A LIGHTWEIGHT TEA BUD DETECTION METHOD FOR ONLINE GRADING OF FRESH TEA LEAVES BASED ON IMPROVED YOLOV11N

/

一种基于改进 YOLOv11n 的面向茶鲜叶在线分级的轻量化茶芽检测方法

Wenguang ZHENG *), Xinyong SHI, Rongyang WANG
School of Mechanical Engineering & Automation, University of Science and Technology Liaoning, Anshan 114000, China
School of Intelligent Manufacturing and Elevator Mechanics, Huzhou Vocational and Technical College, Huzhou 313000, China
Tel: +8618641268671; E-mail: zhengwg@ustl.edu.cn
Corresponding author: Wenguang Zheng
DOI: https://doi.org/10.35633/inmateh-79-12

Keywords: *object detection, fresh tea leaves, YOLOv11, lightweight model*

ABSTRACT

To address the limitations of computational resources and the stringent real-time stability requirements in post-harvest online grading of fresh tea leaves, this study proposes a lightweight object detection method oriented toward efficient deployment. The YOLOv11n network was selected as the baseline model. A StarNet backbone was introduced to enhance nonlinear inter-channel interactions while reducing the complexity of feature extraction. In addition, a DySample dynamic upsampling module was employed to predict content-adaptive sampling locations, thereby improving multi-scale feature reconstruction under lightweight constraints. Furthermore, a wavelet-based pooling structure was designed to perform structure-aware frequency-domain decomposition, preserving critical edge and texture information while reducing redundant computation. The Inner-MPDIoU loss function was also adopted to calculate overlap consistency within a compact core region, thereby improving bounding-box localization stability for fine-grained structures. Based on these integrated improvements, the lightweight YOLOv11-SDWI detection model was developed. Experimental results demonstrated that the proposed model achieved an accuracy of 89.5%, with only 4.9 GFLOPs and 1.89 M parameters. To further verify its engineering applicability, the algorithm was encapsulated into a complete visual software system specifically designed for fresh tea leaf sorting. The developed system provides a functional interface for real-time detection, effectively bridging theoretical algorithm design and practical agricultural applications, and establishing a solid software foundation for future deployment on physical sorting equipment.

摘要

为了解决采后鲜茶在线分级中计算资源有限以及对高实时稳定性的严格要求，本文研究了一种面向高效部署的轻量化目标检测方法。以 YOLOv11n 网络为基线模型。引入 StarNet 主干网络以增强非线性通道间交互并降低特征提取复杂度。采用 DySample 动态上采样模块来预测内容自适应的采样位置，从而在轻量化约束下改善多尺度特征重建。进一步设计了基于小波的池化结构来执行结构感知的频域分解，在保留关键边缘和纹理信息的同时减少冗余计算。此外采用 Inner MPDIoU 损失在紧凑的核心区域上计算重叠一致性，增强了细粒度结构的边界框定位稳定性。基于这些系统性改进，构建了 YOLOv11 SDWI 轻量化检测模型。实验结果表明，所提模型实现了 89.5%的精度，同时仅需 4.9 G 浮点运算数和 1.89 M 参数量。为了验证其工程应用潜力，该算法被封装成专为鲜茶分拣量身定制的完整视觉软件系统。这种系统级实现为实时检测提供了功能性界面，成功弥合了理论算法设计与实际农业应用之间的差距，并为未来在物理分拣设备上的部署奠定了坚实的软件基础。

INTRODUCTION

Tea grading has become a crucial process in modern tea production, as stable and efficient quality classification directly affects product value, consumer acceptance, and processing efficiency.

In practical production, fresh tea leaves are typically graded according to external characteristics such as bud–leaf morphology, tenderness, integrity, and size. Common grading categories include single bud, one bud with one leaf, one bud with two leaves, and broken leaves. Early studies on tea quality analysis and classification mainly relied on traditional image processing and machine learning methods.

Wenguang Zheng, Assoc. Prof. Ph.D. Eng; Xinyong Shi, M.S. Stud. Eng.; Rongyang Wang, Prof. Ph.D. Eng.

For example, multispectral features were extracted using wavelet transforms, and Chinese famous tea classification was achieved via a multi-class least square support vector machine *(Li X. et al., 2011)*. Hyperspectral imaging was employed to model and quantitatively characterize the evolution of color and category during the tea drying process *(Xie C. et al., 2014)*. In addition, tea image classification was achieved based on handcrafted texture features, such as local binary patterns *(Tang Z. et al., 2015)*. These methods typically depend on manually designed features combined with shallow classifiers, and can achieve reasonable performance under controlled experimental conditions. However, due to the lack of end-to-end learning capability, their generalization ability is limited, making it difficult to meet the comprehensive requirements for robustness, accuracy, and scalability in practical applications.

With the rapid development of computer vision and deep learning technologies, convolutional neural network (CNN)–based object detection methods have demonstrated significant advantages in target recognition and localization under complex backgrounds by automatically learning hierarchical feature representations from large-scale data. The Faster R-CNN framework was applied to tea bud recognition and detection, demonstrating significant effectiveness in handling complex poses, occlusions, and background interference *(Xu G. et al., 2020; Zhu H. et al., 2022)*. In addition, a variety of deep learning paradigms have been explored for tea-related visual perception tasks. These include discriminative pyramid networks for semantic segmentation *(Hu G. et al., 2021)*, Mask R-CNN for joint detection and picking-point localization *(Wang T. et al., 2023)*, SSD-based feature fusion models for enhanced bud detection *(Chen B., Yan J., 2021)*, and Transformer-based frameworks such as RT-DETR-Tea to strengthen global feature modeling and multi-scale semantic fusion in complex environments *(Chen Y. et al., 2024)*. These studies indicate that, compared with traditional approaches, deep learning–based object detection algorithms achieve higher detection accuracy and better efficiency in complex environments.

The YOLO family of models represents a typical class of single-stage object detection methods. Through continuous optimization of network architectures and feature fusion strategies, these models have achieved a favorable balance between detection accuracy and computational efficiency and have gradually been introduced into complex agricultural vision scenarios. In tea bud detection research, existing studies have mainly focused on suppressing background interference and modelling scale variation. A two-level fusion framework, combining YOLOv3-based detection with DenseNet201-based classification, was constructed to improve the recognition accuracy of tea bud targets *(Xu W. et al., 2022)*. In addition, a refined TBC-YOLOv7 algorithm was developed by integrating Transformer modules, a bidirectional feature pyramid, and coordinate attention mechanisms, while employing the SIOU loss function to accelerate convergence and enhance detection accuracy *(Wang S. et al., 2023)*. Regarding the YOLOv8 framework, various optimization strategies have been explored. For instance, an occlusion-aware dual-attention mechanism was introduced to enhance model robustness under complex backgrounds and partial occlusions for tea pest and disease detection *(Wu Y. et al., 2025)*. Furthermore, YOLOv8-based models have been improved through attention mechanisms, multi-scale feature fusion, enhanced convolutional structures, Transformer modeling, and regression loss optimization, effectively increasing tea leaf detection and grading accuracy under leaf stacking conditions in complex environments *(Tang X. et al., 2025; Xia Y. et al., 2024; Zhao X. et al., 2024)*. In summary, existing studies indicate that YOLO-series models are capable of effectively handling target scale variation and complex background interference while maintaining real-time performance, thereby providing a solid technical foundation for the development of online tea leaf detection and grading systems.

