An Improved YOLO Algorithm for Kiwifruit Detection

Yi Xia, Minh Nguyen, Wei Qi Yan
Auckland University of Technology, 1010 New Zealand

ABSTRACT

Agriculture is paramount to food industry, with Kiwifruit significantly contributing to New Zealand's exports. Traditional Kiwifruit harvesting methods predominantly rely on human labour. In this book chapter, we address the challenges of fruits overlapping and lossing in Kiwifruit detection by proposing a fruit detection model that integrates an attention mechanism into the YOLOv8 framework. This integration enhances the model in detecting overlapping and small Kiwifruits. Additionally, we replace the YOLOv8 loss function with the Focal-EIoU loss function to mitigate the loss value oscillation caused by using low-quality samples and the incorrect enlargement of edge lengths. Based on the dataset we collected, our proposed method demonstrates superior detection accuracy, as evidenced by comparative and ablation experiments, achieving a mAP0.5@IoU up to 0.942. The results validate the effectiveness of our approach and provide valuable insights for implementing agricultural automation for Kiwifruit producers.

Keywords: Deep learning, YOLOv8, Attention mechanism, Object detection

INTRODUCTION

One of the most recognizable fruits is kiwifruit, which is now known around the world. However, the quick development has brought forth serious problems, such as a lack of workers that has cost money (Ferguson, 2004). To address these issues, it is essential to improve mechanization and establish an efficient workflow for activities such as picking, sorting, cleaning, and packaging to enhance efficiency within the current Kiwifruit industry. Modern artificial intelligence (AI), particularly deep learning, have been suggested as a means of boosting Kiwifruit yield estimation, improving picking effectiveness, and lowering labour expenses. The models must be accurate and effective in order to count kiwifruit (Massah et al., 2021). Traditional detection algorithms extract the physical features of the fruits and take use of image segmentation algorithms to detect the fruit region (Xiao et al., 2024) (Olaniyi et al., 2017). However, these algorithms are less reliable, and the elements like illumination and fruit placement have a significant impact on the outcomes of the detection, making them less useful when deployed in real-world orchard settings (Xia et al., 2023).

Deep learning (Goodfellow et al., 2017) (Yan, 2021), a subject belonging to machine learning, employs deep neural networks for modelling and addressing intricate problems. In the field of visual object detection, deep learning models were designed to identify and locate specific visual objects within a given image or video. These models were trained by using a dataset of labelled images, where the objects of interest are marked by using bounding boxes. Once the training process is completed, the model can be applied to new images or videos for the detection and localization of objects of interest. The deep learning models for visual object detection include YOLO (Bochkovskiy et al., 2020) (Ge et al., 2021) (Redmon et al., 2016), R-CNN (Girshick et al., 2014), and DETR (Carion et al., 2020) (Tang & Yan, 2024). These models have demonstrated marvellous performance in comparison to conventional object detection methods and have been utilized in a variety of domains including self-driving cars, surveillance systems, industry, and agriculture (Liu et al., 2022) (Liu et al., 2021) (Novelero & Dela, 2022).

Visual object detection in agriculture is a research topic in computer vision that aims to develop algorithms for automatic identification and location of objects of interest in images and videos captured in agricultural environments (Bazame et al., 2021). The use of Convolutional Neural Networks (CNNs) is a popular approach for visual object detection in agriculture and has been widely employed in a range of applications such as crop counting, plant disease detection and weed detection due to its effectiveness in object identification. Besides CNNs, other object detection methods such as YOLO and Faster R-CNN are also used in the field of agriculture and have proven to be effective in detecting objects from digital images and videos with high accuracy and fast processing speed (Wang et al., 2021). YOLO, a widely employed visual object detection algorithm, is well-known for its fast processing speed and high accuracy in detecting visual objects in images and videos, not just in agriculture but in various fields (Wang et al., 2021).

A proposed algorithm for improving visual object detection in the field of agriculture is the TIA-YOLOv5 model. In order to create synthetic images, this approach takes use of the method of pixel-level synthetization data augmentation, which entails adding new pixels into original photos. The TIA-YOLOv5 model contains a channel feature fusion with involution (CFFI) technique to successfully merge channel features with the least amount of information loss, as well as a transformer encoder block in its backbone to increase its sensitivity to weeds. To enhance the integration of features at various scales in the prediction head, the adaptive spatial feature fusion (ASFF) technique has also been included. Evaluations utilising a publicly available sugar beetroot dataset showed that the TIA-YOLOv5 network performed better than the YOLOv5 baseline model, with 90% mAP@0.5 (Wang et al., 2022). The YOLOMuskmelon algorithm is characterized by its accuracy and speed for fruit detection. The algorithm incorporates different backbone networks and loss functions, contributing to its high accuracy with an average score of 89.6%, as reported in (Lawal, 2021). YOLO has also been utilized in various agricultural applications beyond object detection, including fruit counting (Xia et al., 2023), seedling detection, and plant growth monitoring. YOLO is a widely used object detection algorithm in the field of agriculture and has demonstrated its effectiveness in a range of tasks, including fruit ripeness detection (Xiao et al., 2021), fruit freshness grading (Fu et al., 2022), and UAV detection (Novelero & Dela, 2022), with high accuracy and rapid processing speed (Wang & Yan, 2021).

