ingegneria sismica
International Journal of Earthquake Engineering
Anno XLIII - Num. 3 - 2026

Counting Optimization of a Spatio-Temporally Decoupled YOLOv8 Model in Scenes with Dense Pods

Han Yang[1,*]

[1] School of Liberal Arts and Sciences, Northeast Agricultural University, Harbin 150030, Heilongjiang, China

SUMMARY: *To improve counting accuracy in dense soybean pod scenes under small-object occlusion, overlap, and repeated response conditions, this study proposes a spatiotemporal feature-decoupled improved YOLOv8 model that differs from detection-then-tracking counting methods. In this paper, "spatiotemporal decoupling" is defined as encoding soybean pod boundaries, contour gradients, and neighborhood occlusion relationships within the detection network as spatial structural representations, while encoding cross-frame center displacement, scale fluctuation, and short-term visibility variation as temporal association representations. Before the detection head, gated fusion is used to calibrate candidate box confidence and constrain counting bias. Unlike post-processing methods such as DeepSORT and ByteTrack, which rely on detection results for trajectory association, the temporal branch in the proposed method directly participates in candidate generation, candidate filtering, and quantity regression, allowing dense target responses to be corrected before NMS over-suppression and short-term missed detections occur. To address the susceptibility of conventional YOLOv8 to single-frame texture interference, weakened slender pod boundaries, and candidate drift in highly overlapping regions, the model constructs a spatial structural branch and a temporal association branch, and further introduces a P2 fine-grained fidelity branch, multi-scale semantic fusion, candidate target filtering constraints, and repeated-counting and missed-counting bias correction methods. On this basis, the model establishes a joint optimization strategy using localization loss, quantity regression loss, and temporal consistency loss. Experimental results show that the improved model achieves MAE/RMSE/F1 values of 4.2/6.8/0.91, 3.1/5.0/0.94, and 6.4/8.9/0.88 on the self-built soybean field dataset, PlantCrop subset, and occlusion-enhanced synthetic sequence, respectively, significantly reducing counting errors compared with the YOLOv8n baseline.The model operates at 51.7 FPS with a single-frame inference time of 19.3 ms on an NVIDIA RTX 4090 platform, meeting the real-time requirements of field counting.*

KEYWORDS: *pod counting; YOLOv8; spatio-temporal feature decoupling; small object detection; counting optimization*

1 Introduction

Soybean pod count is an important trait for yield prediction, selection, and in-field monitoring. Soybean pods in-field typically feature long edges, similar appearance, local overlap, and clustering. In high-resolution images, the effective width of an individual pod can be shrunk to a small number of pixels, and the center-to-center distances between nearby pods can be

*zoe_y12131419@163.com
https://doi.org/10.65102/is20261241
Published: 10 June , 2026

very short, which may cause overlapping candidate boxes, incorrect NMS suppression, and multiple responses. As such, soybean pod counting is not a typical object detection problem, but a counting optimization problem that is constrained by small-object detection, instance discrimination, and short-term temporal consistency.

The current research has primarily improved the detection of crop organs and small objects in three ways. The first direction is to enhance YOLO models for crop phenotypic objects (soybean pods, flowers, and seeds). Jia et al. (2025) developed a modified version of YOLOv8 for soybean pod detection and counting, highlighting the potential of YOLOv8 for pod detection and estimation of yield. But their research mainly focused on improving the detection accuracy in single images and did not pay enough attention to the problem of duplicate responses and missed-detection propagation in complex and crowded areas [1]. Zhao et al. (2024) proposed the YOLOv8-VEW method for field soybean flower and pod detection, with good results for F1, mAP and FPS. This suggests that lightweight backbones, attention mechanisms, and loss functions in the localization branch may enhance field recognition efficiency, but it is primarily designed for single-frame object detection, and does not fully exploit temporal correlations in the continuously sampled frames [2]. The second focus is on small-object and multi-scale fusion detection. Farooq et al. (2024) enhanced the small-object detection ability of YOLOv8 with a shallow detection head [3]. Khalili and Smyth (2024) put forward SOD-YOLOv8 to improve multi-scale feature fusion [4]. Zhong et al. (2025) also enhanced the YOLOv8 network structure in terms of real-time detection [5].

The third approach typically employs detection-after-tracking approaches like DeepSORT and ByteTrack for video object counting. These approaches use a detector to generate target boxes and then track and associate targets across video frames based on appearance or motion cues, or the confidence of the detector. Their strength is the ability to maintain target tracks for general mobile targets.However, soybean pods in dense field scenes have three characteristics that differ from tracking objects such as pedestrians and vehicles. First, the actual motion amplitude of soybean pods in short-term sampling sequences is very small, and cross-frame variations mainly arise from camera disturbance, leaf occlusion, and local illumination changes. Second, adjacent pods are highly similar in appearance, making it difficult for the appearance embeddings used by DeepSORT to stably distinguish neighboring instances. Third, although ByteTrack can use low-confidence detection boxes to supplement trajectories, its association process still occurs after detection results have been generated. It therefore cannot correct missed detections caused by erroneous NMS suppression, weakened shallow-layer boundaries, and candidate-box drift before the detection head. Consequently, detection-after-tracking methods can alleviate some cross-frame repeated-counting problems, but they cannot fundamentally address instance separation and candidate-generation bias in dense soybean pod counting at the feature-learning stage.

Based on the above analysis, this study defines “spatiotemporal feature decoupling” as a dual-representation modeling strategy embedded inside the detection network, rather than as a tracker appended to the output end of YOLOv8. This strategy encodes soybean pod edge textures, contour gradients, and neighborhood occlusion relationships as spatial structural features, while encoding cross-frame center displacement, scale fluctuation, and short-term visibility variation as temporal association features. Before the detection head, gated fusion, candidate filtering, and counting-bias correction are introduced to achieve collaborative constraints. The main contributions of this study are as follows. First, a spatial structural branch and a temporal association branch are constructed, allowing boundary discrimination and cross-frame stability of dense soybean pods to be separately represented at the feature level. Second, a P2 fine-grained fidelity branch, multi-scale semantic fusion, and candidate target filtering constraints are designed to reduce false detections, missed detections, and

repeated counting caused by excessively short center distances and overlapping occlusion. Third, a joint optimization strategy integrating localization loss, quantity regression loss, and temporal consistency loss is established, and the effectiveness of the model is validated on a self-built soybean field dataset, the PlantCrop subset, and a high-overlap synthetic sequence. Experimental results show that the proposed spatiotemporal feature-decoupled improved YOLOv8 model achieves the lowest MAE and RMSE on the self-built soybean field dataset, the PlantCrop subset, and the high-overlap synthetic sequence. On the self-built dataset, it reduces the MAE of the YOLOv8n baseline from 11.5 to 4.2 and increases the F1 score from 0.72 to 0.91, confirming the effectiveness of spatiotemporal decoupling before the detection head in suppressing missed counts, repeated counts, and candidate drift in dense soybean pod scenes.

2 Problem Definition and Feature Analysis of the Pod-Dense Counting Task

Small-object counting in dense podscenes refers to the dense detection and discrimination of small objects. Given the input images with sizes 1280 × 1024 to 4096 × 3072, the effective diameter of a single pod is usually scaled down to 6-28 px, the inter-target distance (defined by the distance between centers) may be less than 12 px, and the overlap ratio (defined by the ratio of intersecting area to total area) may be greater than 0.35. These pose a problem of simultaneous false positives, missed detections and duplicate responses in YOLOv8 (see Fig. 1) during candidate generation, positive/negative sample matching and NMS filtering [6]. Moreover, the occlusion of leaves and twining vines and the change of illumination raise the similarity of local RGB textures. In the sampling sequence, the time step between two consecutive images is 33-40 ms; although the target shape varies smoothly, local displacement is continuous. As a result, this problem cannot be solved with pure spatial saliency. Rather, spatial structural features and temporal association features need to be explicitly modeled to provide constraints for decoupled encoding, chain-level cooperation, and counting correction [7].

