Multiscale Kiwifruit Detection from Digital Images

Yi Xia, Minh Nguyen, Raymond Lutui, Wei Qi Yan

Auckland University of Technology, 1010 Auckland, New Zealand

Abstract. In this paper, we propose an improved YOLOv8-based Kiwifruit detection method using Swin Transformer, aiming to address challenges posed by significant scale variation and inaccuracies in multiscale object detection. Specifically, our approach embeds the encoder from Swin Transformer, based on its sliding-window design, into the YOLOv8 architecture to capture contextual information and global dependencies of the detected objects at multiple scales, facilitating the learning of semantic features. Through comparative experiments with the state-of-the-art object detection algorithms on our collected dataset, our proposed method demonstrates efficient detection of objects at different scales, significantly reducing false negatives while im-proving precision. Moreover, the method proves to be versatile in detecting objects of various sizes in different environmental settings, fulfilling the real-time requirements in complex and unknown Kiwifruit cultivation scenarios. The results highlight the potential practical applications of the pro-posed approach in Kiwifruit industry, showcasing its suitability for addressing real-world challenges and complexities.

Keywords: Object detection · Transformer · Multiscale object detection · YOLOv8.

1 Introduction

The detection of visual objects from digital images is a fundamental and challenging problem in computer vision, with numerous applications ranging from autonomous driving to precision agriculture [18]. In particular, accurate and efficient detection of Kiwifruits from multiscale images is of utmost value for Kiwifruit industry, enabling better monitoring, assessment, and management of Kiwifruit cultivation processes [15]. However, this task poses salient challenges, including significant scale variation, occlusion, and the presence of complex backgrounds, which can lead to inaccuracies and increased computational requirements in traditional object detection methods [3, 27].

To overcome these challenges, in this paper we propose an advanced Kiwifruit detection method that combines the strengths of Swin Transformer and YOLOv8 frameworks together. Swin Transformer has demonstrated the state-of-the-art performance in various computer vision tasks, particularly in capturing long-range dependencies and contextual information within given images [12]. Meanwhile, YOLOv8, an evolution of the popular YOLO (You Only Look Once) architecture, is renowned for its capabilities and versatility of visual object detection in real time [19]. By integrating Swin Transformer into YOLOv8, we aim to enhance the model ability to accurately and

efficiently detect Kiwifruits across a variety of scales and complex environments. The main contributions of this research are:

- Enhanced multiscale detection. Swin Transformer augmented YOLOv8 effectively handles the challenge of multiscale Kiwifruit detection. By leveraging the hierarchical transformer architecture, the model efficiently captures context and dependencies across various image scales, enabling it to detect Kiwifruits accurately regardless of their size.
- Improved precision and recall. The proposed method reduces false negatives and false positives, achieving a more balanced precision-recall trade-off. The integration of Swin Transformer enhances the model's understanding of complex scenes, leading to more reliable and precise Kiwifruit detection results.
- Versatility in real world settings. Our approach demonstrates strong adaptability to diverse environmental conditions, making it well-suited for real-time Kiwifruit detection in complex and unknown cultivation scenarios. This feature is crucial for practical applications in Kiwifruit industry, where unpredictable conditions are prevalent.

To evaluate the effectiveness of our proposed method, we have collected a comprehensive dataset of multiscale Kiwifruit images from geographical locations and cultivation setups. We conduct extensive experiments and compare the performance of our approach against the state-of-the-art object detection methods, including traditional YOLOv8 and other transformer-based models.

The rest of this paper is organized as follows: We provide a detailed review of related work in object detection, transformer-based models, and their applications in agricultural contexts in Section 2. In Section 3, we present the methodology, explaining the integration of Swin Transformer and YOLOv8 for multiscale Kiwifruit detection. In Section 4, we describe experimental setup, dataset, and evaluation metrics to assess the proposed method and the results of the experimental outcomes, demonstrating the superiority of our approach over existing methods. Finally, in Section 5, we conclude the paper and discuss the contributions, limitations, and future directions for research in this area.

2 Literature Review

Fruit detection is a crucial task in precision agriculture and automated harvesting systems [22, 23]. Accurate and efficient fruit detection is essential for yield estimation, crop monitoring, and fruit quality assessment. In recent years, with the advent of deep learning, a number of approaches have been proposed for fruit detection using Convolutional Neural Networks (CNNs), YOLO model, and transformer-based models [11, 24, 25]. Fruit detection using deep learning methods has shown promising results. CNNs have been widely employed for visual object detection tasks, including fruit detection. The YOLO model, known for its real-time object detection capabilities, has been adapted for fruit detection tasks. YOLO models have been extended to improve its accuracy and handle multiscale fruit detection. Transformer-based models,

originally developed for natural language processing, have also been explored for fruit detection [4].

