THE PUBLISHING HOUSE
OF THE ROMANIAN ACADEMY

PROCEEDINGS OF THE ROMANIAN ACADEMY, Series A
Volume 26, Number 3/2025, pp. 311–322
DOI: 10.59277/PRA-SER.A.26.3.11

SCL-YOLO: ENHANCED YOLOv11 ALGORITHM FOR MULTI-CLASS TRADITIONAL CHINESE MEDICINAL HERB DETECTION

Yuqi ZENG[1], Yongcheng ZHOU[2]

[1] Chengdu Seventh People's Hospital, 610200, China
[2] Chongqing University, School of Automation, 400044, China
Corresponding author: Yongcheng ZHOU, E-mail: zhouyongcheng@stu.cqu.edu.cn

Abstract. To address class imbalance, morphological diversity, and missed detections of small or densely distributed targets in complex backgrounds, this paper proposes SCL-YOLO, an efficient and accurate model for multi-class Chinese medicinal herb detection based on the lightweight YOLOv11n architecture. The C3k2_WTConv module expands receptive fields while controlling parameter growth via wavelet transform. LSK-attention enhances discriminative feature learning, and a BiFPN fused with the Semantic Consistency Enhancement Module (SDI) improves cross-scale feature interaction. Additionally, the C3k2_SCConv module reduces redundancy and strengthens the distinction between similar herbs. An Inner-CIoU loss function further improves localization accuracy by generating auxiliary bounding boxes. Evaluated on a custom dataset of 9,709 images across 50 categories, SCL-YOLO achieves 89.5% accuracy (+12.8%), 79.1% mAP_{50} (+7.1%), and a 5.1% increase in F_1 score, while reducing parameter count by 9.3% and boosting inference speed to 129.3 FPS (+3.6%). These results demonstrate that SCL-YOLO offers superior detection performance and generalization capability for complex herb recognition tasks.

Keywords: Chinese medicinal herb, YOLOv11 algorithm, object recognition, SCL-YOLO.

1. INTRODUCTION

Traditional Chinese Medicinal (TCM) herbs have been widely used for centuries as natural remedies in preventive and therapeutic healthcare. With the growing global demand for herbal products, the need for accurate and efficient identification of TCM herbs has become increasingly important. However, this task remains challenging due to significant morphological variability, class imbalance, and complex backgrounds in real-world scenarios – including dense clustering, occlusion, and inconsistent lighting. Moreover, traditional manual inspection is labor-intensive, time-consuming, and susceptible to human error, particularly when dealing with rare, fragmented, or visually similar specimens.

To address these challenges, computer vision-based detection systems have emerged as promising tools for automating herb recognition. Among these, deep learning models, particularly the You Only Look Once (YOLO) series, have demonstrated exceptional performance in real-time object detection tasks [1, 2]. YOLO's single-stage architecture balances speed and accuracy, making it suitable for applications requiring rapid processing, such as industrial quality control or large-scale herb sorting. However, deploying YOLO-based models for TCM herb detection remains challenging. Small herb sizes, occlusions, and intricate background textures in field or storage environments often lead to missed detections or false positives. Furthermore, the inherent class imbalance in herb datasets – where common species dominate over rare ones – exacerbates model bias, reducing generalization capabilities. Among single-stage object detection algorithms, the YOLO series stands out due to its excellent performance in real-time, end-to-end detection, making it well-suited for deployment on edge devices. Consequently, various YOLO-based models have emerged [3].

To address challenges in TCM herb detection, computer vision-based systems, especially YOLO models, have shown strong potential due to their balance of speed and accuracy [1–3]. Santos *et al.* [4] proposed KochDet, a BiFPN-based model for early tuberculosis diagnosis. Das *et al.* [5] compared YOLO versions v1–v7, demonstrating YOLOv7's effectiveness for medicinal leaf recognition. Zhang *et al.* [6] combined ResNet, CBAM, and Local Binary Pattern features with transfer learning, achieving 96.80% F1-score in Chinese herbal classification. Sathiya *et al.* [7] developed a hybrid soft computing framework with optimized

segmentation and an FDB-DNN classifier for real-time herb disease detection. Antunes *et al.* [8] built a YOLO-based model for aromatic herb identification, reaching precision above 0.7. Yang *et al.* [9] used graph convolutional networks with herb knowledge graphs, improving TCM prescription recommendations. Zhang *et al.* [10] proposed a lightweight CGC-YOLOv8 model by integrating conditional convolution, GSConv+VoVGSCSP slim-neck, and coordinate attention, achieving 85.70% precision, and 94.69% mAP50 in ginseng appearance quality recognition, while reducing parameters and FLOPs for mobile deployment. Kumar [11] showed CNNs outperform SVM and KNN in large-scale medicinal plant identification. Peng *et al.* [12] reviewed pesticide residue detection approachs in herbal medicines, emphasizing molecular and biosensor-based techniques. More recently, EAFNet [13] enhanced tiny crack detection with feature amplification and fusion, while [14] leveraged model compression and knowledge distillation for efficient edge-based pattern recognition.Together, these studies highlight the versatility of deep learning in herb identification, disease prevention, and quality assurance.

The motivation for SCL-YOLO arises from critical limitations of YOLOv11 in herb detection: its backbone struggles to capture fine-grained textures (e.g., root/stem patterns) under complex backgrounds; the standard feature pyramid loses details of small/dense herbs; and its loss function lacks discriminative power for morphologically similar species (e.g., Angelica sinensis vs. dahurica). These deficiencies impede robustness against class imbalance, occlusion, and scale variation, necessitating architectural enhancements to address domain-specific challenges.