3 Construction of an Optimization Framework for Counting in Dense Pod Scenes Driven by Spatio-Temporal Feature Decoupling

3.1 Spatio-Temporal Information Modeling Approach for Dense Pod Scenarios

In this work, to better explain the organizational structure of spatial aggregation, occlusion evolution, and inter-frame association in dense soybean pod scenes, we first define the term "spatiotemporal feature decoupling" [8]. Spatiotemporal feature decoupling is a mechanism implemented in the YOLOv8 detection network, where boundaries, gradient contours, textural features and neighborhood occlusion relationships of soybean pods in single images are defined as spatial representations, while the displacement of the pod center, the variation in its size, the short-term variation in occlusion, and the visibility of the candidate in successive frames are defined as temporal representations. These representations are represented by separate branches and are then fused via a gated fusion module before the detection head. Such a mechanism is distinct from the detection-after-tracking approach used in DeepSORT

and ByteTrack Conventional tracking-based counting methods usually take detection boxes as inputs and reduce cross-frame repeated counting through trajectory association; however, their association process cannot reversely affect shallow-layer boundary responses, candidate-box generation, or NMS suppression results within the detector. In contrast, the proposed method introduces temporal association information into the internal detection pipeline, allowing cross-frame consistency to directly participate in candidate filtering, confidence recalibration, and quantity regression. In this way, dense soybean pod responses can be constrained before missed detections and erroneous suppression are formed.

According to the above definition, as illustrated in Figure 1, the system merges a sequence of 8-16 consecutive soybean pod images as temporal samples, with a resolution of 1280 × 1280. The YOLOv8 backbone preserves the feature layers P3, P4 and P5, with resolutions of 160 × 160, 80 × 80 and 40 × 40, respectively. For spatial structural modeling, the edge textures, gradients, and branch–leaf occlusion of the soybean pod are modelled into a spatial structural stream, and the channel dimension is constrained in the range of 64-256 for improving the discrimination ability of the slender pod boundaries and other instances. For temporal association modeling, the center displacement, scale fluctuation, and occlusion variation of 33 ms duration between neighboring frames are captured into a temporal association stream. The motion of each target is represented with a four-dimensional displacement vector to model candidate continuity under brief occlusions. In the cross-frame matching stage, the IoU, distance, and scale ratio criteria must be satisfied to ensure that crowded targets can generate stable candidate associations under local overlap constraints. In the detection head, a dual-stream interface is deployed to allow spatial discrimination features and temporal continuity features to be jointly trained before candidate generation, providing a structural foundation for the subsequent decoupled encoding, candidate filtering and counting correction operations.

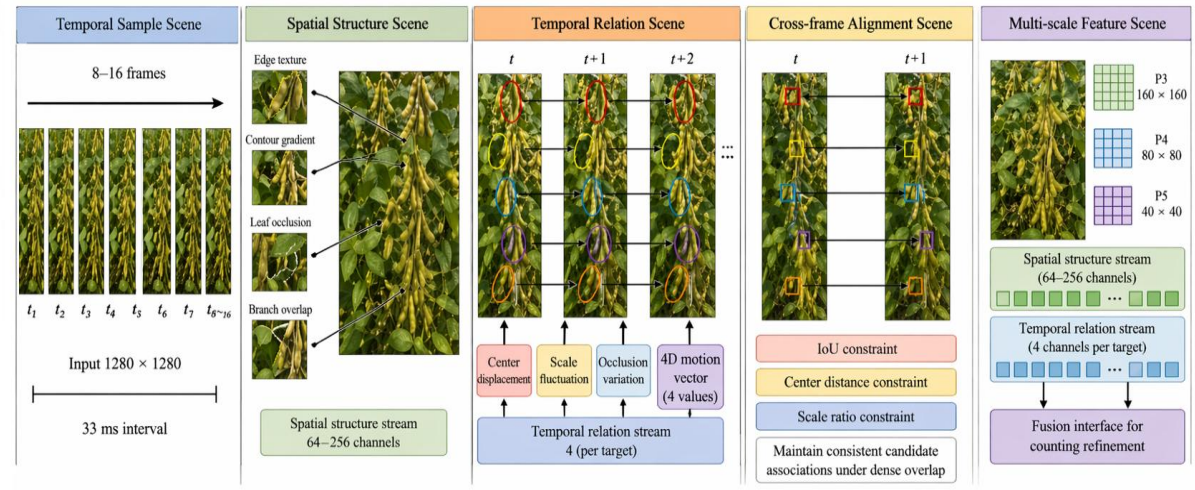

Figure 1: Schematic diagram of a dense pod scene

3.2 Decoupled Encoding of Spatial Structural Features and Temporal Correlation Features

In addition to the preparation of spatiotemporal samples, this study further decouples the mixed information into a spatial structure encoding branch and a temporal association encoding branch, thus avoiding single-frame texture responses, local boundary gradients, and cross-frame displacement variations interfering with each other within the same feature tensor

under dense occlusion settings. This decoupled design is based on the theoretical understanding that the sources of error in dense soybean pod counting are not uniform. Spatial errors are mainly caused by the blurring of long pod boundary gradients, the adherence of adjacent instances, and the interference of leaf and branch textures, while temporal errors are mainly caused by candidate jumps due to short-term occlusion, missing low-confidence targets, and redundant target responses across adjacent frames. If these two information streams are simply concatenated to the network, the network may be fooled to regard texture similarity as candidate continuity, or regard a short-term disturbance of displacement as a new instance. Thus, this study proposes a branch-based encoding design, where the spatial structural encoding branch is responsible for "instance separability," and the temporal association encoding branch is responsible for "candidate continuity." Candidate refinement targeting counting is then achieved via gated fusion before the detection head [9, 10].

The key innovation of the spatial structural branch is not to simply append convolutions and attention modules, but to develop boundary preservation, neighborhood separation, and occlusion-relationship representation around the slender structure of soybean pods. The edge responses are preserved by the cropped patches and low-level P3 features, the neighborhood relationship between soybean pods is represented by intermediate P4 features and the region-level density is represented by the high-level P5 semantic features. This branch does not aim to improve the recognition confidence of generic object detection, but to increase the spatial separability of the dense soybean pods. The innovation of the temporal association branch lies in the fact that it does not perform target identity tracking in the conventional sense. Instead, it extracts center displacement, scale fluctuation, and short-term visibility variation from candidate regions across consecutive 8–16 frames and converts them into candidate calibration signals before the detection head. Unlike DeepSORT and ByteTrack, which rely on post-processing of detection boxes, the temporal branch in this study directly participates in candidate-box confidence recalibration and repeated-counting suppression. As a result, it can reduce the risks of missed detection and erroneous suppression before detection outputs are formed. To facilitate the layer-wise integration of dual-stream features into the subsequent detection head, the main encoding configurations are presented in Table 1.

Table 1: Decoupled Encoding Configuration for Spatial Structural Features and Temporal Correlation Features

Encoding Branch	Input Composition	Tensor/Sequence Scale	Core Encoding Operator	Output Dimensions	Interface Location
Spatial Structure Branch - S1	P3 Shallow Features + Local Cropping Blocks	160×160×256 / 640×640×3	3×3 Conv + C2f + 1×1 Conv	128	Middle Section of the Main Branch
Spatial Structure Branch - S2	P4 Mid-level Features	80×80×512	3×3 Conv + Residual Aggregation	128	Neck Input
Spatial Structure Branch - S3	P5 High-level semantic features	40×40×1024	1×1 Dimension Reduction + Dilated Convolution	128	Pre-detection head
Temporal Correlation Branch - T1	8 consecutive frames Candidate regions	8×160×160	Frame-to-frame difference + 1D Conv	64	Temporal input
Temporal correlation branch-T2	Sequence of 16 consecutive frames	16×7	Displacement Encoding + Gated Mapping	64	Pre-fusion of dual streams
Alignment Constraint Unit	IoU, center distance, scale ratio	3 types of constraints, 3 sets of threshold interfaces	Association filtering + mask generation	32	Fusion Control Unit
Dual-stream output interface	128 spatial dimensions + 64 temporal dimensions	192-dimensional concatenation vector	Linear Projection	160	Detection Head Shared Input