Visual object detection in the field of agriculture is an active area of research that has seen significant progress in recent years. CNNs and other deep learning-based approaches have been found to be highly effective at identifying objects in images and videos that have been broadly employed in a variety of applications, including crop counting, plant disease detection, and weed detection. Other object detection approaches, such as YOLO and Faster R-CNN, have also proven to be effective in detecting objects in images and videos with high accuracy and fast processing speed, which may be useful in real-world applications in the field of agriculture. Despite early successes in the field of computer vision, the advancements in deep learning have enabled the improved performance of the YOLOv8 algorithm. The YOLOv8 algorithm can achieve more accurate results in a faster time frame, demonstrating its superiority over prior models.

In this book chapter, we propose a modified YOLOv8 algorithm for Kiwifruit detection that incorporates Convolutional Block Attention Module (CBAM) and Focal-EIoU loss function (Zhang et al., 2022). The integration of CBAM into YOLOv8 enhances the accuracy in detecting small, densely positioned, and multiple kiwifruits. By integrating CBAM into the backbone of the YOLOv8 network, the model assigns weights to both channel and spatial features in the feature map, increases its sensitivity to relevant features and reducing attention to irrelevant features to enhance accuracy. The Focal-EIoU loss function provides adjusted weight to challenging instances of difficult-to-detect objects, focusing on these challenging instances, thereby to prevent overfitting and improve overall detection performance. Additionally, we validate the performance improvement of our proposed model through comparative experiments and ablation experiments.

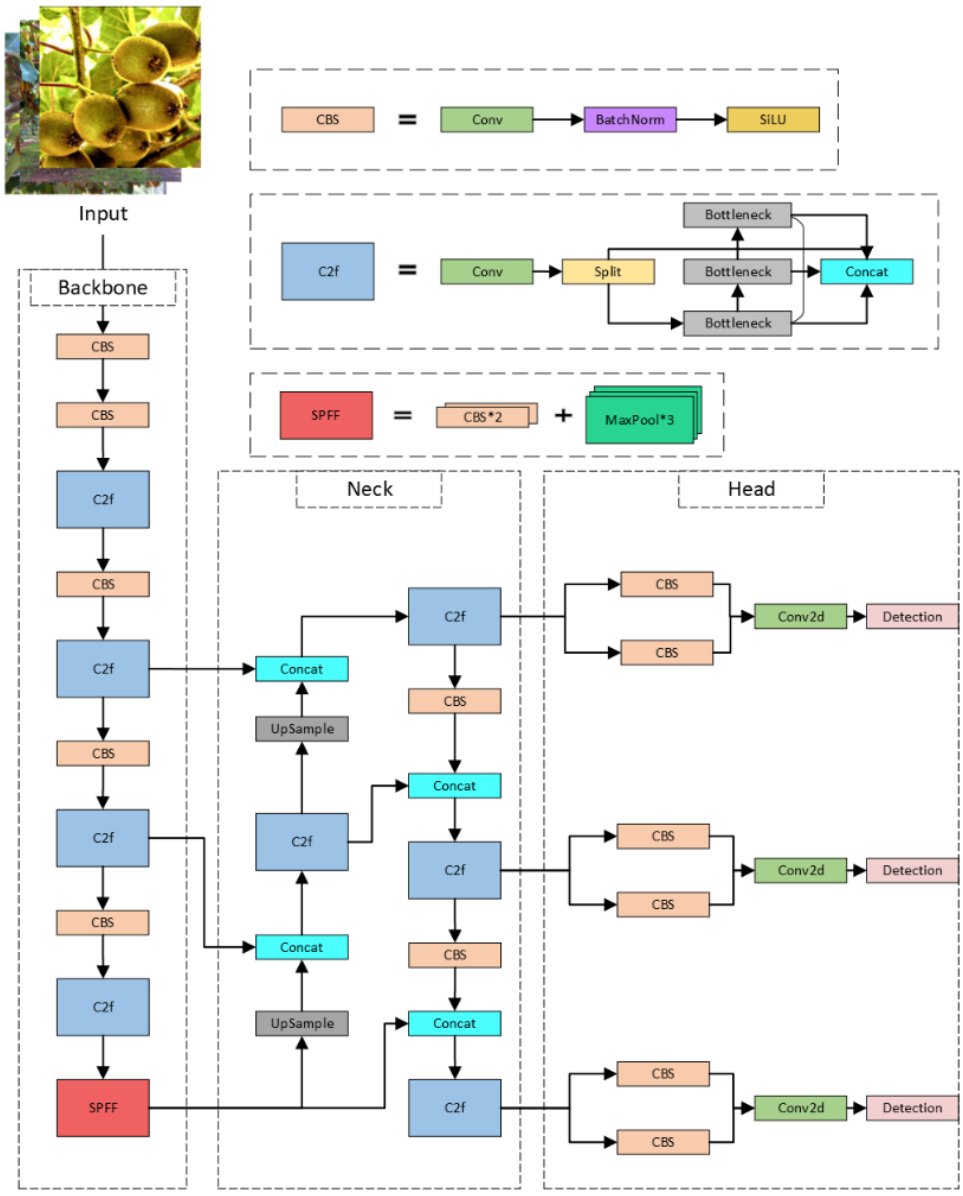

Figure 1. YOLOv8 neural network architecture.