To overcome these limitations, this study proposes SCL-YOLO, an enhanced YOLOv11-based framework specifically designed for multi-class TCM herb detection. The proposed model introduces four key innovations:

1) SCVanillaNet Backbone: We redesign the YOLOv11 backbone using SCVanillaNet, a lightweight structure that eliminates redundant skip connections and merges early dual convolutions at inference for better stability and efficiency. This reduces parameters by 9.3% while maintaining strong feature extraction. WIoUv3 Loss: For bounding box regression, WIoUv3 reweights gradients based on anchor quality, improving convergence and addressing class imbalance by suppressing outliers and emphasizing high-quality matches.

2) LSK-Attention: A large-kernel spatial attention module is integrated to enhance long-range dependency modeling. It decomposes large kernels into cascaded depthwise and dilated convolutions, enabling efficient capture of irregular herb structures with low computational cost.

3) BiFPN Neck: The original FPN+PAN is replaced by BiFPN with weighted bidirectional connections, improving multi-scale fusion and preserving fine details essential for small-object herb detection.

Compared with recent YOLO-based herb detection studies, these contributions represent several novel aspects. Unlike prior works that primarily focused on single-module improvements (e.g., attention-only mechanisms or loss-function refinements), our approach integrates four complementary innovations: a SCVanillaNet backbone with C3k2-WTConv for receptive field expansion, an LSK-Attention module for long-range spatial dependency modeling, a BiFPN with Semantic Decoupling Integration for enhanced cross-scale fusion, and an Inner-CIoU loss for precise localization of dense and morphologically similar herbs. Together, these advancements enable SCL-YOLO to achieve robust performance under class imbalance, occlusion, and scale variation, which were not effectively addressed in recent related manuscripts.

The remainder of this paper is organized as follows. Section 2 introduces the baseline YOLOv11 architecture. Section 3 presents the proposed SCL-YOLO model, including the WTConv module, LSK-attention mechanism, and Inner-CIoU loss function. Section 4 describes the experimental setup, ablation studies, and performance comparisons. Finally, Section 5 concludes the paper and outlines potential directions for future research.

2. MODEL ARCHITECTURE AND PRINCIPLES

The YOLO (You Only Look Once) algorithm, proposed by Redmon et al. in 2016 [15], was designed to address the trade-off between real-time performance and detection accuracy in the field of object detection. Through continuous iterative optimization, the YOLO series has achieved outstanding performance in computer vision. Among its latest iterations, YOLOv11 stands out by combining high recognition accuracy with rapid detection speed. Compared to YOLOv8, YOLOv11 demonstrates improvements in detection accuracy,

processing efficiency, and parameter optimization, making it better suited for diverse data scales and application scenarios. In the Backbone of YOLOv11, the newly introduced C3k2 structure replaces the C2f module from YOLOv8. This architecture enhances overall model performance through deeper feature extraction and optimized information flow. Additionally, the innovative C2PSA (Context-Aware Position-Sensitive Attention) mechanism embeds multi-head attention modules within the Backbone, further strengthening the network's focus on critical regions and improving detection accuracy in complex scenes. Compared to YOLOv8, YOLOv11 adopts depthwise separable convolutions (DWConv) in its decoupled Head structure, significantly reducing computational complexity and parameter count while maintaining high accuracy, thereby boosting computational efficiency. Meanwhile, YOLOv11 retains the CIoU [16] and Distribution Focal Loss (DFL) loss functions, while introducing Binary Cross-Entropy (BCE) for classification tasks. Through an improved task-aligned assignment strategy [17], it enables more flexible dynamic sample allocation during detection, further enhancing model robustness [18].

In multi-category detection tasks for 50 types of Chinese medicinal materials, it is essential to balance fine-grained classification accuracy and robustness to complex backgrounds. However, YOLOv11 exhibits the following limitations in practical applications: First, its backbone network demonstrates limited capability in capturing the unique texture features of Chinese medicinal materials (e.g., root/stem patterns and leaf venation), as standard convolutional modules struggle to effectively extract local discriminative features from medicinal materials. Second, Chinese medicinal images often suffer from uneven illumination, stacking/occlusion, and morphological similarities (e.g., Angelica sinensis vs. Angelica dahurica), while existing feature pyramid structures tend to lose detailed information of small-scale targets (e.g., wolfberry, chrysanthemum) during cross-scale feature fusion. Additionally, traditional loss functions lack sufficient discrimination for inter-class similarities among multi-category medicinal materials (e.g., Fritillaria vs. Lilium), resulting in blurred classification boundaries. To address these issues, the model requires integration of attention mechanisms to enhance local feature responses, reconfiguration of multi-scale feature fusion pathways to adapt to varying sizes of medicinal materials, and development of more discriminative metric learning loss functions.

3. SCL-YOLO MODEL

To enhance the detection accuracy of multi-category Chinese medicinal materials images, an improved SCL-YOLO model is proposed. In the backbone network, the C3k2_WTConv module replaces the original C3k2 structure by substituting traditional convolutions with Wavelet Transform Convolution (WTConv [19]). This modification leverages wavelet transforms to expand the receptive field while minimizing parameter growth. Additionally, the Large Kernel-Attention (LSK-attention [20]) mechanism is integrated to dynamically adjust weights of large convolutional kernels, thereby expanding the receptive field and suppressing background noise interference simultaneously. The neck architecture innovatively combines Bidirectional Feature Pyramid Network (BiFPN [21]) with Semantic Decoupling Integration (SDI [22]) module. BiFPN strengthens cross-scale feature interactions through bidirectional connections across different levels, while SDI enhances semantic consistency via hierarchical feature map fusion. Furthermore, the C3k2_SCConv module is introduced, incorporating a spatial-channel dual reconfiguration mechanism that reduces computational redundancy and significantly improves discrimination efficiency for similar medicinal materials [23]. For loss function optimization, the Inner_CIoU [24] loss is adopted to enable dynamic gradient allocation, with a focus on refining anchor box regression accuracy for easily confused samples. The structural design of the SCL-YOLO model is illustrated in Figure 3.

