scientific reports

Check for updates

OPEN

Lightweight and efficient deep learning models for fruit detection in orchards

Xiaoyao Yang[1], Wenyang Zhao[1], Yong Wang[1], Wei Qi Yan[2] & Yanqiang Li[1]✉

The accurate recognition of apples in complex orchard environments is a fundamental aspect of the operation of automated picking equipment. This paper aims to investigate the influence of dense targets, occlusion, and the natural environment in practical application scenarios. To this end, it constructs a fruit dataset containing different scenarios and proposes a real-time lightweight detection network, ELD(Efficient Lightweight object Detector). The EGSS(Efficient Ghost-shuffle Slim module) module and MCAttention(Mix channel Attention) are proposed as innovative solutions to the problems of feature extraction and classification. The attention mechanism is employed to construct a novel feature extraction network, which effectively utilizes the low-latitude feature information, significantly enhances the fine-grained feature information and gradient flow of the model, and improves the model's anti-interference ability. Eliminate redundant channels with SlimPAN to further compress the network and optimise functionality. The network as a whole employs the Shape-IOU loss function, which considers the influence of the bounding box itself, thereby enhancing the robustness of the model. Finally, the target detection accuracy is enhanced through the transfer of knowledge from the teacher's network through knowledge distillation, while ensuring that the overall network is sufficiently lightweight. The experimental results demonstrate that the ELD network, designed for fruit detection, achieves an accuracy of 87.4%. It has a relatively low number of parameters (4.3×10^5), a GLOPs of only 1.7, and a high FPS of 156. This network can achieve high accuracy while consuming fewer computational resources and performing better than other networks.

Keywords Recognition of apple, Lightweight network, Attention mechanism, Object detection, Deep learning

The advent of scientific and technological advancement has led to the proliferation of automated fruit-picking robots in the domain of intelligent agriculture, which helps to reduce production costs and minimise errors caused by manual labour and inexperienced workers. In harvesting systems, two principal components are typically required: a vision detection module and a mechanical harvesting module[1]. The vision module is used to accurately identify and locate the fruits and is essential to guide the robotic arm accurately and efficiently to increase productivity. Consequently, the ability to efficiently identify and accurately locate fruits is a pivotal aspect that determines the efficacy of automatic picking robot applications[2].

The complexity of agricultural environments is such that the YOLO algorithm was selected for use in agricultural applications, primarily on account of its superior efficiency and adaptability to complex environments. In recent years, scholars have made significant advancements in the field of fruit detection in complex environments. In their study,Liu X et al.[3] employed external characteristics of fruit colour and shape for the purpose of object detection. Tian Y et al.[4] have developed an enhanced version of the You Only Look Once (YOLO) algorithm for the purpose of detecting fruits with varying growth cycles. B Xiao et al.[5] sought to enhance the precision of fruit detection by augmenting the model complexity through the utilisation of Transformer. The advancement of deep learning technology has highlighted the necessity to optimise neural networks through a variety of techniques, which has become a pivotal focus in the field of target detection and target tracking[6–10] research. Nevertheless, despite the sophistication of algorithms such as YOLOv5, the application of such technologies in agriculture encounters considerable obstacles. The intricate network structure results in suboptimal real-time efficiency, while the elevated computational power demands render deployment a challenging undertaking.

Scholars have directed increased attention towards enhancing target detection models through the use of an attention mechanism, with the objective of focusing on target features. Xu et al.[11] enhanced the instance

[1]Institute of Automation, Qilu University of Technology (Shandong Academy of Sciences), Jinan 250014, China. [2]Auckland University of Technology, Auckland, New Zealand. ✉email: Liyq@sdas.org

segmentation performance by incorporating channel attention into the segmentation network. Zhang et al.[12] employed channel attention to optimise the crack. M. Jaderberg et al.[13] put forth a Transformer-based channel attention mechanism for STNs, whereby the input data is spatially transformed through learning, thereby enhancing the network's resilience to geometric transformations such as image distortion and rotation. Furthermore, the optimisation of the bounding box loss function is of paramount importance. The most commonly employed methods include IoU[14], GioU[15], Diou[16], CIoU[16], Eiou[17] and SioU[18]. This is achieved primarily through the introduction of geometric constraints to the bounding box, thereby ensuring the accuracy of target detection.Nevertheless, despite the enhanced outcomes achieved by these techniques, there are still substantial challenges associated with their deployment in agricultural applications. The intricacy of the network structure results in suboptimal real-time performance and a high demand for computational resources.

Recently, numerous scholars have been engaged in research aimed at integrating the YOLO algorithm with lightweight detection modules, including the ensemble GhostNet[19], MobileDets[20], PP-PicoDet[21], and EdgeNeXt[22], to reduce the number of parameters. Notwithstanding the enhancement in detection velocity, these lightweight algorithms are deficient in their capacity for feature extraction. Plus this, Knowledge distillation is a frequently employed lightweighting method in target detection networks[23]. In a recent study, Lan et al.[24] demonstrated that the application of this method to the detection of autopilot yielded superior results.

In an effort to salute these challenges, this paper puts forward a proposal for a lightweight object detection network for the purpose of detecting apples in complex environments.The BASELINE model used in this paper is the YOLOv5 algorithm. The principal contributions of this research are outlined below:

1. This study presents a novel approach to data integration, resulting in the construction of a high-quality dataset. The reliability of this dataset is then verified through a series of experiments. Notably, all experiments conducted in this paper are based on this dataset, thereby addressing the dearth of data in this field.
2. In order to achieve an optimal balance between processing speed and accuracy, In this study, we put forward a novel lightweight feature extraction module, designated Efficient Shuffle Slim (EGSS).
3. To enhance the capability of feature extraction module. This study presents an innovative plug-and-play lightweight hybrid attention mechanism that simultaneously integrates channel, spatial, local and global information.
4. In order to maximise the compression model, this paper proposes SlimPAN to eliminate redundant channels through channel compression.
5. In order to compensate for the lack of the traditional loss function, Shape-IOU is used in this paper by taking into account the effect of the bounding box itself.
6. The knowledge distillation approach guarantees that the network will perform the required object detection tasks with a minimal number of parameters and computations, thereby facilitating the rapid and efficient operation of the model.
7. In order to ensure the feature extraction capability of the model while reducing the complexity.A novel detection network, the Efficient Lightweight object Detector (ELD), is proposed. It integrates the EGSS module and combines with SlimPAN. The network in question requires only a minimal amount of computing resources.This paper is structured as follows: Section II is dedicated to an in-depth examination of the high-quality dataset constructed in the course of this research project. Part III provides a comprehensive analytical account of the innovative lightweight feature extraction module, hybrid attention mechanism, and the novel object detection network theme. The fourth section presents the findings of the analysis of the experimental data collected throughout the research process. Parts V, VI and VII, respectively, present a discussion of the research process, an analysis of the limitations and a summary of the conclusions.

Related work

YOLOv5

The most prevalent target detection network is YOLO[25–28], which is a single-stage network designed for expeditious and efficacious one-shot detection[29]. The YOLOv5 algorithm is a popular choice in the field of object detection due to its efficient performance. In this study, YOLOv5 is further developed in the base framework with the objective of reducing the complexity of the code and improving the performance of the network. This constitutes a benchmark for this paper, with the aim of improving the implementability of the work. As illustrated in Fig. 1, the YOLO network partitions the input image into a grid of S Ã- S cells. The grid cells predict Bgo bounding boxes, each of which contains a target. Each bounding box comprises five parameters: (x,y,h,w,c), which denote the centre coordinates, width, height and confidence level of the bounding box, respectively. Ultimately, a C-dimensional vector is generated for each bounding box, indicating the probability that the bounding box belongs to each category. In essence, the output tensor of YOLOv5 comprises predictive information for each grid cell, with a total output dimension of S Ã- S Ã- (B Ã- 5 + C). This encompasses the parameter and category predictions for the B bounding boxes of each grid cell.

Knowledge distillation algorithm

Both knowledge distillation algorithms and neural network pruning[30] are effective techniques for reducing the computational burden of deep learning models. However, neural network pruning is particularly well-suited to networks with larger models. Q Lan et al.[24] reweighed each instance and each scale for distillation based on the teacher's loss, thereby facilitating the selective transfer of the teacher's knowledge to the students. Sangwoo Park et al.[31] presents a Cosine Similarity-Based Knowledge Distillation (CSKD) for robust, lightweight object detectors.Jike et al.[32] proposed the multi-constraint molecular generation (MCMG) approach that can satisfy multiple constraints by combining conditional transformer and reinforcement learning algorithms through knowledge distillation. Kim et al.[33] enabled the model to exhibit high performance by knowledge distillation

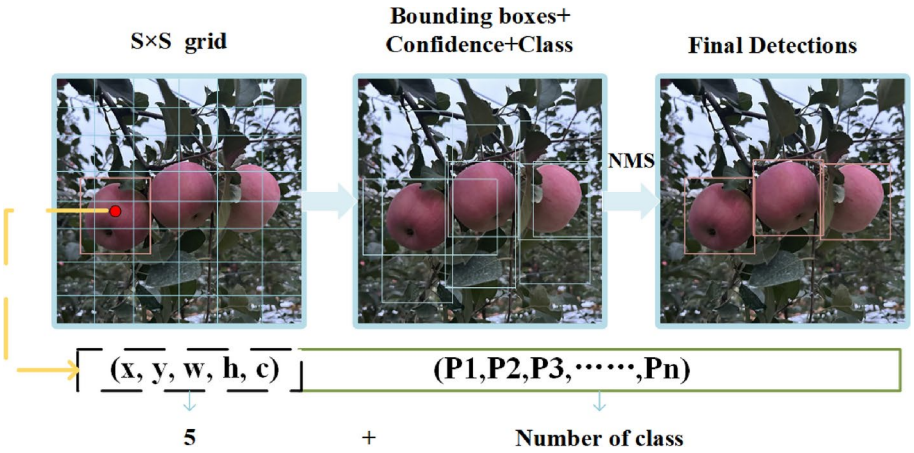

Fig. 1. YOLO model for apple detection from images.

without charging feature spoofing caused by modal differences. These scholars have demonstrated proficiency in the application of knowledge distillation, which serves as a valuable reference point for this study.

Feature extraction module

The feature extraction module constitutes the core of the target detection network. A popular approach to enhancing its feature extraction capability is to add an attention mechanism to the feature extraction module. Zhao et al.[34] integrated the convolutional block attention mechanism with ShuffleNetv2 as a feature extraction module, thereby enhancing the network's capacity to discern object features.Wang et al.[35] employed a self-attention mechanism for the extraction and integration of multi-scale features in both the feature extraction module and the feature fusion module. This approach demonstrated some improvement in accuracy but did not yield effective lightweighting.Guo et al.[36] devised a lightweight feature extraction module to enhance the precision of safflower detection in agricultural settings through a synergistic integration of Ghostconv and CBAM attention mechanisms. Lin et al.[37] devised a novel cross-attention mechanism based on the Transformer structure, which is deployed between image blocks derived from single-channel features to capture global information. They demonstrated that this module has the potential to serve as a general-purpose feature extraction network. However, it necessitates greater computational resources compared to convolution-based networks.J Fu et al.[38] proposed a dual attention network to adaptively integrate local features with global dependencies. Its efficacy was validated on multiple datasets, yielding favourable results. However, the network's intricate structure necessitates a considerable computational burden, which may prove challenging to accommodate within the feature extraction module. The research conducted by the aforementioned scholars serves as a valuable reference for this study. However, their proposed methods still require significant improvement in terms of lightweighting. Consequently, this paper proposes a feature extraction module that is more efficient and lightweight, and incorporates a lightweight hybrid attention mechanism. This results in a reduction in computation and makes it more suitable for deployment on devices with limited computing power.