2.1 Traditional CNN Models

Traditional CNN models have served as the backbone for a slew of deep learning-based object detection tasks, including fruit detection [11, 13]. AlexNet, VGG, and ResNet are among the most influential CNN architectures that have significantly contributed to advancements in computer vision [4, 5].

AlexNet was one of the pioneering CNN architectures that achieved a breakthrough in the ImageNet competition. It comprises multiple convolutional and pooling layers, followed by fully connected layers. AlexNet's success motivated the widespread adoption of CNNs in various vision tasks, including fruit detection [8].

VGG employs a deep network with small (3×3) convolutional filters. The use of smaller filters enables a deeper exploration of spatial information, leading to improve feature representation. VGG has been broadly applied to fruit detection tasks, achieved high accuracy in identifying various fruit categories.

ResNet introduced the concept of residual connections to address the vanishing gradient problem in very deep networks [7]. By introducing skip connections that enable the direct flow of gradients, ResNet allowed training significantly deeper networks. In fruit detection applications, ResNet as a feature extractor has achieved competitive performance.

2.2 YOLO Model

The initial version of YOLO models, YOLOv1, provided real-time object detection but faced limitations in detecting small objects, such as tiny fruits, due to its single scale approach. Subsequent versions (YOLOv2 and YOLOv3) addressed these limitations by introducing improvements such as anchor boxes, multiscale detection, and feature extraction across different network layers [19]. Since then, a series of versions of YOLO have been proposed, including YOLOv2, YOLOv5, YOLOX, YOLOv7 and YOLOv8, through the modifications in network architecture and the addition of data augmentation modules. Currently, YOLO has been successfully applied to fruit detection tasks, achieved real-time and accurate fruit detection results in various agricultural settings [27]. The versatility and real-time capabilities of YOLO models in handling multiscale objects have made it a popular choice for fruit detection in automated agricultural systems [16, 20].

2.3 Transformer-Based Models

The overall framework of Transformer-based image detection consists of three main components. Firstly, the input image undergoes visual feature extraction using a CNN backbone network, such as VGG, ResNet, or others. Subsequently, visual features are encoded and decoded by using Transformer architecture, which includes multi-head self-attention and encoder-decoder attention mechanisms. Finally, the object classes and bounding boxes are predicted using a feed-forward network [1].

DEtection Transformer (DETR) is a pioneering object detection method that adopts the Transformer architecture. It incorporates a CNN backbone network, Transformer

encoder-decoder structure, and a feedforward network (FFN). The CNN backbone extracts image features, which are then transformed into one-dimensional feature maps and processed by Transformer encoder. However, DETR has limitations in terms of slow training convergence, high computational complexity, and relatively poor performance in detecting small objects. In response, Deformable DETR was proposed by utilizing the deformable attention module to improve small object detection and training efficiency [21].

YOLOS is another series of widely applied object detection models based on Vision Transformer (ViT) architecture [2]. YOLO replaces the image classifier in ViT with a bipartite matching loss, enabling it to handle arbitrary-sized object detection tasks without requiring precise spatial or geometric structures [10]. YOLO stands for its adaptability to different Transformer structures, offering flexibility in object detection tasks.

Swin Transformer is a novel approach that leverages the Transformer architecture for computer vision tasks. It has gained attention in image segmentation and visual object detection domains. Swin Transformer takes use of shifted window-based self-attention effectively reduces computational complexity while maintaining desirable performance and making it advantageous for dense prediction tasks downstream. However, further enhance of the object detection is due to the occluded objects.