Although the YOLO family has demonstrated strong performance in real-time object detection, there remains considerable room for optimization in specific application scenarios. In recent years, lightweight optimization of YOLO-based models has become an active research topic, with improvements explored in backbone design, feature fusion, sampling strategies, and training objectives. For instance, the YOLO-PEFL model was developed by replacing the YOLOv4 backbone with ShuffleNetV2 and adjusting anchor sizes, reducing model size by approximately 80% while maintaining 96.71% accuracy for efficient pear flower detection *(Wang C. et al., 2022)*. Similarly, the PAG-YOLO architecture was introduced, incorporating spatial and channel attention mechanisms with a specialized loss function to improve small-ship detection accuracy with negligible computational overhead *(Hu J. et al., 2021)*. To further reduce complexity, the GhostNet-YOLOv5 framework was proposed by integrating a GhostNet backbone, achieving 76.31% accuracy with only 5.94 million parameters and a 10 MB weight size *(Cao M. et al., 2022)*. Another approach involved adopting ShuffleNetV2 as the YOLOv5 backbone combined with channel pruning in the neck and detection head, compressing the model to 27% of its original size *(Zhang S. et al., 2023)*.

Additionally, a lightweight tea bud detection model was constructed by introducing GhostConv, bottleneck attention modules, and weighted feature fusion into YOLOv5, which increased mAP by 9.66% while significantly reducing FLOPs and parameters *(Gui Z. et al., 2023)*. Recent studies have also focused on optimizing bounding box regression loss functions to enhance training stability and performance under constrained model capacities.

However, existing studies have mainly focused on object detection during tea garden picking or in complex natural scenes, while relatively less attention has been paid to post-harvest grading, which is a critical stage in the processing pipeline. Therefore, this paper focuses on the research scenario of post-harvest fresh tea leaf grading equipment. Under this operating condition, fresh tea leaves, after mechanical pre-screening, are fed into intelligent sorting equipment, within which multi-stage single-row conveying devices are used to orderly transport and spread the tea leaves, so that they present a low-occlusion and relatively regular arrangement during the visual detection stage. Therefore, under this application background, how to construct a lightweight detection model that balances computational efficiency, detection accuracy, and engineering deplorability remains an urgent problem to be solved. Based on this, this paper proposes an improved lightweight object detection framework, YOLOv11-SDWI. The proposed method performs coordinated optimization from both the network architecture and training strategy levels, enhancing feature representation capability and regression stability while maintaining a compact model structure. At the same time, a fresh tea leaf dataset oriented toward post-harvest online grading scenarios is constructed to provide data support for subsequent model training and evaluation. Experimental results show that the proposed method can maintain competitive detection performance while significantly reducing computational complexity and parameter scale, verifying its engineering application potential in practical tea sorting systems.

MATERIALS AND METHODS

YOLOv11 network improvements

The YOLO family has continuously evolved in the field of object detection, becoming a representative technical framework with significant influence and broad practical adoption *(Ge Z. et al., 2021; Redmon J., Farhadi A., 2017)*. Multi-scale feature representation and extraction capabilities have been further enhanced in YOLOv11 through improved backbone and neck architectures, as well as the introduction of modules such as C3K2, SPPF, and C2PSA, demonstrating superior performance for small-object detection in complex scenes *(Khanam R., Hussain M., 2024)*. Given that tea buds are small targets and traditional methods often suffer from insufficient accuracy or excessive computational overhead, YOLOv11n was selected as the baseline model for further design and optimization in this study.

To address the real-time requirements and computational constraints in fresh tea leaf sorting, this study develops the YOLOv11-SDWI model based on the YOLOv11n baseline. This architecture integrates the StarNet backbone, the DySample dynamic upsampling module, the Wavelet Pooling downsampling module, and the Inner-MPDIoU bounding box regression loss function. The structural details of the proposed YOLOv11-SDWI network are illustrated in Figure 1.

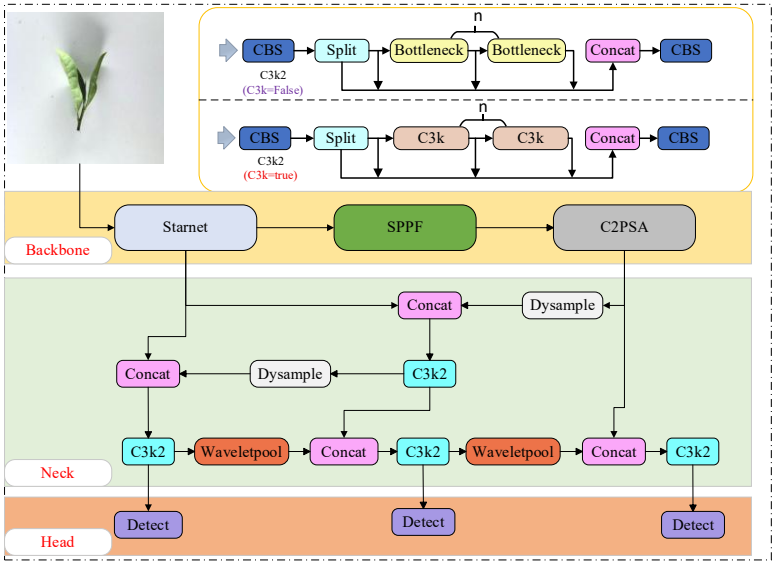


Fig. 1 - Architecture of the proposed YOLOv11-SDWI lightweight fresh tea leaf detection network

Lightweight Backbone Optimization Using StarNet

In post-harvest fresh tea leaf online grading scenarios, detection models are typically deployed on edge devices, where computational and storage resources are limited and long-term stable real-time operation is required. Therefore, the backbone network must maintain sufficient feature representation capability under low computational cost and a small parameter budget. Traditional convolutional backbones usually enhance representation by increasing network depth or channel width, which significantly increases computational and storage overhead, while channel interactions are mainly modelled through linear aggregation, limiting their ability to capture complex feature combinations. To address these issues, this study introduces StarNet into the YOLOv11n backbone *(Ma X. et al., 2024)***Error! Reference source not found.**, leveraging star-shaped operations to enhance nonlinear inter-channel interactions. This design improves the effective feature output per unit of computation while preserving a compact architecture, thereby enabling a more efficient lightweight feature extraction backbone.

As illustrated in Fig. 2, StarNet adopts a stage-wise hierarchical architecture. After the initial convolutional stem, feature maps are progressively processed by multiple stages (Stage1–Stage4). Each stage consists of a down-sampling convolution followed by several stacked StarBlocks, enabling multi-scale semantic feature extraction with gradually increasing receptive fields. StarBlock is the fundamental building unit of StarNet, and its internal data flow is explicitly shown in the figure. Given an input feature with channel dimension d, the feature is first processed by a depth wise convolution (DW-Conv) to extract spatial information with low computational overhead. The resulting feature is then fed into two parallel fully connected (FC) branches that expand the channel dimension to 4d.One branch is activated by ReLU6, and the outputs of the two branches are fused through an element-wise multiplication operation. This operation introduces nonlinear inter-channel interactions without explicitly increasing the output channel dimension. The fused feature is subsequently projected back to dimension d through an FC layer and further refined by a DW-Conv. A residual connection is finally applied to combine the block input and output, ensuring stable feature propagation across stacked StarBlocks.