The structure of this book chapter is listed as follows: A review of the relevant literature is given in Section 2, and Section 3 describes the methodology of our suggested approach. In Section 4, the outcomes of our research are discussed and presented. Section 5 brings the work to a close and outlines our goals for additional research.

RELATED WORK

YOLOv8 (You Only Look Once) is a state-of-the-art object detection algorithm. The algorithm is an improvement over the previous version of YOLOv7 and other object detection algorithms (Wang et al.,

2023), which aims to improve the accuracy and processing speed of object detection from digital images and videos. YOLOv8 algorithm is based on a single shot multibox detector (SSD) architecture, which allows the model to predict bounding boxes and class probabilities for visual objects in an image in a single forward pass. The algorithm also is use of anchor boxes and a concept of "mosaic data augmentation" to improve the ability of the model to detect visual objects of different scales. One of the key innovations of YOLOv8 is the use of a "scale-aware training" method, which allows the model to better handle visual objects having different sizes in an image. This is achieved by training the model on a diverse set of images, including images with visual objects of different scales by using a "mosaic data augmentation", which combines multiple images to form a single training image. In addition, YOLOv8 takes use of an efficient implementation of the architecture which allows it to process images at a higher frame rate and make it suitable for real-time applications. The YOLOv8 network architecture is shown in Figure 1. This version of YOLO takes advantage of a more complex network architecture than its predecessors, which allows the model to detect visual objects with more accuracy and generalization, and more classes. In this book chapter, YOLOv8 is harnessed to detect Kiwifruits in real-time. The scale-aware training method allows the model to better handle objects of different sizes in an image, the efficient implementation allows it to process images at a higher frame rate, making it suitable for real-time applications.

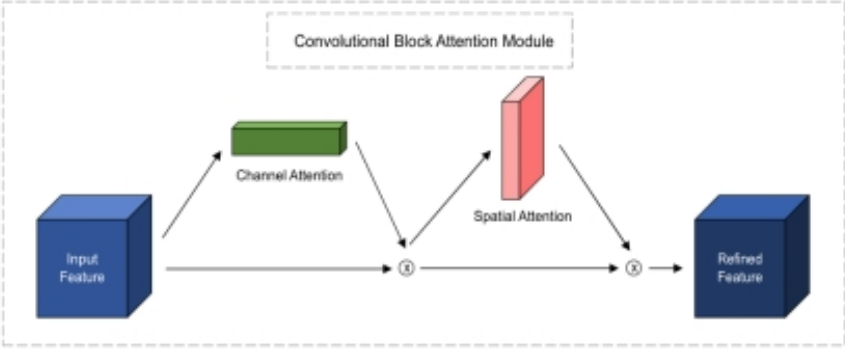

Figure 2. The main structure of CBAM module.

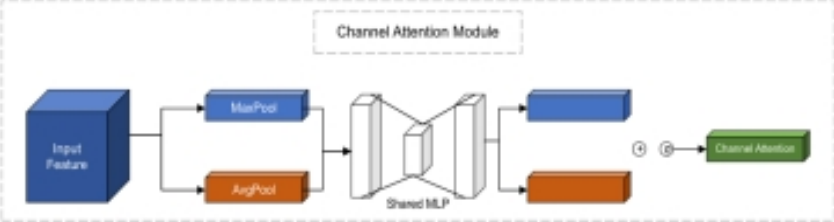

Figure 3. The structure of CA module.

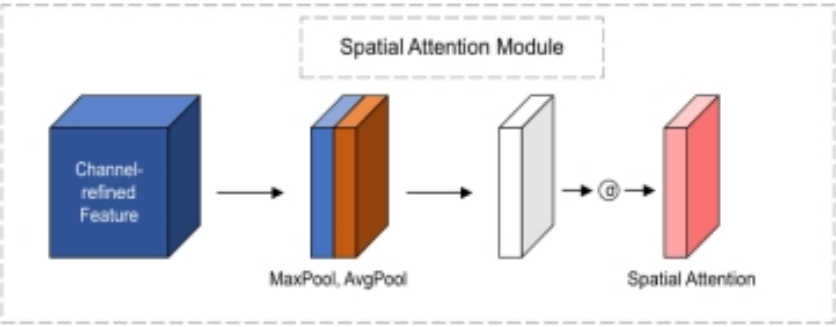

Figure 4. The structure of SA module.

Attention mechanisms have also been widely adopted in the field of computer vision. In this context, attention mechanisms provide a way for models to selectively focus on certain regions of an image when making predictions (Vaswani et al., 2017). This is particularly meaningful in cases where the image contains multiple visual objects or regions of interest that are relevant to the task at hand. Attention mechanisms in computer vision were implemented in a similar fashion related to NLP (Shan & Yan, 2021). They are often implemented as a layer within a neural network architecture and take in a set of feature maps as input.

The attention layer then produces a new set of output feature maps, which are a weighted sum of the input feature maps. The weights are determined by the similarity between the query and key vectors, with the most similar pairs receiving the highest weights. This allows the model to selectively focus on the most relevant regions of the image when making predictions. Attention mechanisms have been shown to improve the performance of computer vision models on a variety of tasks, such as visual object detection, image segmentation and image captioning. Additionally, attention mechanisms have been found to be useful for understanding the decisions made by neural network models, as the attention weights can provide insight into which regions of the image the model is using to make its predictions (Hu et al., 2018) (Wang et al., 2020).