3 Methodology

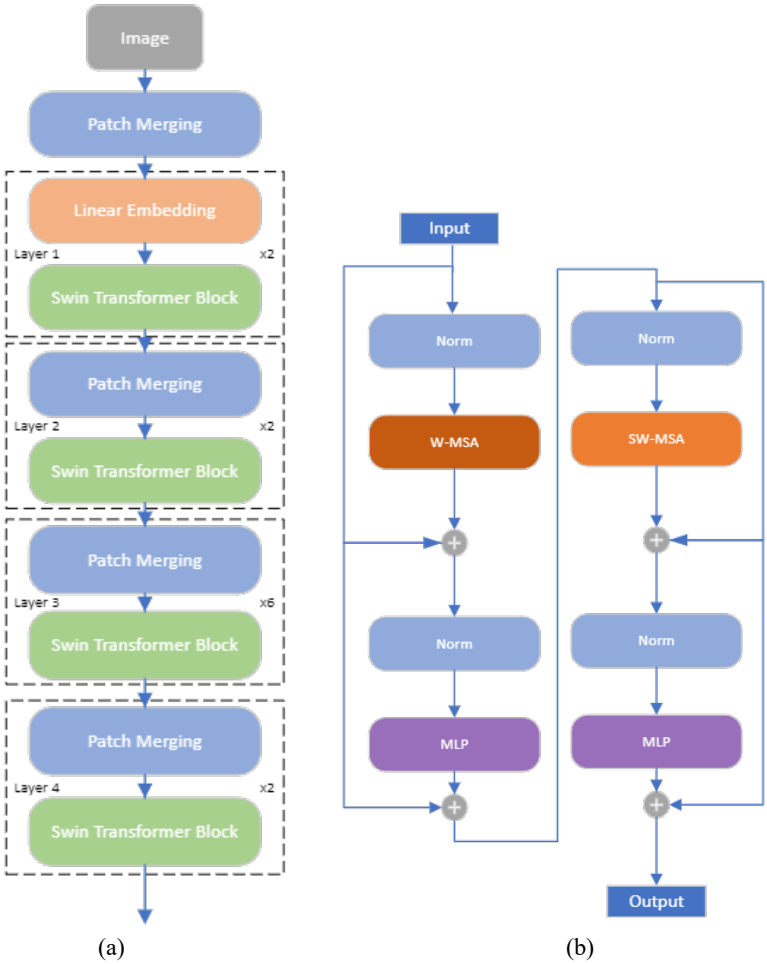
Swin Transformer model exhibits the capability to interact between local and global information across different network layers, thereby extracting hierarchical features. However, this model comes with a drawback of having a large number of parameters and high sensitivity, leading to high computational demands and training complexity. On the other hand, YOLOv8 model offers the advantage of having a smaller number of model parameters, resulting in faster training speed. However, its feature extraction ability is relatively weaker compared to Swin Transformer model.

In light of these considerations, we propose a novel approach that combines the feature extraction strengths of Swin Transformer with the practicality of YOLOv8, aiming to enhance the feature extraction capability of YOLOv8 and improve the accuracy and speed of multiscale object detection. By leveraging the advantages of both models, we aim to address the challenges posed by real-world scenarios involving multiscale targets, particularly in the context of Kiwifruit detection. This model is tailored to meet the demands of diverse target scales encountered in real-world settings, striking a balance between feature extraction efficiency and detection performance.

3.1 Swin Transformer

In Swin Transformer, Microsoft proposes Transformer as a versatile backbone for computer vision tasks, attracting significant attention in various domains such as image segmentation and object detection. The overall structure of Swin Transformer is depicted in Fig. 1 [13]. Similar to the hierarchical structure of the feature pyramid, Swin-Transformer model [21] is a multiscale fusion-based Transformer model that extracts features at different scales using a design with non-overlapping movable windows,

enabling cross-window connections for information interaction between local and global features. As shown in Fig. 1(a), Swin-Transformer encoder comprises a patch partition module and four consecutive stages. Each stage includes two types of attention modules: Window Multi-Head Self-Attention (W-MSA) module and Shifted Window Multi-Head Self-Attention (SW-MSA) module. The W-MSA module divides the feature map into non-overlapping windows and employs multi-head self-attention mechanism (MSA) to compute attention scores for each individual window.



(a) (b)

Fig. 1. The architecture of Swin Transformer

However, the W-MSA module lacks global correlation among windows. To address this, the SW-MSA module modifies the window partitioning by cyclically shifting windows through Shift Window, thus fusing features from multiple windows while preserving the relative positional relationship using a Mask mechanism to incorporate context information at different scales. In Fig. 1(b), alternating use of W-MSA and SW-

MSA modules in each stage combines hierarchical local attention with global self-attention mechanism, resulting in features at different levels. Stage 1 contains a Linear Embedding layer that linearly transforms the channel dimension of each pixel, mapping it to C dimensions. The remaining stages utilize the patch merging layer for downsampling and merging information from multiple windows. As a result, Swin Transformer exhibits excellent scalability, making it well-suited for handling objects of different scales and dense targets.