3.1. C2f-Wavelet Transform Convolution (C3k2-WTC)

In the field of object detection, the C3k2 module serves as the core feature extraction component in the YOLOv11 architecture, enabling deep integration of multi-level features through its cross-stage partial connection mechanism. However, in Chinese medicinal material detection scenarios, the C3k2 module exhibits limitations in feature extraction due to its local convolution operations, particularly when processing complex

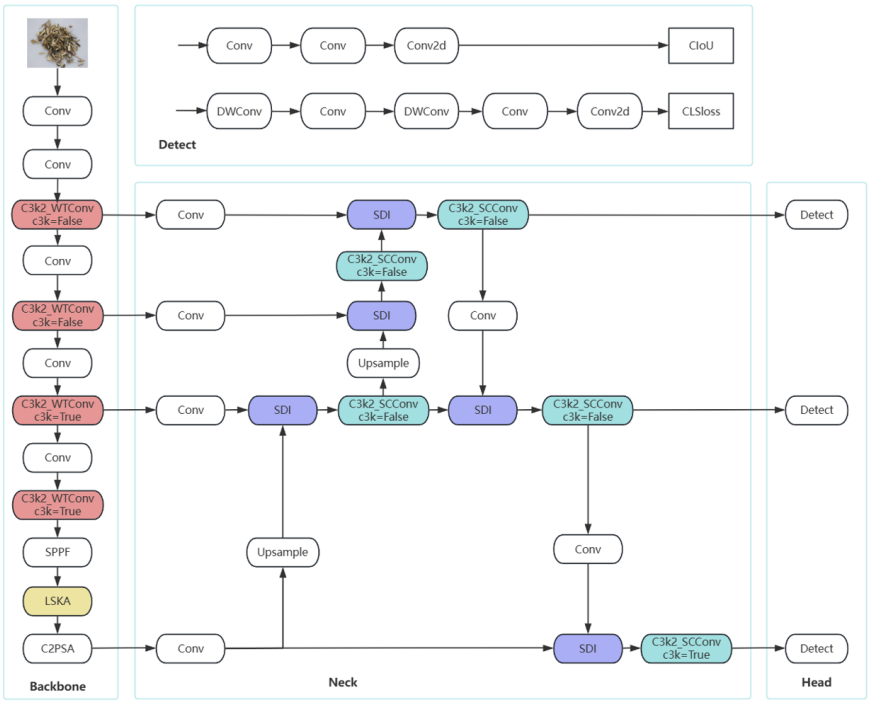

Fig. 1 – SCL-YOLO network structure.

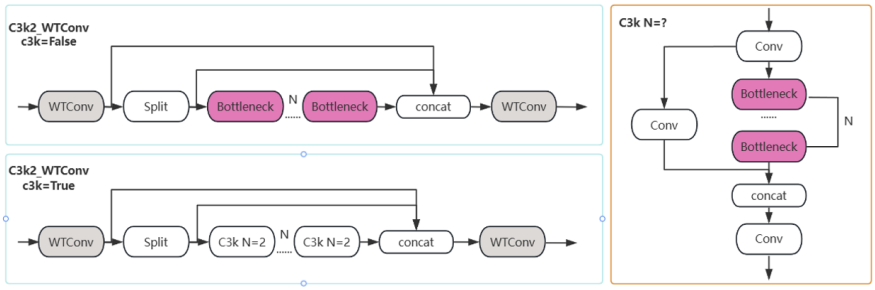

Fig. 2 – C3k2-WTC Network structure.

images. Meanwhile, the module's parameter count demonstrates a quadratic relationship with the convolutional kernel size, leading to increased computational resource consumption and compromised model generalization performance [25]. To address these challenges, the C3k2-WTC (Wavelet Transform Convolution) module is introduced, which utilizes wavelet transforms to enhance the receptive field while minimizing parameter growth. Utilizing the simplicity of Haar wavelets, an image X of size $N \times N$ is decomposed into four subbands: LL (low-frequency), LH, HL, and HH (high-frequency). The transformation process focuses on extracting relevant features at different scales. The Haar wavelet transform can be expressed as:

$$\begin{cases} \mathbf{LL}_{i+1} = \frac{1}{2}\left(x_{2i,2j} + x_{2i,2j+1} + x_{2i+1,2j} + x_{2i+1,2j+1}\right), \\ \mathbf{LH}_{i+1} = \frac{1}{2}\left(x_{2i,2j} - x_{2i,2j+1} + x_{2i+1,2j} - x_{2i+1,2j+1}\right), \\ \mathbf{HL}_{i+1} = \frac{1}{2}\left(x_{2i,2j} + x_{2i,2j+1} - x_{2i+1,2j} - x_{2i+1,2j+1}\right), \\ \mathbf{HH}_{i+1} = \frac{1}{2}\left(x_{2i,2j} - x_{2i,2j+1} - x_{2i+1,2j} + x_{2i+1,2j+1}\right), \end{cases} \tag{1}$$

where i denotes the current decomposition level. By recursively applying the Haar wavelet transform to the LL subband, we create a cascade of wavelet transformations that effectively capture both low-frequency and high-frequency components of the input image, thereby facilitating detailed feature extraction.

Fig. 3 – The process and structure of the LSKAmechanism.

In the WTConv methodology, the input X undergoes transformation via wavelet techniques, followed by convolution operations with a small kernel W applied to each frequency subband. This process is represented as:

$$\mathbf{Y} = \mathrm{IWT}(\mathbf{W} * \mathbf{X}_{\mathrm{subband}}) \tag{2}$$

where IWT (Integer Wavelet Transform) denotes the inverse wavelet transform, recombining the transformed frequency components back into the spatial domain and $*$ denotes the convolution operation performed on each wavelet-transformed frequency subband. This technique enables the model to effectively integrate multi-scale features, enhancing its sensitivity to both fine and coarse details.

