Algorithm for Wheat Spike Contour Extraction and Recognition in Complex Field Backgrounds

Xin Xu
College of Information and Management Science, Henan Agricultural University,
Haiyang Zhang
College of Information and Management Science, Henan Agricultural University,
Jiangchuan Lu
College of Information and Management Science, Henan Agricultural University,
Ziyi Guo
College of Information and Management Science, Henan Agricultural University,
Juanjuan Zhang
College of Information and Management Science, Henan Agricultural University,
Jibo Yue
College of Information and Management Science, Henan Agricultural University,
Yuanyuan Fu
College of Information and Management Science, Henan Agricultural University,
Xinming Ma
wheatdoctor@163.com

College of Information and Management Science, Henan Agricultural University,

Research Article

Keywords: Contour Extraction, Depthwise Separable Network, YOLOv9, Functional Enhancement, Attention Mechanism

Posted Date: April 24th, 2025

DOI: https://doi.org/10.21203/rs.3.rs-6224545/v1

Additional Declarations: No competing interests reported.

Version of Record: A version of this preprint was published at Plant Methods on August 22nd, 2025. See the published version at https://doi.org/10.1186/s13007-025-01433-1.

Abstract

Background: Spikelet number, a core phenotypic parameter for wheat yield composition, requires precise estimation through accurate spike contour extraction and differentiation between grain surfaces and spikelet surfaces. However, technical challenges persist in precise spike segmentation under complex field backgrounds and morphological differentiation between grain/spikelet surfaces.

Method: Building on two-year multi-angle wheat spike imagery, we propose an enhanced YOLOv9-LDS multi-scale object detection framework. The algorithm innovatively constructs a lightweight depthwise separable network (LDSNet) as backbone, balancing computational efficiency and accuracy through channel re-parameterization strategy; incorporates an ELA local attention module to build feature enhancement networks, and employs dual-path feature fusion mechanisms to strengthen edge texture responses, significantly improving discrimination of overlapping spikes and complex backgrounds. Further optimizes the loss function system by replacing traditional IoU with SIoU metric, enhancing bounding box regression through dynamic focus factors, and adding high-resolution small-object detection layers to mitigate dense spikelet feature loss.

Results: Independent test set validation shows the improved model achieves 83.9% contour integrity recognition rate and 92.4% mAP@0.5, exceeding baseline by 3.2 and 5.3 percentage points respectively. Ablation studies confirm LDSNet-ELA integration reduces false positives by 27.6%, while the enhanced loss function system improves small-object recall by 19.4%.

Conclusions: The proposed framework demonstrates superior performance in complex field scenarios with dense targets and dynamic illumination. The multi-scale feature synergy enhancement mechanism overcomes traditional models' limitations in detecting overlapping spikes. This method not only enables precise spike phenotyping but also provides robust algorithmic support for intelligent field spikelet counting systems, advancing translational applications in crop phenomics.

Introduction

Wheat is widely cultivated around the world and is one of the three major cereal crops. Wheat is of great significance to worldwide food production, food security, and the transformation of industrial patterns; accurate scientific prediction of wheat yield helps to guarantee national food supply security and maintain social stability [1]. The yield of wheat per unit area is determined by the number of spikes per unit area, the grain number per spike, and the thousand-grain weight of the variety [2]. While algorithms for counting spikes per unit area have become relatively mature and the thousand-grain weight can be obtained by consulting varietal information, the automated counting of grains per spike has emerged as the primary focus in automated yield estimation [3]. Conventional wheat yield assessments are generally carried out post-heading, involving manual counting of wheat grains on spikes; such a human-based survey method is both laborious and prone to inaccuracy [4].

The use of deep learning and computer vision technologies enables fast and precise estimation of wheat yield. Xu et al. [5] introduced a CBAM-HRNet model utilizing a convolutional quick attention mechanism that fuses neural network and image processing technologies to segment and count wheat spikes during the grain filling period, achieving segmentation accuracy above 85%. Geng et al. [6] introduced MATransUNet, a wheat spike grain segmentation model founded on multiple attention mechanisms, that first performs an initial segmentation of wheat spike grains, followed by the incorporation of an adaptive erosion module, SGCountM, thereby integrating image processing with the geometric and texture features of wheat spikes for segmentation and counting. However, in actual estimation procedures, because different wheat spike phenotypes (namely, the spike grain face and the spikelet face, as illustrated in Fig. 1) require distinct counting methods, quickly and accurately locating the complete spike contour while identifying both the spike grain face and the spikelet face is essential for precise spike grain counting [6].