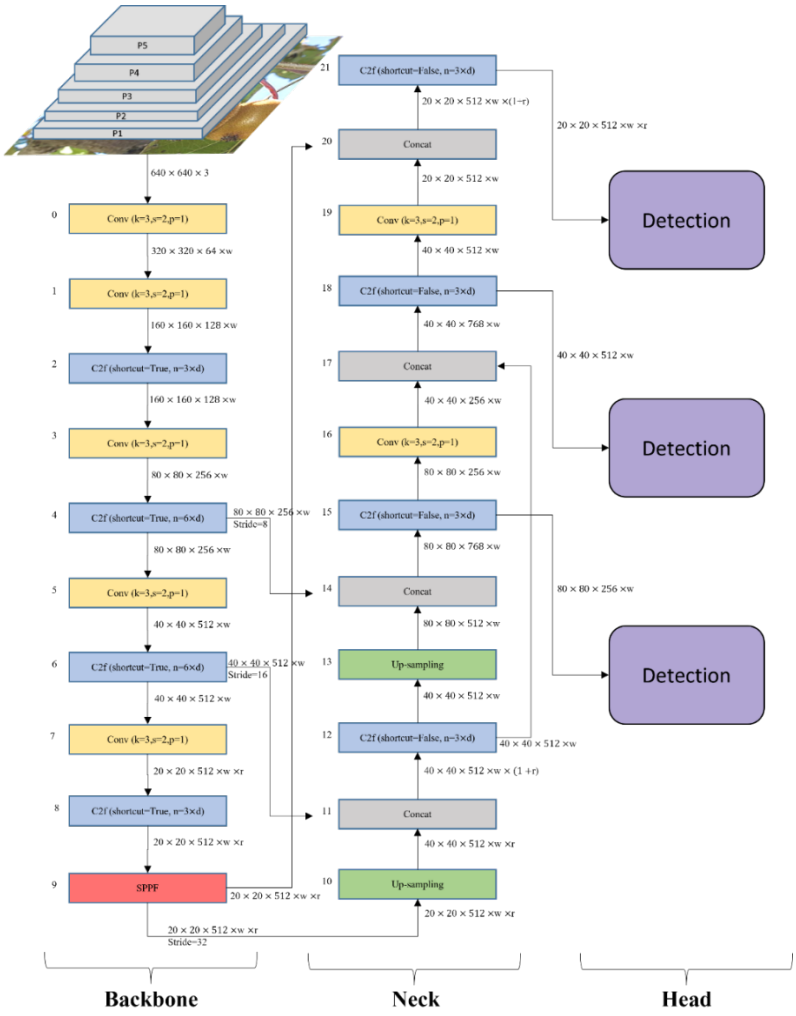


Fig. 2. The architecture of YOLOv8 model

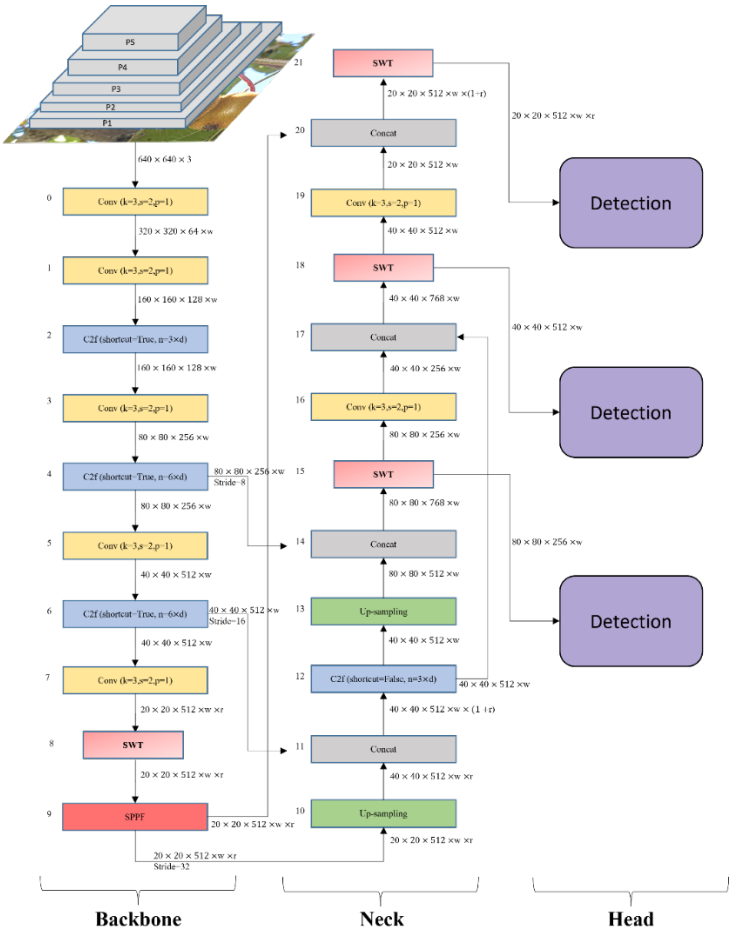


Fig. 3. The architecture of enhanced Kiwifruit detection model

3.2 YOLOv8

YOLOv8 architecture is composed of four main components, as illustrated in Fig. 1, including Input, Backbone, Neck, and Detection Head [20]. The input images undergo a series of data augmentation processes such as cropping, adaptive scaling, and Mosaic, before being fed into the Backbone. The Backbone is responsible for extracting features from the preprocessed images and generates three sets of feature maps at different scales, which are then forwarded to the Neck for further processing. In contrast to YOLOv5, the Neck of YOLOv8 replaces the CSP module with the C2f module and directly incorporates the feature outputs from different stages of the Backbone through upsampling operations. The prediction section decouples the tasks of object classification and bounding box regression, conducting separate predictions. The three detection heads are designed to handle objects of different sizes, thus accelerating model convergence speed and improving detection accuracy.