3.2. Introducing a dynamic large-kernel attention mechanism

The adaptive kernel selection technique elfiectively enhances the model's ability to focus on contex-tual regions, serving as a benelicial complement to channel or spatial attention mechanisms [20]. Given theirregular, elongated shapes of medicinal herbs, a dynamic large convolution kernel selection mechanism,LSK-attention, is incorporated into the backbone network of the original model. LSK-attention captures extensive contextual information from the image by utilizing large andseparable convolutional kernels along with spatial dilated convolutions [26]. This approach generates anattention map that highlights important regions, which is then used to weight the original features.allowing the network to focus more on critical features, This process enhances the model's performance by strengthening its focus on essential characteristics, The process and structure of the LSKAmechanism are illustrated in Figure 5.

3.3. Neck optimization for multi-scale feature fusion

To address the issues of shallow detail loss and deep semantic misalignment in traditional feature fusion networks for Chinese medicinal material detection tasks, an optimized neck structure scheme based on Bidirectional Feature Pyramid Network (BiFPN) and Semantic Decoupling Integration (SDI) module is proposed. This approach constructs a multi-level feature interaction network through a bidirectional cross-scale connection architecture. While maintaining the original topological structure of the feature pyramid, it introduces a learnable feature weight allocation mechanism to enable dynamic fusion of features across different scales.

To improve cross-scale connectivity, we apply three optimizations: (1) removing single-input edge nodes that lack fusion capability; (2) introducing skip connections between same-level input-output nodes to enhance interaction with minimal cost; and (3) abstracting each bidirectional path into a stackable feature layer for deep fusion. As shown in Fig. 6, these strategies collectively enhance feature propagation and utilization efficiency. To mitigate semantic inconsistency in multi-scale fusion, we introduce the Hierarchical Semantic Enhancement (SDI) module. It adopts a dual-path design: the low-level branch uses 3×3 convolutions for spatial-aware processing to retain fine-grained morphology, while the high-level branch leverages SE attention to enhance semantic selectivity. During integration, low-level features ($l = 2$) are upsampled via bilinear interpolation, and high-level features ($l = 4$) are downsampled via max-pooling, with residual identity mapping preserved. As shown in Fig. 8, SDI effectively aligns semantic representations across scales.

3.4. Loss function improvement

To meet the robustness requirements of traditional Chinese medicinal herb detection, the loss function is optimized. The original YOLOv11 model uses CIoU Loss as the bounding box regression loss function, which

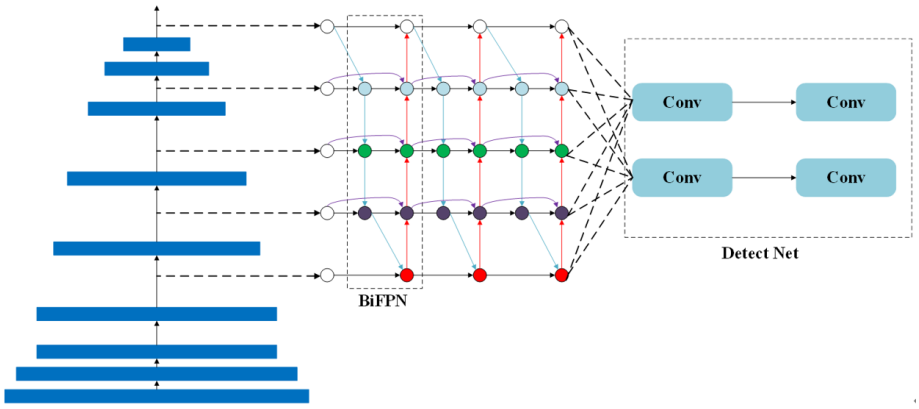

Fig. 4 – BiFPN multi-scale fusion structure.

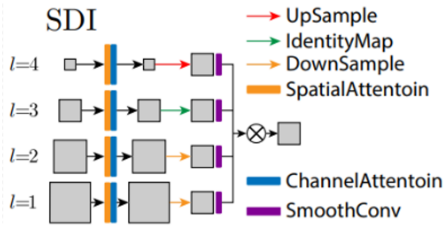

Fig. 5 – SDI network structure.

does not adequately balance difficult and easy samples. In the proposed improvement, an additional aspect ratio penalty term is introduced as part of the loss function to better reflect the true differences between predictions and ground truth when aspect ratios differ significantly. This adjustment increases the model's computational overhead slightly but improves localization accuracy and robustness. The improved CIoU Loss function is given in Eq. (3):

$$\begin{cases} L_{\mathrm{CIoU}} = L_{\mathrm{IoU}} + \dfrac{(x-x^{gt})^2+(y-y^{gt})^2}{(w^{gt})^2+(h^{gt})^2} + \partial v, \\ L_{\mathrm{IoU}} = 1-\mathrm{IoU} = 1-\dfrac{\mathrm{Intersection}}{\mathrm{Union}} = 1-\dfrac{W^i H^i}{w^{gt}h^{gt}+wh-W^iH^i}, \\ \partial = v/L_{\mathrm{IoU}}+v, \\ v = \dfrac{4}{\pi^2}\left(\tan^{-1}\dfrac{w}{h}-\tan^{-1}\dfrac{w^{gt}}{h^{gt}}\right)^2 \end{cases} \tag{3}$$

In Eqs. (3), L_{IoU} represents the overlap between the predicted and ground truth boxes, ∂ is a balance parameter, and v denotes the aspect ratio consistency. Other parameters are defined as shown in Fig. 10.