Dataset construction

The intricate nature of the orchard environment presents a significant challenge for traditional object detection networks, particularly in terms of accurately extracting features from apple fruits within this context. In order to facilitate the network's ability to discern a greater number of fruit features in authentic settings, this study has constructed a comprehensive dataset, designated "Appledatas," which encompasses 3,151 images. The majority of these images were sourced from a contemporary orchard in Zibo, Shandong Province, while a modest proportion were gathered from a traditional orchard in Tai'an, Shandong Province. The rationale behind collecting data from disparate geographical regions is to mitigate the impact of regional variations in fruit morphology on the performance of the network in feature extraction. By encompassing data from both standardised and non-standardised orchards, the network is enabled to learn fruit features from a diverse range of environments, thereby enhancing its generalisation capabilities. The dataset encompasses the majority of object detection scenarios encountered throughout the project, predominantly comprising mainstream situations such as light influence, view angle influence, occlusion influence, and dense fruit influence. Illustrative examples can be observed in Fig. 2. The dataset is divided into three distinct sets: training, validation, and test. The ratio of these sets is 8:1:1, respectively. The data is labelled in YOLO format using Labelme. The publicly available datasets in the industry are composed of fruit images in a single background without any interference. This situation is far from the actual orchard environment and does not satisfy the industry's needs. Consequently, this paper addresses this gap in the literature. The dataset enhances the target features, enabling the model to improve its ability to extract and generalise features, thereby avoiding overfitting and underfitting. The dataset has been validated through experimentation, demonstrating its efficacy as a reliable source of feature inputs for the network, which significantly improves the reliability and accuracy of the object detection network in practical applications. The

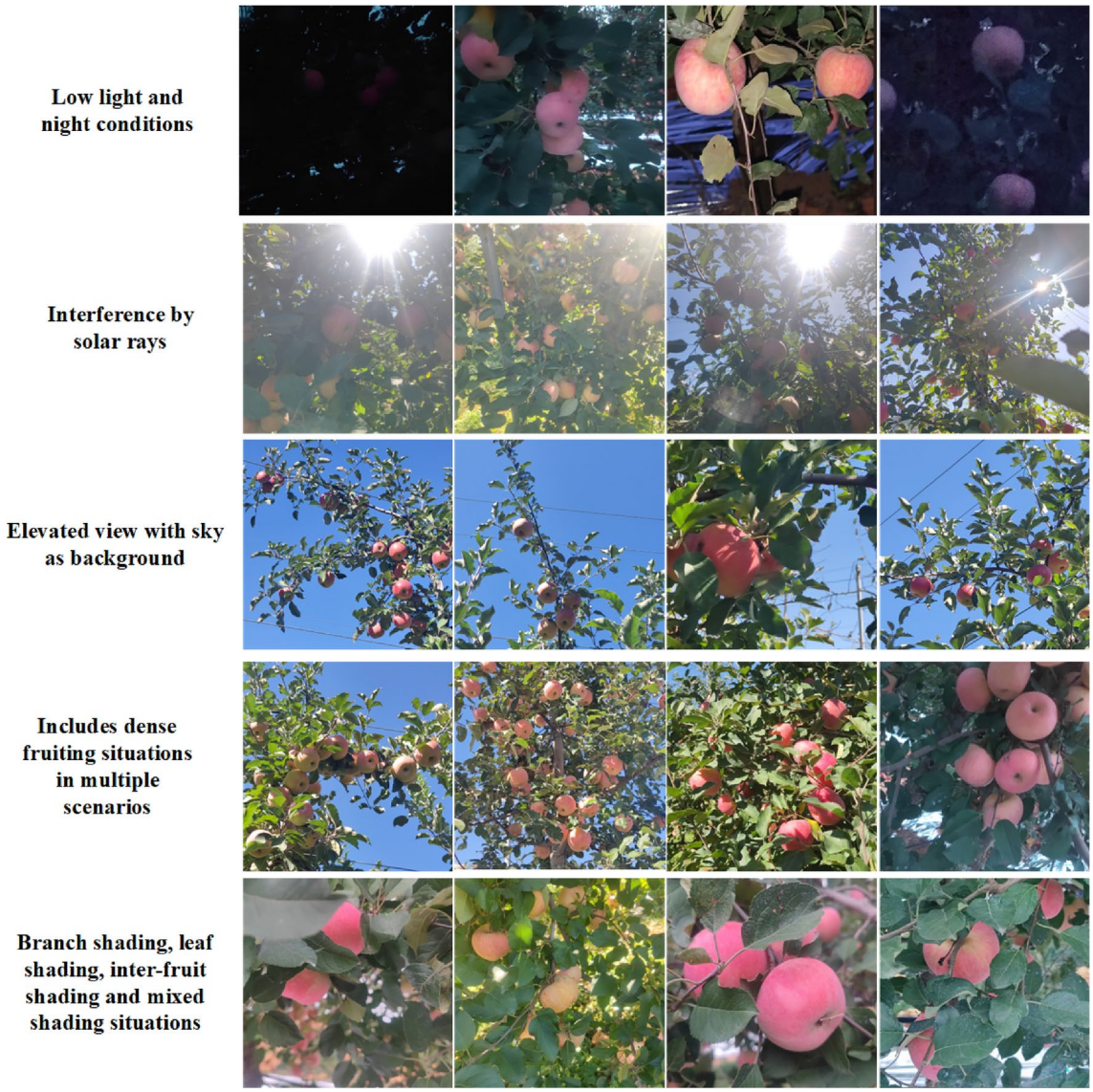

Fig. 2. Typical scenarios for appledatas data centres.

datasets used and/or analysed during the current study available from the corresponding author on reasonable request.

Methods

Overall network architecture

Figure 3 illustrates the lightweight object detection network, designated as the ELD, which is the subject of this study. The objective of the network is to reduce the consumption of computational resources while maintaining high object detection accuracy, thereby enhancing its practicality. The specific network parameters are detailed in Table 1.

Images captured in practical applications contain a considerable amount of superfluous background information. In order to eliminate the interference and maintain optimal network performance, a novel lightweight yet efficient backbone network is proposed. The module initially preserves the essential characteristics of the fruits with high efficiency through the Conv_bn_SiLU module. Subsequently, the underlying feature extraction is carried out by the EGSS module, including steps 1 and 2. The EGSS module prioritises the reuse of initial features as well as feature refinement and fusion through the MACttention concern mechanism. As the EGSS module is more capable of processing feature information at the end of the backbone, the traditional SPPF structure is cancelled at the end of the backbone.

The SlimPAN methodology employs channel compression and point-by-point convolution at the neck network to effectively conserve resources. This approach entails the initial reduction of the number of channels and their subsequent unification, which optimizes the utilization of network resources. Given the enhanced feature fusion capacity of the C3 configuration, the substructure does not result in significant resource wastage

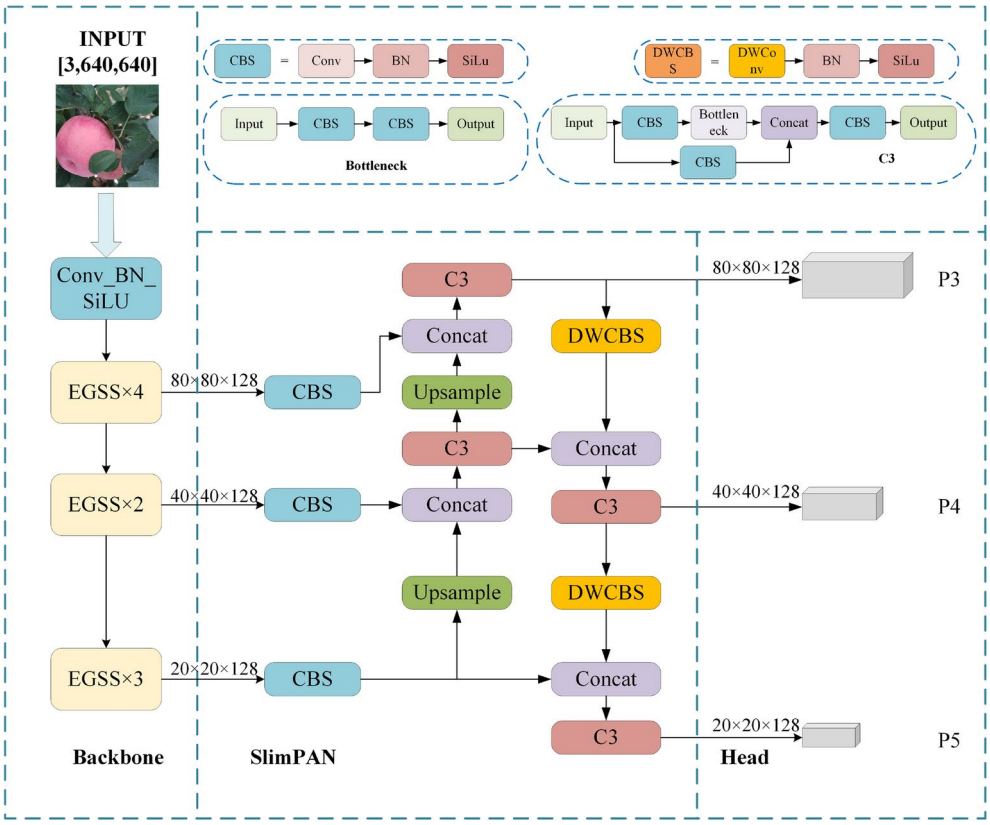

Fig. 3. Architecture of ELD network(The EGSS lightweight feature extraction module employs MCAttention, an attention mechanism, to construct a novel backbone network in an innovative way. Furthermore, it has been lightweighted by SlimPAN to achieve a compressed model. Subsequent incorporation of the Shape-IOU loss function and utilisation of knowledge refinement serve to enhance the accuracy and stability of target detection).

Layers	Module/step	Input size	Kernel size	Layers	Module/step	Input size	Kernel size
1	CBS/2	640×640 ×3	3 ×3	14	Upsample	20×20 ×512	
2	EGSS/2	320×320 ×32		15	Concat		
3	EGSS/1	160×160 ×64		16	C3/1	40×40 ×128	
4	EGSS/2	160×160 ×128		17	Upsample	40×40 ×128	
5	EGSS/1	80×80 ×128		18	Concat	80×80 ×128	
6	CBS/1	80×80 ×128	1 ×1	19	C3/1	80×80 ×128	
7	EGSS/2	80×80 ×128		20	DWConv/2	80×80 ×128	5 ×5
8	EGSS/1	40×40 ×256		21	Concat		
9	CBS/1	40×40 ×256	1 ×1	22	C3/1	40×40 ×128	
10	EGSS/2	40×40 ×256		23	DWConv/2	40×40 ×128	
11	EGSS/1	20×20 ×512		24	Concat		5 ×5
12	EGSS/1	20×20 ×512		25	C3/1	20×20 ×128	
13	CBS/1	20×20 ×512	1 ×1	26	Detect		

Table 1. Detailed configuration table of each module in the ELD network.

following channel compression, and thus it is retained for utilisation in this investigation. Furthermore, depth-separable convolution with a smaller parameter number is employed in lieu of the conventional convolution for downsampling. The convolution kernel size can be modified to regulate the sensory field size. Subsequently, the Shape-IOU loss function is employed to address the limitations of the traditional loss function, thereby enhancing the network's resilience. Ultimately, a lightweight network with high accuracy and straightforward deployment is developed through knowledge distillation.