With the continuous development of deep learning algorithms, they have demonstrated significant advantages in the field of wheat spike image segmentation [7]. Sébastien Dandrifosse et al. [8] proposed a wheat spike segmentation and counting model based on unsupervised learning and the DeepMAC segmentation algorithm, using a small amount of annotated RGB images from the heading to the maturity stage, and achieved an 86% segmentation accuracy through active learning. Wang et al. [9] introduced a modified EfficientDet-D0 object detection model for wheat spike identification, specifically targeting occlusion problems in wheat spike segmentation. They simulated occlusions in real wheat images by choosing and erasing rectangles according to the count and size of spikes present, ensuring that the model could capture additional features and boosting the accuracy to 94%. Wang et al. [10] implemented wheat spike segmentation under complex field background conditions based on a multi-stage convolutional neural network, accelerating the segmentation speed by using a new activation function, the Positive Rectified Linear Unit (PrLU). Li et al. [11] proposed a network model based on YOLOv7 that is capable of tracking wheat spike contours across consecutive frames, achieving a detection accuracy of 93.8%.

The development of neural networks has led to a qualitative improvement in the accuracy of wheat spike phenotype recognition [12–14]. Liu et al. [15] collected three-dimensional information of wheat spikes using radar, and classified wheat spike types by employing a kernel prediction convolutional neural network (KP-CNN) in conjunction with density-based spatial clustering and Laplacian-based region growing techniques. Arpan K. et al. [16] proposed an algorithmic model named SlypNet based on Mask R-CNN and U-Net, which is capable of extracting morphological features of wheat spikes from images to achieve classification and recognition. Wen et al. [17] proposed a method that integrates multi-scale features to recognize different varieties of wheat spikes, achieving an identification accuracy of 92.62%.

However, current research on the extraction and recognition of wheat spike contours faces limitations in terms of application and real-time performance, lacking practical implementation [8, 18–20]. The primary objective of this study is to achieve wheat spike contour extraction and phenotype classification in natural field environments. This study conducted a two-year natural field experiment using three wheat varieties, Xinmai26, Kexing3302, and Zhengmai136, collecting wheat spike images with a mobile terminal, and constructed an improved YOLOv9-LDS model to achieve complete extraction of wheat spike contours and classification of their types. The study modified the basic structure of the backbone network to mitigate the influence of complex backgrounds on the training process, and employed data augmentation to enhance the extraction of wheat spike features; it also embedded an Efficient Local Attention (ELA) module to strengthen target features at the spike edges, added a local small target detection layer to expand the receptive field and improve the recognition accuracy of overlapping spike contours, and replaced the loss function to achieve more accurate prediction boxes. This approach provides a viable solution for wheat spike contour extraction and serves as a reference for wheat spike phenotype classification in complex farmland scenarios.

Materials and methods

Experimental design

The experimental cultivation site was chosen at the Modern Agricultural Science and Technology Research Experimental Base in Yuanyang, Henan Province, China (35°6'46"N, 113°56'51"E). Varieties Xinmai26, Kexing3302, and Zhengmai136 were chosen for a two-year experiment [21]. Among them, the sowing rate was 187.5 kg/hm^2 for Xinmai26, 185.5 kg/hm^2 for Kexing3302, and 187.5 kg/hm^2 for Zhengmai136; the sowing dates were October 12, 2023, and October 15, 2024, for the respective years. For each variety, four fertilization approaches were employed: Nitrogen Fertilizer 10 (21.74 kg of urea and 8 kg of triple superphosphate per mu), Nitrogen Fertilizer 15 (32.61 kg of urea and 8 kg of triple superphosphate per mu), organic fertilizer (chicken manure), and straw return; with a basal-to-topdressing ratio of 6:4, topdressing was applied at the jointing stage (March of the following year). Other measures were consistent with standard high-yield fields.

Data acquisition

Data was collected from April 2023 to the end of May 2023, and from April 2024 to early June 2024. The collection was conducted under clear, windless conditions, with photography carried out from 9:00 to 16:30. The collection devices included the XiaoMi MAX 2S smartphone (with a 12-megapixel dual-camera system featuring wide-angle and telephoto lenses) and the RedMi K40 Pro smartphone (equipped with a 64-megapixel main camera, an 8-megapixel ultra-wide lens, and a 5-megapixel tele-macro lens). The raw images were in JPG format, having resolutions of 4032×3024 and 9248×6944.

In order to ensure the complexity and rigor of the training samples and to eliminate the offsets in prediction box boundaries and positions during background and group recognition, three sampling methods were adopted in the experiment: The first approach involves in-situ field sampling, directly collecting group wheat spikes and ensuring that each image includes three or more plants with distinctly visible contours and clarity adequate for identifying both the spikelet face and the spike grain face phenotypes; The second approach consists of in-field detached sampling, capturing images of isolated single wheat spikes in a natural field setting; The third approach is laboratory-based detached sampling, capturing images of individual isolated wheat spikes under controlled laboratory conditions. As illustrated in Fig. 2.

In order to reduce errors resulting from the shooting distance, a consistent 10 cm distance was maintained during photography. Additionally, random sampling points were arranged in each experimental area, where 5 to 6 images were taken at each point to ensure a minimum of 42 images per area; in all, this study obtained 3024 original images. Refer to Table 1.