Existing IoU-based regression approaches primarily aim to accelerate convergence through additional loss terms, yet overlook the inherent limitations of IoU itself. In traditional Chinese medicinal herb detection, the fixed regression strategy fails to adapt to different detectors or tasks, resulting in reduced accuracy under dense small-target or complex scene conditions.

To address this issue, we introduce the Inner-CIoU loss function. Based on the original CIoU loss function, Inner-CIoU incorporates an additional aspect ratio control factor, ratio, to adjust the size of auxiliary bounding boxes. This factor varies according to the IoU and scale of the detection targets, applying different auxiliary boxes for calculating loss based on target size and scale. This approach accelerates convergence, produces more efficient regression results, and improves the accuracy of predicted bounding boxes. The definition of

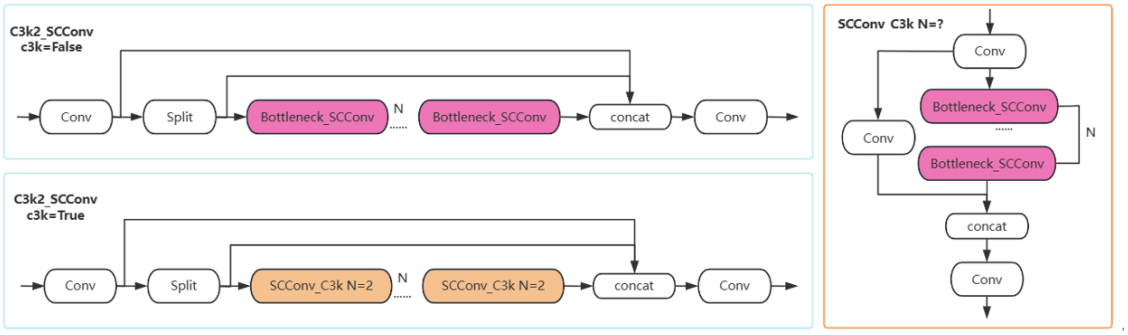

Fig. 6 – Schematic diagram of the meaning of loss function parameters.

Inner-IoU is provided in Eqs. (4).

$$\begin{cases} b_{gt}^{l} = x_{gt}^{c} - \frac{w^{gt}\cdot\rho}{2}, & b_{gt}^{r} = x_{gt}^{c} + \frac{w^{gt}\cdot\rho}{2}, \\ b_{gt}^{t} = y_{gt}^{c} - \frac{h^{gt}\cdot\rho}{2}, & b_{gt}^{b} = y_{gt}^{c} + \frac{h^{gt}\cdot\rho}{2}, \\ b_{l} = x_{c} - \frac{w\cdot\rho}{2}, & b_{r} = x_{c} + \frac{w\cdot\rho}{2}, \\ b_{t} = y_{c} - \frac{h\cdot\rho}{2}, & b_{b} = y_{c} + \frac{h\cdot\rho}{2}. \end{cases} \tag{4}$$

In Eqs. (4), x represents element-wise multiplication (Hadamard product), which is used to scale the width and height components of the bounding box coordinates by the factor ρ.

$$\text{Intersection}_{\text{inner}} = \left(\min(b_r^{gt}, b_r) - \max(b_l^{gt}, b_l)\right) \times \left(\min(b_b^{gt}, b_b) - \max(b_t^{gt}, b_t)\right) \tag{5}$$

$$\text{Union}_{\text{inner}} = \left(w^{gt}\cdot h^{gt}\cdot\rho^2\right) + \left(w\cdot h\cdot\rho^2\right) - \text{Intersection}_{\text{inner}} \tag{6}$$

$$\text{IoU}_{\text{inner}} = \frac{\text{Intersection}_{\text{inner}}}{\text{Union}_{\text{inner}}}. \tag{7}$$

The variable ratio corresponds to a scaling factor, typically ranging from $[0.5, 1.5]$. Traditional IoU calculations consider the entire overlap region between the predicted and ground truth bounding boxes, whereas InnerIoU focuses on the core part of the bounding box to make a more precise overlap judgment.

Compared to the IoU loss, when ratio < 1, the auxiliary bounding box is smaller than the actual bounding box, meaning its effective area is smaller than that of IoU loss. However, its gradient magnitude is absolutely larger than that of IoU loss, which can accelerate convergence for IoU samples. In contrast, when ratio > 1, the auxiliary bounding box is larger, expanding the effective area of the regression and providing benefits for low-IoU regression. For the dataset used in this paper, the ratio value is set to 1.2. The Inner-IoU is applied to the original model loss function, as defined in Eq. (8):

$$L_{\text{Inner-CIoU}} = L_{\text{CIoU}} + \text{IoU} - \text{IoU}_{\text{inner}}. \tag{8}$$

The extracted features are divided into a classification branch and a localization branch. The classification path is supervised by the Binary Cross-Entropy (BCE) loss, while the localization branch is optimized using the Inner-CIoU loss. This decoupled head design enables task-specific learning during training.

To clarify the role of the classification algorithm in the SCL-YOLO framework, we describe its integration with the detection head and loss functions. The BiFPN-SDI neck provides semantically consistent features to a decoupled head, where the classification branch predicts categories using BCE loss and the regression branch refines bounding boxes with Inner-CIoU loss. This design links feature extraction, classification, and localization for robust herb detection.

3.5. Overall training pipeline

To provide a comprehensive overview of the proposed SCL-YOLO framework, we summarize the entire training procedure in Algorithm 1. The pipeline integrates our four major innovations, aimed at addressing the challenges of class imbalance, morphological diversity, and dense small-object distribution in traditional Chinese medicinal herb detection. Specifically, the SCVanillaNet backbone combined with the C3k2-WTConv module expands the receptive field via wavelet transforms, enabling effective capture of both coarse and fine-grained features while controlling parameter growth. The LSK-Attention mechanism further enhances the ability to model long-range dependencies, allowing the network to emphasize irregular herbal structures and suppress background noise. At the neck stage, a BiFPN equipped with the Semantic Decoupling Integration (SDI) module ensures more effective cross-scale feature fusion, improving semantic consistency and enhancing the representation of small and densely distributed herbs. Finally, the Inner-CIoU loss introduces auxiliary bounding box constraints to optimize localization accuracy, especially for morphologically similar classes.