Lightweight feature extraction module

Images acquired during apple detection in orchard environments typically exhibit a relatively complex background and are frequently contaminated with a considerable amount of noise. This has a considerable detrimental effect on the feature extraction performed by the network, which in turn affects the stability of the real-time object detection by the network. In order to address these challenges and enhance the utilisation of useful feature information in complex scenes, we have designed a lightweight feature extraction module, EGSS(Efficient Ghost-Shuffle Slim moudle), which is available in two different structures. Figure 4a illustrates a structure with a step size of 1, whereby the original input is segmented in consideration of the richness of the original image features. One part of the image retains the original features, while the other part undergoes feature extraction to eliminate unnecessary noise. Subsequently, the features are processed through operations such as the attention mechanism. Subsequently, the original features are superimposed, thereby establishing direct information interaction and enabling direct supervision of the earlier layers. Finally, effective features are fully mixed through shuffling operations, thereby promoting the effective reuse of low-latitude feature information. The structural design ensures both lightness and effective use of low and high latitude features, thereby overcoming the limitations of the majority of existing models. Figure 4b illustrates the structure with a step size of 2, which performs a complete feature reuse and feature extraction of the original features to satisfy the requisite feature richness. Following the initial splicing step, additional feature processing mechanisms are incorporated to mitigate the network degradation caused by the complex uncertainty of the original features.

During the phase of feature extraction, the model size and model feature extraction capability are balanced by two different raw feature utilisation methods, which are combined with MCAttention. The function of the MCAttention attention mechanism is to process multiple feature information, such as channel information, with the aim of extract high-level feature information and capture dependencies between features. It is essential to that when employing cascade operations for feature map expansion, there is a potential for the loss of positional information. To address this, partial convolution operations can be utilized after cascading to enhance the perception of the information and further optimize the model's capability in feature processing.

In the feature extraction module, depth-separable convolution was employed on multiple occasions. With respect to computational efficiency in comparison to the parameter count K Ã- K Ã- C1 Ã- C2 of the standard convolution, the parameter count (denoted as K Ã- K Ã- C1 + C1 Ã- C2) of the depth-separable convolution is significantly reduced, as illustrated in Fig. 5 for a comparison of the two convolutions. Despite the reduction in computational cost, the computational complexity of deep separable convolution during training is still higher than that of traditional convolution, particularly in deep networks. Furthermore, the lack of acceleration support for deep separable convolution on some hardware platforms may also reduce its efficiency. To address this issue, DWConv before MCAttention in the feature extraction module is only used for the teacher network in knowledge distillation. In general, the module guarantees lightweight operation while ensuring the effective utilisation of low-latitude feature information for feature extraction within the network.

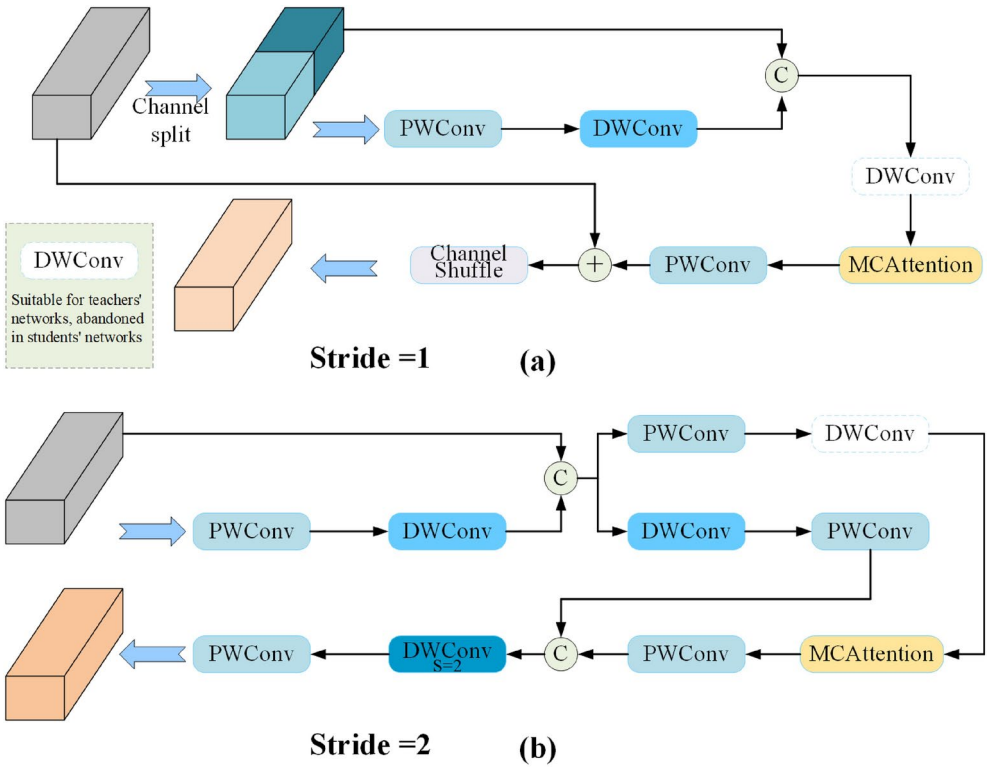

Fig. 4. EGSS module.

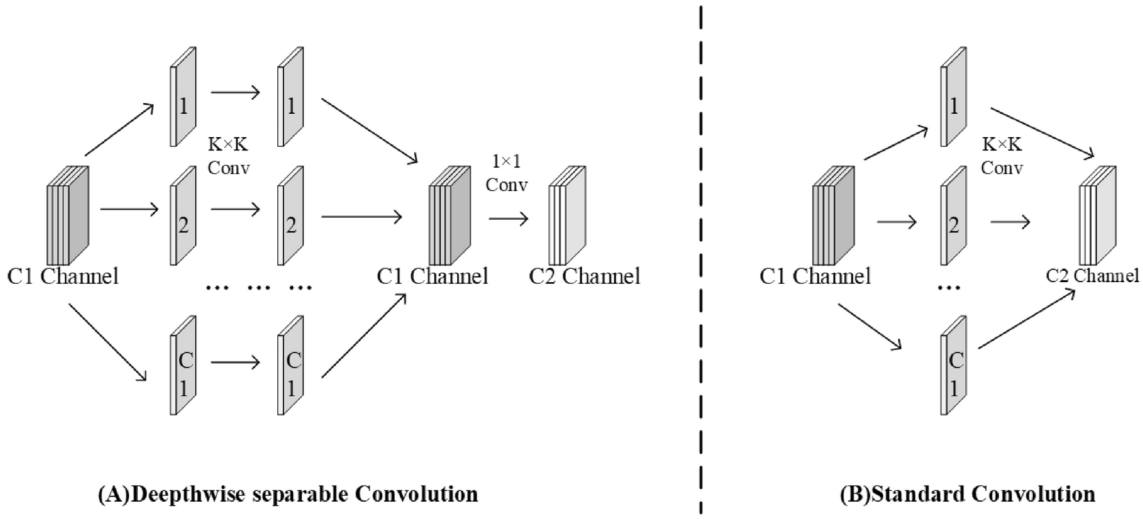

Fig. 5. Contrast between depthwise separable convolution and standard convolution.

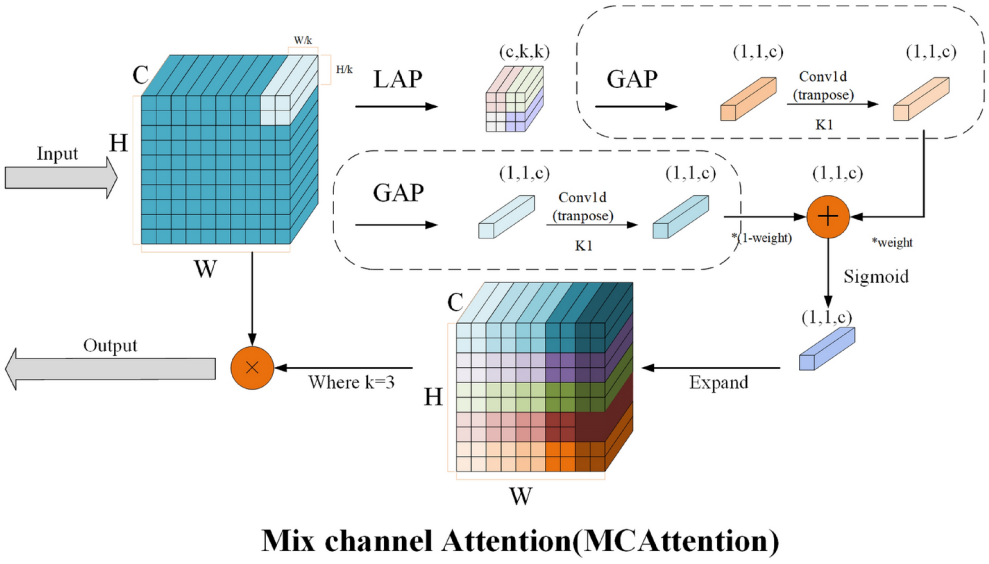

Fig. 6. MCAttention module.

Mix channel attention

The prevailing channel attention mechanism is currently limited in its scope, focusing exclusively on the overall information content and failing to account for the spatial differences between channels. When the entire channel is subjected to feature compression without due consideration of the spatial information, some crucial data may be discarded, potentially leading to an inaccurate prediction of the categories. Furthermore, the pixel-level spatial attention mechanism is susceptible to feature redundancy due to the equal distribution of weights across specific local regions. It is therefore necessary to introduce spatial information in a way that avoids information redundancy. This can be achieved by first processing advanced chunking on one branch and then processing the other branch globally.

As illustrated in Fig. 6, which depicts the principle of MCAttention, the input for this module is processed in parts and subsequently merged. The initial stage transforms the input into a 1*C*K*K vector, enabling the extraction of local spatial information through local pooling. This is followed by a conversion to one-dimensional features, which forms the basis for the second stage, in which global information refinement is conducted directly. Subsequently, the two parts of information are merged to achieve hybrid attention. To circumvent the issue of accuracy deterioration resulting from channel reduction, a one-dimensional convolution is employed to capture the interaction data between each channel and its neighbouring channels. This is achieved through the utilisation of a hybrid channel attention mechanism[39].

The structure in question can be processed by weighting the two-part operation, where weight is (0,1). The experimental results demonstrate that the method of multiplying the two branches by 0.5 and subsequently

adding them together yields superior outcomes compared to the direct addition of the branches. This is primarily due to the suppression of redundant negative information through the operation of weighted sum, as illustrated in Fig. 7. In this figure, MCAttention-0.5 represents the result of addition by a factor of 0.5, while MCAttention-none denotes the result of direct addition. The size of the weighting factor or the operation of direct summing may be selected based on the specific circumstances. This paper employs a weighting factor of 0.5. In conclusion, MCAttention is an effective method for partitioning and utilising the key information of each channel without causing computational redundancy.