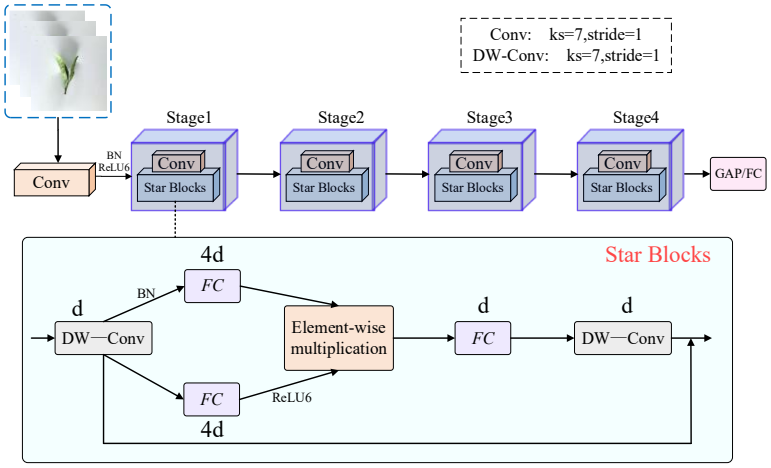


Fig. 2 - Schematic diagram of the StarNet backbone architecture

By stacking multiple StarBlocks within each stage, StarNet progressively enhances feature representation from low-level details to high-level semantics while maintaining a compact structure. When integrated into the YOLOv11 backbone, the input tea leaf image is sequentially processed by the stem convolution, StarNet backbone stages, and the subsequent neck for multi-scale feature fusion. This design provides compact yet discriminative feature representations for the detection head, which is particularly beneficial for lightweight deployment under resource-constrained edge environments.

DySample Upsampling Module

During multi-scale feature fusion, the upsampling module directly determines the spatial reconstruction quality of high-level semantic features. Conventional nearest-neighbor or bilinear interpolation relies on fixed sampling over regular grids that is independent of feature content; while computationally efficient, such methods often fail to accurately recover spatial structures near object boundaries, fine-grained textures, or regions with significant structural variation. To address this limitation, this paper introduces the DySample *(Liu W. et al., 2023)* upsampling module into the Neck of YOLOv11.

By predicting content-adaptive sampling locations, DySample enables learnable feature reconstruction, thereby enhancing the representation capability of multi-scale feature fusion under lightweight constraints.

As illustrated in Fig. 3, the overall workflow of DySample consists of two components: sampling coordinate construction and feature resampling. Given an input feature map $X \in R^{C \times H \times W}$, the module first predicts a set of two-dimensional offset information through a lightweight linear mapping, and constructs a high-resolution sampling coordinate set S based on a regular initial grid. Subsequently, differentiable bilinear resampling is performed on the input feature X according to S, generating an output feature X' whose spatial resolution is increased to $sH \times sW$. Since the resampling process is differentiable with respect to the sampling coordinates, the prediction of sampling locations can be jointly optimized with the detection objectives during training, thereby enabling end-to-end adaptive upsampling.

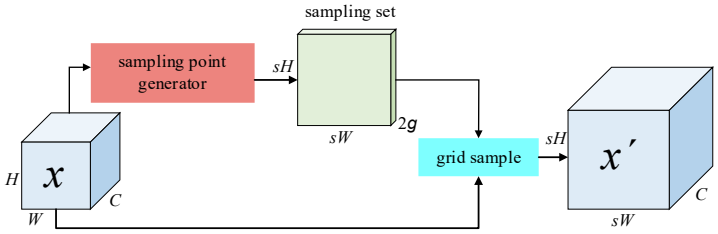


Fig. 3 - Overall workflow of the DySample upsampling module

During the construction of sampling coordinates, DySample achieves a balance between representation capability and computational complexity by adopting a grouped sampling + scope control strategy. As shown in the upper part of Fig. 4, under the static scope control mode, the module predicts an offset map O in the low-resolution feature domain via a linear mapping and applies a fixed scaling factor to constrain its magnitude, so that the initial sampling points are distributed around a regular grid. The offsets are then mapped to the high-resolution sampling grid through a spatial–channel rearrangement operation and added to the regular initial grid to obtain the final sampling coordinate set S. By constraining the maximum displacement of sampling points with a fixed scope, this strategy ensures structural stability of the sampling pattern while avoiding severe resampling instability caused by excessive offsets, making it suitable for regions with relatively smooth structural variations.

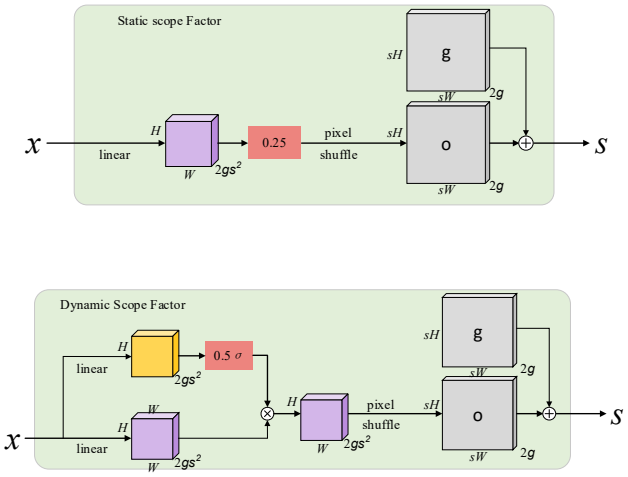


Fig. 4 - Static and dynamic scope control mechanisms in DySample

As illustrated in the lower part of Fig. 4, under the dynamic scope control mode, DySample further introduces an input-adaptive scale modulation factor on top of offset prediction. This factor is jointly applied with the predicted offsets to dynamically adjust the sampling displacement magnitude at each spatial location. Specifically, in flat regions or structurally simple areas, the displacement magnitude is automatically suppressed, causing sampling positions to remain close to the regular grid; in contrast, in edge regions or areas with significant structural variations, the displacement magnitude is adaptively amplified, thereby enhancing the sensitivity of sampling locations to spatial structure changes. Through this mechanism, DySample enables region-aware adaptive sampling without introducing a significant increase in computational cost, effectively accommodating spatial heterogeneity across different regions of the feature map.

As shown in Fig. 5, DySample provides the LP (Low-resolution to Pixel Shuffle) and PL (Pixel Shuffle to Low-resolution) offset prediction schemes. The LP structure predicts offsets directly in the low-resolution feature domain and maps them to the high-resolution sampling grid via spatial-channel rearrangement. The PL structure projects features into the high-resolution domain first to predict offsets, which are then reshaped to match the sampling coordinates. Both structures ultimately generate a sampling coordinate set *S* for the subsequent differentiable resampling process.

