into a pair of feature vectors along the horizontal coordinate direction and vertical coordinate direction, we added spatial feature vectors along the channel

direction to implement the exchange of information between channels. This set of spatial feature vectors could selectively enlarge the valuable feature channels and restrain the useless feature channels, thus improving the performance of the network. Each feature map was aggregated into 1x1 pixels, which were reweighted with horizontal and vertical feature vectors to achieve an effective aggregate perception of spatial and channel features. In order to reduce the loss of original information and alleviate the problem of gradient disappearance, we artificially made some layers of the network skip the connection of the next layer of neurons, making the non-adjacent layers connected, weakening the strong connection between each layer, and alleviating the degradation of network caused by the excessive depth. By making full use of spatial and channel information, it was expected to improve the ability to capture local and global information in the network.

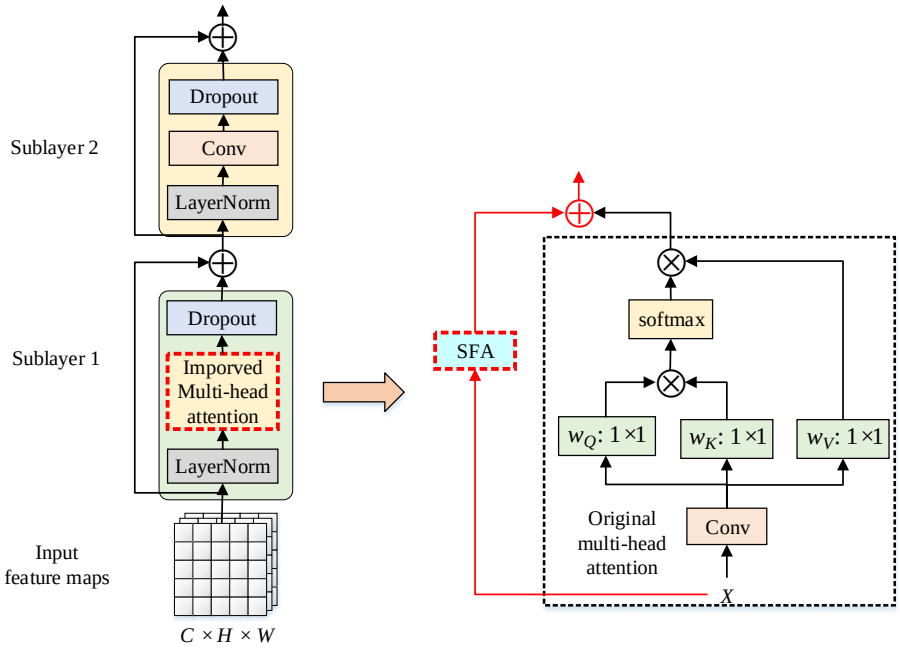

Fig. 3. The structure of space-friendly transformer encoder (SF-TE), where the red marks the improvements, SFA corresponds to Figure 2

Space-friendly transformer encoder (SF-TE) Inspired by the idea of Visformer to introduce convolution into transformer encoders, we designed space-friendly transformer encoders. The structure of improved space-friendly transformer encoder is illustrated in Figure 4, which could be divided into two sublayers, with the first layer being a multi-head attention layer and the second

sublayer being a fully connected layer. Specifically, the input feature maps first passed through a multi-head attention sublayer, in which multiple attention heads performed attention computations synchronously to establish long-range dependencies in the spatial domain and preserve non-local features. Then they passed through the fully connected layer, where the feature maps were re-stored to their original size by multiple convolution operations. Finally, the result of residual connection was output after the Dropout operation, contributing to better convergence and overfitting prevention. To further achieve the proper perception of high-density targets, occluded targets, and small targets, the space-friendly attention mechanism was embedded into the multi-head attention mechanism. On the basis of the original multi-head attention mechanism, we added a path for feature extraction through SFA and connected it with the feature map extracted by the original multi-head attention mechanism in a cross-layer way. This expanded the channel dimension of the network, thus making full use of the channel information and achieving the aggregate perception in both channel and spatial domains.