Lightweight neck SlimPAN

Figure 8 illustrates the SlimPAN structure, where the green circles represent C3, Concat operations or DWSBS, Concat operations. The connection process between these circles represents the fusion process between different features in the neck network. Following the extraction of features from the backbone network, the input image generates three distinct feature maps of varying dimensions, which are then transferred to the neck network for feature fusion. Upon entering the SlimPAN, the data is initially unified through a 1Ã-1 convolution operation, which serves to eliminate redundant and unimportant channels while simultaneously reducing the cost of feature computation. The depth-separable convolution[40] is employed for all convolutions, with the exception of the initial 1Ã-1 convolution. The depth-divisible convolution extends the perceptual range by employing

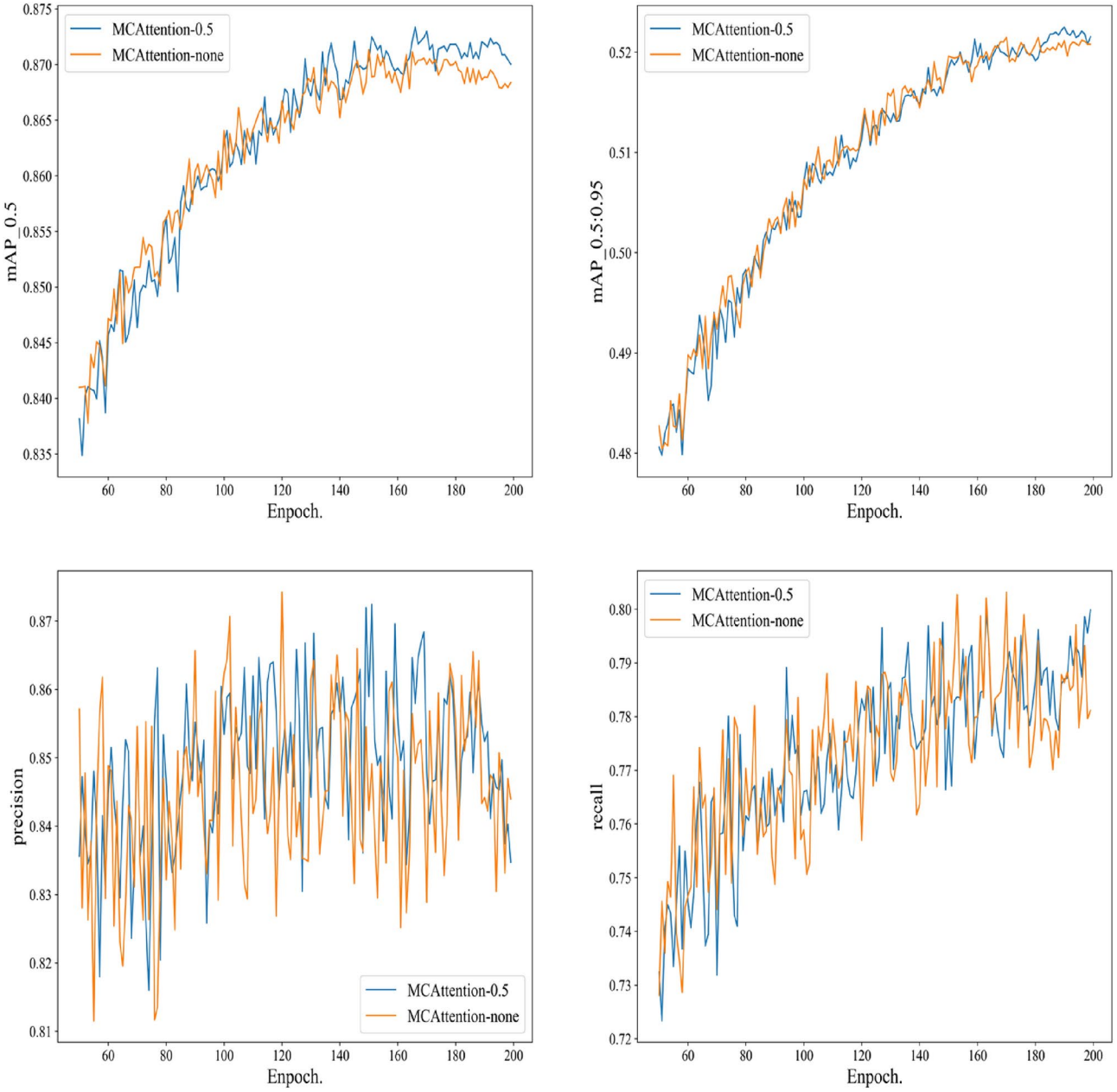

Fig. 7. Comparison of the experimental results of the two MCAttention branch superposition methods.

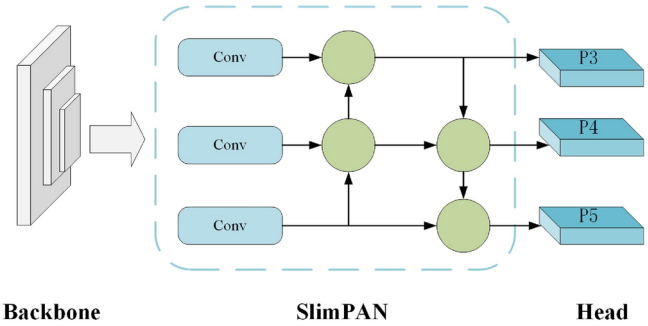

Fig. 8. Lightweight feature fusion module.

5Ã-5 convolutions, thereby enhancing the perceptual range and generalisation ability of the neural network. The results of the experiment are presented in Fig. 9. The objective of this design is to facilitate more efficient information fusion with a reduced number of parameters.

Shape-IOU loss function

This paper introduces a new boundary regression loss function, called Shape-IoU[41]. As shown in Fig 10. This novel approach seeks to rectify the limitations of conventional loss functions by placing particular emphasis on the regression problem of the bounding box. The definition of this function is presented in Eqs. (1–5).

$$ww = \frac{2 \times (w^{gt})^{scale}}{(w^{gt})^{scale} + (h^{gt})^{scale}} \tag{1}$$

$$hh = \frac{2 \times (h^{gt})^{scale}}{(w^{gt})^{scale} + (h^{gt})^{scale}} \tag{2}$$

$$distance^{shape} = hh \times (x_c - x_c^{gt})^2/c^2 + ww \times (y_c - y_c^{gt})^2/c^2 \tag{3}$$

$$\Omega^{shape} = \sum_{t=w,h} (1 - e^{-\omega_t})^{\theta}, \theta = 4 \tag{4}$$

$$\begin{cases} |\omega_w = hh \times \frac{|w - w^{gt}|}{max(w, w^{gt})} \\ |\omega_h = ww \times \frac{|h - h^{gt}|}{max(h, h^{gt})} \end{cases} \tag{5}$$

Where scale, ww and hh denote the scaling factor, horizontal and vertical weighting factors; depending on the size of the target and the size of the gtound truth (GT), respectively. Equation (6) represents the total Shape-IoU loss function.

$$L_{\text{Shape-IoU}} = 1 - \text{IoU} + \text{distance}^{shape} + 0.5 \times \Omega^{shape} \tag{6}$$

Knowledge distillation

Figure 11 illustrates how knowledge distillation (KD) employs a more intricate teaching network to educate a more streamlined and lightweight learner network. This enables the smaller network to obtain a greater quantity of useful knowledge without imposing an additional burden on the network.

The teacher network, as outlined in this paper, comprises a fine-tuned EGSS module and an extended ELD network, designated as ELD-Teacher. The fine-tuned EGSS module is referred to as EGSS-Teacher, which serves to enhance the feature. By incorporating depth-separable 3Ã-3 convolutions and expanding the depth and width of the network, the representation and abstraction capabilities of the teacher network can be enhanced. This allows for the extraction of more complex high-level features, which can be employed to educate the student network. It is unlikely that the teacher and student networks will exhibit significant structural differences. If the networks are too dissimilar, it will be challenging for the student network to acquire useful knowledge, even if the teacher network demonstrates robust learning capabilities and feature information. This process can be interpreted as the student's learning capacity being constrained, rendering it impossible to learn new knowledge even if this limit is surpassed.

The pivotal stage in this process is the dynamic alteration of the softmax probability distribution through the modification of the temperature coefficient T, as illustrated in Eq. (7).

$$q_i = \frac{exp(\frac{z_i}{T})}{\Sigma_j exp(\frac{z_j}{T})} \tag{7}$$

where T=1 corresponds to the original softmax function,T<1 steepens the probability distribution, and T>1 makes the probability distribution smoother. During the process of knowledge distillation, the output results of

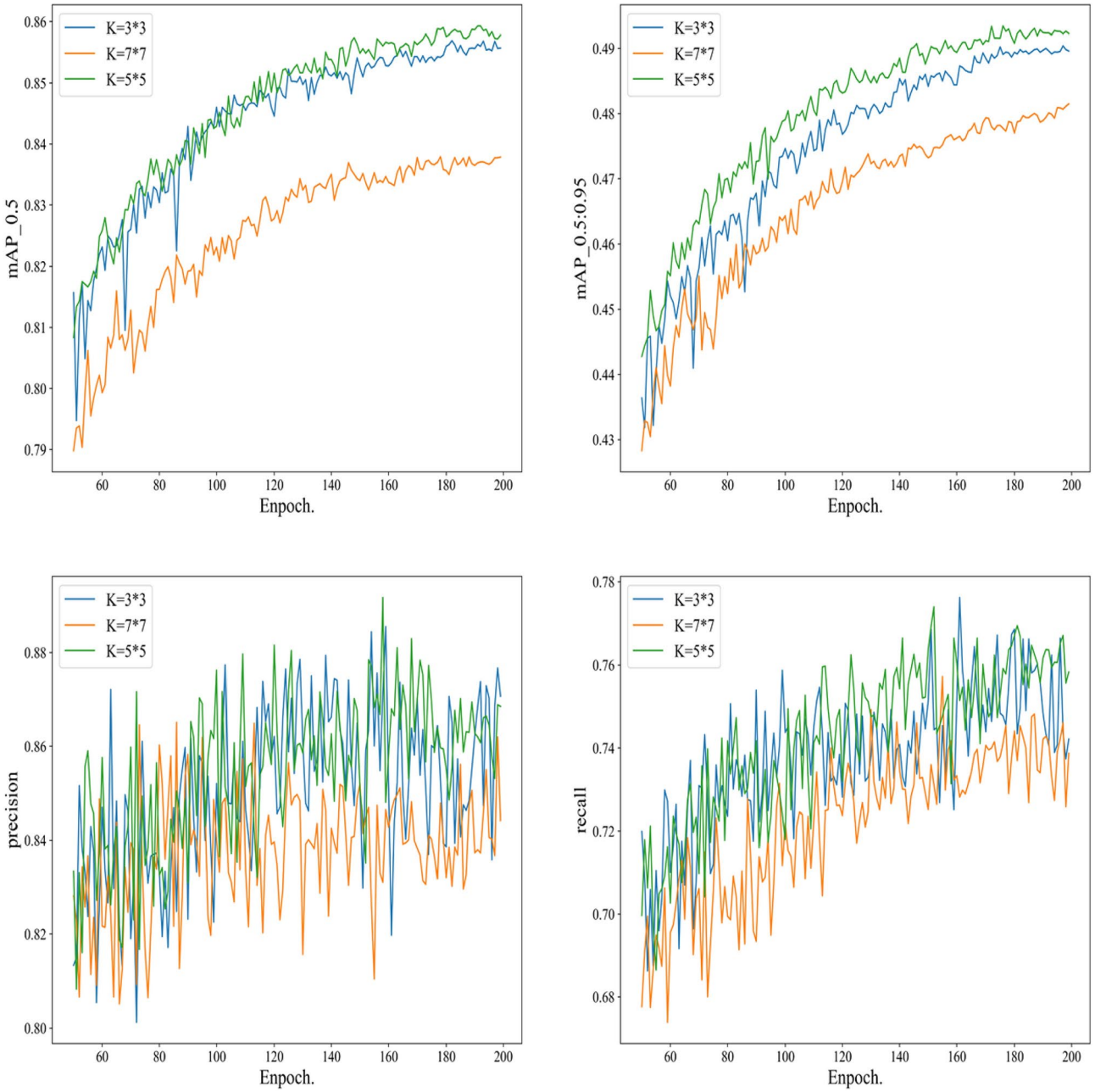

Fig. 9. Comparison of experimental results with different receptive fields under three convolution kernels.

different classes are typically balanced by increasing the temperature (T>1). This operation allows the student model to consider the output results of different classes equally, resulting in more information being obtained during the training process.

Soft labels and soft predictions are generated when the teacher and student networks make predictions following a softmax warm-up. The discrepancy between the soft labels and the soft predictions is calculated in order to guarantee that the predictions of the student network are analogous to those of the teacher network. The calculation process is illustrated in Fig. 12.