The classification loss adopts the VFL Loss, characterized by asymmetric positive-negative sample weighting to emphasize on positive samples as primary instances, effectively addressing the issue of class imbalance. For regression loss, the DFL module is introduced to enable the network to quickly focus on the distribution of positions close to the target locations. Notably, YOLOv8 demonstrates exceptional performance in terms of speed and accuracy, surpassing current state-of-the-art object detectors. The ability of this model achieved a balance between computational efficiency and detection precision positions as a promising solution for various practical applications in the field of object detection.

3.3 Enhanced Kiwifruit Detection Model

As illustrated in Fig. 3, to address the issue of imprecise multiscale object detection caused by the semantic information of Convolutional Neural Networks (CNNs) in real-world Kiwifruit images, we have integrated Swin Transformer model into the YOLOv8 backbone network, specifically replacing the top-level C2f module with the Swin-Transformer module. This modification allows us to perform global pixel-level operations on the low-resolution feature maps extracted by C2f. Thus, we can leverage the advantages of the self-attention mechanism while effectively reducing computational complexity and conserving memory space [22].

Furthermore, we have incorporated Swin Transformer module into the Neck structure to capture correlations and importance across different regions. This enhancement contributes to improve the adaptability of this model to various object sizes, enhance visual object detection accuracy, and achieve a better balance between speed and precision under parallel computation. The backbone network of this improved Kiwifruit detection model demonstrates robust modeling capabilities for capturing context information related to target backgrounds, edge shapes, and other contextual factors. These capabilities effectively guide downstream tasks of classification and localization based on semantic information. Moreover, the enhanced model exhibits superior scalability and practical applicability.

By adopting Swin Transformer model and integrating it into the YOLOv8 backbone network, our enhanced Kiwifruit detection model achieves much precise and efficient multiscale object detection [14]. The effective combination of self-attention mechanisms and computational optimizations results in a robust and efficient detection framework. The ability of this model to leverage semantic information for downstream tasks makes it well-suited for real-world Kiwifruit detection scenarios. The integration of Swin Transformer into YOLOv8 framework represents a novel and promising approach for achieving better detection performance and maintaining a balance between speed and accuracy, thus advancing the state-of-the-art in Kiwifruit detection [6].

4 Results

4.1 Dataset and Evaluation Metrics

We collected a comprehensive dataset of Kiwifruit images to conduct our experiments. The dataset was obtained by downloading Kiwifruit orchard videos from the internet

and segmenting them into individual frames. Additionally, we sourced Kiwifruit images from various online platforms to enhance the dataset's robustness. The dataset comprises of 3,000 original Kiwifruit images, gathered from diverse sources, including videos from different orchards, images at different ripeness stages, and images with varying sizes due to different camera distances. Our aim was to cover a wide range of Kiwifruit object scales to meet the requirements of multiscale Kiwifruit detection in real-world scenarios. To ensure data quality and consistency, the dataset underwent rigorous data cleaning procedures. We employed the Roboflow tool for efficient data labeling and ensure accurate object detection annotations for each image. Augmentation techniques, including mirror flipping and horizontal/vertical axis flipping, were applied to augment the dataset and enhance the generalization capability of this model. The final dataset consisted of 3,700 training images, 1,057 testing images, and 530 validation images. The diversity of this training set enabled the model to learn robust features across various scenarios, while the testing and validation sets served as critical benchmarks to evaluate the generalization performance on previously unseen data effectively.

Fig. 4. Dataset of Kiwifruit images at multiple scales

In this paper, the evaluation criterion of this model is the mean Average Precision (mAP). For a specific class of objects, its detection accuracy can be obtained from the Precision-Recall (PR) curve, where Precision (P) represents the probability of a positive prediction being correct, and Recall (R) shows the probability of correctly identifying positive samples [17]. The calculations are defined as follows:

$$Precision = \frac{TP}{TP + FP} = \frac{TP}{Total\ Positive\ Results} \tag{1}$$

$$Recall = \frac{TP}{TP + FN} = \frac{TP}{Total\ Ground\ Truths} \tag{2}$$

$$AP_i = \int_0^1 p(r)dr \tag{3}$$

$$mAP = \frac{1}{N}\sum_{i=1}^{N} AP_i \tag{4}$$

where *TP* denotes the number of true positives (actual positive samples correctly predicted as positive), *FP* shows the number of false positives (actual negative samples incorrectly predicted as positive), and *FN* indicates the number of false negatives (actual positive samples incorrectly predicted as negative).