Improved neck with transformer encoders To enhance the feature extraction ability of the network for target objects of different scales, YOLOv5s generally consisted of five stages referred to as [P1, P2, P3, P4, P5]. The output feature maps of these stages had distinct scales and were used to extract features from objects of different sizes, finally, the feature maps of P3, P4, and P5 were used for the detection task. These feature maps at different scales provided extensive multi-scale feature information for the target detection task. To enhance the feature extraction ability of the network for small targets, we attempted to embed SF-TE (space-friendly transformer encoder) into the neck. Since SF-TE greatly increased the parameters and computation cost of the network, applying it on low-resolution feature maps instead of high-resolution feature maps would lead to the increment of the expensive computation and memory cost, which was an obstacle for implementing real-time target detection tasks with limited resources. Therefore, we only embedded SF-TE into P3, P4, P5 in the neck of YOLOv5s. The improved Neck effectively enriched the information content of the feature maps, significantly enhancing the over perception results of in-field pears. The eventual output feature map contained denser non-local and local information, which allowed for the detection of small and occluded pears in natural environments.

Decoupled anchor-free detectors The detector of object detection network can be divided into anchor-based and anchor-free, and the former is usually used in traditional target detection network. However, the detection performance of anchor-based depends on the design of anchor frame to some extent, and it is very sensitive to the size, aspect ratio and quantity of anchor box. However, a large number of hyperparameters are used in the initialization design of the anchors, which makes it difficult to adjust these hyperparameters and cost a lot of time and computation to optimize. Considering that there were a large

number of small pears in this application scene, which was different from the size distribution of the target in the general scene, it could not well match the preset anchor size. Therefore, we chosen to use the non-anchor detector instead of the anchor-based detector. Some scholars argued that there is a spatial misalignment problem in localization and classification tasks of object detection [26]. This meant that the two tasks have different focuses and degrees of interest, with classification focusing more on the features extracted that are most similar to existing categories, while localization focuses more on the location coordinates with the anchor points and thus on the bounding box parameter correction. Therefore, we chosen to use a decoupling head structure that can independently complete classification and positioning tasks by using different decoupled head branches, which is beneficial to the final detection effect and accuracy.

5 Experiments

5.1 Evaluation metrics

Since the images collected in the natural environment contained many pears of different sizes, we used AP and F1-score to evaluate the performance of the network. AP refers to the average value of all ten intersections (IoU) thresholds with a uniform step size of 0.05 in the range of [0.50, 0.95]. F1-score is the harmonic average of precision and recall, with a value ranging from 0 to 1. Among them, a large F1-score value indicates good detection accuracy. The formula for the F1-score is as follows:

$$P = TP/(TP + FP) \tag{1}$$

$$R = TP/(TP + FN) \tag{2}$$

$$F = \frac{2 \times P \times R}{(P + R)} \tag{3}$$

Where TP (True Positive) denotes the number of predicted positive samples,FP (False Positive) denotes the number of predicted positive but negative samples, and FN (False Negative) denotes the number of predicted negative but positive samples.

5.2 Implementation details

The network was trained on the ground dataset with eight images as a batch and the loss was updated once per iteration for a total of 200 epochs on a single NVIDIA GTX 2080Ti. The detection of pears was carried out on a single NVIDIA GTX 1650, which simulated the limited computing environment in the natural environment. Using SGD as the optimizer, the initial learning rate was set to 0.01, the weight decay rate was set to 0.00048, and the momentum factor was set to 0.937. It gradually decayed to 1E-4 as the iterations proceeded. This experimental network was trained using transfer learning, and further training was performed on the pre-trained weights of the MS COCO dataset.