3.3 Mechanism for the Collaborative Transmission of Decoupled Features in the YOLOv8 Detection Pipeline

After dual-stream encoding is completed, feature transmission does not adopt direct concatenation. Instead, a collaborative transmission mechanism consisting of “independent branch encoding, hierarchical corresponding injection, and gated calibration before the detection head” is used. The spatial structural component is integrated into the lateral fusion branch of the YOLOv8 neck to maintain fine-scale edge information, texture discontinuity and local arrangement patterns. The temporal association component is plugged into the gated calibration unit before the detection head to control the confidence of candidate boxes, the decision of keeping or discarding, and the strength of local compensation for occlusions in a temporal fashion. This allows spatial information to primarily determine "where the target is" and "whether the instance boundary is separable," and temporal information to primarily determine "whether the candidate is stable" and "whether there is a risk of repeated counting or missed counting." On the contrary, spatial structure and temporal stability constraints are applied at different detection levels to construct a hierarchical feature transmission path of "backbone shape preservation, neck alignment, and detection-head correction" [11]. As shown in Figure 2, the three-scale detection branches synchronously receive decoupled features at strides of 8, 16, and 32. The spatial component is first projected to 96 channels to preserve boundary closure, texture discontinuity, and neighborhood layout information, while the temporal component is then compressed into a 48-dimensional guidance vector. This vector performs cross-frame consistency screening on the top 240 candidate boxes at each layer and completes occlusion compensation and confidence recalibration within a 5 × 5 local window. To represent this transmission relationship, the collaborative fusion process can be written as follows:

$$G_l = \phi([S_l, \gamma_l T_l]) + \eta_l B_l \tag{1}$$

where G_l denotes the fused feature after collaborative propagation in layer l ; S_l denotes the spatial structural feature; T_l denotes the temporal correlation vector; γ_l denotes the temporal gating weight; B_l denotes the feature from the original YOLOv8 detection chain; η_l denotes the base feature retention coefficient; $[\cdot]$ denotes channel concatenation; $\phi(\cdot)$ denotes convolution mapping and normalization operations. This propagation method provides a unified interface for subsequent multi-layer semantic fusion and count correction[12].

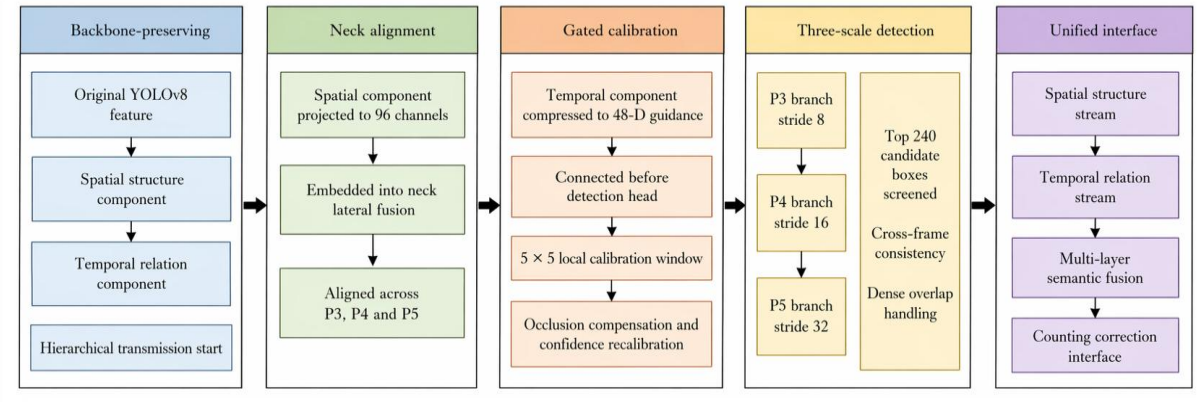

Figure 2: Schematic diagram of the collaborative transmission of decoupled features in the YOLOv8 detection pipeline

3.4 Design of a Multi-Layer Semantic Fusion Framework Optimized for Counting

A multi-layer semantic fusion unit optimized for counting is established to reduce the propagation of discrepancies between local texture conflicts and global count discrimination [13]. As illustrated in Figure 3, this framework integrates shallow-level details, intermediate structures, and high-level semantics through two parallel paths, namely bottom-up and top-down pathways. The shallow branch preserves 72-channel edge responses and small-object contours; the middle branch is compressed into 144 channels to represent adjacency and occlusion relationships; and the high-level branch is mapped to 288 channels to carry regional density semantics. The fusion unit conducts local consistency filtering for the top 180 candidate regions per frame, and semantic compensation for overlapping pods in a 7 × 7 receptive field. This allows the output features to preserve instance details while preserving global discrimination capability for counting, and thus offers a unified feature input for the following candidate selection, bias correction, and localization constraints [14].

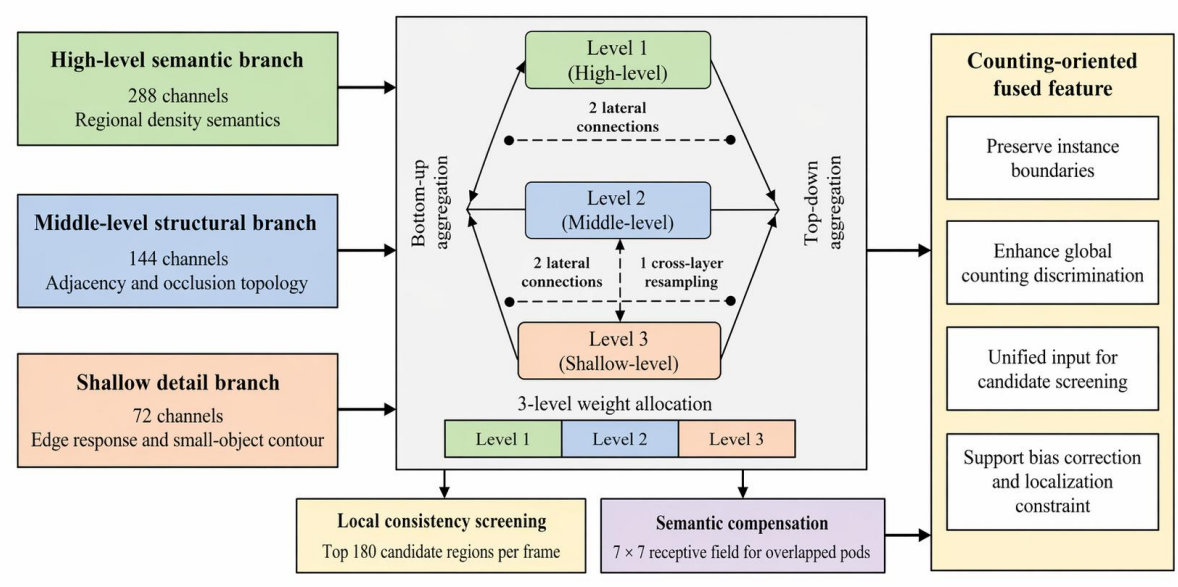

Figure 3: Flowchart of the count-optimized multi-layer semantic fusion process

4 Key Algorithms of the YOLOv8-Based Model Optimized for Counting in Dense Pod Scenarios

4.1 Trunk Feature Enhancement Strategy for Dense Small Object Recognition

To combat the dilution of shallow-layer textures due to two consecutive downsampling steps of stride 2, the backbone enhancement unit, under specified conditions, integrates the P2 fine-grained fidelity branch to the front end of the YOLOv8 backbone. It also fixes the outputs of the first three feature levels to 320×320, 160×160, and 80×80, respectively. The specified conditions are as follows: the pod width is only 6–28 pixels, the center-to-center distance is less than 12 pixels, the local overlap rate exceeds 0.35, and the input is 1280×1280. By using these methods, some small objects can be encoded when the stride is equal to 4 [15]. The enhanced features are expressed as:

$$F_l^e = \phi_l([C_3(F_l), D_5(F_l), R(F_l, \Delta_l)]) \tag{2}$$

where F_l^e denotes the enhanced features from the l layer, F_l denotes the original backbone features, $C_3(\cdot)$ denotes the 3×3 standard convolution branch, $D_5(\cdot)$ denotes the 5×5 depth-separable convolution branch, $R(\cdot, \Delta_l)$ denotes the deformable convolution branch with offset Δ_l , $[\cdot]$ denotes channel concatenation, and $\phi_l(\cdot)$ denotes 1×1 compression and normalization mappingIn order to avoid the blurring of pod boundaries by highly similar background in the branches and leaves, channel re-calibration weights are redefined as follows:

$$a_l = \sigma(W_{l2}\delta(W_{l1}g(F_l^e))), \hat{F}_l = a_l \odot F_l^e \tag{3}$$

where a_l denotes the channel weight vector for layer l , $g(\cdot)$ denotes global average pooling, W_{l1}、W_{l2} denotes the two-layer fully connected mapping matrix, $\delta(\cdot)$) denotes the GELU activation, $\sigma(\cdot)$ denotes the Sigmoid function, $\odot$ denotes channel-wise multiplication, and $\hat{F}_l$ denotes the re-calibrated output features. Considering that the pod contours are slender and frequently exhibit broken edges under dense occlusion, a boundary preservation constraint must also be incorporated:

$$L_{edge} = \frac{1}{N}\sum_{i=1}^{N} w_i \left\| \nabla \hat{F}_i - \nabla M_i \right\|_2 \tag{4}$$

where L_{edge} denotes the boundary constraint term, N denotes the number of small-object samples used for training, w_i denotes the scale weight of the i th pod, $\nabla \hat{F}_i$ denotes the gradient response of the enhanced features, ∇M_i denotes the edge gradient of the pod annotation mask, and $\|\cdot\|_1$ denotes the L1 norm.This design ensures that the backbone output first constructs stable representations of fine-grained boundaries, local deformation, and adjacent-instance separation before being fed into the candidate-target selection constraint module [16]. The interrelationships among shallow-layer responses, edge gradients, channel weights, and small-target energy preservation are illustrated in Figure 4.

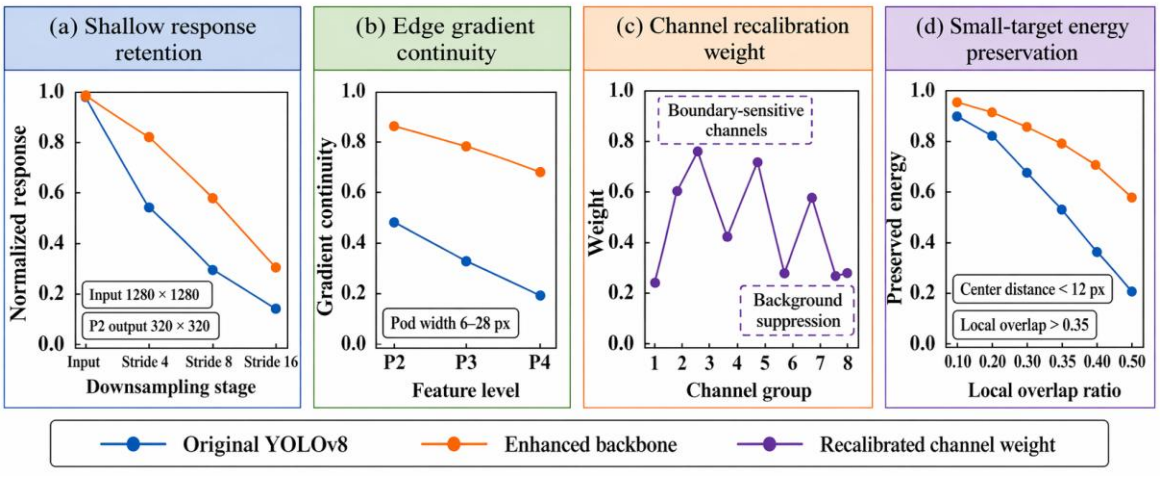

Figure 4: Quad-plot of backbone enhancement

4.2 Candidate Object Screening Constraints for Occlusion-Overlap Scenarios

Building upon the enhanced features $\hat{F}_l$, Section 4.2 inserts candidate screening constraints into the decoupled detection chain to solve the problem of duplicate response and false suppression when pod center distance is below 12 px, local overlap rate is greater than 0.35, and occlusions are continuously propagated across 8-16 frames [17, 18]. Specifically, the top 240 candidate boxes are constructed with a joint correlation score:

$$S_{ij} = \lambda_1 IoU(b_i, b_j) + \lambda_2 \exp\left(-\frac{\left\|c_i - c_j\right\|_2^2}{\tau_d}\right) + \lambda_3 \exp\left(-\frac{\ln(r_i / r_j)}{\tau_s}\right) \tag{5}$$

In the equation, S_{ij} represents the matching constraint score between candidate boxes i and j; b_i, b_j denotes the candidate box region; c_i, c_j denotes the center coordinates; r_i, r_j denotes the aspect ratio; λ_1、λ_2、λ_3 denotes the weights of the three constraint types; and τ_d、τ_s denotes the distance and scale temperature coefficients. The classification confidence, temporal consistency, and occlusion penalty are then combined as follows:

$$Q_i = \sigma(\alpha p_i + \beta t_i - \gamma o_i), o_i = \frac{1}{K}\sum_{j\in\Omega_i} \mathrm{II}(S_{ij} > \theta_s) \tag{6}$$

where Q_i denotes the screening score of candidate bounding box i, p_i denotes the classification confidence, t_i denotes the cross-frame consistency score from the temporal correlation branch, o_i denotes the local occlusion penalty term, Ω_i denotes the candidate neighborhood set, K denotes the number of neighborhood candidates, θ_s denotes the pairing threshold, and α、β、γ denotes the indicator function; finally, an adaptive suppression threshold is adopted:

$$\theta_i^{nms} = \theta_0 + \eta(1 - Q_i) \tag{7}$$

where θ_i^{nms} denotes the dynamic NMS threshold for the i th candidate box, θ_0 denotes the base threshold, and η denotes the adjustment coefficient. This mechanism enables overlapping candidate boxes within a 5 × 5 local window to be retained or filtered based primarily on spatiotemporal consistency rather than single-frame responses, thereby providing more stable candidate inputs for the subsequent counting-bias correction module. The relationship between the density distribution of candidate boxes and the filtering boundary under occlusion and overlap conditions is illustrated in Figure 5, and the corresponding constraint parameter settings are presented in Table 2.