A Convolution Block Attention Module (CBAM) is a specific type of attention mechanism that has been proposed for computer vision tasks (Woo et al., 2018). As shown in Figure 2, CBAM combines both channel-wise attention and spatial attention mechanisms to selectively focus on important regions of an image (Zhu et al., 2019). Figure 3 shows that the channel attention mechanism in CBAM is employed to emphasize on the most important channels in the feature maps. This is accomplished by taking the average value and the maximum value of each channel and then passing them through a fully connected layer to produce a weight for each channel. These weights are then harnessed to weight the channels in the feature maps. As shown in Fig. 4, the spatial attention mechanism in CBAM is employed to emphasize important regions of the image. This is completed by passing the feature maps through a convolutional layer to produce a 2D attention map. This attention map is then accommodated to weight the feature maps. By combining both channel-wise and spatial attention mechanisms, CBAM allows the model to selectively focus on both the most important channels and regions of the image. This can improve the performance of the model based on a variety of computer vision tasks, such as visual object detection, semantic segmentation, and image classification.

The essence of loss functions is to choose an appropriate function to measure the proximity between the output distribution of the model and the sample label distribution. It can adjust the weight parameters in the model, guiding the direction of learning in the convolutional neural network model. Therefore, the selection and design of the loss functions can reflect the characteristics of the current model.

YOLOv8 loss function takes use of a multitask loss that was designed to optimize the performance of the object detection model. YOLOv8 adopts Vari Focal (VFL) loss as the classification loss function, Dostronition Focal Loss (DFL) and CIOU loss as the regression loss function (Zheng et al., 2020).

Improving the performance of classifiers is a key part of optimizing detectors. Focal loss modifies the traditional cross-entropy loss to address the class imbalance between positive and negative samples or between difficult and easy samples. In order to address the inconsistent use of quality estimation and classification between training and inference, Quality Focal Loss (QFL) further extends Focal Loss with a joint representation of classification scores and localization quality for classification supervision (Zheng et al., 2020). As shown in equation (1), where p is the predicted value, l is the label, and α is the hyperparameter. VariFocal Loss (VFL) is derived from Focal Loss, but it treats positive and negative samples asymmetrically (Zhang et al., 2022). By considering positive and negative samples with different levels of importance, it balances the learning signals from both samples. Therefore, the YOLOv8 model makes use of VFL as a classification loss function.

$$VFL(p,l) = \begin{cases} -l(l\,log(p) + (1-l)\,log(1-p)) & l > 0 \\ -\alpha p^{\gamma}\,log(1-p) & l = 0 \end{cases} \quad (1)$$

$$\mathcal{L}_{CIoU} = 1 - IoU + \frac{\rho^2(b,b^{gt})}{c^2} + \alpha v \quad (2)$$

$$\alpha = \frac{v}{(1-IoU)+v,} \quad (3)$$

DFL reduces the underlying continuous distribution of box positions to a discretized probability distribution. It takes account of the ambiguity and uncertainty in the data without introducing any other prior, which helps to improve the box localization accuracy, especially if the boundaries of GT boxes are ambiguous. In YOLOv8 model, DFL mainly models the box location as a general distribution, which allows the network to quickly focus on the distribution of locations that are close to the target location and increase the probability of detecting the object. The CIoU loss function corresponds to three geometric parameters which are overlapping area, center point distance, and aspect ratio. α and v refer to the aspect ratio and the formula is shown in equation (2) and equation (3), where α is a positive trade-off parameter, v measures the consistency of aspect ratio (Zheng et al., 2020).

In the previously introduced CIoU loss function, the penalty term includes the distance and relative proportion of the bounding boxes (Zheng et al., 2020). Zhang et al. proposed the Focal-EIoU loss function to solve the problem of severe oscillation of loss values caused by low-quality samples (Zhang et al., 2022). This loss function differs from the CIoU loss function in YOLOv8, as the Focal-EIoU loss function directly adopts the side length as a penalty term.

The EIoU loss function is an improvement of the CIoU loss function, which addresses the issue of the penalty term becoming ineffective in certain situations where the predicted bounding box width and height satisfy specific conditions in the CIoU loss function. The definition of the EIoU loss function is given by Equation 4, where c_w and $c_h$are defined as the width and height of two rectangular anchor boxes. $\mathcal{L}_{IoU}$, $\mathcal{L}_{dis}$, and $\mathcal{L}_{asp}$ are the IoU loss, distance loss, and aspect ratio loss, respectively (Zheng et al., 2020).

The FocalL1 loss is a modified version of the focal loss that addresses the imbalanced problem in regression problems. In object detection, most of the predicted boxes based on anchor boxes have low Intersection over Union (IoU) values with the ground truth, leading to high fluctuations in loss values when training on such low-quality samples. The purpose of FocalL1 is to resolve the imbalance between high- and low-quality samples. By assigning a smaller gradient to low-quality samples, FocalL1 loss suppresses the impact of these samples (Zhang et al., 2022). As shown in Equation 5, FocalL1 loss calculates the regression loss by summing up the deviations of $x, y, w, and\ h$.