5.3 Quantitative experimental results compared with YOLO series

In addition to the proposed SFYOLO, the typical or recent members of YOLO series, such as YOLOv3-SPP, YOLOv4s, YOLOv5s, YOLOv6s and YOLOv7s were employed for making performance comparison based on the ground dataset. The detection accuracy, detection time, memory usage was taken into consideration. From Table 1, it could be seen that the proposed SFYOLO achieve on AP (average precision) and F1-score of 93.12% and 86.73%, which was much better than the results acquired by YOLOv3-SPP, YOLOv4s and YOLOv5s. Contrastively, YOLOv6s and YOLOv7s could offer closer detection results to the proposed SFYOLO. However, compared to the aforementioned two networks, the proposed SFYOLO reduced the detection time by 27% and 40%, meanwhile, the FLOPs decreased by 58% and 82.6%. In addition, in comparison to YOLOv5s, the AP of SFYOLO increased by 2.03% with slightly higher detection time and a larger memory usage. Therefore, it could be drawn that the proposed SFYOLO had the ability of offering better results in pear detection.

Table 1. Comparison among SFYOLO and other widely used and novel networks of YOLO series in terms of detection accuracy and efficiency on the ground dataset.

Networks	AP(%)	F1-score(%)	Detection time (ms)	Memory usage(MB)	FLOPs(G)
YOLOv3-SPP	89.51	83.72	16.1	120.32	157.1
YOLOv4s	91.98	85.62	20.2	246.34	137.2
YOLOv5s	91.09	84.35	**11.2**	**13.70**	**16.4**
YOLOv6s	92.93	85.92	18.2	285.78	44.2
YOLOv7s	93.02	86.23	22.3	64.23	104.7
SFYOLO	**93.12**	**86.73**	13.2	54.52	18.2

5.4 Qualitative experiments results on ground dataset

When detecting pears in practical application scenes, the collected images often contain a large number of pears with different scales, serious occlusion and cluttered density. It was difficult to detect these pears, which was often the main factor leading to low detection accuracy. If the improvements of the proposed network could effectively enhance the detection accuracy of these pears, it could provide a feasible scheme for the application of deep neural network in the field of agriculture. In order to verify that the proposed network could meet the requirements of the natural environment, we focus on its detection performance in the areas with uneven pear size and serious background noise. The quantitative test results of SFYOLO with YOLOv5, YOLOv6, and YOLOv7 is shown in Figure 4. In the comparison of columns a and b, the detection performance of the networks for dense areas containing a large number of small pears was mainly concerned. The results showed that the cases of missed detection and false detection of SFYOLO were less than those of YOLOv6 and YOLOv7, and

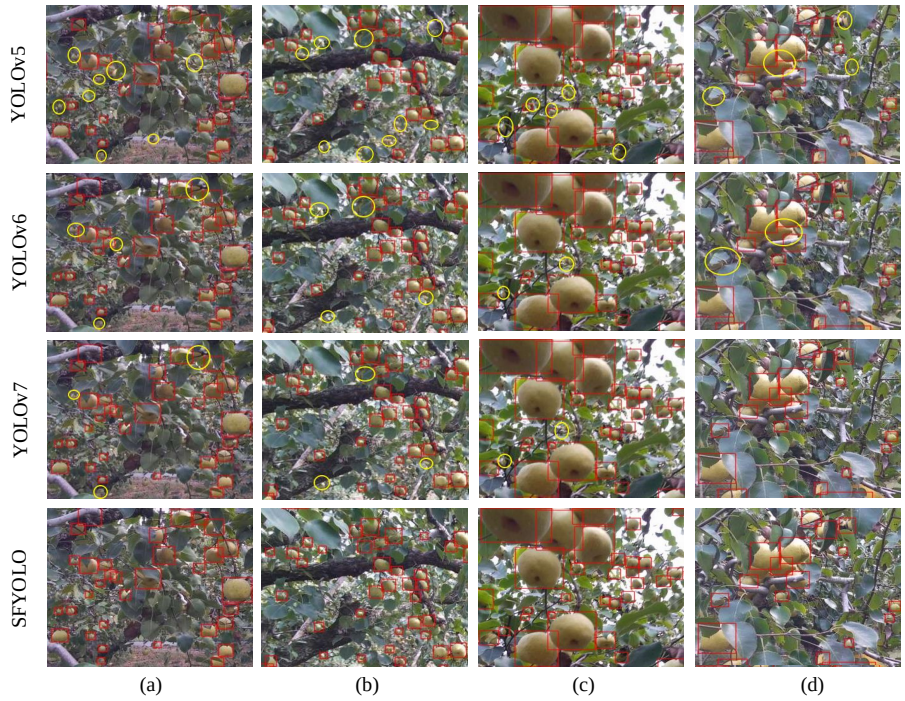

Fig. 4. Comparison of pear detection on the ground dataset, with YOLOv5s, YOLOv6, YOLOv7 and SFYOLO. The cases of missed detection are highlighted in yellow