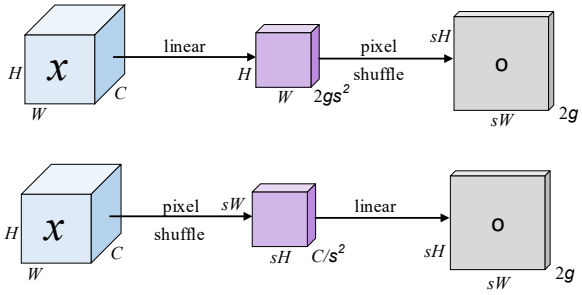


Fig. 5 - Offset prediction structures of DySample: LP and PL schemes

In summary, DySample reformulates the traditional rule-based upsampling process into a learning framework that integrates regular priors with content-adaptive offsets, enabling the upsampling module to actively adapt to object structures and scale variations during end-to-end training. In this work, simply replacing the original fixed-interpolation upsampling module in the YOLOv11 neck with DySample can significantly improve feature reconstruction quality with almost no increase in computational complexity.

Structure-Aware Downsampling Method Based on Wavelet Transform

In convolutional neural networks, downsampling operations are commonly employed to enlarge the receptive field and reduce feature resolution. However, traditional downsampling methods based on strided convolution or pooling essentially perform direct spatial subsampling, which easily introduces aliasing effects during resolution reduction and leads to irreversible loss of high-frequency structural information such as edges and textures. For post-harvest fresh tea leaf grading tasks, bud contours, leaf-edge morphology, and local structural variations constitute critical discriminative cues; degradation of such structural information at the downsampling stage can directly impair subsequent feature fusion and detection accuracy. To address this issue, this study introduces the WaveletPool downsampling module *(Williams T., Li R., 2018)*, which replaces conventional spatial downsampling with a structure-aware frequency-domain decomposition mechanism.

As shown in Fig. 6, WaveletPool applies a two-dimensional Discrete Wavelet Transform (DWT) to the input feature map. Given an input feature map *X*, WaveletPool does not perform spatial downsampling directly; instead, it first applies fixed wavelet filtering operations independently to each channel. Specifically, the module employs four fixed two-dimensional Haar wavelet kernels to convolve the input features in parallel, thereby decomposing the original feature representation into four sub-band components with clear physical interpretations: LL, LH, HL, and HH. Among them, LL represents the low-frequency approximation component, which mainly preserves global structure and contour information, while LH, HL, and HH correspond to high-frequency detail components in different directions, characterizing edges, textures, and local structural variations, respectively.

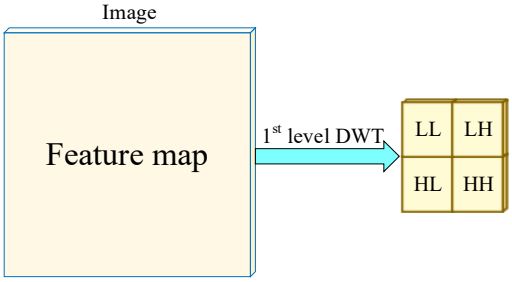


Fig. 6 - Schematic diagram of WaveletPool-based feature downsampling

This frequency-domain reorganization ensures structural integrity across scales. By preserving essential discriminative cues, WaveletPool facilitates precise feature reconstruction and target localization, effectively mitigating the information loss common in conventional lightweight designs.

Bounding Box Regression Constraint Based on Internal Region Consistency

In the post-harvest fresh tea leaf detection task, bud–leaf targets are often slender, small in scale, and subject to pronounced boundary variations. When bounding box regression is optimized solely by global overlap, it may suffer from boundary deviations that are difficult to further converge once the predicted and ground-truth boxes enter a high-overlap regime. To address this issue, this work introduces an Inner-IoU *(Wang D. et al., 2025)* based regression strategy: it computes overlap consistency on a more compact core region defined inside both boxes, and incorporates key geometric position offset constraints to strengthen the joint alignment of the target body and boundary structure under a lightweight setting.

As illustrated in Fig. 7, let the spatial resolution of the input feature map be $h \times w$. The predicted bounding box width and height in the feature space are denoted as w_{pred} and h_{pred}, respectively, while the corresponding ground-truth width and height are denoted as w_{gt} and h_{gt}. The top-left and bottom-right coordinates of the predicted bounding box are (x_1^{pred}, y_1^{pred}) and (x_2^{pred}, y_2^{pred}), respectively, and the corresponding ground-truth coordinates are (x_1^{gt}, y_1^{gt}) and (x_2^{gt}, y_2^{gt}).

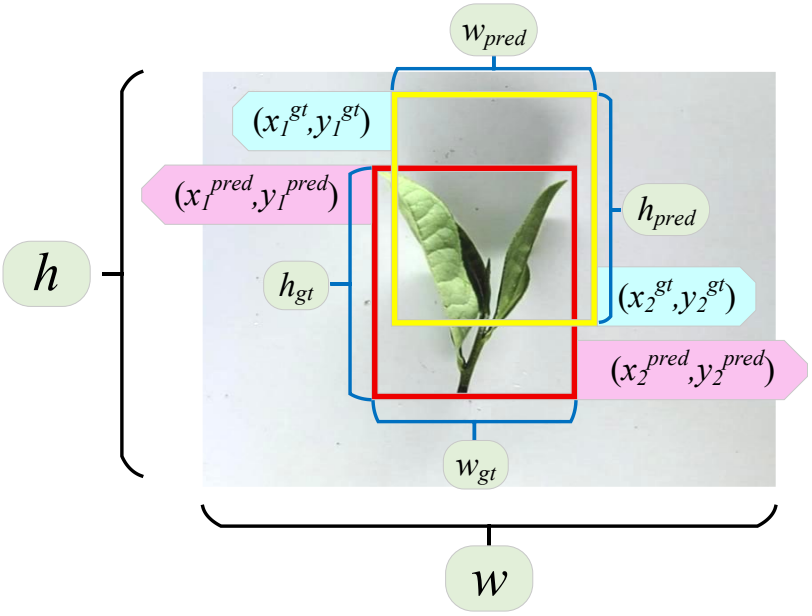


Fig. 7 - Illustration of Bounding Box Regression under Inner-IoU Constraint

At the level of region constraint, the regression supervision is not directly imposed on the complete predicted box and ground-truth box. Instead, a more compact core region is constructed inside both boxes, focusing exclusively on the effective area covered by the target body. This internal region is spatially constrained jointly by the predicted box and the ground-truth box, such that the overlap computation is concentrated on the target itself. As a result, interference from background regions and boundary redundancy is effectively reduced during the regression process. Under scenarios where the target size is small or the two boxes already exhibit a high degree of overlap, this constraint strategy can still maintain stable optimization signals, thereby improving the continuity and convergence of the regression process.

At the level of geometric position constraint, explicit modelling of the positional offsets between the predicted box and the ground-truth box is further introduced. As illustrated in Fig. 7, the spatial discrepancies at both the top-left and bottom-right corners of the predicted and ground-truth boxes are jointly incorporated into the regression constraint, enabling coordinated correction of boundary geometry across position, width, and height dimensions. To prevent imbalance in optimization strength across targets of different scales during regression, these positional offsets are normalized using the minimum enclosing region jointly formed by the two boxes. This normalization ensures that targets of varying sizes are subject to a consistent geometric constraint scale during training.

Through the combined effect of internal region consistency constraints and key positional distance awareness, the regression process no longer focuses solely on whether the predicted box covers the target, but further strengthens fine-grained alignment in terms of boundary geometry and spatial structure. Without introducing additional inference overhead, this strategy effectively improves the localization accuracy and stability of detection boxes in scenarios involving fine-grained structural details.