The Focal-EIoU loss function is a combination of the EIoU loss function and the FocalL1 loss function mentioned above (Zhang et al., 2022). As shown in equation (6), the hyperparameter γ is employed to control the curvature of the curve.

$$\mathcal{L}_{EIoU} = \mathcal{L}_{IoU} + \mathcal{L}_{dis} + \mathcal{L}_{asp} = 1 - IoU + \frac{\rho^2(\boldsymbol{b},\boldsymbol{b^{gt}})}{c^2} + \frac{\rho^2(w,w^{gt})}{c_w^2} + \frac{\rho^2(h,h^{gt})}{c_h^2} \quad (4)$$

$$\mathcal{L}_{FocalL1} = \sum_{i\in\{x,y,w,h\}} \mathcal{L}_{Focal}(\left|\boldsymbol{B}_i - \boldsymbol{B}_i^{gt}\right|) \quad (5)$$

$$\mathcal{L}_{Focal-EIoU} = IoU^{\gamma}\mathcal{L}_{EIoU} \quad (6)$$

METHODOLOGY

To the best of our knowledge, there is currently a lack of publicly available, labelled datasets for training Kiwifruit detection. In shed light on this, we sourced images and videos of real kiwifruit from the internet and segmented the videos into individual frames. The resulting dataset consisted of 1,500 images. To enhance the efficiency of model training, we manually removed images that did not feature kiwifruit and

duplicate images, resulting in a dataset of 1,224 images. We utilized the Roboflow tool to label the dataset, as well as to auto-orient the images. Additionally, to address the issue of variations in image size, we resized the original images to 640×640 pixels by using the Roboflow tool.

(a) Original *(b) Horizontal* *(c) vertical*

Figure 5. An example of original image and flip augmentation image.

Data augmentation is a crucial method that can significantly improve the performance of deep learning models. These models are sensitive to the pixel order of the features of the object being detected and may make incorrect detections if the object is mirrored or flipped in the image. Kiwifruit can appear in various orientations and scales in an image, making it challenging for the model to accurately detect and count them. In this book chapter, we apply geometric transformations to the dataset in order to avoid overfitting and non-convergence. As shown in Figure 5, image enhancement method primarily consists of random horizontal flipping and random vertical flipping. Random horizontal flipping refers to flipping the image about its vertical centre line, whereas random vertical flipping refers to flipping the image about its horizontal centre line. Through this method, the original dataset was augmented to 1,885 images. The augmented dataset was then divided into training, validation, and test dataset of 1516, 246, and 123 images respectively.

In this book chapter, we conduct experiments for a Kiwifruit detection task using YOLOv8, with a dataset collected and pre-processed manually. Unlike the original YOLOv8 model, we propose a method that firstly inserts a CBAM module into the main structure of YOLOv8 and then replaces the CIoU loss function in YOLOv8 with the Focal-EIoU loss function. The Convolution Block Attention Module assigns weights to channel and spatial features in the feature map, enables the model to focus on the target object and suppress attention to non-targets. The CBAM module is split into two parts: The channel attention module and the spatial attention module. As shown in Figure 6, the CBAM module is inserted to the convolution layer in the main network and the model inputs a 640×640 size image into the main network and outputs the prediction results to complete the object detection. The CIoU loss function shown in the YOLOv8 model is described in detail in the relevant work section, which considers not only the centre distance and overlapping region of the bounding box but also the aspect ratio of the bounding box. However, the difference in aspect ratio reflected in the function is not the true difference of the length and width of the anchor box and its confidence, while hindering the effective optimization of similarity. Additionally, to address the issue of sharp fluctuations in the loss value caused by low-quality samples, the Focal-EIoU loss function integrates the EIoU loss function and the FocalL1 loss function. Therefore, in this experiment, the CIoU loss function is replaced by using the better-performing Focal-EIoU loss function.

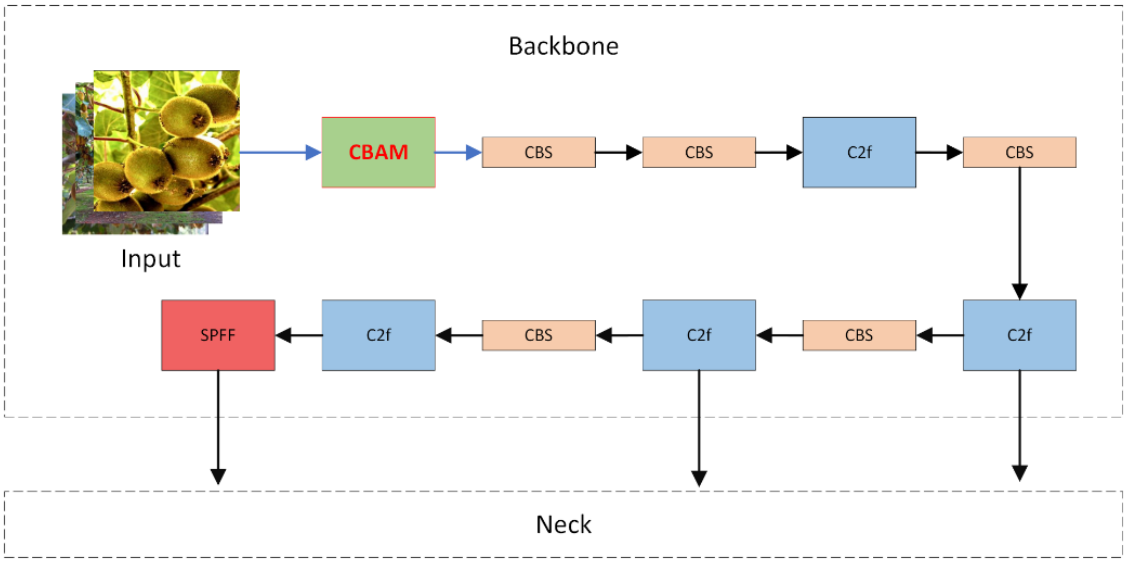

Figure 6. The position of the CBAM module inserted in the backbone of YOLOv8.