Experimental results and analyses

Experimental platform and parameter settings

In our experiments, we utilize Microsoft Windows 10 Operating System, NVIDIA3060 graphics card, Python3.8, Pytorch, CUDA, and cudnn to build a deep learning frame work. The epoch uniform setting is set to 200, and the knowledge distillation operation epoch is set to 300. The optimizer uses SGD by default. The evaluation of network performance in the experiment is primarily based on the trained networkâ€™s perfor mance in the validation set. This includes the metrics such as mAP, P (Precision), R (Recall), model file size, relevant model

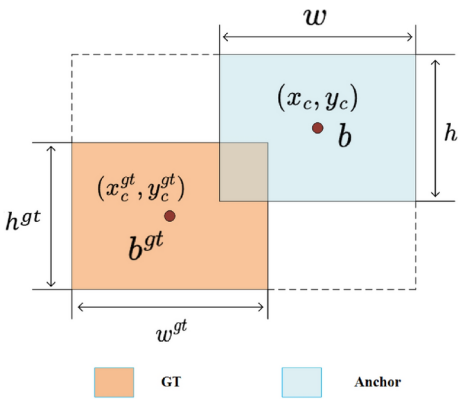

Fig. 10. Principle of shape-iou loss function.

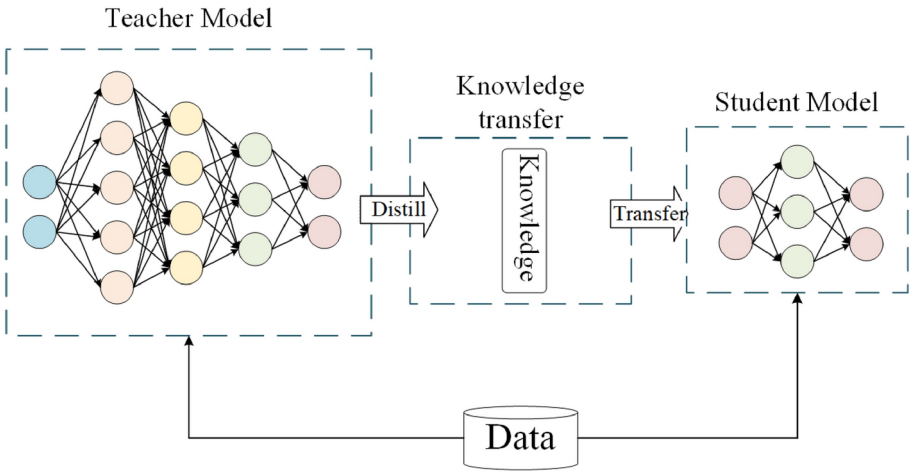

Fig. 11. Knowledge distillation principle.

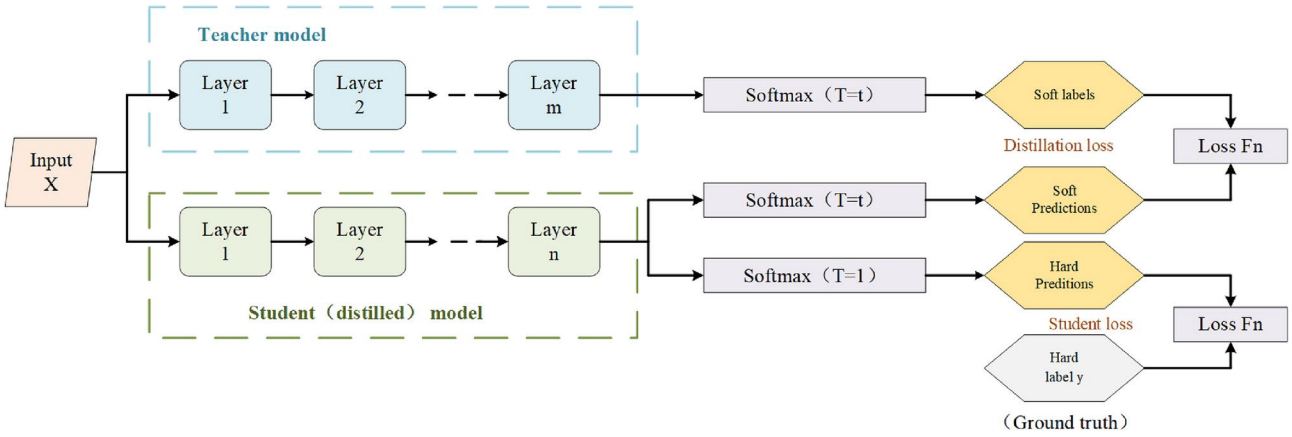

Fig. 12. Knowledge distillation loss calculation.

parameters, GFLOPs, infer ence time, and GPU speed. GPU speed is calculated as 1000/(preprocessing time +inference time+post-processing time). The experimental results are quantified using the aforementioned references. Table 2 displays the hardware configuration and model parameters relevant to the experiment.

Name	Configuration	Name	Param.
CPU	I5-12400F	Learning rate	0.01
GPU	NVIDIA3060	Optmizer	SGD
CUDA	12.2	Batch size	16
torch	2.1.0	enpoch	200

Table 2. Hardware configuration and model parameters.

Evaluation standard

In this study, precision (P), recall (R), mean accuracy (mAP),Average accuracy for relatively small targets in the data(mAPsmall) and FPS are mainly used to evaluate the model performance. As shown in Eqs. 8 and 9. where TP and FN represent true positives and false negatives in each original sample, respectively. mAP denotes the mean accuracy across all categories, where i denotes a specific category and n denotes the total number of categories. In this study, the focus was limited to the identification of fruits, hence n = 1. FPS = 1000/(pre-process+NMS+inference).

$$\begin{aligned} P &= \frac{\mathrm{TP}}{\mathrm{TP}+\mathrm{FP}} \times 100\%\cdot \\ R &= \frac{\mathrm{TP}}{\mathrm{TP}+\mathrm{FN}} \times 100\%\cdot \leftarrow \end{aligned} \tag{8}$$

$$AP = \int_0^1 PR\mathrm{d}R \times 100\%^{\leftrightarrow} \qquad mAP = \frac{1}{n}\sum_i^n AP(i) \times 100\%^{\leftrightarrow} \tag{9}$$

Comparison of attention mechanisms

In order to facilitate a comparative analysis of diverse attention mechanisms, this study incorporates a selection of representative approaches, including ECA[42], SE[43], CA[44], CBAM[45], SimAM[46], and MCAttention, into the EGSS module for experimental comparison under the same dataset.

Figure 13 and Table 3 present a comparative analysis of the performance of various attention mechanisms, as measured by mAP@0.5, mAP@0.5:0.95, precision, and recall. The designation "None" signifies the absence of an attention mechanism within the model. In contrast, the model utilising MCAttention exhibits the highest mAP@0.5, mAP@0.5:0.95 and recall values, thereby indicating that the model employing MCAttention is more efficient in comparison to the models that do not utilise or employ alternative attention mechanisms. This suggests that the proposed attention method is an effective approach. In particular, the CA spatial attention mechanism exhibited a high degree of instability, indicating that the study is more applicable to channel attention. Furthermore, the incorporation of superfluous attentional confederates has the potential to negatively impact the outcomes.

In conjunction with Fig. 14, the impact of diverse attentions can be more effectively conveyed, thus providing a more comprehensive understanding of the subject matter. As illustrated in the heat map, CA, CBAM and a spatial attention model demonstrated the capacity to focus on specific target characteristics. Although SimAM may focus on more information regarding large fruits, it is unable to recognise occlusion and the target is not readily apparent. In contrast, SE and ECA, which represent traditional forms of attention, demonstrate superior results. However, they are unable to extract features from the entire fruit. In contrast, MCAttention is capable of not only focusing on the detailed parts of the image but also of performing comprehensive attention to the features of the large target. This is sufficient to demonstrate the superiority of this module.

Comparison of loss functions

This paper employs the use of Shape-IoU and compares it with other commonly utilised IoUs. The P-value of Shape-IoU is significantly enhanced with the same detection speed, exhibiting a 0.26-point improvement over that of CIOU. Additionally, the map0.5 demonstrates a slight improvement, indicating that the model's stability has been augmented. The results of the experiment are presented in Table 4. It should be noted that the target may require different specifications depending on the device and application scenario. In such cases, the scale factor can be adjusted to align with the actual requirements. In this study, the scale factor employed was 1.0.

Knowledge distillation series comparison experiments

The efficacy of the teacher network in knowledge distillation directly impacts the learning outcome of the student network. A comparative and analytical examination of the three teacher networks, namely YOLOV5s, YOLOV5x and EGSS-tercher, reveals that the utilisation of EGSS-teacher as the teacher network enables the student network to obtain a more comprehensive range of information. This indicates that the most efficient method of knowledge refinement is achieved when the teacher and student networks have similar structures, as evidenced by the experimental results presented in Table 5.

It was established through empirical investigation that the optimal refining effect is attained by employing the L2 loss function and setting the refining temperature at T=20. This finding was derived from an experimental comparison that also demonstrated the pivotal role of the refining loss function in this process. These results are depicted in Figure 15. A comparison of the experimental results is presented in Table 6 for the reader's

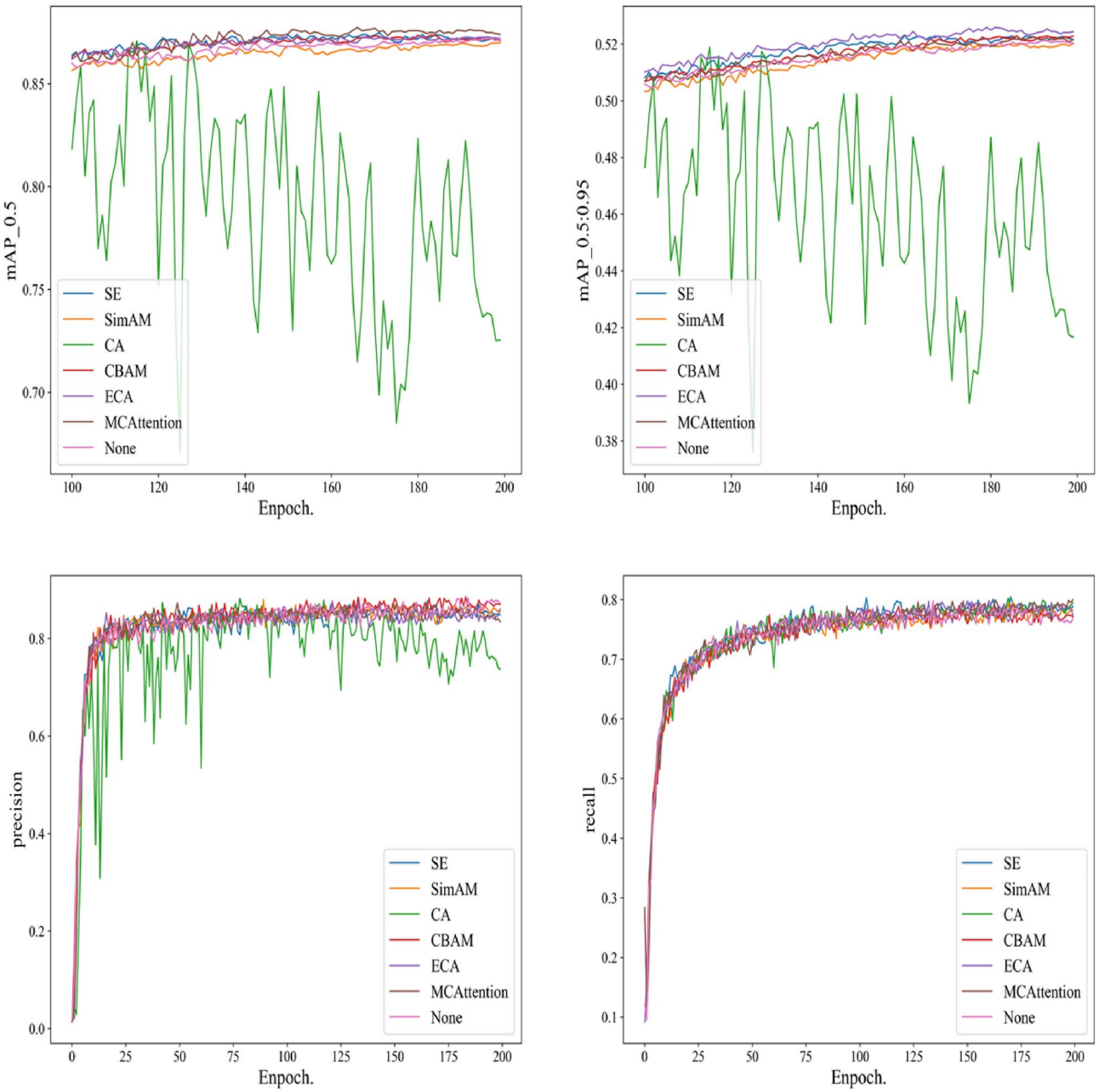

Fig.13. Comparison of different attention modules in mAP@0.5, mAP@0.5:0.95, precision and recall.