Algorithm 1 Training Pipeline of the Proposed SCL-YOLO

Initialize model parameters θ
for epoch $= 1$ to E **do**
  for each mini-batch $\{(x_i, y_i)\}_{i=1}^{N}$ from $\mathscr{D}$ **do**
    Extract features F using SCVanillaNet backbone with C3k2-WTConv:
    $Y = IWT(W * X_{\text{subband}})$
    Apply LSK-Attention to obtain weighted feature maps F_{att}
    Fuse cross-scale features with BiFPN-SDI:
    $F_{fused} = \text{BiFPN-SDI}(F_{att})$
    Predict classification scores $\hat{y}$ and bounding boxes $\hat{b}$
    Compute classification loss with BCE:
    $\mathscr{L}_{cls} = -\sum_{c=1}^{C} [y_c \log \hat{y}_c + (1 - y_c) \log(1 - \hat{y}_c)]$
    Compute localization loss with Inner-CIoU:
    $\mathscr{L}_{loc} = L_{CIoU} + IoU - IoU_{inner}$
    $IoU_{inner} = \frac{Intersection_{inner}}{Union_{inner}}$
    Total loss:
    $\mathscr{L}_{total} = \mathscr{L}_{cls} + \mathscr{L}_{loc}$
    Update parameters:
    $\theta \leftarrow \theta - \eta \nabla_{\theta} \mathscr{L}_{total}$
  end for
end for
return θ

As illustrated in Algorithm 1, the overall pipeline explicitly demonstrates how the proposed modules co-operate within the SCL-YOLO framework, ensuring both accurate classification and robust localization. This structured pseudocode improves the clarity and reproducibility of our approach, directly addressing the reviewer's suggestion.

To validate the proposed modules, we conducted experiments on the TCM herb dataset. Section 4 presents the settings, ablation studies, and comparative evaluations to demonstrate the effectiveness of SCL-YOLO.

4. EVALUATION AND RESULTS

4.1. Comparison with mainstream algorithms

In this section, we describe the experimental setup, evaluation metrics, and results, providing evidence for the effectiveness of the proposed SCL-YOLO framework. The experiment utilized a self-built TCM dataset containing 9,709 images across 50 herb types (e.g., wolfberry, ginseng, honeysuckle). All images were resized

to 416×416 due to computational constraints. The dataset was randomly split into training, validation, and test sets with a 7:2:1 ratio; the test set was excluded from training. Four data augmentation techniques were applied to improve generalization: image stitching (40%), flipping (20%), brightness/contrast adjustment (20%), and motion blur (20%). Training was conducted over 100 epochs with a batch size of 64, learning rate of 0.001, and 4 data loader workers. Mosaic augmentation was disabled in the final 10 epochs. To address the sensitivity of the proposed approach to parameter settings, a two-stage hyperparameter tuning strategy was adopted. In the coarse search stage, we explored key parameters within broad ranges: WTConv kernel size (3–9), LSK-attention dilation rate (1–4), BiFPN depth (2–5 layers), and Inner-CIoU ratio (0.8–1.5). In the fine-tuning stage, smaller increments were applied within these ranges. The final settings – WTConv kernel size = 5, LSK dilation rate = 2, BiFPN depth = 3 layers, and Inner-CIoU ratio = 1.2 – were selected based on the highest mAP on the validation set while preventing overfitting. This ensured stable convergence and a balanced trade-off between accuracy and efficiency. To ensure fairness, all baseline models and the proposed SCL-YOLO were trained and evaluated under identical conditions, using the same dataset of 9,709 herb images (50 categories). The training setup (416×416 input size, identical augmentations, batch size 64, learning rate 0.001, Adam optimizer) was kept consistent so that performance gains stem solely from the proposed architectural enhancements.

To further address the reviewer's concern regarding the connection between the literature review and our experiments, we additionally compared our approach with several representative approaches discussed in introdction. The results are reported in Table 1.

Table 1

Comparison with representative approaches discussed in the literature review

Model	Precision (%)	Recall (%)	mAP50 (%)	mAP50:95 (%)	FPS	FLOPs (G)	Params (MB)
Das et al [5]	45.6	36.5	35.8	26.7	161	8.2	3.1
Zhang et al [6]	44.8	35.6	34.4	25.5	102	11.2	4.6
Sathiya et al [7]	43.9	35.2	33.7	24.9	83	14.3	5.8
Antunes et al [8]	46.1	37.7	36.8	28.1	120	8.1	3.2
Yang et al [9]	44.6	35.8	34.6	25.4	65	15.4	6.5
Zhang et al [10]	47.1	39.6	39.8	27.3	181	8.2	3.6
Ours	**50.3**	**42.2**	**43.5**	**32.6**	**215**	**6.5**	**2.64**

As shown in Table 1, our approach achieves the best performance across all metrics, particularly improving mAP50:95 and FPS while reducing FLOPs and parameter size, demonstrating its robustness and efficiency compared with representative approaches from the literature review.