RESULT ANALYSIS

In this book chapter, the experimental environment is shown in Table 1.

Table 1. Training environment parameters

GPU Name	PyTorch	CUDA
Tesla T4 (15110MiB)	Version 1.13	Version 11.6

Precision, recall, and mean average precision (mAP) were employed as assessment measures in this book chapter to evaluate the efficacy of our proposed kiwifruit detection model. Precision is defined as the ratio of the actual positive samples to all positive samples in the projected samples, as stated in equation (7). Recall is defined as the ratio of the actual positive samples to all expected positives in the predicted samples, as stated in equation (8). Equation (9) demonstrates that mAP is the total of the average precision divided by all categories. mAP is identical to AP because there is only one type of kiwifruit discussed in this research project. When the IoU is set to 0.5 and the average mAP when the IoU ranges from 0.5 to 0.95 with a step size of 0.05, they are referred to as mAP@0.5 and mAP@0.95, respectively.

$$Precision = \frac{TP}{TP+FP} \tag{7}$$

$$Recall = \frac{TP}{TP+FN} \tag{8}$$

$$mAP = \frac{1}{N}\sum_{i=1}^{N}\int_{0}^{1} PdR \tag{9}$$

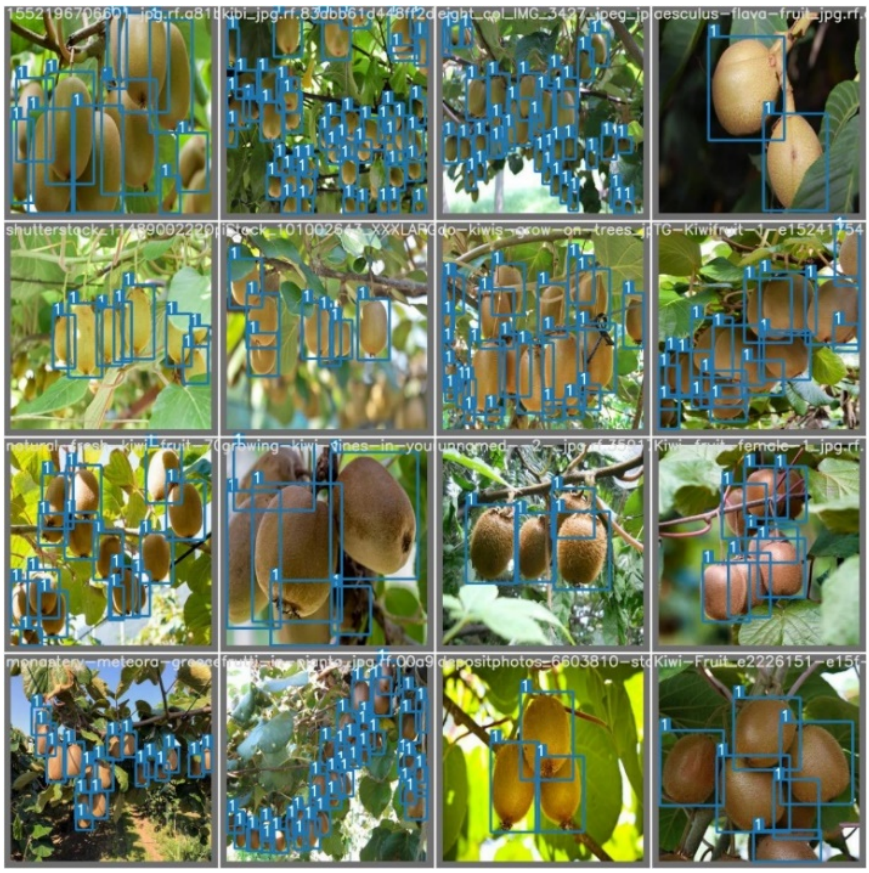

Figure 7. The prediction results of our proposed model.

Figure 7 depicts the detection capability of our proposed Kiwifruit model, indicating that the model is much proficient in detecting densely packed, overlapping, and small objects. We conducted comparison and ablation experiments to validate the superiority of our proposed model. Firstly, we validated the performance of different YOLO models on the dataset we collected by comparing their results. The comparison results are shown in Table 2, where the latest proposed YOLOv8 model outperforms other models under the same experimental conditions. Compared to the widely used YOLOv5 model and the previous YOLO v7 model, the accuracy of the v8 model improved by 2.9% and 0.4% respectively, and the mAP improved by 4.8% and 0.4% respectively when IoU is set to 0.5. This demonstrates that our selected baseline model has already acquired good performance and provides a foundation for subsequent improvement.