Intelligent Sorting Software System Design

To apply the proposed YOLOv11 SDWI model to actual fresh tea leaf sorting scenarios, a visual software system is designed in this study. The complete operational workflow of the software system is illustrated in Figure 8.

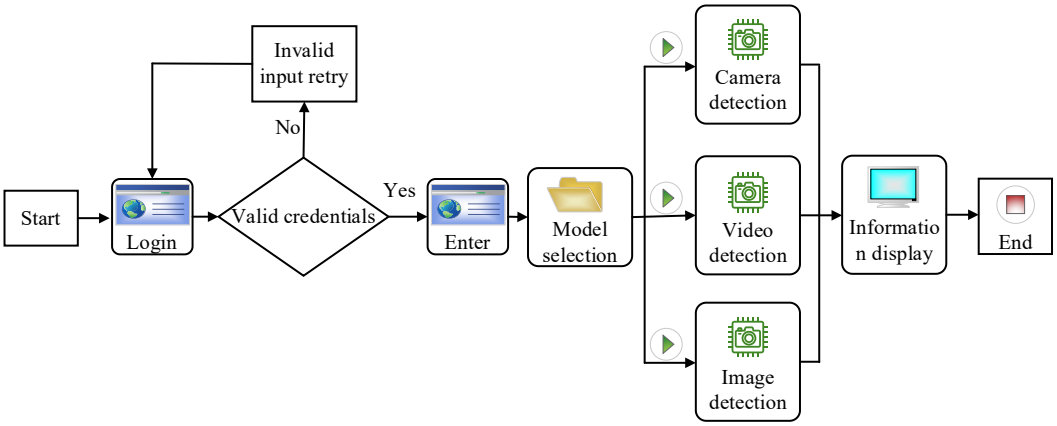


Fig. 8 - Operational workflow of the visual software system for fresh tea leaf sorting

The process begins with a user authentication stage where users input their account and password to log in. If the credentials are incorrect, the system prompts an error and requires re login. Upon successful authentication, the user enters the main interface and proceeds to the model selection module. Here the core inference engine loads the trained lightweight model weights. To accommodate diverse practical needs, the system offers 3 distinct detection modes including camera detection, video detection, and image detection. Once a specific mode is selected and visual data is input, the backend immediately executes feature extraction and bounding box regression. The system then continuously outputs target categories and confidence scores and bounding box coordinates to the front end information display module before ending the process. This logic ensures a seamless transition from data input to result visualization. The initial login interface of the developed intelligent system is presented in Figure 9, demonstrating a user friendly design tailored for practical agricultural applications.

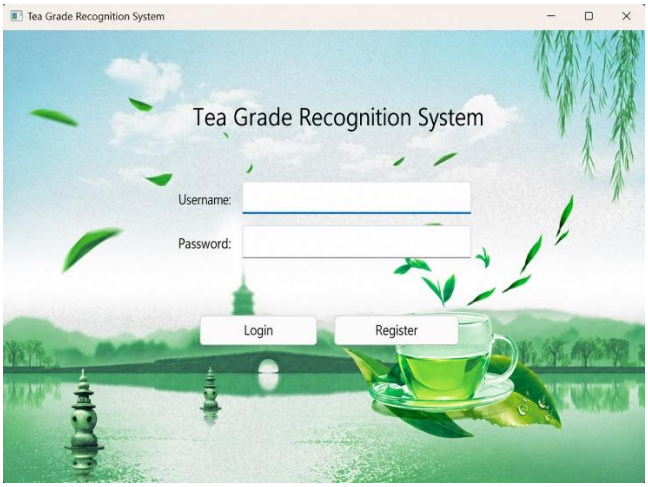


Fig. 9 - Initial graphical user interface of the developed intelligent sorting software

Image Acquisition and Dataset Construction

This study constructs an image dataset of post-screened fresh tea leaves for tea grade recognition and online grading. To address the stacking and occlusion that can occur during conveyance in intelligent sorting equipment and consequently degrade visual recognition, the multi-stage single-file conveying unit of the sorting system is used to spread and separate the leaves, so that they present a low-occlusion and relatively orderly arrangement on the conveyor belt, enabling the acquisition of image samples with clearer contours and texture details. Images are captured frame by frame using a Canon 6D II camera at a fixed viewpoint, covering diverse poses, scales, and local detail variations of fresh tea leaves during conveyance to enhance data diversity and representativeness.

During dataset construction, the raw images are first screened and organized, and samples with obvious blur, abnormal exposure, missing targets, or severe occlusion are removed. Ultimately, 1,563 valid images of fresh tea leaves of different grades are obtained to build the grading dataset. To ensure objective and reproducible training and evaluation, all images are randomly split into training, validation, and test sets at a ratio of 7:2:1.

Data Screening and Image Annotation

In this study, the collected fresh tea leaf images were manually annotated using the bounding-box labelling tool Labellmg to obtain object category labels and spatial localization information. According to the main picking morphology, fresh tea leaves were divided into four classes: one bud and one leaf, one bud and two leaves, one bud and three leaves, and one bud with multiple leaves, encoded as 0, 1, 2, and 3, respectively. To ensure a comprehensive dataset representation, a total of 8,267 instances were annotated across 1,563 images. According to the picking morphology, the instances were categorized into four classes: one bud with one leaf (1,923 instances), one bud with two leaves (2,196 instances), one bud with three leaves (1,747 instances), and one bud with multiple leaves (2,401 instances). The class definitions and index mapping are listed in Table 1.

Table 1

Categories of fresh tea leaves and corresponding codes

Category	Code
One bud with one leaf	0
One bud with two leaves	1
One bud with three leaves	2
One bud with multiple leaves	3

After annotation, labels were saved as XML files in the PASCAL VOC format. Each bounding box corresponds to a target instance, and a single image may contain multiple instances; the number of boxes varies with the number of fresh tea leaf targets in the image. Representative appearances and poses of the four classes are shown in Figure 10.

To support subsequent YOLO-based training and inference, a Python script was further used to convert the annotations from the VOC format to the YOLO format. After conversion, each image is associated with a TXT label file, where each target is represented as n, x, y, w, h. Here, n denotes the class index, x and y are the normalized horizontal and vertical coordinates of the bounding-box center, and w and h are the normalized width and height of the bounding box. To enhance the generalization ability of the model, data augmentation is applied to the training set during training to increase sample diversity.

Fig. 10 - Examples of fresh tea leaf samples from different grade categories

RESULTS

Experimental Environment and Training Strategy

The experimental platform in this study is based on the Ubuntu 20.04 operating system with Python 3.8. The deep learning framework used is PyTorch 1.10.0, with CUDA 11.3 configured to enable GPU-accelerated computation. In terms of hardware, the platform is equipped with a single NVIDIA GeForce RTX 3090 GPU with 24 GB of memory, an AMD EPYC 7642 48-Core Processor (20 vCPU), and 90 GB of system memory.