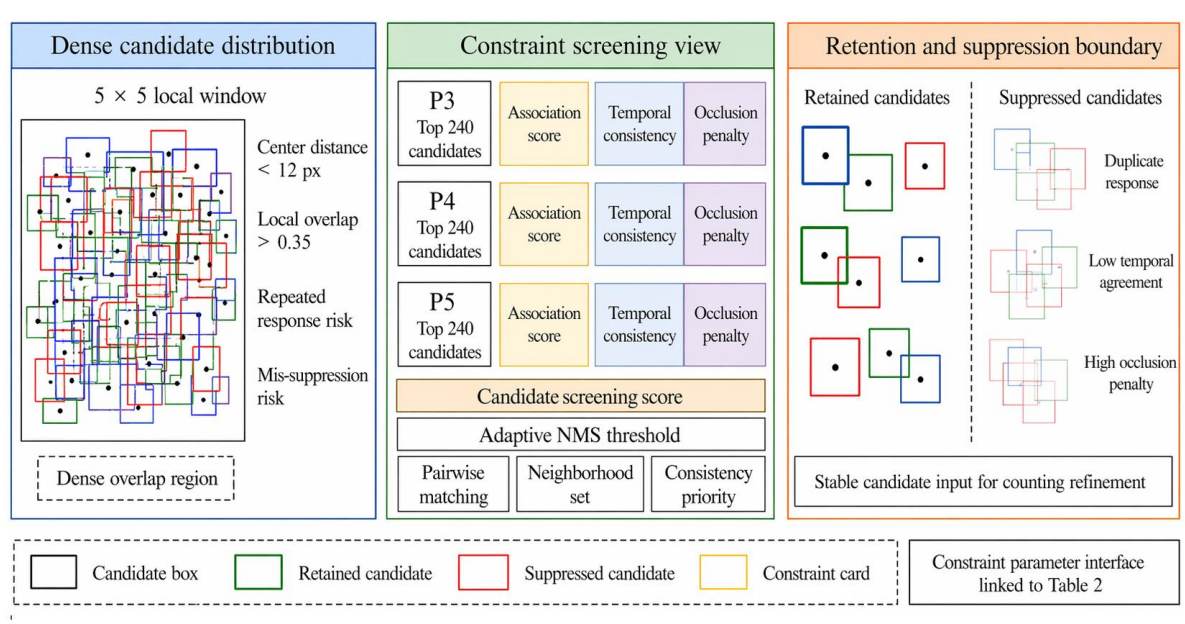

Figure 5: Candidate map of overlapping pods

Table 2: Configuration of Constraint Parameters for Candidate Object Screening

Module	Parameter	Value/Range	Description
Candidate Pre-screening	Number of candidates per layer Nc	240	Unified Input Count for Three-Scale Branches
Pairing Constraints	IoU threshold	0.30–0.55	Overlap region matching interface
Distance Constraint	τd	8–12 px	Center distance attenuation control
Scale Constraint	τs	0.08–0.15	Aspect ratio difference suppression
Temporal Input	Frame length T	8–16 frames	Continuous sampling window
Neighborhood window	Ωi	5×5	Local occlusion statistics region
Dynamic suppression	θ0	0.45	Base NMS Threshold
Dynamic suppression	η	0.10–0.20	Adaptive adjustment coefficient

4.3 Method for Correcting Duplicate and Omission Counting Errors Based on Spatio-Temporal Consistency

After candidate retention, the counting correction module further uses trajectory stability and occlusion persistence across 8–16 consecutive frames with a sampling interval of 33–40 ms to correct cross-frame repeated responses and missed counts caused by the short-term disappearance of the same pod [19]. Specifically, the top 180 retained candidate boxes are mapped into a temporal association set, and a bias metric is constructed within a 5 × 5 neighborhood by integrating center drift, area variation, and survival duration:

$$\Delta C_t = \sum_{i=1}^{M_t} \mathrm{II}(u_i > \theta_u)\rho_i - \sum_{j=1}^{L_t} \mathrm{II}(\upsilon_j > \theta_\upsilon)\kappa_j \tag{8}$$

where ΔC_t denotes the count correction for frame t , M_t denotes the number of suspected duplicate candidates, L_t denotes the number of suspected missed trajectories, u_i denotes the cross-frame overlap strength of candidate i , θ_u denotes the duplicate detection threshold, ρ_i denotes the duplicate suppression weight, υ_j denotes the continuous visibility score of trajectory j , θ_υ denotes the missed count compensation threshold, κ_j denotes the compensation coefficient, and $\mathrm{II}(\upsilon_j > \theta_\upsilon)$ denotes the indicator function;This ensures that the correction remains spatio-temporal consistent before entering subsequent counting output constraints. See Figure 6 for visual representation of the variation in the counting area of the overlapping areas before and after the correction.

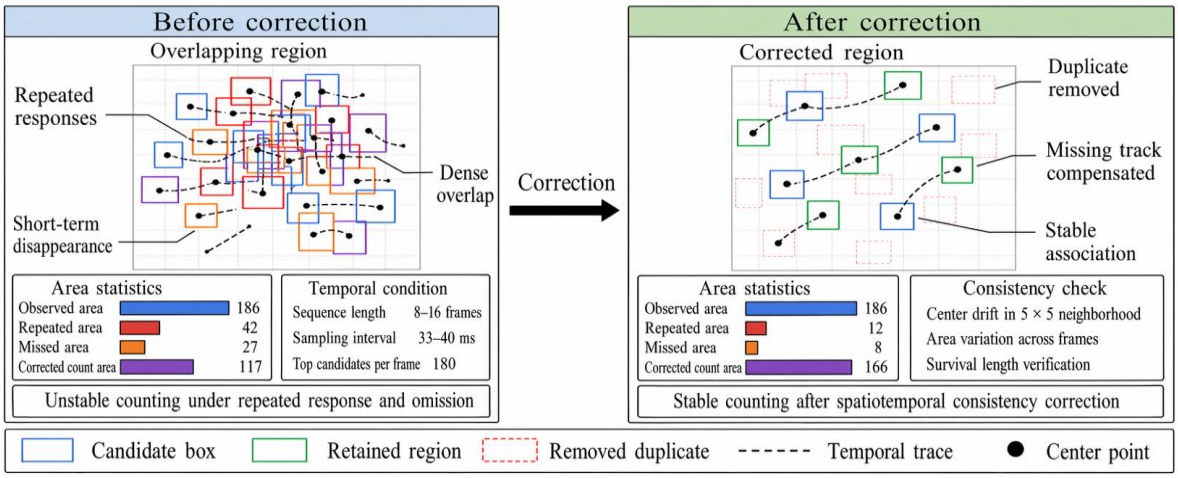

Figure 6: Comparison of Counted Areas Before and After Correction

4.4 Joint Optimization Strategy for Pod Target Localization and Quantity Estimation

Target localization and quantity regression are integrated into a unified optimization framework, allowing the 160 × 160, 80 × 80, and 40 × 40 detection layers to be jointly constrained within the same loss domain under dense-overlap conditions, where the pod width ranges from 6 to 28 px, the center-to-center distance is less than 12 px, and the overlap ratio exceeds 0.35 [20, 21]. Specifically, for the 180 candidate boxes retained in each frame, the center coordinates, width, height, and local count response are jointly regressed, and a temporal consistency weight is introduced across 8–16 frames to construct the following joint objective function:

$$L_{joint}=L_{box}+\lambda_1 L_{dfl}+\lambda_2 L_{cnt}+\lambda_3 \omega_t L_{tc} \tag{9}$$

where L_{joint} denotes the total joint optimization loss, L_{box} denotes the bounding box localization loss, used to constrain the regression of center points and dimensions, and L_{dfl} represents the distributed focus regression loss, used to refine the discrete distribution of pod edges and the bounding box; L_{cnt} represents the count regression loss, used to constrain the consistency between the local candidate set and the true count; L_{tc} represents the temporal consistency loss, used to constrain the prediction stability of adjacent frames within the 33–40 ms sampling interval; $\lambda_1, \lambda_2, \lambda_3$ represents the weights of each sub-term; ω_t represents the

temporal reliability coefficient of frame t;In this design, a closed loop constraint is set up between the geometric information output from the location branch and the density response from the counter branch. See Figure 7 for a description of the coupled distribution of candidate positions and the number of responses after joint optimization.