4.2. Ablation study

Based on the proposed improvements, the C3k2_WTConv module and LSK-attention were introduced into the backbone to expand the receptive field and control parameter growth. In the neck, BiFPN with Semantic Decoupling Integration (SDI) enhanced multi-scale feature fusion, while the C3k2_SCConv module reduced redundancy and improved recognition of similar herbs. The loss function was further optimized. The improved SCL-YOLO was compared with YOLOv11n under identical settings. As shown in Fig. 21, SCL-YOLO achieved faster convergence, lower stabilized loss, and earlier training stability. Further evaluations (Fig. 22) confirmed higher Precision, Recall, and mAP, with P–R curves (Fig. 23) closer to the top-right corner, indicating stronger detection performance. The model reached stable convergence at ≈ 65 epochs, where validation loss and mAP plateaued. The SCL-YOLO model are available in the images folder of the supplementary repository: `https://github.com/zhouyongcheng024/SCL-YOLO`.

To assess the contribution of each component, ablation studies were conducted with YOLOv11n as the baseline. Parameters were computed via the PyTo rch summary tool, and results are reported in Table 2, showing that each module incrementally improved Precision, Recall, and mAP, with the full model achieving the best performance.

In addition to accuracy and recall, we further evaluated the cost-effectiveness of SCL-YOLO. The backbone with WTConv and LSK-attention has an asymptotic complexity of $O(N \times k^2)$, where N is the number of input

features and k the kernel size. Compared with YOLOv11n, our model reduces parameters by 9.3% while maintaining higher accuracy. On an NVIDIA RTX 4060 Laptop GPU, it processes a 416×416 image in 7.7 ms (129.3 FPS), indicating that the proposed improvements achieve a favorable trade-off between accuracy and efficiency for real-time deployment scenarios.

Table 2

Ablation study results of different modules on model performance

Module Configuration	**Precision (%)**	**Recall (%)**	**mAP (%)**
Baseline	85.2	78.6	82.1
+ LSKA	87.4	80.3	84.5
+ BiFPN + SDI	88.1	81.7	85.8
+ C3K2_SCConv	89.0	82.9	86.7
Full Model (All)	**90.5**	**84.2**	**88.1**

The analysis revealed that each improvement point effectively enhanced the overall performance of the model. The LSKA module significantly improved precision and other indicators, the BiFPN+SDI module enhanced detection effect by fusing multiple features, and the C3k2_SCConv module ensured a balanced improvement in the model. The combined effect of all improvement strategies was significant.

These module-wise improvements are further illustrated by heatmaps and visualizations in Figs. 26–30, where LSKA and BiFPN enhance focus on key herb structures (e.g., ginseng roots, wolfberry clusters). In dense or occluded scenes, they effectively reduce false positives, demonstrating that our improvements not only boost metrics but also better capture herbal features in complex conditions.

To validate the superiority of the improved model compared to other mainstream algorithms, a unified dataset was used to compare the SCL-YOLO model with various classical models. First, the changes in the loss function during the training process of SCL-YOLO and other mainstream algorithms were analyzed, with the comparison results shown in Figure 24.png. During the training process, SCL-YOLO demonstrated a faster convergence speed, a significantly lower loss function value after stabilization, and quicker model convergence, proving the significant effect of the loss function improvement in SCL-YOLO. Further comparisons of precision, recall, and mAP indicators were conducted to validate the overall performance optimization. As shown in Figures 24.png and 25.png, the SCL-YOLO model exhibited high performance in all main indicators among models of similar scale. The evaluative results indicated that the improved model had significant improvements in accuracy, recall, mAP, and other key indicators, outperforming YOLOv11n and other mainstream object detection models, and demonstrating higher robustness and reliability in data.

After analyzing the evaluative indicators, to further observe the optimization effects of each improvement on the model, a visualization analysis of key evaluative aspects was conducted to explore the specific optimization effects brought about by the relevant improvements. First, to validate the optimization of the receptive field by the w3k2_WTconv module and the dynamic large-kernel attention mechanism, heatmaps of the model's convolution process were exported for comparison, as shown in Figure 26.png. The color characteristics of different regions in the images in Fig. 6 reflect the model's attention to targets before and after improvement. The SCL-YOLO model had a more complete perception range and more concentrated attention on different types of TCM targets, indicating the good effect of the improved w3k2_WTconv module and dynamic large-kernel attention mechanism. Next, to validate the optimization of the neck structure by multi-scale feature fusion, the network features in the detection layer before and after improvement were visualized, as shown in Figures 27.png to 32.png. The 80×80, 40×40, and 20×20 detection layers of the TCM detection model before and after improvement are visualized. From subfigures Figure 27.png, Figure 29.png, and Figure 31.png, it can be seen that the target information in the feature detection layers before improvement was not rich and complete enough and was greatly affected by background interference. The SCL-YOLO model, which fused multi-scale feature semantic information, had more distinct target feature semantic information and background region segmentation, as shown in subfigures Figure 28.png, Figure 30.png, and Figure 32.png, with richer semantic information and feature representation. To more intuitively demonstrate the detection effect of the SCL-YOLO

model and validate its practical engineering application effect, YOLOv11n and the improved SCL-YOLO were used to detect TCM targets in a unified test set, with detection results in different scenarios shown in Figure 33.png.The improved algorithm improved the missed detection and false detection problems of the original YOLOv11n model. According to the practical application comparison, the SCL-YOLO model effectively improved missed and false detection phenomena, effectively validating the detection effect and generalization ability of the improved model. It meets the practical application requirements for TCM detection in different scenarios.

5. CONCLUSION

To address class imbalance and morphological diversity in TCM herb detection, this study proposed the SCL-YOLO model, which achieves notable performance gains through multi-level architectural innovations. The C3k2_WTConv module, based on wavelet transforms, enhances texture feature extraction, while the LSK attention mechanism improves discriminative capability via adaptive receptive fields. The BiFPN-SDI fusion structure strengthens multi-scale feature interaction, enabling accurate detection of densely distributed small herbs. Evaluation on a 50-class dataset shows significant improvements in accuracy and generalization. SCL-YOLO demonstrates strong practical value under challenging conditions such as lighting variation and partial occlusion, supporting the digitalization of TCM. Despite the improved performance of SCL-YOLO, several limitations remain. Detection accuracy may decline for rare herb classes due to limited samples, suggesting the need for data augmentation or semi-supervised learning. Although the model achieves high accuracy and speed, its GPU memory requirements may hinder deployment on ultra-low-power edge devices. Moreover, being trained on a static dataset, the model may struggle with varying lighting and camera conditions in real-world applications. Future work will explore domain adaptation, continual learning, and knowledge distillation to enhance robustness and adaptability.