Furthermore, we validated the impact of two improvements on the model performance through ablation experiments. As shown in Table 3, we took use of YOLOv8 model as the baseline to verify the effectiveness of the improvements. The results of the experiments indicate that the addition of the CBAM module and the replacement of the Focal-EIoU loss function both have a positive effect on the performance of the model. Compared to the baseline, the addition of the CBAM module and the replacement of the Focal-EIoU loss function improved the mAP by 1.2% and 1.6% respectively when IoU is set to 0.5. The method of simultaneously inserting the CBAM module and replacing the Focal-EIoU loss function that we proposed achieved a more significant improvement. Compared to the baseline, the mAP improved by 2.1% when IoU is set to 0.5 and by 1.9% when IoU is set to 0.95. The results of the comparison experiments and ablation experiments indicate that the performance of our proposed method has been improved based on YOLOv8, thus it verifies the validity of the proposed method in this book chapter.

Table 2. Comparison results on our dataset

Method	Epoch	Size	Precision	Recall	mAP@0.5	mAP@0.95
YOLOv4	150	640	0.861	0.813	0.854	0.513
YOLOv5	150	640	0.892	0.833	0.873	0.585
YOLOv6	150	640	0.904	0.876	0.906	0.609
YOLOv7	150	640	0.917	0.897	0.917	0.649
YOLOv8	150	640	0.921	0.905	0.921	0.658
Ours	150	640	**0.932**	**0.909**	**0.942**	**0.677**

Table 3. Ablation studies on our dataset

CBAM	Focal-EIoU	Baseline	mAP@0.5	mAP@0.95
		√	0.921	0.658
√		√	0.933	0.661
	√	√	0.937	0.671
√	√	√	**0.942**	**0.677**

CONCLUSION

The automation of agriculture based on fruit detection can reduce the dependence on manual labor and decrease the cost of agricultural production. Efficient and accurate fruit detection algorithms lay the technical foundation for the popularization of automated harvesting robots. In this book chapter, we propose a YOLOv8-based method for detecting Kiwifruit in orchards. We combine the YOLOv8 model with the CBAM module to increase the model performance based on effective information, enhance the performance of the current YOLOv8 model. Furthermore, we replace the loss function in YOLOv8 with a more comprehensive Focal-EIoU loss function. We demonstrate through comparative experiments and ablation experiments that the improved YOLOv8 model is superior in Kiwifruit detection compared to the original YOLOv8. However, there is still much room for improvement in our work. Our future work includes reducing the size of the model and increasing the size of the kiwifruit database. Additionally, we will realize the tracking and counting of kiwifruit in the orchards through multi-object tracking, providing theoretical and technical support for future practical applications (Mi & Yan, 2024).

REFERENCES

Al-Sarayreha, M. (2020) Hyperspectral Imaging and Deep Learning for Food Safety. PhD Thesis. Auckland University of Technology, New Zealand.

Bazame, H. C., Molin, J. P., Althoff, D., & Martello, M. (2021). Detection, classification, and mapping of coffee fruits during harvest with computer vision. Computers and Electronics in Agriculture, 183, 106066.

Bochkovskiy, A., Wang, C. Y., & Liao, H. Y. M. (2020). Yolov4: Optimal speed and accuracy of object detection. arXiv preprint arXiv:2004.10934.

Carion, N., Massa, F., Synnaeve, G., Usunier, N., Kirillov, A., & Zagoruyko, S. (2020). End-to-end object detection with transformers. In European Conference on Computer Vision (pp. 213-229).

Ferguson, A. R. (2004). 1904—the year that kiwifruit (Actinidia deliciosa) came to New Zealand. New Zealand Journal of Crop and Horticultural Science, 32(1), 3-27.

Fu, Y., Nguyen, M., & Yan, W. (2022). Grading methods for fruit freshness based on deep learning. SN Computer Science, 3(4), 264.