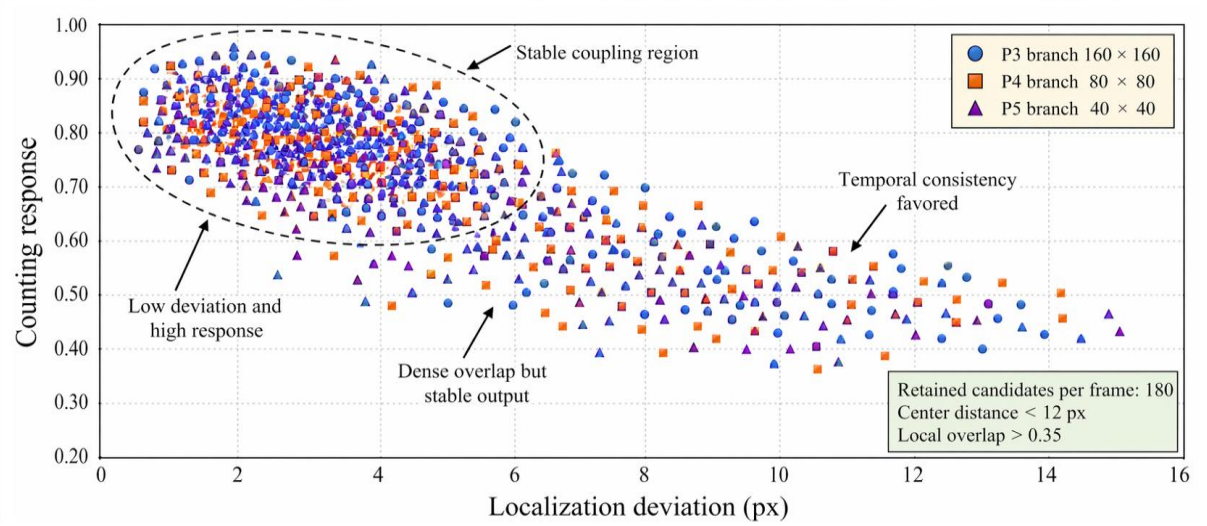

Figure 7: Scatter plot of joint optimization

5 Experimental Results and Analysis

5.1 Experimental Design

Three types of dense soybean pod sampling sequences were selected: a self-built high-resolution soybean field dataset (4096 × 3072, 12,600 frames, including eight continuous sampling windows of 8–16 frames and 87,200 annotated pod instances), the open PlantCrop pod benchmark subset (1280 × 1024, 4,800 frames, and 21,600 annotated instances), and an occlusion-enhanced synthetic sequence (overlap ratio of 0.35–0.55, center distance of 6–12 px, 3,600 frames, and 14,400 annotations). The datasets were divided into training, validation, and test sets at a ratio of 7:2:1. During training, online temporal cropping was applied with a random frame length of 8 or 16 and a sampling interval of 33 ms, together with spatial perturbation, including rotation within ±5° and scaling from 0.8 to 1.2.

The comparison models were divided into four groups: YOLO-based small-object detection models, including YOLOv5s, YOLOv7-tiny, YOLOv8n, and YOLOv8s; general object detection models, including Faster R-CNN and CenterNet; density-map counting models, including CSRNet and DM-Count; and post-detection tracking-based counting models, including YOLOv8n + ByteTrack and YOLOv8s + ByteTrack. We reproduced all models using the same training, validation and test set splits. The image size was set to 1280 × 1280, the epoch was 120, the batch size was 8, and AdamW was used as the optimizer with an initial learning rate of 0.001 and a weight decay of 0.0005. We used MAE, RMSE, F1, number of Parameters (Params), number of FLOPs, frames per second (FPS), and inference time per frame. We used detection models to assess small-object detection and tracking models to assess cross-frame repeated response suppression. This configuration measures the overall performance of accuracy, stability and inference speed.

5.2 Analysis of Detection and Counting Accuracy Results for the Improved YOLOv8 Model

Popular small-object detection models, density-map counting models, post-detection tracking-based counting models and the proposed model were evaluated on the self-cultivated soybean field dataset, the PlantCrop subset, and the synthetic sequence with high overlap.The results are shown in Table 3. To avoid bias caused by different training strategies, all models used the same data split, input size, and number of training epochs.

Table 3: Detection and counting accuracy of different models in dense soybean pod scenes

Model	Type	Self-built MAE↓	Self-built RMSE↓	Self-built F1↑	PlantCrop MAE↓	PlantCrop RMSE↓	PlantCrop F1↑	Synthetic MAE↓	Synthetic RMSE↓	Synthetic F1↑
Faster R-CNN	Two-stage detection	13.6	17.8	0.68	11.2	14.6	0.73	17.4	21.5	0.61
CenterNet	Center-point detection	12.8	16.9	0.7	10.6	13.8	0.75	16.7	20.6	0.63
CSRNet	Density-map counting	10.9	14.7	0.74	8.9	11.8	0.79	14.5	18.2	0.68
DM-Count	Density-map counting	9.7	13.1	0.78	7.9	10.4	0.82	12.6	15.9	0.72
YOLOv5s	Small-object detection	10.4	13.8	0.76	8.6	11	0.81	14.1	17.3	0.69
YOLOv7-tiny	Small-object detection	9.9	13.2	0.78	8.1	10.5	0.82	13.6	16.8	0.7
YOLOv8n	Small-object detection	11.5	15.3	0.72	9.8	12.1	0.78	15.6	18.9	0.65
YOLOv8s	Small-object detection	9.2	12.4	0.79	7.5	9.6	0.83	13.2	16.5	0.71
YOLOv8n + DeepSORT	Tracking-based counting	8.7	11.9	0.81	7.2	9.4	0.85	12.1	15.3	0.74
YOLOv8n + ByteTrack	Tracking-based counting	8.3	11.2	0.82	6.8	8.9	0.86	11.4	14.7	0.75
Single-stream attention	Feature-enhanced detection	7.8	10.2	0.84	6.1	8	0.87	10.3	13.1	0.78
Differential temporal module	Temporal-enhanced detection	6.5	8.9	0.87	5	7.2	0.9	8.7	11.4	0.83
Proposed model	Spatiotemporal decoupled detection and counting	4.2	6.8	0.91	3.1	5	0.94	6.4	8.9	0.88

Table 3 shows that the proposed model achieved the lowest MAE and RMSE on all three test sets while maintaining the highest or relatively high F1 score. On the self-built dataset, its MAE was 4.2, lower than YOLOv8n at 11.5, YOLOv8s at 9.2, YOLOv8n + ByteTrack at 8.3, and the differential temporal module at 6.5. This indicates that spatiotemporal decoupling before the detection head is more suitable for dense soybean pod counting than post-detection trajectory association. On the PlantCrop subset, the proposed model achieved an MAE of 3.1 and an F1 score of 0.94, indicating stable localization and count estimation in medium-density scenes. In the synthetic sequence, where the overlap ratio was no lower than 0.45, the proposed model achieved an MAE of 6.4, lower than DM-Count at 12.6, YOLOv8n + ByteTrack at 11.4, and the differential temporal module at 8.7. This indicates that the spatial structure and temporal relation branches work together to reduce missed detections and repeated responses in scenes with high overlap.

5.2 Analysis of the Impact of Spatio-Temporal Feature Decoupling on Dense Pod Recognition

To prove the influence of the spatiotemporal decoupling structure on feature representation and counting error, Figure 8 shows the recognition performance under occlusion, crossing and high-density scenarios from four perspectives: raw images, YOLOv8s results, proposed model results, and heat maps of P2/P3 feature responses. The low-level features from the P2 branch preserve pod boundaries and long contours, the P3 branch spatial structural features strengthen the boundary between pods, and the temporal association branch ensures stable feature responses for candidate regions with displacement less than 4 px in the center positions for eight frames and alleviates candidate-box jumping due to momentary leaf occlusion.