REFERENCES

[1] Tan S, Lu G, Jiang Z, et al. Improved YOLOv5 network model and application in safety helmet detection. In: 2021 IEEE International Conference on Intelligence and Safety for Robotics (ISR). IEEE; 2021, pp. 330–333.

[2] Tan S, Yan J, Jiang Z, et al. Approach for improving YOLOv5 network with application to remote sensing target detection. Journal of Applied Remote Sensing. 2021; 15(3): 036512.

[3] Chen H, Tan S, Xie Z, et al. A new approach based on YOLOv5 for remote sensing object detection. In: 2022 China Automation Congress (CAC). IEEE; 2022, pp. 605–610.

[4] Dos Santos P R S, de Carvalho Brito V, de Carvalho Filho AO, et al. KochDet: BiFPN-based deep architecture for tuberculosis diagnosis. Biomedical Signal Processing and Control. 2024; 91: 106056.

[5] Das S, Chatterjee M, Stephen R, et al. Unveiling the potential of YOLO v7 in the herbal medicine industry: A comparative examination of YOLO models for medicinal leaf recognition. Int. J. Eng. Res. Technol. 2024; 13(11).

[6] Zhang C, Pan X, Jiang Y. Attention-enhanced ResNet and feature fusion for Chinese herbal classification. In: Proc. 2023 4th Int. Symp. Artif. Intell. Med. Sci. 2023.

[7] Sathiya V, Josephine MS, Jeyabalaraja V. An automatic classification and early disease detection technique for herbs plant. Comput. Electr. Eng. 2022; 100: 108026.

[8] Antunes SN, et al. Model development for identifying aromatic herbs using object detection algorithm. AgriEngineering. 2024; 6(3): 1924–1936.

[9] Yang Y, et al. Multi-layer information fusion based on graph convolutional network for knowledge-driven herb recommendation. Neural Netw. 2022; 146: 1–10.

[10] Zhang LJ, et al. Accurate recognition of ginseng appearance quality based on improved YOLOv8n lightweighting. Trans. Chin. Soc. Agric. Eng. 2024; 40(24): 274–282.

[11] Kumar P, Kumar V. CNN and edge-based segmentation for the identification of medicinal plants. In: Proc. 2024 5th Int. Conf. Intell. Commun. Technol. Virtual Mobile Netw. (ICICV). IEEE; 2024.

[12] Peng B, Xie Y, Lai Q, et al. Pesticide residue detection technology for herbal medicine: current status, challenges, and prospects. Anal. Sci. 2024; 40(4): 581–597.

[13] Zhou Z, Zhao W, Song K, et al. EAFNet: Extraction-amplification-fusion network for tiny cracks detection. Eng. Appl. Artif. Intell. 2024; 134: 108691.

[14] Jing N. Neural network-based pattern recognition in the framework of edge computing. Sci. Technol. 2024; 27(1): 106–119.

[15] Redmon J. You only look once, unified, real-time object detection. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. Las Vegas. IEEE; 2016, pp. 779–788.

[16] Zheng Z, Wang P, Liu W, et al. Distance-IoU loss: Faster and better learning for bounding box regression. In: Proceedings of the AAAI Conference on Artificial Intelligence. 2020, 34(07): 12993–13000.

[17] Feng C, Zhong Y, Gao Y, et al. Tood: Task-aligned one-stage object detection. In: 2021 IEEE/CVF International Conference on Computer Vision (ICCV). IEEE Computer Society; 2021, pp. 3490–3499.

[18] Aristoteles A, Syarif A, Sutyarso S, Lumbanraja FR. Identification of human sperm based on morphology using the you only look once version 4 algorithm. International Journal of Advanced Computer Science and Applications (IJACSA). 2022; 13(7): 424–431.

[19] Finder SE, Amoyal R, Treister E, et al. Wavelet convolutions for large receptive fields. In: European Conference on Computer Vision. Cham: Springer Nature Switzerland; 2024, pp. 363–380.

[20] Lau KW, Po LM, Rehman Y A U. Large Separable Kernel Attention: Rethinking the Large Kernel Attention design in CNN. Expert Systems with Applications. 2024; 236: 121352.

[21] Tan M, Pang R, Le QV. EfficientDet: scalable and efficient object detection. In: Proceedings of the IEEE/ CVF Conference on Computer Vision and Pattern Recognition. Piscataway. IEEE; 2020, pp. 10781–10790.

[22] Peng Y, Chen DZ, Sonka M. U-Net v2: Rethinking the skip connections of U-Net for medical image segmentation. In: 2025 IEEE 22nd International Symposium on Biomedical Imaging (ISBI). IEEE; 2025.

[23] Li J, Wen Y, He L. SCConv: spatial and channel reconstruction convolution for feature redundancy. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. 2023; pp. 6153–6162.

[24] Zhang H, Xu C, Zhang S. Inner-IoU: more effective intersection over union loss with auxiliary bounding box. arXiv preprint. 2023; arXiv:2311.02877.

[25] Shi Z, Yin Z, Chang S, et al. Efficient oriented object detection with enhanced small object recognition in aerial images. arXiv preprint. 2024; arXiv:2412.12562.

[26] Li Y, Hou Q, Zheng Z, et al. Large Selective Kernel Network for remote sensing object detection. arXiv preprint. 2023; arXiv:2303.09030.

Received June 15, 2025