Since Precision and Recall are measured on different dimensions, Average Precision (*AP*) is introduced, which represents the average precision values at different recall levels. A higher AP indicates fewer detection errors. The mAP is obtained by taking the average of *AP* values across all class categories, providing an overall measure of the model's detection performance.

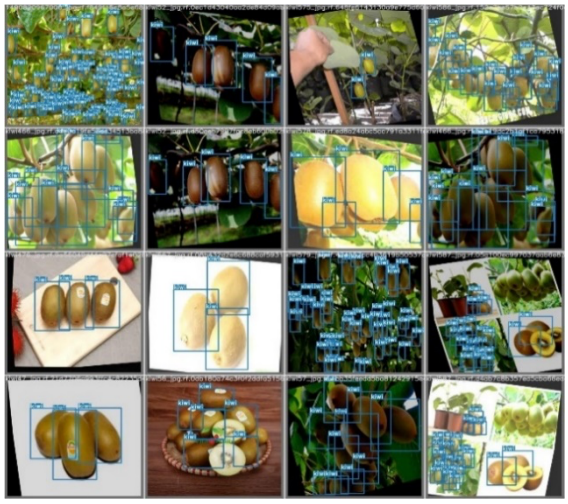


Fig. 5. Visual results of multiscale Kiwifruit detection model

4.2 Experimental Parameter Configuration

The experiments in this paper were conducted on a system running Windows 11 Operating System, equipped with a Geforce RTX 3060 GPU, an AMD R7-5800 CPU, and 16 GB of RAM. The CUDA version used was 12.0. The experiment was conducted using Python 3.8 and PyTorch 2.0, a deep learning framework, to build the model.

The training process involved setting the number of epochs to 150, indicating the number of complete iterations over the entire training dataset. The batch size was configured to 16, defining the number of samples processed in each iteration. The stochastic gradient descent (SGD) optimization algorithm was adopted for weight updates during training, with a weight decay of 0.0005 applied to regulate the magnitude of weight updates and prevent overfitting.

4.3 Analysis of Experimental Results

In Fig. 5, we present visual representation of the prediction results obtained from our proposed multiscale Kiwifruit detection model, which was trained on the custom dataset specifically created for this study. The displayed images vividly demonstrate the robust and accurate detection capabilities of our model, as it successfully identifies Kiwifruits having a diversity of colors, sizes, and shapes. The model exhibits a remarkable ability to discern Kiwifruits from cluttered backgrounds and handle variations in appearance, enabling it to effectively adapt to real-world scenarios. The obtained results highlight the efficacy of our novel model in addressing the challenges posed by multiscale Kiwifruit detection, thereby reaffirming the practical relevance of our model for applications in automated agricultural systems and other computer vision tasks.

Table 1. The comparison of various object detection models on our dataset

Model	Epoch	Size	Precision	Recall	mAP@0.5	mAP@0.5:0.95
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
CornerNet	150	640	0.772	0.694	0.764	0.462
DETR	150	640	0.671	0.625	0.619	0.415
Swin Transformer	150	640	0.836	0.794	0.809	0.511
Our proposed (SWT+YOLOv8)	150	640	0.951	0.932	0.947	0.712

The comparison of visual object detection models on our dataset is presented in Table 1. We evaluated the state-of-the-art models, including YOLOv4, YOLOv5, YOLOv6, YOLOv7, YOLOv8, CornerNet, DETR, and Swin Transformer, along with our proposed model (SWT+YOLOv8) [9]. Among all models, our proposed model demonstrates the best overall performance, achieving an impressive precision of 0.951, recall of 0.932, mAP@0.5 of 0.947, and mAP@[.5:.95] of 0.712. Furthermore, compared with the performance of our proposed model with other models, it is evident that fusion of Swin Transformer with YOLOv8 results in a substantial enhancement in both precision and recall. The mAP scores also show a significant boost, indicating the effectiveness of our proposed approach in handling multiscale object detection tasks.

To further demonstrate the effectiveness of Swin Transformer in our proposed Kiwifruit detection model, we conducted a series of ablation experiments. In these experiments, we investigated the impact of integrating SWT at a scale of components of the model: 1) Adding SWT to the backbone; 2) Adding SWT to the neck; 3) Adding SWT to both the backbone and neck. We set the YOLOv8 model as the baseline model for conducting ablation experiments as shown in Table 2.