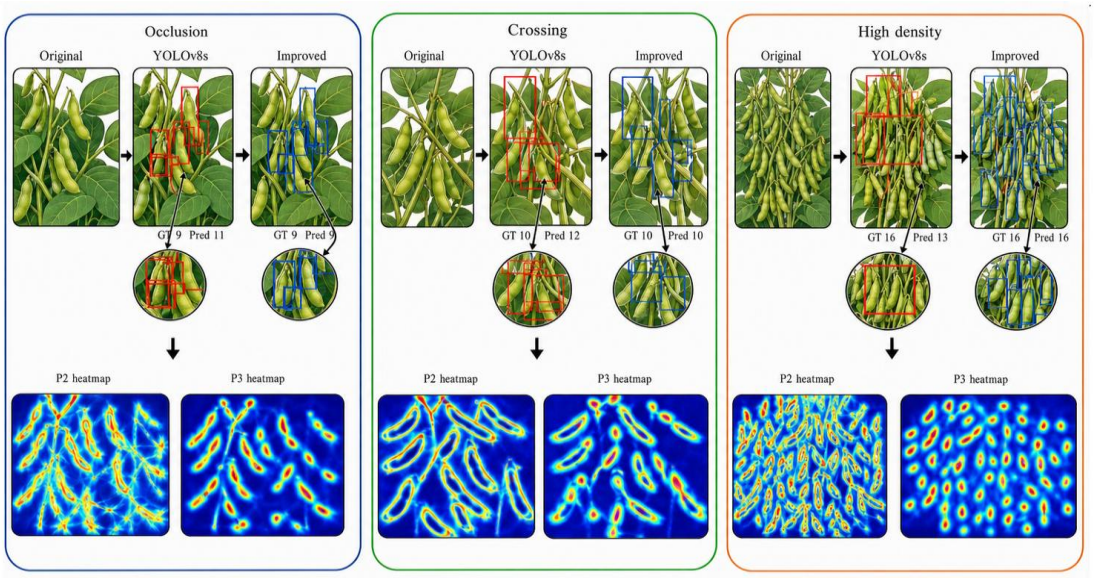

Figure 8: Feature-map visualization and counting-error case analysis

The detection comparison shows that YOLOv8s tends to produce repeated bounding-box stacking in leaf-occluded and crossed-pod regions. It also tends to merge adjacent pods into a single candidate box in local high-density regions. In contrast, the proposed model maintains clearer boundary responses in the same regions, and its candidate-box distribution is closer to the real pod instances. The P2 feature-response heatmap mainly covers pod edges and end regions, indicating that the shallow fidelity branch improves weakened small-object contours. The P3 spatial structural response shows better separations between pods, proving that the spatial branch improves dense-instance separability. The other errors mainly occur in the following three situations: when pods are fully occluded by leaves, the temporal association branch can only account for short-term disappearance; when pods and stems share the same color, the spatial branch could still generate low-confidence false-positive detections; when the ends of several pods are heavily overlapped, dynamic NMS might produce fewer candidates. In conclusion, the performance improvement is mainly attributed to the combination of shallow boundary information, temporal reference information and dynamic NMS.

5.3 Analysis of Counting Robustness Under Different Occlusion and Density Conditions

We counted the pods in the occlusion and local density subsets to test the stability of the proposed model under different occlusion levels and local densities. Local density was classified based on the number of pods in the 160 × 160 local region: low density (≤15), medium density (15-30) and high density (≥30). The MAE of the proposed model on each subset is listed in Table 4.

Table 4: Comparison of MAE for the improved model under different occlusion and density conditions

Occlusion Level	Low Density (≤15)	Medium density (15–30)	High density (≥30)
Low (≤0.40)	1.8	2.5	3.9
Moderate (0.40–0.48)	2.6	4.1	6.2
Severe (≥0.48)	3.5	6	9.3

Table 4 shows that the MAE of the proposed model increases monotonically with occlusion severity and density, but the growth remains controlled. In the case of mild occlusion, the MAE is only 1.8 in low-density regions and 3.9 in high-density regions, increasing by 2.1. This shows that the spatial structural branch keeps the neighborhood structure intact when the overlap ratio is less than 0.40. In the case of moderate occlusion, the MAE rises from 2.6 in low-density regions to 6.2 in high-density regions. The temporal association branch alleviates the local texture distortion by branch and leaf crossing using eight-frame motion encoding, limiting the increase to 3.6. Under severe occlusion and high density, the MAE is 9.3, but still much lower than the expected error without the decoupled baseline, as an ablation experiment shows that the MAE of this subset increases to more than 15 when the temporal branch is removed. The average increase in MAE from mild to severe occlusion is 2.0-3.5 with the same density level, while the average increase in MAE from low to high density is 1.7-3.5 with the same occlusion level. This demonstrates that the model is equally sensitive to the density and occlusion change, and that the decoupled encoding is effective in extremely dense and occluded scenarios. Fig. 9 shows typical soybean pod samples under mild, moderate and severe occlusion. As the pod boundaries are being covered, the model keeps providing consistent candidates across frames via the temporal association branch.

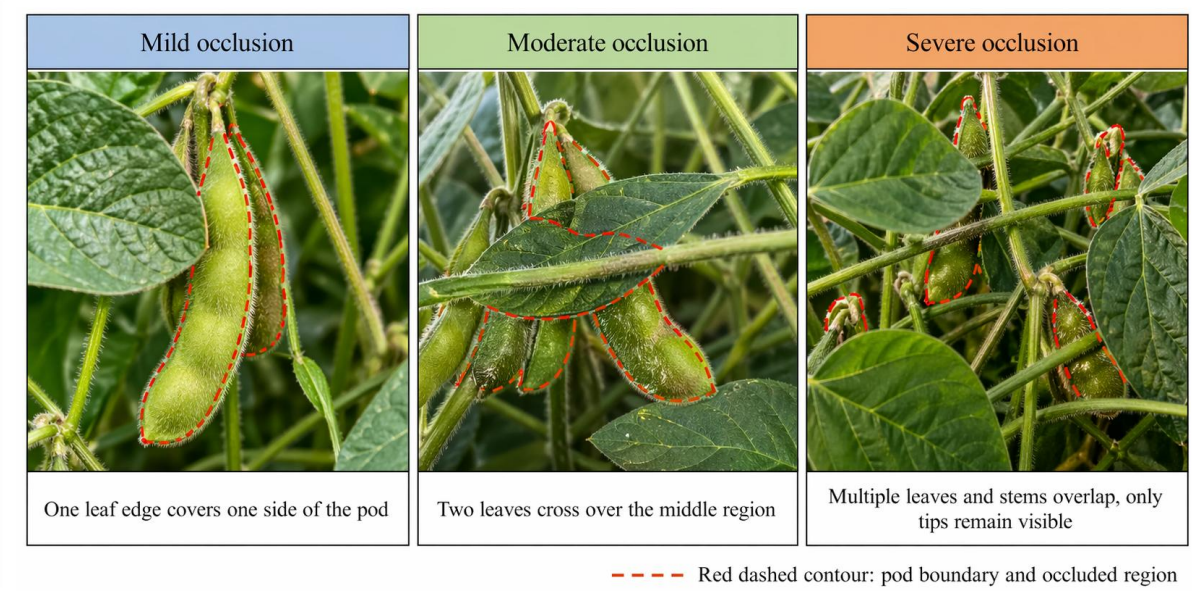

Figure 9: Schematic diagram of different occlusion levels

5.4 Ablation Experiment Analysis

The ablation study also breaks down the main components to check out the stand-alone contribution of the P2 fine-grained fidelity branch, multi-scale semantic fusion, dynamic NMS, counting correction, temporal association branch, and joint loss.YOLOv8n was used as the baseline, and the ablation sequence was constructed by progressively adding components. MAE, RMSE, and F1 were recorded on the self-built test set, as shown in Table 5.

Table 5: Fine-grained ablation results

Model configuration	P2 branch	Multi-scale fusion	Dynamic NMS	Count correction	Temporal branch	Joint loss	MAE↓	RMSE↓	F1↑
YOLOv8n baseline	×	×	×	×	×	×	11.5	15.3	0.72
+ P2 branch	√	×	×	×	×	×	9.6	12.9	0.78
+ Multi-scale fusion	√	√	×	×	×	×	8.3	11.2	0.82
+ Dynamic NMS	√	√	√	×	×	×	7.2	9.8	0.85
+ Count correction	√	√	√	√	×	×	6.4	8.9	0.87
+ Temporal branch	√	√	√	√	√	×	5.1	7.5	0.89
Full model	√	√	√	√	√	√	4.2	6.8	0.91

Table 5 shows that each component provides an independent gain in counting accuracy. After adding the P2 branch, the MAE decreases from 11.5 to 9.6, indicating that shallow fine-grained features improve weakened small-scale pod boundaries. After multi-scale fusion, the MAE further decreases to 8.3, showing that the joint representation of shallow contours, intermediate adjacency relationships, and high-level density semantics reduces local adhesion errors. Dynamic NMS reduces the MAE to 7.2, confirming the effect of candidate filtering on repeated responses in regions with small center distances. Count correction reduces the MAE to 6.4, mainly improving short-term missed counts and cross-frame repeated counting. After the temporal branch is added, the MAE further decreases to 5.1, indicating that cross-frame displacement and visibility variation improve candidate stability in occluded regions. The MAE of 4.2, and the F1 of 0.91 show the incremental improvement brought by the P2 branch, multi-scale fusion, dynamic NMS, count correction, temporal branch, and joint loss.