Attention	mAP@0.5	mAP@0.5:0.95	P	R
None	0.872	0.521	0.862	0.777
SE	0.871	0.523	0.857	0.781
SimAM	0.869	0.520	0.842	0.787
ECA	0.873	0.525	0.846	0.769
CA	0.871	0.519	0.871	0.771
CBAM	0.871	0.523	0.876	0.765
MCAttention	0.874	0.528	0.864	0.792

Table 3. Comparison of addition attention mechanisms in EGSS modules.

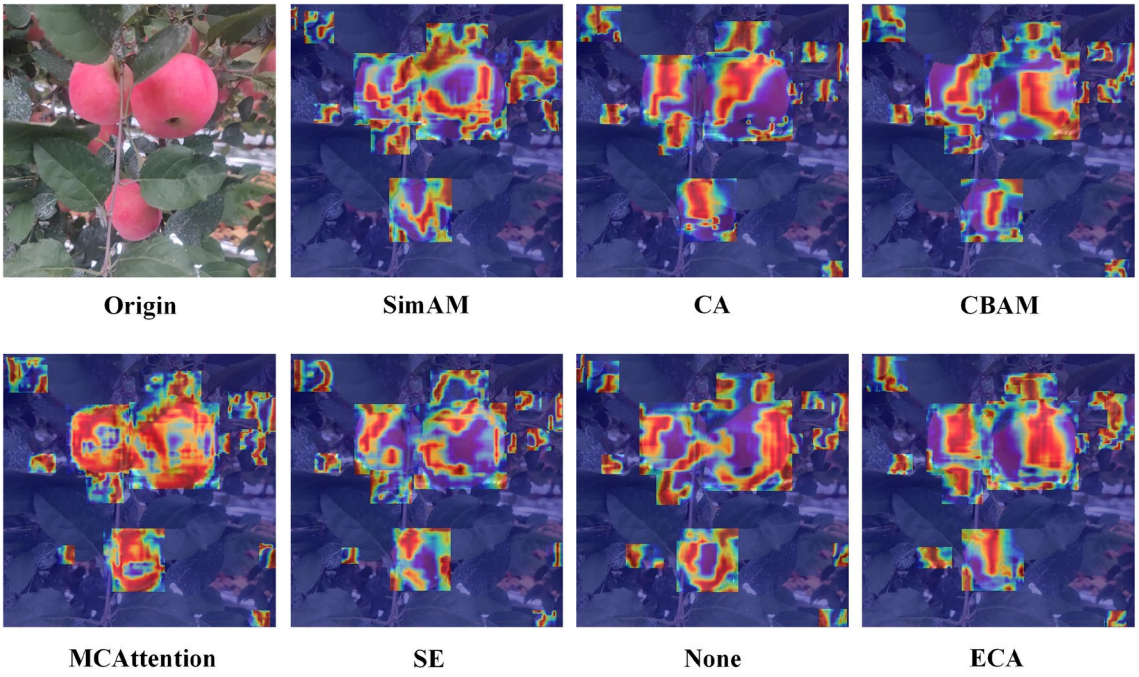

Fig. 14. Heat maps of different attention mechanisms.

Loss	mAP@0.5	*P*
Ciou	0.858	0.846
Shape-iou	0.860	0.872
Siou	0.853	0.845
Giou	0.856	0.841
Diou	0.855	0.857
Eiou	0.856	0.856

Table 4. Loss function comparison experiment.

Teacher	Student map@0.5
YOLOv5s	0.872
YOLOv5x	0.871
EGSS-Tarcher	0.874

Table 5. Comparison of average precision after distillation of student networks by different teacher networks.

convenience. It can be seen that the distillation loss function and temperature adjustment must be adjusted according to the specific circumstances in order to achieve the most effective outcome.

Network improvement ablation experiments

Ablation experiments are a standard tool in the analysis of complex neural networks, allowing for the examination of the impact of specific substructures and training methods on the model. This is a crucial aspect of the structural design of neural networks. In order to verify the effectiveness of the EGSS feature extraction module and the MCAttention attention mechanism, ablation experiments were conducted, the results of which are presented in Table 7.

Table 7 demonstrates that the incorporation of the EGSS module results in a 61% reduction in parameters and a 57% reduction in computation compared to the original model. The mAP@0.5 value remains largely unaltered, while the object detection speed is significantly accelerated. The application of MCAttention on this basis results in an elevation of mAP@0.5 to the standard of the original model, while the amount of computation remains unaltered. Subsequently, further model compression is achieved through lightweight feature fusion, resulting in a reduction of 94% in parameters and 89% in computation compared to the original model. This is accompanied by a slight decline of 1.9% in mAP@0.5, which is deemed to be an acceptable trade-off. To further enhance the

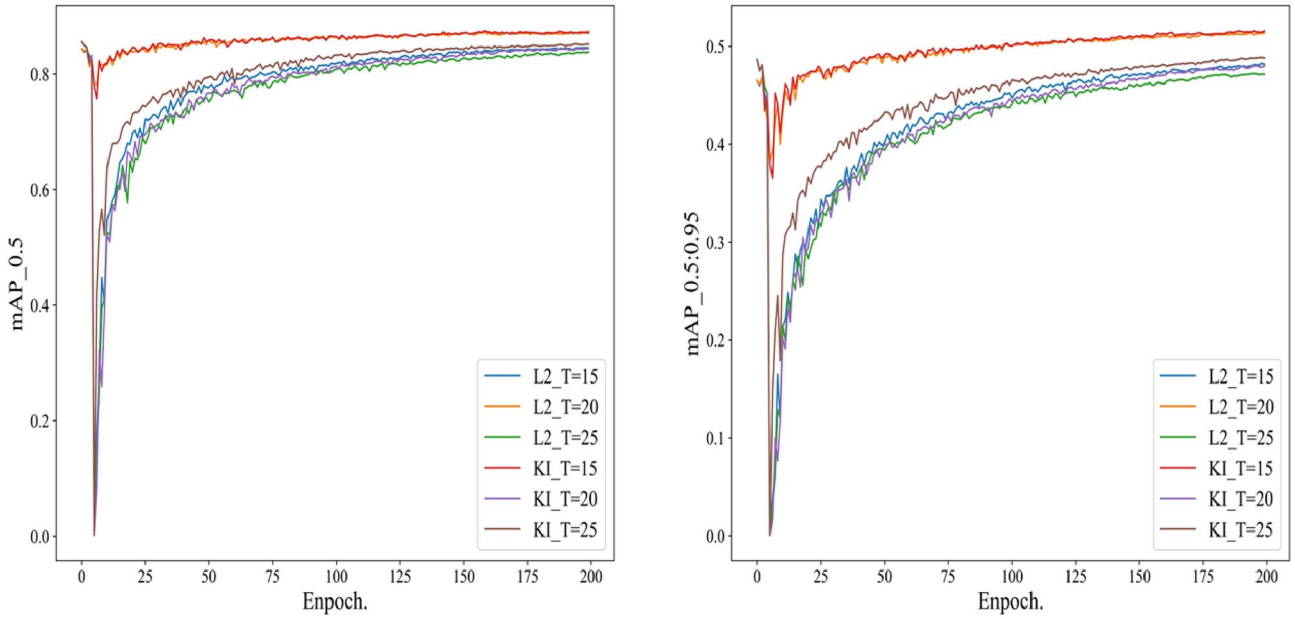

Fig. 15. Comparison of experimental results on distillation loss and distillation temperature.

Loss	Temperature	mAP0.5	P	R	FPS
L2	25	0.856	0.864	0.751	154
L2	20	0.874	0.884	0.772	156
L2	15	0.851	0.860	0.749	119
Kl	25	0.855	0.849	0.763	156
Kl	20	0.851	0.863	0.739	111
Kl	15	0.874	0.876	0.777	143

Table 6. Comparison experiment of distillation loss and distillation temperature.

Model	mAP@0.5	GFLOPs	Parameters	Weight(M)	Inference time(ms)	FPS
Baselime	0.877	15.8	7.0×10^6	14.4	7.3	88
+EGSS	0.872	6.9	3.1×10^6	6.7	4.7	123
+MCAttention	0.877	7.2	3.1×10^6	6.7	4.7	138
+SlimPAN	0.858	1.7	4.3×10^5	1.3	3.8	156
+Shape-IOU	0.860	1.7	4.3×10^5	1.3	3.8	156
+Distillation	0.874	1.7	4.3×10^5	1.3	3.4	156

Table 7. Ablation experiments.

object detection efficiency, knowledge distillation was employed to elevate the model accuracy to 0.874, while the detection speed was increased to 1.77 times that of the original model. In summary, the integration of these modules has led to the enhancement of accuracy, a substantial reduction in both computational and parametric quantities of the new model, as well as notable improvements in inference time and frames per second. The network demonstrates an equilibrium between a streamlined configuration and network performance.