For the first ablation experiment (1), we incorporated SWT into the backbone of the YOLOv8 model. The results show that the addition of SWT to the backbone significantly improved the model's performance. Specifically, the precision, recall, mAP@0.5, and mAP@[.5:.95] scores all demonstrated notable enhancements, validating the effectiveness of SWT in enhancing the feature extraction capability at the backbone level.

In ablation experiment, we introduced SWT to the neck component of YOLOv8 model. Similar to the first experiment, this integration also led to remarkable improvements in the detection results. The precision, recall, mAP@0.5, and mAP@[.5:.95] scores all exhibited substantial increases, which affirm the positive impact of SWT in refining the feature fusion process at the neck level.

Lastly, in the third ablation experiment, we simultaneously added SWT to both the backbone and neck of the YOLOv8 model. This comprehensive integration of SWT at both levels further boosted the model's performance, resulting in the highest precision, recall, mAP@0.5, and mAP@0.5:0.95 scores among all the ablation settings. The combined effect of SWT in both Backbone and Neck demonstrated its complementary nature, which led to superior detection results.

Table 2. The ablation experiments of Kiwifruit detection models based on our dataset

Model	Epoch	Size	Precision	Recall	mAP@0.5	mAP@0.5:0.95
Baseline	150	640	0.921	0.905	0.921	0.658
(1)	150	640	0.928	0.911	0.924	0.673
(2)	150	640	0.939	0.916	0.940	0.689
(3)	150	640	0.951	0.932	0.947	0.712

From the ablation experiments presented in Table 2, it can be observed that inserting Swin Transformer modules at different locations within the YOLOv8 model has varying degrees of impact on its performance. After Swin-Transformer modules are inserted in the Backbone, there is a slight improvement in the performance of our proposed model. However, if Swin Transformer modules are inserted in the Neck, the mAP shows a significant improvement compared to the baseline. Notably, if Swin Transformer modules are inserted at multiple positions, such as in both the Backbone and Neck, the model exhibits further improvement in all evaluated metrics.

4.4 Discussion

In this paper, we presented a novel approach for multiscale Kiwifruit detection by combining the strengths of Swin Transformer and YOLOv8 models. The integration of Swin Transformer and YOLOv8 model is proved to be a powerful strategy for multiscale object detection. The hierarchical and multiscale feature extraction capabilities of Swin Transformer effectively captured contextual information, resulting in robust performance in real-world Kiwifruit detection scenarios. Leveraging YOLOv8 as the baseline model provided a strong foundation to showcase the advantages of our proposed method. The ablation experiments further confirmed the significance of

incorporating Swin Transformer modules at different positions within the YOLOv8 architecture. Notably, the improvements in model performance were most notable after inserting Swin Transformer modules in the Neck, underscoring the importance of utilizing Swin Transformer's capabilities in feature fusion and enhancing the representation of different scales. The achieved results of our proposed (SWT+YOLOv8) model, surpassing the state-of-the-art models in terms of precision, recall, and mAP, highlight the effectiveness of our approach in tackling multiscale object detection challenges. The substantial gains in detection accuracy demonstrate the potential applicability of our method not only in Kiwifruit detection but also in other object detection tasks.

5 Conclusion

In this paper, we address limitations of the existing models in achieving accurate multiscale Kiwifruit object detection. To overcome these limitations, we propose a novel approach that combines the hierarchical and multiscale feature extraction capabilities of Swin Transformer with the practicality of YOLOv8, which has demonstrated excellent performance in handling multiscale object detection tasks. By enhancing the feature extraction capabilities of the model, our approach improves the accuracy of multiscale object detection. Specifically, our model effectively captures contextual information and demonstrates robustness in real-world Kiwifruit detection scenarios. The experimental results validate the effectiveness of our proposed method and achieve the state-of-the-art performance on our Kiwifruit dataset. Through comprehensive evaluation metrics, we measure the precision, recall, and mAP of the model, confirming its superior detection accuracy [5, 26].

To sum up, the achievements of this research project is the advancement of Transformer-based object detection models and demonstrate the potential in addressing the challenges of multiscale object detection in real-world scenarios. The proposed method shows promise in various computer vision tasks, further will drive the development of research work related to visual object detection.