obviously less than YOLOv5. This showed that SFYOLO could effectively detect small pears, and was better than YOLOv5, YOLOv6 and YOLOv7 qualitatively. In the comparison of columns of c and d, the detection performance of the networks for areas containing pears with different scales and blocking each other was mainly concerned. The results showed that SFYOLO still had a beneficial effect in detecting pears occluded by leaves or by each other, and the effect was slightly better than that of YOLOv6 and YOLOv7, and obviously better than that of YOLOv5. In summary, SFYOLO could effectively detect pears in the natural environment, especially for small and occluded ones. This helped to provide a lightweight vision system for portable devices and provided auxiliary information for subsequent decision-making such as picking path planning.

5.5 Qualitative experimental results on aerial dataset

In agricultural monitoring, three-dimensional monitoring was of great significance to promote orchard automation. The aerial images obtained by aircraft could be used for pear growth monitoring, intelligent yield estimation and orchard spraying. In order to verify the performance of the proposed network in

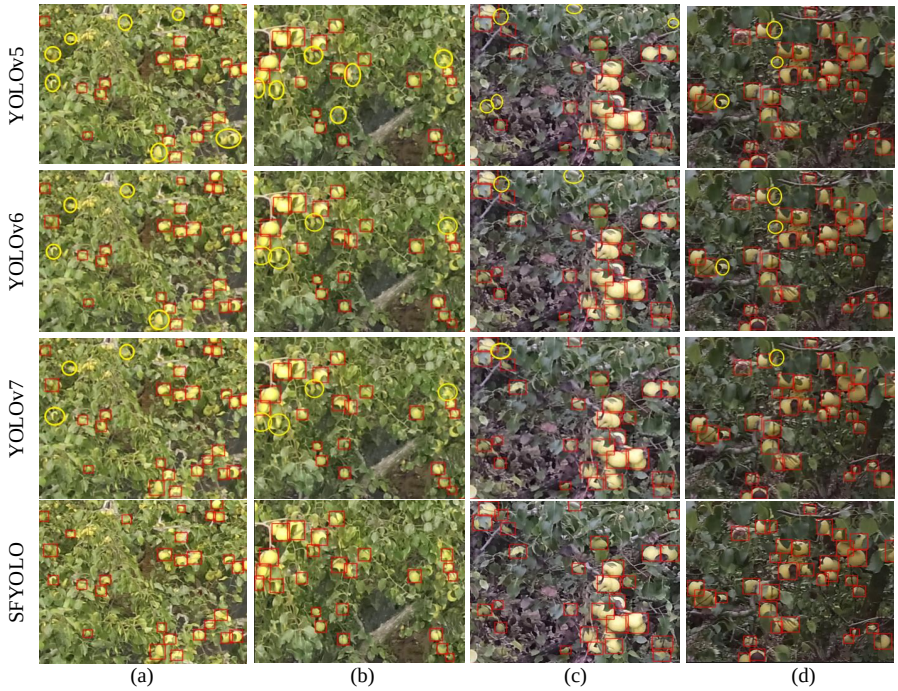

Fig. 5. Comparison of pear detection in the aerial dataset captured by drones from the bird's-eye viewpoint and side viewpoint. The cases of missed detection are highlighted in yellow