Experimental comparison of different target

This study compares the performance of the proposed ELD algorithm with that of established object detection algorithms to assess its effectiveness. The image resolution employed in the dataset is 640Ã-640, and all the comparative experimental algorithms are based on the PyTorch open-source framework. The experimental results are presented in Table 8, which includes the following algorithms: YOLOv5[47], YOLOv7-tiny[48], YOLOv6[49], YOLOv8[50], Gold-YOLO[51], YOLOv9[52], and YOLOv10[53]. These networks were evaluated on the basis of their relatively lightweight structure. Furthermore, networks that combine YOLO with lightweight models were evaluated, including Shufflenetv2[54], Ghostnetv2[55], Mobilenetv3[56], and FasterNet[57]. The latest lightweight

Model	mAP@0.5	mAPsmall	GFLOPs	Parameters	Weight/M	Inference time(ms)	FPS	CPU FPS
YOLOv5s	0.877	0.228	15.8	7.0×10^6	14.4	7.3	88	10
YOLOv5n	0.872	0.192	4.1	1.8×10^6	3.9	3.6	149	20
YOLOv6s	0.854	0.209	44.2	17.2×10^6	36.2	13.53	68	5
YOLOv6n	0.854	0.198	11.1	4.7×10^6	9.3	6.18	146	12
YOLOv7-tiny	0.884	0.233	13.2	6.2×10^6	12.3	7.1	114	15
YOLOv8s	0.873	0.201	28.4	11.1×10^6	22.5	5.8	161	18
YOLOv8n	0.870	0.193	8.1	3.0×10^6	6.2	4.8	185	30
YOLOv9n	0.880	0.230	18.3	4.2×10^6	9.2	15.0	52	5
YOLOv10s	0.881	0.228	24.4	8.0×10^6	16.6	6.9	141	16
Gold-YOLO-N	0.858	0.196	12.1	5.6×10^6	12	4.8	156	10
ShuffleNetV2-YOLO	0.838	0.143	1.8	8.4×10^5	2.1	2.8	172	50
GhostNetV2-YOLO	0.860	0.187	4.5	2.7×10^6	6.2	8.6	84	10
MobileNetV3-YOLO	0.845	0.173	2.5	1.4×10^6	3.2	3.2	164	43
FasterNet-YOLO	0.864	0.182	2.2	1.1×10^6	2.4	3.0	147	44
YOLOv5-DE	0.868	0.196	5.2	1.9×10^6	4.2	6.0	133	38
GGS-PF-YOLOv5	0.855	0.188	2.2	4.1×10^5	1.1	5.1	113	23
YOLOv5-Lite-S	0.770	0.005	3.7	1.5×10^6	3.4	3.3	78	11
DETR	0.767	0.003	225	41×10^6	—	—	—	—
ELD	0.874	0.222	1.7	4.3×10^5	1.3	3.4	156	47

Table 8. Comparison experiment of different target detection algorithms.

networks include YOLOv5-DE[58],GGS-PF-YOLOv5[59],YOLOv5-Lite[60]. DETR[61] represents the inaugural object detection framework to utilise Transformer as a fundamental component of the object detection model, thereby providing a valuable reference point. The Transformer model features a sophisticated architectural design that incorporates a multi-head attention mechanism, positional encoding, residual connectivity, and other advanced components. This results in a large model size. The capture of long-distance dependencies in input sequences requires the model to have sufficient parameters and capacity to handle long sequences of information. Training on large-scale datasets is usually required to obtain better performance. The self-attention mechanism necessitates the computation of the correlation between all features, which results in a relatively large number of parameters. In contrast, convolutional neural networks typically possess the property of shared weights, and the number of parameters is relatively small.Challenges to meet the real-time requirements of edge devices.

In conjunction with Fig. 16, this provides a more detailed and sophisticated perspective on the performance of each network. The diameter of the circles in the figure is indicative of the number of parameters. The smaller the circle and the closer it is to the upper right corner, the superior the results. The ELD model is the smallest, fastest, and most accurate of the networks, indicating that this network achieves an optimal balance between these two aspects. In comparison to the benchmark model, YOLOv5s, the proposed model exhibits a mere 0.3% reduction in accuracy, yet is nearly twice as rapid, accompanied by a considerable decrease in the number of parameters and computational cost. The results demonstrate that the network exhibits a more comprehensive performance at a minimal cost in terms of accuracy.

Discussion

In practice, the fruit ripening period is distinguished by the presence of lush branches and leaves, as well as the viewing angle of image acquisition, which gives rise to complex scenarios characterised by overlapping and occluded targets, as well as dense targets. Figure 17 compares the detection performance of the improved model and the mainstream detection network in these scenarios. The YOLOv5, YOLOv7, and YOLOv8 models demonstrate a certain degree of leakage and insensitivity to the feature regions, in contrast to the improved model, which exhibits more accurate recognition and localization properties.

Figure 18 illustrates the outcomes of the model proposed in this paper in comparison to those produced by the widely utilized object detection network. In cases (a) and (b), which are largely free of interference and relatively straightforward to detect, the networks demonstrate satisfactory performance. However, YOLOv7 exhibits a slight tendency to under-detect, while YOLOv5 displays a proclivity for re-detection. In Case (c), all three networks (YOLOv5, YOLOv7, and YOLOv8) demonstrate a tendency to miss detection of small target occlusion, with YOLOv5 exhibiting a particularly high rate of retries. The interference in Case (d) is more pronounced, posing a significant challenge to the network. All networks exhibit a degree of missed detection, with YOLOv5 displaying the most severe performance. The proposed model demonstrates the strongest interference ability.

Limitations of the study

The limitations of this questioned study can be summarised in these points:

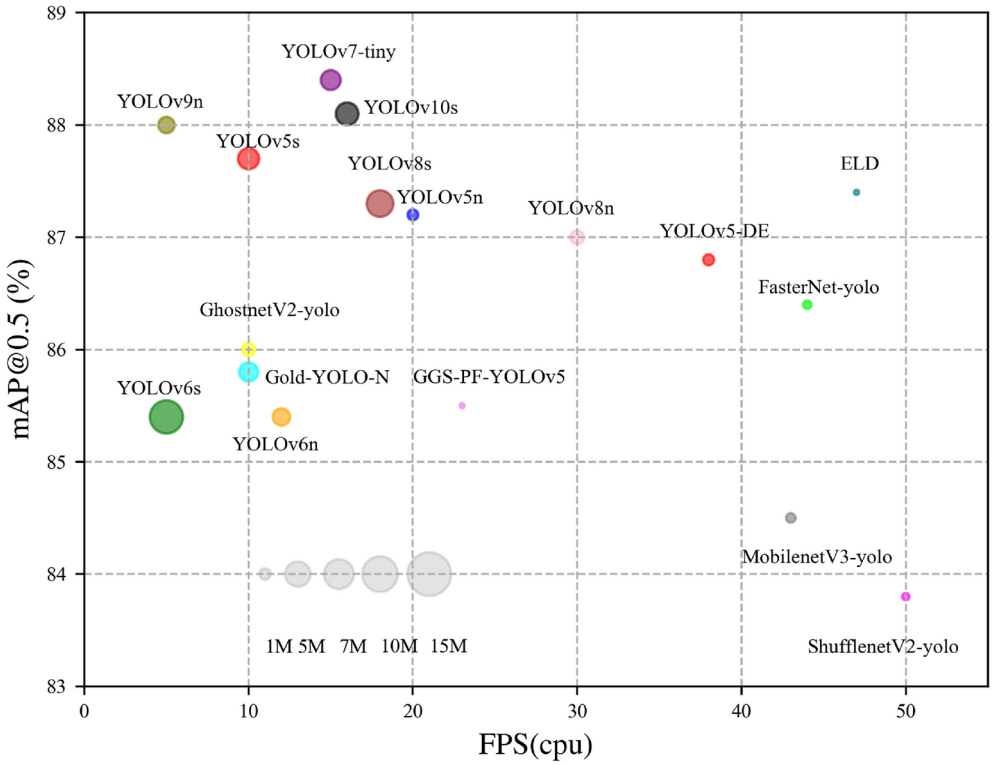

Fig. 16. Different object detection networks compared in terms of mAP@0.5, FPS, and model size.

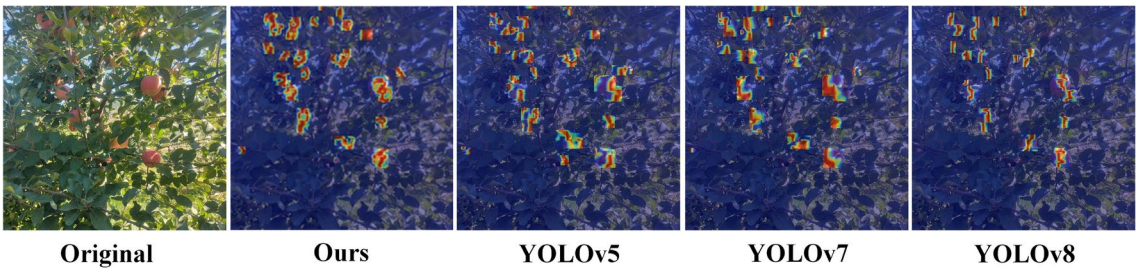

Fig. 17. Heat map of different detection networks.

(1) The current dataset, which contains apple species as well as natural environment condition numbers can be further increased.
(2) Although the model works better than conventional mainstream object detection algorithms in terms of accuracy and lightness, there is still a certain leakage rate under strong sunlight.

Conclusion

In the context of automated harvesting devices, the accuracy of target identification is of paramount importance. This paper addresses the limitations of existing methods and proposes a lightweight object detection model that maintains good object detection accuracy. This study considers real-world application scenarios in a comprehensive manner and proposes an EGSS feature extraction module and a hybrid attention mechanism for MCAttention in an innovative manner. The results demonstrate the efficacy of these methods in extracting target features for dense , even in the presence of strong interference. Both modules are designed as plug-and-play components that can be fully integrated into other object detection networks. Furthermore, the accuracy of the network is enhanced through the implementation of a knowledge refinement strategy. While the network and method values presented in this paper are specific to apple detection, they can be generalized to other fruit detection scenarios.

The final network achieves an accuracy of 87.4% with a mere 1.7 GFLOPs of computation, a parameter count that constitutes a mere 6.1% of the baseline model. It demonstrates the ability to reduce computational overhead, enhance object detection speed, and guarantee object detection accuracy. Subsequent to this, it will be deployed

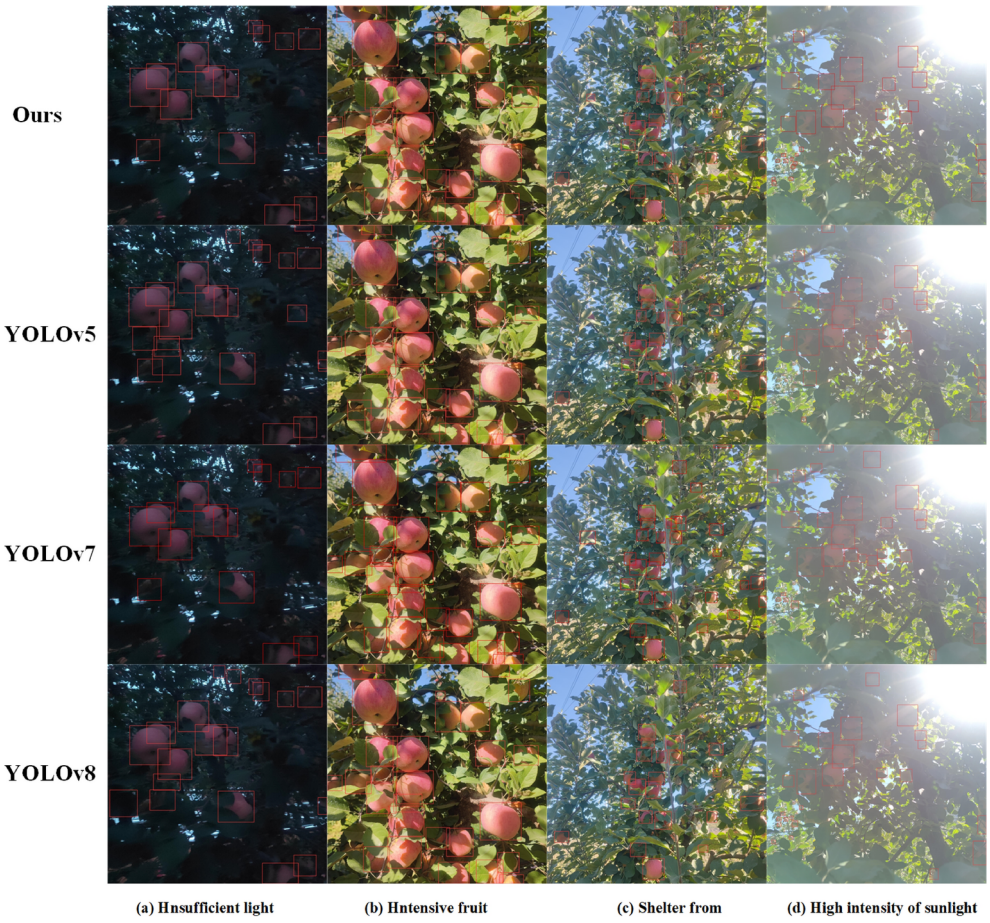

Fig. 18. Comparison of classical network detection results in typical real-world scenarios.

to edge devices to further validate its effectiveness and integrate a object detection and localization system for practical engineering applications.