To ensure a rigorous and fair comparison, all models evaluated in this study, including the proposed YOLOv11-SDWI and various baseline architectures, were trained from scratch for 300 epochs without utilizing any pre-trained weights. Furthermore, a uniform input resolution was maintained across all experiments, and the standard data augmentation strategies provided by YOLOv11 were adopted. No individual hyperparameter tuning was performed for any specific model, as all hyperparameters were kept strictly consistent with the default configurations of the baseline YOLOv11 model.

Evaluation Metrics

This study evaluates model performance from two aspects: detection accuracy and computational efficiency. Commonly used metrics in object detection are adopted, including *Precision (P), Recall (R), and Mean Average Precision (mAP)*, as well as indicators reflecting model complexity and inference cost, such as the number of parameters (Params), floating-point operations (FLOPs), real-time inference speed measured in Frames Per Second (FPS), and model weight file size. These metrics are used to comprehensively and quantitatively assess the performance of the proposed improved model, and their definitions and meanings are described as follows.

$$P = \frac{TP}{TP + FP} \tag{1}$$

$$R = \frac{TP}{TP + FN} \tag{2}$$

$$AP = \int_0^1 (P \times R) dx \tag{3}$$

$$mAP = \frac{1}{N} \sum_{i=1}^{N} AP(i) \tag{4}$$

In the above definitions:

True Positives (TP) denote the number of actual positive samples that are correctly predicted as positive, that is, the number of targets correctly detected by the model;

False Negatives (FN) denote the number of actual positive samples that are incorrectly predicted as negative, that is, the number of targets missed by the model;

False Positives (FP) denote the number of actual negative samples that are incorrectly predicted as positive, that is, the number of background regions mistakenly detected as targets.

Ablation Experiments

To verify the effectiveness of the proposed lightweight improvement strategies, YOLOv11 is adopted as the baseline model, and multiple modifications are introduced from both network architecture and training strategy perspectives. Ablation experiments are conducted to analyze the impact of different module combinations on model performance. At the network architecture level, improvements are applied to the backbone network, feature fusion process, and downsampling strategy; at the training strategy level, the bounding box regression loss function is optimized. For clarity in ablation analysis, the symbols a, b, c, and d are used to denote the four modules StarNet, DySample, Wavelet Pooling, and Inner-MPDIoU, respectively, with their corresponding relationships listed in Table 2.

Table 2

Module Notation Used in Ablation Experiments

Module	StarNet	DySample	Wavelet Pooling	Inner-MPDIoU
Symbol	a	b	c	d

The ablation results are shown in Table 3. Introducing StarNet (a) effectively reduces the model parameters from 2.58 M to 1.95 M and the FLOPs from 6.3 G to 5.2 G, while improving the detection accuracy to 87.9%, demonstrating efficient model compression with stable performance. Further incorporating DySample (a+b) or Wavelet Pooling (a+c) leads to additional accuracy gains and reduced computational cost, indicating that these lightweight modules contribute positively when combined with an optimized backbone. When StarNet, DySample, and Wavelet Pooling are jointly integrated (a+b+c), the FLOPs and parameter count are reduced to 4.9 G and 1.89 M, corresponding to reductions of 22.2% and 26.7% compared with the baseline,

while achieving an accuracy of 89.1%. By further introducing Inner-MPDIoU to form the complete YOLOv11-SDWI model, the detection accuracy is increased to 89.5% without additional computational overhead, demonstrating effective synergy among the proposed components.

Table 3

Results of network ablation experiments

Model	Mean Average Precision mAP@0.5(%)	Floating Point Operations (FLOPs)/G	Number of Parameters (Params)(×10⁶)
YOLOv11	87.0	6.3	2.58
YOLOv11-SDWI	89.5	4.9	1.89
YOLOv11+a	87.9	5.2	1.95
YOLOv11+a+b	88.5	5.0	1.94
YOLOv11+a+c	88.1	5.2	2.11
YOLOv11+a+d	88.5	5.0	1.94
YOLOv11+b+c	88.7	6.3	2.53
YOLOv11+a+b+c	89.1	4.9	1.89

Backbone Network Comparison

To verify the applicability and advantages of the selected backbone network in lightweight object detection tasks, comparative experiments were conducted on several mainstream lightweight backbones under the same detection framework and unified training strategy. The compared backbones include StarNet, EfficientViT, FasterNet, MobileNetV4, and UniRepLKNet. The performance of different backbone networks was comprehensively evaluated from three aspects: detection accuracy, floating-point operations (FLOPs), and parameter scale. The corresponding results are presented in Table 4. Among these metrics, detection accuracy is measured using the mean average precision at an IoU threshold of 0.5.

Table 4

Performance Comparison of Different Lightweight Backbone Networks

Module	Mean Average Precision mAP@0.5(%)	Floating Point Operations (FLOPs)/G	Number of Parameters (Params)(×10⁶)
Starnet	87.9	5.2	1.95
efficientViT	87.3	7.9	3.73
Fasternet	86.8	9.2	3.90
Mobilenetv4	85.7	21.0	5.43
unireplknet	86.4	14.1	5.80

As shown in Table 4, StarNet achieves a mean average precision of 87.9% at an IoU threshold of 0.5, which is comparable to EfficientViT (87.3%) and higher than FasterNet, UniRepLKNet, and MobileNetv4. Despite similar detection accuracy, StarNet exhibits significantly lower computational complexity, requiring only 5.2G FLOPs, which is markedly less than EfficientViT (7.9G), FasterNet (9.2G), UniRepLKNet (14.1G), and MobileNetv4 (21.0G). In addition, StarNet contains only 1.95M parameters, approximately half of those in EfficientViT and FasterNet, and substantially fewer than UniRepLKNet and MobileNetv4.

Overall, these results indicate that while some backbone networks achieve competitive accuracy, they do so at the cost of increased computation or model size. In contrast, StarNet provides a more favorable balance between detection performance and efficiency, making it better suited for real-time, resource-constrained tea leaf sorting applications. Accordingly, StarNet is selected as the backbone network of the proposed lightweight detection framework.

Comparative Experiments with Different Detection Models

To comprehensively evaluate the potential and efficiency of the proposed YOLOv11-SDWI model for intelligent post-harvest tea leaf sorting, an extensive comparative analysis is conducted against a diverse range of mainstream object detection baselines. These baselines encompass classic two-stage detectors, conventional single-stage architectures, and multiple representative iterations within the YOLO series, thereby ensuring a thorough evaluation across different network paradigms. To guarantee fairness, all models are trained and validated on the identical dataset using strictly unified evaluation protocols. The assessment

focuses on the critical trade-off between detection accuracy and computational complexity, which is a key consideration for agricultural equipment with limited computing resources: the mean Average Precision at an IoU threshold of 0.5 (mAP@0.5) is utilized to quantify detection precision, while the parameter scale (Params) and floating-point operations (FLOPs) are simultaneously reported to gauge the lightweight characteristics and theoretical deployment feasibility. The detailed quantitative results are summarized in Table 5.

Table 5

Comparative evaluation of mainstream object detection models.