cross-domain detection tasks, aerial dataset including bird's-eye viewpoint and side viewpoint were used to further explore the robustness of the network to the change of viewpoints. Figure 5 shows the qualitative test results of YOLOv5, YOLOv6, YOLOv7, and SFYOLO on the aerial dataset using the weights of network trained on the ground dataset. Columns a and b were images taken by drones from the bird's-eye viewpoint, and columns c and d were images taken by drones from side viewpoint. These images contain a large number of small and occluded pears, which made it much more difficult to detect. Under the influence of light and background noise, the pears in the bird's-eye viewpoint were closer to the color of leaves and soil, so they were more difficult to detect. As could be seen from the figure, the detection performance of SFYOLO was slightly better than that of YOLOv6 and YOLOv7 with less cases of missed detection. Among them, the improvement in the bird's-eye viewpoint was more obvious, indicating that the SFYOLO can effectively reduce the interference of background noise, so as to identify the target more accurately. Compared with the detection ability of YOLOv5, the detection ability of SFYOLO was significantly improved, and the cases of missed detection were greatly reduced in both viewpoints. This benefited from the construction of SFA mechanism, the embedding of space-friendly

transformer encoder and the employment of decoupling head, which not only improved the ability of local and global feature extraction, but also enhanced the transfer and generalization ability of the network.

To sum up the above results, due to the effective design of network and the novel improvement of structure, SFYOLO was endowed with sufficient accuracy and robustness improvement, and could carry out cross-domain detection efficiently and accurately. Additionally, SFYOLO could be directly transferred for pear detection in aerial images, which not only verified the generalization ability, but also reduced the workload, improved the production efficiency, bringing potential improvement of production benefit.

5.6 Ablation studies

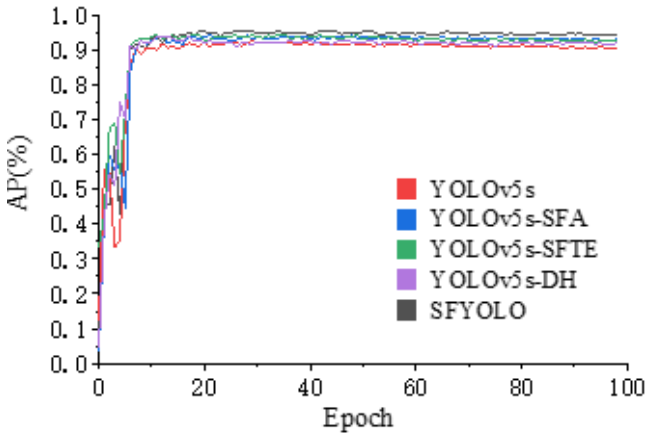

Fig. 6. Line graph of AP metrics for ablation experiments on the ground dataset

To further explore the impact of the improvements of the original YOLOv5s, three sets of ablation experiments were designed in this paper, and the improvement of each improvement on the network performance was discussed based on the results of AP metrics. In order to explore the influence of SFA mechanism on the network, it was added to P3, P4 and P5 in the neck of the original YOLOv5s, which was named as YOLOv5s-SFA. In order to explore the influence of the position of SF-TE in the network, it was applied at the end of backbone, which was named as YOLOv5s-SFTE. In order to explore the impact of decoupled head on performance, decoupled head were applied as head, which was named as YOLOv5s-DH. The results of the ablation experiment are shown in Figure 6 and Table 2. Each module could contribute to the performance improvement in pear detection on the ground dataset.

Table 2. Results in the ablation experiments.

Models	AP(%)	Detection Time(ms)	Memory usage(MB)
YOLOv5s	91.09	11.2	13.7
YOLOv5s-SFA	91.96(+0.87)	11.8(+0.6)	23.2(+9.5)
YOLOv5s-SFTE	92.44(+1.49)	15.7(+4.5)	113.9(+100.2)
YOLOv5-DH	92.12(+1.03)	12.1(+0.9)	17.9(+4.2)
SFYOLO	93.12(+2.03)	13.2(+2.0)	54.5(+40.8)

Effect of space-friendly attention mechanism (YOLOv5s-SFA) The AP of YOLOv5s-SFA on the validation set increased by 0.87%. The memory usage increased by 1.69 times than the original size. In the early stage of training, YOLOv5s-SFA converged rapidly, started to decrease slightly after twenty generations, and finally reached stable status. YOLOv5s made a certain improvement in accuracy with faint computational cost and memory usage improvement, indicating that SFA mechanism played an indispensable role in the feature map and helped in the network convergence rapidly.