Fu, Y. (2020) Fruit Freshness Grading Using Deep Learning. Master's Thesis. Auckland University of Technology, New Zealand.
Ge, Z., Liu, S., Wang, F., Li, Z., & Sun, J. (2021). YOLOx: Exceeding YOLO series in 2021. arXiv preprint arXiv:2107.08430.
Girshick, R., Donahue, J., Darrell, T., & Malik, J. (2014). Rich feature hierarchies for accurate object detection and semantic segmentation. In IEEE Conference on Computer Vision and Pattern Recognition (pp. 580-587).
Goodfellow, I., Bengio, Y., & Courville, A. (2016). Deep Learning. MIT press.
Hu, J., Shen, L., & Sun, G. (2018). Squeeze-and-excitation networks. In IEEE Conference on Computer Vision and Pattern Recognition (pp. 7132-7141).
Lawal, O. M. (2021). YOLOMuskMelon: quest for fruit detection speed and accuracy using deep learning. IEEE Access, 9, 15221-15227.
Liu, G., Hou, Z., Liu, H., Liu, J., Zhao, W., & Li, K. (2022). TomatoDet: Anchor-free detector for tomato detection. Frontiers in Plant Science, 13, 942875.
Liu, Y., Yang, G., Huang, Y., & Yin, Y. (2021). SE-Mask R-CNN: An improved Mask R-CNN for apple detection and segmentation. Journal of Intelligent & Fuzzy Systems, 6715-6725.
Massah, J., Vakilian, K. A., Shabanian, M., & Shariatmadari, S. M. (2021). Design, development, and performance evaluation of a robot for yield estimation of kiwifruit. Computers and Electronics in Agriculture, 185, 106132.
Mi, Z., & Yan, W. (2024). Strawberry ripeness detection using deep learning models. Big Data and Cognitive Computing (pp. 92).
Novelero, J. M., & Cruz, J. C. D. (2022). On-tree mature coconut fruit detection based on deep learning using UAV images. In IEEE International Conference on Cybernetics and Computational Intelligence (CyberneticsCom) (pp. 494-499).
Olaniyi, E. O., Oyedotun, O. K., & Adnan, K. (2017). Intelligent grading system for banana fruit using neural network arbitration. Journal of Food Process Engineering, 40(1).
Redmon, J., Divvala, S., Girshick, R., & Farhadi, A. (2016). You only look once: Unified, real-time object detection. In IEEE Conference on Computer Vision and Pattern Recognition (pp. 779-788).
Shan, T., & Yan, J. (2021). SCA-Net: A spatial and channel attention network for medical image segmentation. IEEE Access, 9, 160926-160937.
Tang, S., & Yan, W. (2024). Utilizing RT-DETR model for fruit calorie estimation from digital images. Information (pp. 469).
Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., ... & Polosukhin, I. (2017). Attention is all you need. Advances in Neural Information Processing Systems, 30.
Wang, A., Peng, T., Cao, H., Xu, Y., Wei, X., & Cui, B. (2022). TIA-YOLOv5: An improved YOLOv5 network for real-time detection of crop and weed in the field. Frontiers in Plant Science, 13, 1091655.
Wang, C. Y., Bochkovskiy, A., & Liao, H. Y. M. (2023). YOLOv7: Trainable bag-of-freebies sets new state-of-the-art for real-time object detectors. In IEEE/CVF Conference on Computer Vision and Pattern Recognition (pp. 7464-7475).
Wang, L., & Yan, W. Q. (2021). Tree leaves detection based on deep learning. In International Symposium on Geometry and Vision (pp. 26-38). Springer International Publishing.
Wang, Q., Wu, B., Zhu, P., Li, P., Zuo, W., & Hu, Q. (2020). ECA-Net: Efficient channel attention for deep convolutional neural networks. In IEEE/CVF Conference on Computer Vision and Pattern Recognition (pp. 11534-11542).

Woo, S., Park, J., Lee, J. Y., & Kweon, I. S. (2018). CBAM: Convolutional block attention module. In European Conference on Computer Vision (ECCV) (pp. 3-19).
Xiao, B., Nguyen, M., & Yan, W. Q. (2021). Apple ripeness identification using deep learning. In International Symposium on Geometry and Vision (pp. 53-67). Springer International Publishing.
Xiao, B., Nguyen, M., & Yan, W. Q. (2023). Fruit ripeness identification using transformers. Applied Intelligence (pp. 22488-22499). Springer International Publishing.
Xiao, B., Nguyen, M., & Yan, W. Q. (2024). Apple ripeness identification from digital images using transformers. Multimedia Tools and Applications (pp. 7811-7825).
Xiao, B., Nguyen, M., & Yan, W. Q. (2024). Fruit ripeness identification using YOLOv8 model. Multimedia Tools and Applications (pp. 28039-28056).
Xia, Y., Nguyen, M., Lutui, R., & Yan, W. Q. (2023). Multiscale Kiwifruit detection from digital images. In Pacific-Rim Symposium on Image and Video Technology (pp. 82-95).
Xia, Y., Nguyen, M., & Yan, W. Q. (2022). A real-time kiwifruit detection based on improved YOLOv7. In International Conference on Image and Vision Computing New Zealand (pp. 48-61).
Xia, Y., Nguyen, M., & Yan, W. Q. (2023). Kiwifruit counting using KiwiDetector and KiwiTracker. In SAI Intelligent Systems Conference (pp. 629-640).
Xue, Y. Yan, W. (2023) YOLO models for fresh fruit classification from digital videos. Handbook of Research on AI and ML for Intelligent Machines and Systems, pp. 421-435, Chapter 17, IGI Global.
Yan, W. (2023). Computational Methods for Deep Learning. Springer.
Yan, W. (2019). Introduction to Intelligent Surveillance. Springer.
Yang, X., Zhao, W., Wang, Y., Yan, W., Li, Y. Lightweight and efficient deep learning models for fruit detection in orchards. Scientific Reports 14, 26086
Zhang, Y. F., Ren, W., Zhang, Z., Jia, Z., Wang, L., & Tan, T. (2022). Focal and efficient IOU loss for accurate bounding box regression. Neurocomputing, 506, 146-157.
Zhao, K., Yan, W. (2021) Fruit detection from digital images using CenterNet. International Symposium on Geometry and Vision.
Zhao, K., Nguyen, M., Yan, W. (2024) Evaluating accuracy and efficiency of fruit image generation using generative AI diffusion models for agricultural robotics. IEEE IVCNZ'24
Zheng, Z., Wang, P., Liu, W., Li, J., Ye, R., & Ren, D. (2020). Distance-IoU loss: Faster and better learning for bounding box regression. In AAAI Conference on Artificial Intelligence (Vol. 34, No. 07, pp. 12993-13000).
Zhu, X., Cheng, D., Zhang, Z., Lin, S., & Dai, J. (2019). An empirical study of spatial attention mechanisms in deep networks. In IEEE/CVF International Conference on Computer Vision (pp. 6688-6697).