Model	Mean Average Precision mAP@0.5(%)	Floating Point Operations (FLOPs)/G	Precision *P*	Recall *R*	Number of Parameters (Params)(×10⁶)	MB
YOLOv5	83.5	4.1	77.4	78.3	1.76	77.3
YOLOv7-tiny	86.3	13.0	81.9	85.4	6.01	44.6
YOLOv8n	86.3	8.1	82.2	80.4	3.06	39.5
YOLOv9t	86.9	11.0	80.8	79.3	2.66	90.5
YOLOv10n	85.6	8.2	78.4	78.8	2.67	71.2
YOLOv11n	87.0	6.3	82.1	81.2	2.58	23.8
YOLOv12n	84.9	6.5	76.9	77.4	2.50	25.2
YOLOv13n	83.1	5.9	76.6	75.8	2.64	16.4
Fast RCNN	87.7	370.3	81.3	80.9	13.7	241.5
SSD	84.3	62.7	80.2	80.7	26.4	76.4
RTDETR	78.5	100.6	73.8	72.6	28.4	113.6
Ours	89.5	4.9	85.4	84.9	1.89	19.3

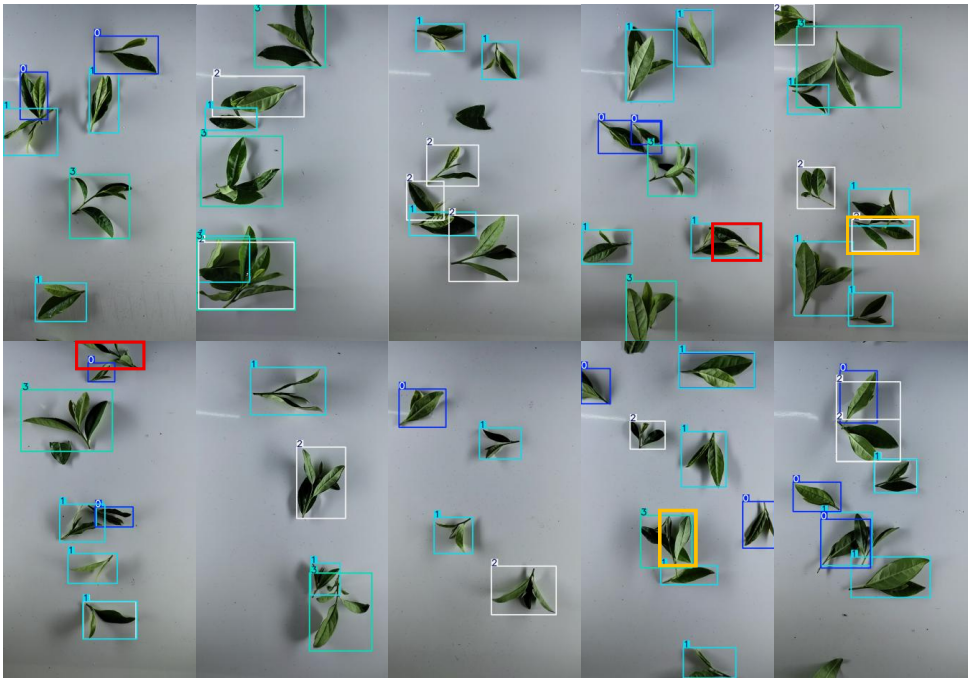
Based on the detailed data in Table 5, the proposed YOLOv11-SDWI model significantly outperforms mainstream detection networks across multiple core performance indicators. Compared to the baseline YOLOv11n model, our network not only reduces the parameter count from 2.58M to 1.89M but also increases the mean average precision from 87.0% to 89.5%, achieving a dual optimization of performance and computational complexity. This leading position remains robust during horizontal comparisons with other lightweight versions. For instance, YOLOv12n has a computational overhead of 6.5G floating point operations with a precision of only 84.9%, meaning our model delivers a detection capability 4.6% higher while maintaining a lower computational cost. Similarly, YOLOv13n exhibits a higher computational cost of 5.9G but only achieves a mean average precision of 83.1%. When compared with heavy two stage networks like Fast RCNN, the latter reaches an accuracy of 87.7% but imposes a computational burden over 75 times that of our model, making it entirely unsuitable for real-time processing in hardware restricted environments. On the other end of the spectrum, YOLOv5 possesses a slightly smaller calculation volume leading by 0.8G, yet its detection precision is 6.0% lower than our model. This substantial precision gap renders it inadequate for fresh tea processing tasks that demand high sorting quality.

Synthesizing the experimental results above, the YOLOv11-SDWI model discovers an ideal balance between detection accuracy and lightweight design. It provides a high mean average precision of 89.5% and a precision rate of 85.4% using a minimal storage space of 19.3 MB. The data fully validate that the model effectively eliminates redundant computations and strengthens the recognition ability for tiny tea bud targets without sacrificing feature representation. This architecture combining high efficiency and accuracy demonstrates strong competitiveness in the specific agricultural application of intelligent fresh tea sorting. It offers solid technical support for portable agricultural equipment with limited computing resources, presenting significant theoretical value and practical deployment potential.

Model Visualization Analysis

Beyond computational efficiency, the proposed YOLOv11-SDWI demonstrates enhanced localization stability in complex backgrounds. Figure 11 presents a comparative visualization of 10 randomly selected representative test samples. Quantitative observation of these specific images reveals that the baseline model produced 4 false negatives (missed detections) and 3 false positives (misclassifying background leaves as buds). Following the structural optimizations, YOLOv11-SDWI successfully reduced the false negatives to 2 and the false positives to 2. This targeted reduction in visual detection errors confirms that the lightweight modifications effectively preserve critical structural features, minimizing both target omission under occlusion and background interference in practical grading scenarios.

YOLOv11 Detection Results



YOLOv11-SDWI Detection Results

Fig. 11 - Comparison of actual detection effect before and after model improvement

Intelligent Sorting System Visualization and Functional Testing

Based on the preliminary software architecture design, this study successfully developed the fresh tea grade recognition system. The actual operational interface is illustrated in Figure 12. After importing the test image or video stream, the built in lightweight core inference engine executes target feature extraction and classification localization. The right information statistics panel is designed to monitor operational metrics including processing time and target count. This functional layout ensures that operators can intuitively observe the working status of the detection algorithm.

Fig. 12 - Main detection interface of the fresh tea grade recognition system

In the central main display area, the system precisely marks the spatial position of each tea bud through visualized colored bounding boxes and overlays the corresponding confidence scores above the targets. The detection results and location information data table at the bottom synchronously records the specific attributes of all tracked targets including categories and precise bounding box coordinate ranges. These structured coordinate and category output data are formatted to be directly transmittable as control commands to downstream mechanical sorting execution mechanisms in future applications. The overall software implementation proves that the lightweight model constructed in this study can be effectively packaged into an independent application with comprehensive visualization features. This development provides a necessary and reliable software foundation for the future physical integration of post-harvest fresh tea online grading equipment.

CONCLUSIONS

This paper proposes a lightweight object detection framework YOLOv11 SDWI for post-harvest fresh tea leaf online grading. By systematically optimizing the network architecture and training strategy of the baseline YOLOv11n model the proposed method achieves a comprehensive enhancement in both detection accuracy and computational efficiency. Experimental results demonstrate that the YOLOv11 SDWI model outperforms the baseline model across all accuracy metrics. The mean average precision increases from 87.0 % to 89.5 %. The precision rises from 82.1 % to 85.4 %. The recall also improves from 81.2 % to 84.9 %. Simultaneously the proposed model significantly reduces computational costs. The floating point operations are decreased from 6.3 G to 4.9 G. The parameter count is reduced from 2.58 M to 1.89 M. The model weight size is compressed from 23.8 MB to 19.3 MB. This proves that the improved model successfully strikes an optimal balance between high performance and lightweight design.