Effect of space-friendly transformer encoder (YOLOv5-SFTE) Compared with the YOLOv5 network, YOLOv5-SFTE increased the AP on the validation set by 1.49%. In the early stage of training, the accuracy of the network fluctuates greatly and then keeps stable gradually. Compared with SFYOLO, the reason for its lower overall accuracy might be that transformer encoder was more difficult to fit on small-size datasets. Although the memory usage and detection time increased a lot, it was still a useful part of the network compared with other one-stage object detection networks.

Effect of decoupled head (YOLOv5s-DH) Compared with the YOLOv5 network, YOLOv5s-DHhad 1.03% higher AP on the verification set. In the early stage of training, the accuracy of the network fluctuated greatly and then gradually kept relatively stable. Compared with YOLOv5s, the reason for its higher overall accuracy might be that the decoupled anchor-free detectors had almost no manual preset hyperparameters, which could be well migrated to the application scene in this paper to satisfy the object detection requirements. While SFYOLO-DH has achieved effective accuracy improvement, there was almost no significant improvement in detection time and memory usage, which was very suitable for completing high-precision detection tasks in an environment with limited computing resources.

6 Conclusion

Improving the detection ability of the network for target fruits was the basis of implementing automated agronomic management. To address the problem of

target fruit detection in real complex scenes, this paper proposed SFYOLO with the inherent characteristics of the aggregation attention of space domain and channel domain. The construction of the Space-Friendly Attention mechanism and the modification of the original transformer encoder enabled the network to aggregate spatial feature and channel feature as well as capture both local and global information in an effective and efficient way. The decoupled anchor-free head enabled the network to complete the tasks of classification, location and regression independently and better, which further enhanced the detection performance. The experimental results showed that our network enhances the feature extraction ability and improved the detection accuracy of small pears and occluded pears. The proposed SFYOLO achieved an average accuracy of 93.12% on the ground dataset and outperformed the typical or recent members of the YOLO series such as YOLOv6 and YOLOv7 on both the ground dataset and the aerial dataset in terms of speed and accuracy. In the future, we will further investigate the ways of implementing pear detection at different growth stages, as well as the reduction of the training and detection cost to better support real-time detection of other fruits.

References

1. J. Chen, J. Wu, Z. Wang, H. Qiang, G. Cai, C. Tan, and C. Zhao, "Detecting ripe fruits under natural occlusion and illumination conditions," *Computers and Electronics in Agriculture*, vol. 190, p. 106450, 2021.
2. I. Perez-Borrero, D. Marin-Santos, M. E. Gegundez-Arias, and E. Cortes-Ancos, "A fast and accurate deep learning method for strawberry instance segmentation," *Computers and Electronics in Agriculture*, vol. 178, p. 105736, 2020.
3. B. B. Sharma and N. Kumar, "Iot-based intelligent irrigation system for paddy crop using an internet-controlled water pump," *International Journal of Agricultural and Environmental Information Systems (IJAEIS)*, vol. 12, no. 1, pp. 21–36, 2021.
4. Z. Sun, W. Feng, J. Jin, Q. Lei, G. Gui, and W. Wang, "Intelligent fertilization system based on image recognition," in *2021 IEEE 6th International Conference on Computer and Communication Systems (ICCCS)*, pp. 393–399, IEEE, 2021.
5. R. Gai, N. Chen, and H. Yuan, "A detection algorithm for cherry fruits based on the improved yolo-v4 model," *Neural Computing and Applications*, pp. 1–12, 2021.
6. X. Hu, Y. Liu, Z. Zhao, J. Liu, X. Yang, C. Sun, S. Chen, B. Li, and C. Zhou, "Real-time detection of uneaten feed pellets in underwater images for aquaculture using an improved yolo-v4 network," *Computers and Electronics in Agriculture*, vol. 185, p. 106135, 2021.
7. J. Redmon and A. Farhadi, "Yolo9000: better, faster, stronger," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 7263–7271, 2017.
8. J. Redmon and A. Farhadi, "Yolov3: An incremental improvement," *arXiv preprint arXiv:1804.02767*, 2018.
9. A. Bochkovskiy, C.-Y. Wang, and H.-Y. M. Liao, "Yolov4: Optimal speed and accuracy of object detection," *arXiv preprint arXiv:2004.10934*, 2020.
10. Z. Ge, S. Liu, F. Wang, Z. Li, and J. Sun, "Yolox: Exceeding yolo series in 2021," *arXiv preprint arXiv:2107.08430*, 2021.
11. A. S. Glenn Jocher, "Yolov5." https://github.com/ultralytics/yolov5.