References

1. Carion, N., Massa, F., Synnaeve, G., Usunier, N., Kirillov, A., Zagoruyko, S.: End-to-end object detection with Transformers. ECCV, pp. 213-229. Springer (2020).
2. Fang, Y., Liao, B., Wang, X., Fang, J., Qi, J., Wu, R., Niu, J., Liu, W.: You only look at one sequence: Rethinking Transformer in vision through object detection, https://arxiv.org/abs/2106.00666.
3. Ferguson, A.: 1904—the year that Kiwifruit (Actinidia deliciosa) came to New Zealand. New Zealand Journal of Crop and Horticultural Science, 32, 3-27 (2004).
4. Fu, Y., Nguyen, M., Yan, W.Q.: Grading methods for fruit freshness based on deep learning. SN Computer Science. 3, (2022).
5. Girshick, R., Donahue, J., Darrell, T., Malik, J.: Rich feature hierarchies for accurate object detection and semantic segmentation. IEEE Conference on Computer Vision and Pattern Recognition, pp. 580-587 (2014).

6. Gong, H., Mu, T., Li, Q., Dai, H., Li, C., He, Z., Wang, W., Han, F., Tuniyazi, A., Li, H., Lang, X., Li, Z., Wang, B.: Swin-transformer-enabled YOLOv5 with attention mechanism for small object detection on satellite images. Remote Sensing. 14, 2861 (2022).
7. He, K., Zhang, X., Ren, S., Sun, J.: Identity mappings in deep residual networks. Computer Vision – ECCV 2016. 630–645 (2016).
8. Krizhevsky, A., Sutskever, I., Hinton, G.E.: ImageNet classification with deep convolutional Neural Networks. Communications of the ACM. 60, 84–90 (2012).
9. Law, H., Deng, J.: CornerNet: Detecting objects as paired keypoints. International Journal of Computer Vision. 128, 642–656 (2019).
10. Liu, Y., Nand, P., Hossain, M.A., Nguyen, M., Yan, W.Q.: Sign language recognition from digital videos using feature pyramid network with detection transformer. Multimedia Tools and Applications. 82, 21673–21685 (2023).
11. Liu, Y., Yang, G., Huang, Y., Yin, Y.: SE-Mask R-CNN: An improved Mask R-CNN for apple detection and segmentation. Journal of Intelligent Fuzzy Systems, 41, 6715-6725 (2021).
12. Liu, Z., Lin, Y., Cao, Y., Hu, H., Wei, Y., Zhang, Z., Lin, S., Guo, B.: Swin Transformer: Hierarchical vision transformer using shifted windows. IEEE/CVF International Conference on Computer Vision (ICCV). (2021).
13. Liu, Z., Yan, W., Yang, B.: Image denoising based on a CNN model. IEEE ICCAR (2018).
14. Luo, Z., Yan, W., Nguyen, M.: Kayak and sailboat detection based on the improved YOLO with Transformer. International Conference on Control and Computer Vision. (2022).
15. Massah, J., Asefpour Vakilian, K., Shabanian, M., Shariatmadari, S.: Design, development, and performance evaluation of a robot for yield estimation of Kiwifruit. Computers and Electronics in Agriculture, 185, 106132 (2021).
16. Pan, C., Liu, J., Yan, W., et ak.: Salient object detection based on visual perceptual saturation and two-stream hybrid networks. IEEE Transactions on Image Processing (2021).
17. Pan, C., Yan, W.: A learning-based positive feedback in salient object detection. IEEE IVCNZ (2018).
18. Pan, C., Yan, W.: Object detection based on saturation of visual perception. Multimedia Tools and Applications, 79 (27-28), 19925-19944 (2020).
19. Redmon, J., Divvala, S., Girshick, R., Farhadi, A.: You only look once: Unified, real-time object detection. IEEE CVPR. pp. 779-788 (2016).
20. Shen, D., Xin, C., Nguyen, M., Yan, W.: Flame detection using deep learning. IEEE ICCAR (2018).
21. Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A., Kaiser, Ł., Polosukhin, I.: Attention is all you need. Advances in Neural Information Processing Systems 30 (2017).
22. Wang, L., Yan, W.: Tree leaves detection based on deep learning. International Symposium on Geometry and Vision, pp. 25-38 (2021).
23. Xia, Y. Kiwifruit Detection and Tracking from A Deep Learning Perspective Using Digital Videos. Master's Thesis, Auckland University of Technology, New Zealand (2023).
24. Xia, Y., Nguyen, M., Yan, W.Q.: A real-time Kiwifruit detection based on improved YOLOv7. Image and Vision Computing. 48–61 (2023).
25. Yan, W.: Computational Methods for Deep Learning – Theory, Algorithms, and Implementations (2nd Edition). Springer (2023).
26. Yan, W.: Introduction to Intelligent Surveillance (3rd Edition). Springer (2019).
27. Zhao, K., Nguyen, M. Yan, W.: Fruit detection from digital images using CenterNet. International Symposium on Geometry and Vision (2021).