Overall, the proposed method exhibits strong potential for practical deployment in post-harvest tea sorting systems. Building upon the visual software system developed in this study future work will focus on expanding the dataset to include multiple tea varieties and further improving detection accuracy. Most importantly the plan is to integrate the proposed software architecture with mechanical sorting hardware to conduct real world physical deployment and joint debugging on actual production lines ultimately achieving the complete engineering application of the intelligent tea grading equipment.

ACKNOWLEDGEMENT

This work was supported by the Fundamental Research Funds for the Liaoning Universities LJ242410146044 and the Huzhou Public Welfare Application Research Project 2024GZ226.

REFERENCES

[1] Cao, M., Fu, J., Zhu, J., Cai, C., (2022). Lightweight tea bud recognition network integrating GhostNet and YOLOv5. *Mathematical Biosciences and Engineering*, Vol. 19, pp. 12897-12914, USA.

[2] Chen, B., Yan, J., Wang, K., (2021). Fresh Tea Sprouts Detection via Image Enhancement and Fusion SSD. *Journal of Control Science and Engineering*, Vol. 2021, pp. 6614672, United Kingdom.

[3] Chen, Y., Guo, Y., Li, J., Zhou, B., Chen, J., Zhang, M., Cui, Y., Tang, J., (2024). RT-DETR-Tea: A Multi-Species Tea Bud Detection Model for Unstructured Environments. *Agriculture*, Vol.14, pp.2256, Switzerland.

[4] Ge, Z., Liu, S., Wang, F., Li, Z., Sun, J., (2021). YOLOX: Exceeding YOLO series in 2021. *arXiv preprint*, arXiv:2107.08430, USA.
[5] Gui, Z., Chen, X., Wang, Y., Zhang, H., (2023). A lightweight tea bud detection model based on Yolov5. *Computers & Electronics in Agriculture*, Vol. 205, pp. 107636, Netherlands.
[6] Hu, G., Li, S., Wan, M., et al., (2021). Semantic segmentation of tea geometrid in natural scene images using discriminative pyramid network. *Applied Soft Computing*, Vol. 113, pp. 107984, Netherlands.
[7] Hu, J., Zhi, Y., Chen, M., Liu, P., (2021). PAG-YOLO: A Portable Attention-Guided YOLO Network for Small Ship Detection. *Remote Sensing*, Vol. 13, pp. 3059, Switzerland.
[8] Khanam, R., Hussain, M., (2024). YOLOv11: An overview of the key architectural enhancements. *Computer Vision and Pattern Recognition*, Vol. 10, pp. 17725, USA.
[9] Li, X., Nie, P., Qiu, Z., He, Y., (2011). Using wavelet transform and multi-class least square support vector machine in multi-spectral imaging classification of Chinese famous tea. *Expert Systems with Applications*, Vol. 38, pp. 11149-11159, United Kingdom.
[10] Liu, W., Lu, H., Fu, H., Cao, Z., (2023). Learning to Upsample by Learning to Sample. *2023 IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 5027-5037, Paris/France.
[11] Ma, X., Dai, X., Bai, Y., Wang, Z., (2024). Rewrite the Stars. *2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 1213-1224, Seattle/USA.
[12] Redmon, J., Farhadi, A., (2017). YOLO9000: Better, Faster, Stronger. *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 7263-7271, Piscataway/USA.
[13] Tang, X., Tang, L., Li, J., Guo, X., (2025). Enhancing multilevel tea leaf recognition based on improved YOLOv8n. *Frontiers in Plant Science*, Vol. 16, pp. 1540670, Switzerland.
[14] Tang, Z., Su, Y., Er, M., Qi, F., Zhang, L., Zhou, J., (2015). A local binary pattern based texture descriptors for classification of tea leaves. *Neurocomputing*, Vol. 168, pp. 1011-1023, Netherlands.
[15] Wang, C., Wang, Y., Liu, S., He, P., Zhang, Z., Zhou, Y., (2022). Study on Pear Flowers Detection Performance of YOLO-PEFL Model Trained With Synthetic Target Images. *Frontiers in Plant Science*, Vol. 13, pp. 911473, Switzerland.
[16] Wang, D., Lv, H., Zhu, X., Zheng, Y., (2025). Detection and identification of fish skin lesions in the underwater environment based on improved YOLOv8 model for aquaculture. *Aquaculture*, Vol. 608, pp. 742722, Netherlands.
[17] Wang, S., Wu, D., Zheng, X., (2023). TBC-YOLOv7: a refined YOLOv7-based algorithm for tea bud grading detection. *Frontiers in Plant Science*, Vol. 14, pp. 1223410, Switzerland.
[18] Wang, T., Zhang, K., Zhang, W., Liu, H., (2023). Tea picking point detection and location based on Mask-RCNN. *Information Processing in Agriculture*, Vol. 10, pp. 267-275, Netherlands.
[19] Williams, T., Li, R., (2018). Wavelet Pooling for Convolutional Neural Networks. *Proceedings of the International Conference on Learning Representations (ICLR)*, Vancouver/Canada.
[20] Wu, Y., Wan, S., Wu, L., Zhao, Y., (2025). DAONet-YOLOv8: An occlusion-aware dual-attention network for tea leaf pest and disease detection. *arXiv preprint*, arXiv:2511.23222, USA.
[21] Xia, Y., Zhang, L., Wang, Q., Li, H., (2024). Recognition Model for Tea Grading and Counting Based on the Improved YOLOv8n. *Agronomy*, Vol. 14, pp. 1251, Switzerland.
[22] Xie, C., Li, X., Shao, Y., He, Y., (2014). Color Measurement of Tea Leaves at Different Drying Periods Using Hyperspectral Imaging Technique. *PLoS ONE*, Vol. 9, pp. e113422, USA.
[23] Xu, G., Zhang, X., et al., (2020). Recognition approaches of tea bud image based on faster R-CNN depth network (基于 Faster R-CNN 深度网络的茶芽图像识别方法). *Journal of Optoelectronics Laser*, Vol. 31, pp. 1131-1139, Tianjin/China.
[24] Xu, W., Zhao, X., Liu, P., Chen, S., (2022). Detection and classification of tea buds based on deep learning. *Computers & Electronics in Agriculture*, Vol. 192, pp. 106547, Netherlands.
[25] Zhang, S., Chen, H., Wu, J., Yang, L., (2023). Edge device detection of tea leaves with one bud and two leaves based on ShuffleNetv2-YOLOv5-Lite-E. *Agronomy*, Vol. 13, pp. 577, Switzerland.
[26] Zhao, X., Wang, F., Li, C., Sun, Y., (2024). A quality grade classification method for fresh tea leaves based on an improved YOLOv8x-SPPCSPC-CBAM model. *Scientific Reports*, Vol. 14, pp. 4166, United Kingdom.
[27] Zhu, H., Li, X., et al., (2022). Tea Bud Detection Based on Faster R-CNN Network (基于 Faster R-CNN 网络的茶芽检测). *Transactions of the Chinese Society for Agricultural Machinery*, Vol. 53, pp. 217-224, Beijing/China.