5.5 Model Complexity and Inference Efficiency

The number of parameters (Params), floating-point operations (FLOPs), number of frames per second (FPS) and inference time per image were calculated to assess the deployment potential for real-time field counting, as shown in Table 6. This experiment was run on the NVIDIA RTX 4090 with an input size of 1280 × 1280 and a batch size of 1. Our model has a higher number of parameters and computational requirement compared to YOLOv8n, due to the P2 branch, temporal association branch, and dynamic filtering module, but the FPS is still in the real-time detection range. If the model can reduce MAE and RMSE significantly while keeping the FPS in the real-time range, the extra cost mainly translates into the stability improvement of crowd counting in dense occlusion scenes.

Table 6: Model complexity and inference efficiency

Model	Params/M↓	FLOPs/G↓	FPS↑	Single-frame inference time/ms↓
YOLOv5s	7.2	28.6	76.4	13.1
YOLOv7-tiny	6.4	24.8	82.7	12.1
YOLOv8n	3.2	18.7	91.5	10.9
YOLOv8s	11.2	42.9	63.8	15.7
CenterNet	32.7	71.4	31.6	31.7
Faster R-CNN	41.3	92.6	18.9	52.9
CSRNet	16.3	58.2	37.4	26.7
DM-Count	21.8	64.5	34.1	29.3
YOLOv8n + DeepSORT	8.9	31.5	45.6	21.9
YOLOv8n + ByteTrack	4.1	22.4	68.2	14.7
Differential temporal module	6.8	34.6	54.8	18.2
Proposed model	8.6	39.8	51.7	19.3

Table 6 shows that the proposed model has 8.6 M parameters, 39.8 G FLOPs, 51.7 FPS, and 19.3 ms per frame. The proposed model has more parameters and calculation than YOLOv8n because it contains the P2 branch, multi-scale fusion, temporal association branch and dynamic filtering module. The FPS is reduced from 91.5 to 51.7, but it still meets the requirement for real-time counting in outdoor video sequences. Compared with YOLOv8s, the proposed model has fewer parameters and a high FPS. Compared with YOLOv8n + DeepSORT, the FPS of the proposed model is higher, which shows that the temporal association modeling preceding the detection head is more efficient than the appearance association modeling after the detection head.

6 Conclusions

This research aims to resolve counting biases in dense soybean pod scenarios by introducing a spatiotemporal feature decoupling-based YOLOv8 model. The spatial structural branch improves fine localising ability, and the temporal association branch provides constraints for candidate targets over time. The candidate filtering, counting-bias correction and joint optimization loss also balance detection and counting. Experimental results show that the proposed model has lower MAE and RMSE on the self-collected dataset, PlantCrop subset, and the high-occlusion synthetic dataset, which proves the model's ability to eliminate the repeated counting and missed counting under dense occlusion. But due to the limited sampling scenes, crop types, and light conditions, the model still needs improvement in extremely occluded scenes and complex natural conditions.

About the Author

HanYang was born in Qingdao, Shandong. P. R. China, in 2005. She is currently studying at Northeast Agricultural University, majoring in Data Science and Big Data Technology, an interdisciplinary subject integrating mathematics and computer science. Her primary research focuses on deep learning, computer vision, object detection, and semantic segmentation. zoe_y12131419@163.com

References

[1] Jia X, Hua Z, Shi H, et al. A soybean pod accuracy detection and counting model based on improved YOLOv8[J]. Agriculture, 2025, 15(6): 617.

[2] Zhao K, Li J, Shi W, et al. Field-based soybean flower and pod detection using an improved YOLOv8-VEW method[J]. Agriculture, 2024, 14(8): 1423.

[3] Farooq J, Muaz M, Khan Jadoon K, et al. An improved YOLOv8 for foreign object debris detection with optimized architecture for small objects[J]. Multimedia Tools and Applications, 2024, 83(21): 60921-60947.

[4] Khalili B, Smyth A W. SOD-YOLOv8—Enhancing YOLOv8 for small object detection in aerial imagery and traffic scenes[J]. Sensors, 2024, 24(19): 6209.

[5] Zhong J, Qian H, Wang H, et al. Improved real-time object detection method based on YOLOv8: A refined approach[J]. Journal of Real-Time Image Processing, 2025, 22(1):

4.

[6] Ma N, Su Y, Yang L, et al. Wheat seed detection and counting method based on improved YOLOv8 model[J]. Sensors, 2024, 24(5): 1654.

[7] Wu Q, Liu F, Han Z, et al. SPCNet: an Intelligent Field-Based Soybean Seed Counting Algorithm for Salinity Stress Response Evaluation[J]. Journal of Crop Health, 2025, 77(5): 145.

[8] Liu F, Liu H, Wu Q, et al. Pod-pose: an efficient top-down keypoint detection model for fine-grained pod phenotyping in mature soybean[J]. Plant methods, 2025, 21(1): 82.

[9] Cai L, Shou X. PodFormer: An Adaptive Transformer-Based Framework for Instance Segmentation of Mature Soybean Pods in Field Environments[J]. Electronics, 2025, 15(1): 80.

[10] Wu Q, Liu H, Zhu H, et al. YOLO_SSP: An Auto-Algorithm to Detect Mature Soybean Stem Nodes Based on Keypoint Detection[J]. Agronomy, 2025, 15(5): 1128.

[11] Li Y, Teng S, Chen J, et al. FEI-YOLO: a lightweight soybean pod-type detection model[J]. Agronomy, 2024, 14(11): 2526.

[12] Nwankwo E, Onyenanu I, Nwobi-Okoye C. Sustainable Brake Pad Composites: RSM-Based Optimization of African Oil Bean and Palm Fiber for Enhanced Performance[J]. NIPES-Journal of Science and Technology Research, 2025, 7(4): 87-107.

[13] Yang H, Song Y, Liang Y, et al. SDC-YOLOv8: An Improved Algorithm for Road Defect Detection Through Attention-Enhanced Feature Learning and Adaptive Feature Reconstruction[J]. Sensors, 2026, 26(2): 609.

[14] Lei S, Yi H, Sarmiento J S. Synchronous end-to-end vehicle pedestrian detection algorithm based on improved yolov8 in complex scenarios[J]. Sensors, 2024, 24(18): 6116.

[15] Zhai Z L, Niu N W J, Feng B M, et al. An improved YOLOv8 model enhanced with detail and global features for underwater object detection[J]. Physica scripta, 2024, 99(9): 096008.

[16] Du Q, Zhang N, Bi W, et al. SST-YOLO: An Improved Autonomous Driving Object Detection Algorithm Based on YOLOv8[J]. Applied Sciences, 2026, 16(7): 3456.

[17] Zhang Y, Liu H, Dong D, et al. Dpf-yolov8: Dual path feature fusion network for traffic sign detection in hazy weather[J]. Electronics, 2024, 13(20): 4016.

[18] Han J, Lyu D, Xia C. SCEW-YOLOv8 Detection Model and Camera-LiDAR Fusion Positioning System for Whole-Growth-Cycle Management of Cabbage[J]. Applied Sciences, 2026, 16(7): 3510.

[19] Han J, Chen H, Ding Y, et al. You Only Look Once–Aluminum: A Detection Model for Complex Aluminum Surface Defects Based on Improved YOLOv8[J]. Symmetry, 2025,

17(5): 724.

[20] Xu W, Xu J, Ji Y, et al. Research on enhancing road apparent crack detection based on the improved YOLOv8n model[J]. Plos one, 2025, 20(9): e0330218.

[21] Xudong S, Xiumin S. SNF-YOLOv8: A Lightweight PCB Defect Detection Algorithm base on Multiscale Feature Fusion and Attention Scale Sequence Fusion[J]. Journal of Electronic Testing, 2025, 41(4): 561-573.