Data availability

The datasets used and/or analysed during the current study available from the corresponding author on reasonable request.

Received: 31 July 2024; Accepted: 15 October 2024
Published online: 30 October 2024

References

1. Lehnert, C., Sa, I., McCool, C., Upcroft, B. & Perez, T. Sweet pepper pose detection and grasping for automated crop harvesting. In *2016 IEEE International Conference on Robotics and Automation (ICRA)*, 2428–2434 (IEEE, 2016).
2. Wu, L., Ma, J., Zhao, Y. & Liu, H. Apple detection in complex scene using the improved yolov4 model. *Agronomy***11**, 476 (2021).
3. Liu, X., Zhao, D., Jia, W., Ji, W. & Sun, Y. A detection method for apple fruits based on color and shape features. *IEEE Access***7**, 67923–67933 (2019).
4. Tian, Y. et al. Apple detection during different growth stages in orchards using the improved yolo-v3 model. *Comput. Electron. Agric.***157**, 417–426 (2019).
5. Xiao, B., Nguyen, M. & Yan, W. Q. Apple ripeness identification from digital images using transformers. *Multimed. Tools Appl.***83**, 7811–7825 (2024).
6. Xue, Y. et al. Handling occlusion in UAV visual tracking with query-guided redetection. *IEEE Trans. Instrum. Measure.***73**, 5030217 (2024).
7. Xue, Y. et al. Consistent representation mining for multi-drone single object tracking. *IEEE Trans. Circ. Syst. Video Technol.*[SPACE]https://doi.org/10.1109/TCSVT.2024.3411301 (2024).
8. Xue, Y. et al. Mobiletrack: Siamese efficient mobile network for high-speed UAV tracking. *IET Image Proc.***16**, 3300–3313 (2022).
9. Xue, Y., Jin, G., Shen, T., Tan, L. & Wang, L. Template-guided frequency attention and adaptive cross-entropy loss for UAV visual tracking. *Chin. J. Aeronaut.***36**, 299–312 (2023).
10. Xue, Y. et al. Smalltrack: Wavelet pooling and graph enhanced classification for uav small object tracking. *IEEE Trans. Geosci. Remote Sens.***61**, 5618815 (2023).
11. Xu, R. et al. Instance segmentation of biological images using graph convolutional network. *Eng. Appl. Artif. Intell.***110**, 104739 (2022).

12. Zhang, J., Qian, S. & Tan, C. Automated bridge surface crack detection and segmentation using computer vision-based deep learning model. *Eng. Appl. Artif. Intell.***115**, 105225 (2022).
13. Jaderberg, M., Simonyan, K., Zisserman, A. et al. Spatial transformer networks. *Adv. Neural Inform. Process. Syst.***28** (2015).
14. Yu, J., Jiang, Y., Wang, Z., Cao, Z. & Huang, T. Unitbox: An advanced object detection network. In *Proceedings of the 24th ACM international conference on Multimedia*, pp. 516–520 (2016).
15. Rezatofighi, H. et al. Generalized intersection over union: A metric and a loss for bounding box regression. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 658–666 (2019).
16. Zheng, Z. et al. Distance-IOU loss: Faster and better learning for bounding box regression. *Proc. AAAI Conf. Artif. Intell.***34**, 12993–13000 (2020).
17. Zhang, Y.-F. et al. Focal and efficient IOU loss for accurate bounding box regression. *Neurocomputing***506**, 146–157 (2022).
18. Gevorgyan, Z. Siou loss: More powerful learning for bounding box regression. *arXiv preprint*[SPACE]arXiv:2205.12740 (2022).
19. Han, K. et al. Ghostnet: More features from cheap operations. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp 1580–1589 (2020).
20. Xiong, Y. et al. Mobiledets: Searching for object detection architectures for mobile accelerators. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 3825–3834 (2021).
21. Yu, G. et al. Pp-picodet: A better real-time object detector on mobile devices. *arXiv preprint*[SPACE]arXiv:2111.00902 (2021).
22. Maaz, M. et al. Edgenext: Efficiently amalgamated CNN-transformer architecture for mobile vision applications. In *European conference on computer vision* 3–20 (Springer, Berlin, 2022).
23. Hinton, G. Distilling the knowledge in a neural network. *arXiv preprint*[SPACE]arXiv:1503.02531 (2015).
24. Lan, Q. & Tian, Q. Instance, scale, and teacher adaptive knowledge distillation for visual detection in autonomous driving. *IEEE Trans. Intell. Vehic.***8**, 2358–2370 (2022).
25. Redmon, J., Divvala, S., Girshick, R. & Farhadi, A. You only look once: Unified, real-time object detection. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 779–788 (2016).
26. Redmon, J. & Farhadi, A. Yolo9000: better, faster, stronger. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 7263–7271 (2017).
27. Redmon, J. & Farhadi, A. Yolov3: An incremental improvement. *arXiv preprint* [SPACE]arXiv:1804.02767 (2018).
28. Bochkovskiy, A., Wang, C.-Y. & Liao, H.-Y. M. Yolov4: Optimal speed and accuracy of object detection. *arXiv preprint*[SPACE]arXiv:2004.10934 (2020).
29. Patnaik, S. K., Babu, C. N. & Bhave, M. Intelligent and adaptive web data extraction system using convolutional and long short-term memory deep learning networks. *Big Data Min. Anal.***4**, 279–297 (2021).
30. Im Choi, J. & Tian, Q. Visual-saliency-guided channel pruning for deep visual detectors in autonomous driving. In *2023 IEEE Intelligent Vehicles Symposium (IV)*, 1–6 (IEEE, 2023).
31. Park, S., Kang, D. & Paik, J. Cosine similarity-guided knowledge distillation for robust object detectors. *Sci. Rep.***14**, 18888 (2024).
32. Wang, J. et al. Multi-constraint molecular generation based on conditional transformer, knowledge distillation and reinforcement learning. *Nat. Mach. Intell.***3**, 914–922 (2021).
33. Kim, M.-J. et al. Screening obstructive sleep apnea patients via deep learning of knowledge distillation in the lateral cephalogram. *Sci. Rep.***13**, 17788 (2023).
34. Zhao, J. et al. Yolo-granada: A lightweight attentioned yolo for pomegranates fruit detection. *Sci. Rep.***14**, 16848 (2024).
35. Wang, J. et al. Toward surface defect detection in electronics manufacturing by an accurate and lightweight yolo-style object detector. *Sci. Rep.***13**, 7062 (2023).
36. Guo, H., Wu, T., Gao, G., Qiu, Z. & Chen, H. Lightweight safflower cluster detection based on yolov5. *Sci. Rep.***14**, 18579 (2024).
37. Lin, H., Cheng, X., Wu, X. & Shen, D. Cat: Cross attention in vision transformer. In *2022 IEEE International Conference on Multimedia and Expo (ICME)*, 1–6 (IEEE, 2022).
38. Fu, J. et al. Dual attention network for scene segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp 3146–3154 (2019).
39. Wan, D. et al. Mixed local channel attention for object detection. *Eng. Appl. Artif. Intell.***123**, 106442 (2023).
40. Chollet, F. Xception: Deep learning with depthwise separable convolutions. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp 1251–1258 (2017).
41. Zhang, H. & Zhang, S. Shape-iou: More accurate metric considering bounding box shape and scale. *arXiv preprint*[SPACE]arXiv:2312.17663 (2023).
42. Wang, Q. et al. Eca-net: Efficient channel attention for deep convolutional neural networks. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 11534–11542 (2020).
43. Hu, J., Shen, L. & Sun, G. Squeeze-and-excitation networks. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 7132–7141 (2018).
44. Hou, Q., Zhou, D. & Feng, J. Coordinate attention for efficient mobile network design. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 13713–13722 (2021).
45. Woo, S., Park, J., Lee, J.-Y. & Kweon, I. S. Cbam: Convolutional block attention module. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pp. 3–19 (2018).
46. Yang, L., Zhang, R.-Y., Li, L. & Xie, X. Simam: A simple, parameter-free attention module for convolutional neural networks. In *International Conference on Machine Learning*, pp. 11863–11874 (PMLR, 2021).
47. Alexey Bochkovskiy, H.-Y. M. L., Chien-Yao Wang. Yolov5. https://github.com/ultralytics/yolov5 (2021).
48. Wang, C.-Y., Bochkovskiy, A. & Liao, H.-Y. M. Yolov7: Trainable bag-of-freebies sets new state-of-the-art for real-time object detectors. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 7464–7475 (2023).
49. Li, C. et al. Yolov6: A single-stage object detection framework for industrial applications. *arXiv preprint*[SPACE]arXiv:2209.02976 (2022).
50. Alexey Bochkovskiy, C.-Y. W. & Liao, H.-Y. M. Yolov8. https://github.com/ultralytics/ultralytics (2023).
51. Wang, C. et al. Gold-yolo: Efficient object detector via gather-and-distribute mechanism. *Adv. Neural Inf. Process. Syst.***36** (2024).
52. Wang, C.-Y., Yeh, I.-H. & Liao, H.-Y. M. Yolov9: Learning what you want to learn using programmable gradient information. *arXiv preprint*[SPACE]arXiv:2402.13616 (2024).
53. Wang, A. et al. Yolov10: Real-time end-to-end object detection. *arXiv preprint*[SPACE]arXiv:2405.14458 (2024).
54. Ma, N., Zhang, X., Zheng, H.-T. & Sun, J. Shufflenet v2: Practical guidelines for efficient cnn architecture design. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pp. 116–131 (2018).
55. Tang, Y. et al. Ghostnetv2: Enhance cheap operation with long-range attention. *Adv. Neural. Inf. Process. Syst.***35**, 9969–9982 (2022).
56. Howard, A. et al. Searching for mobilenetv3. In *Proceedings of the IEEE/CVF International Conference on computer Vision*, pp. 1314–1324 (2019).
57. Chen, J. et al. Run, don't walk: chasing higher flops for faster neural networks. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 12021–12031 (2023).
58. Ma, X., Li, Y., Yang, Z., Li, S. & Li, Y. Lightweight network for millimeter-level concrete crack detection with dense feature connection and dual attention. *J. Build. Eng.***94**, 109821 (2024).
59. Xiao, Q., Li, Q. & Zhao, L. Lightweight sea cucumber recognition network using improved yolov5. *IEEE Access***11**, 44787–44797 (2023).
60. Chen, X. & Gong, Z. Yolov5-lite: lighter, faster and easier to deploy. *Accessed: Sep***22** (2021).

61. Carion, N. et al. End-to-end object detection with transformers. In *European Conference on Computer Vision*, pp. 213–229 (Springer, 2020).

Acknowledgements
All authors disclosed no relevant relationships.

Author contributions
X.Y. conceived the experiment(s),X.Y. conducted the experiment(s),W.Z. and Y.W. and W.Q.Y. and Y.L analysed the results. All authors reviewed the manuscript.

Funding
This study was funded by Shandong Province Agricultural Machinery R&D, Manufacturing and Promotion Application Integration Pilot Project(Project No. NJYTHSD-202323) and Qilu University of Technology (Shandong Academy of Sciences) Science, Education and Industry Integration Pilot Project Major Innovative Projects(Project No. 2024ZDZX09).

Declarations

Competing interests
The authors declare no competing interests.

Additional information
Correspondence and requests for materials should be addressed to Y.L.

Reprints and permissions information is available at www.nature.com/reprints.

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