12. J. Hu, L. Shen, and G. Sun, "Squeeze-and-excitation networks," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 7132–7141, 2018.
13. S. Woo, J. Park, J.-Y. Lee, and I. S. Kweon, "Cbam: Convolutional block attention module," in *Proceedings of the European conference on computer vision (ECCV)*, pp. 3–19, 2018.
14. Q. Hou, D. Zhou, and J. Feng, "Coordinate attention for efficient mobile network design," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 13713–13722, 2021.
15. A. Dosovitskiy, L. Beyer, A. Kolesnikov, D. Weissenborn, X. Zhai, T. Unterthiner, M. Dehghani, M. Minderer, G. Heigold, S. Gelly, *et al.*, "An image is worth 16x16 words: Transformers for image recognition at scale," *arXiv preprint arXiv:2010.11929*, 2020.
16. N. Carion, F. Massa, G. Synnaeve, N. Usunier, A. Kirillov, and S. Zagoruyko, "End-to-end object detection with transformers," in *European conference on computer vision*, pp. 213–229, Springer, 2020.
17. B. Graham, A. El-Nouby, H. Touvron, P. Stock, A. Joulin, H. Jégou, and M. Douze, "Levit: a vision transformer in convnet's clothing for faster inference," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 12259–12269, 2021.
18. H. Wu, B. Xiao, N. Codella, M. Liu, X. Dai, L. Yuan, and L. Zhang, "Cvt: Introducing convolutions to vision transformers," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 22–31, 2021.
19. Z. Chen, L. Xie, J. Niu, X. Liu, L. Wei, and Q. Tian, "Visformer: The vision-friendly transformer," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 589–598, 2021.
20. Z. Tian, C. Shen, H. Chen, and T. He, "Fcos: Fully convolutional one-stage object detection," in *Proceedings of the IEEE/CVF international conference on computer vision*, pp. 9627–9636, 2019.
21. W. Liu, D. Anguelov, D. Erhan, C. Szegedy, S. Reed, C.-Y. Fu, and A. C. Berg, "Ssd: Single shot multibox detector," in *European conference on computer vision*, pp. 21–37, Springer, 2016.
22. C. Li, L. Li, H. Jiang, K. Weng, Y. Geng, L. Li, Z. Ke, Q. Li, M. Cheng, W. Nie, *et al.*, "Yolov6: A single-stage object detection framework for industrial applications," *arXiv preprint arXiv:2209.02976*, 2022.
23. C.-Y. Wang, A. Bochkovskiy, and H.-Y. M. Liao, "Yolov7: Trainable bag-of-freebies sets new state-of-the-art for real-time object detectors," *arXiv preprint arXiv:2207.02696*, 2022.
24. T.-Y. Lin, M. Maire, S. Belongie, J. Hays, P. Perona, D. Ramanan, P. Dollár, and C. L. Zitnick, "Microsoft coco: Common objects in context," in *European conference on computer vision*, pp. 740–755, Springer, 2014.
25. H. Zhang, M. Cisse, Y. N. Dauphin, and D. Lopez-Paz, "mixup: Beyond empirical risk minimization," *arXiv preprint arXiv:1710.09412*, 2017.
26. G. Song, Y. Liu, and X. Wang, "Revisiting the sibling head in object detector," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 11563–11572, 2020.