Table 1
Wheat ear dataset information. Each side of the wheat ears was photographed to expand the dataset.

Wheat varieties	Shoot data	Weather	Growth period	Image size	Shoot device	Number of images
Kexing3302	16/04/2023	Sunny	Stage grouting	4032*3024	XiaoMi MAX 2S	102
				9248*6944	RedMi K40 pro	133
	15/05/2023	Cloudy	Stage grouting	4032*3024	XiaoMi MAX 2S	98
				9248*6944	RedMi K40 pro	117
	20/04/2024	Sunny	Stage grouting	4032*3024	XiaoMi MAX 2S	119
				9248*6944	RedMi K40 pro	143
	18/05/2024	Sunny	Stage grouting	4032*3024	XiaoMi MAX 2S	137
				9248*6944	RedMi K40 pro	159
Xinmai 26	16/04/2023	Sunny	Stage grouting	4032*3024	XiaoMi MAX 2S	126
				9248*6944	RedMi K40 pro	137
	15/05/2023	Cloudy	Stage grouting	4032*3024	XiaoMi MAX 2S	115
				9248*6944	RedMi K40 pro	130
	20/04/2024	Sunny	Stage grouting	4032*3024	XiaoMi MAX 2S	126
				9248*6944	RedMi K40 pro	129
	18/05/2024	Sunny	Stage grouting	4032*3024	XiaoMi MAX 2S	100
				9248*6944	RedMi K40 pro	145
Zhengmai136	16/04/2023	Sunny	Stage grouting	4032*3024	XiaoMi MAX 2S	116
				9248*6944	RedMi K40 pro	124
	15/05/2023	Cloudy	Stage grouting	4032*3024	XiaoMi MAX 2S	145
				9248*6944	RedMi K40 pro	114
	20/04/2024	Sunny	Stage grouting	4032*3024	XiaoMi MAX 2S	126
				9248*6944	RedMi K40 pro	136
	18/05/2024	Sunny	Stage grouting	4032*3024	XiaoMi MAX 2S	120
				9248*6944	RedMi K40 pro	127

Data processing

Data normalization

Differences in image resolution may impose greater requirements on devices during later stages of model training [22]. To avoid resource wastage, normalization was applied to the original image data. An online image processing tool was used to enhance image quality and compress the dimensions of all original image files, normalizing the resolution to 640×640 while ensuring clarity.

Data enhancement

The number and complexity of images can affect model performance. Data augmentation employs simple processing techniques to expand the original images, addressing the issue of insufficient data [23]. To ensure the diversity and complexity of the dataset, images obtained from in-field on-site sampling were horizontally segmented by setting random pixel points along the X-axis to crop them into 3–5 images. For images acquired via detached sampling, data augmentation was performed through rotations of 90°, 180°, and 270°, as well as horizontal and vertical mirroring and other processing techniques.

Dataset construction

After processing, some images may pose challenges for feature extraction; therefore, in line with the primary objectives of this study, images with indistinct target features or poor quality were manually filtered out during annotation using the LabelImg tool. A total of 6,865 normalized images were obtained after data augmentation. After annotation and screening, a basic dataset comprising 6,572 images was formed and divided into training, validation, and test sets in an 8:1:1 ratio. The training set included 5,258 images, the validation set 655 images, and the test set 659 images. In total, 26,597 samples of the spike grain face and the spikelet face were annotated.

Model construction

In complex backgrounds, wheat spike images face problems of highly similar edge features and pronounced adhesion. Moreover, factors such as spike overlap, leaf occlusion, and interference from awns result in redundant computations and the omission of small edge target features in the extraction network. Thus, the model is required to have enhanced abilities for global information awareness and edge feature differentiation.

Being one of the excellent single-stage object detection algorithms at present, YOLOv9 has shown remarkable potential for fast detection of wheat spike phenotypes [24]. In order to ensure a balance between contour extraction and recognition accuracy, the YOLOv9-C variant—with a moderate model size and superior detection speed—was chosen for deep learning training [24].

To prevent parameter and data redundancy in the backbone model, this study optimized the backbone network based on the MobileNetV4 architecture [25], aiming to enhance feature extraction and expand the model's receptive field. To address the challenge of extracting target features in complex backgrounds, the ELA (Efficient Local Attention) module [26] was introduced, effectively enhancing the model's ability to recognize edge features and improving the integrity of contour extraction. Considering that occlusion between leaves and wheat spikes can affect bounding box regression performance, the SIoU (Scylla Intersection over Union) loss function [27] was used to replace the original function, enhancing the sensitivity of bounding boxes to spatial positions and accelerating model convergence. Additionally, a dedicated small-target detection layer was added to the neck network to address the loss of small target features in edge and overlapping regions, significantly enhancing detection accuracy. The specific network architecture is shown in Fig. 3.

Lightweight network architecture

To ensure real-time performance and accuracy in automated yield estimation, this study proposed a Lightweight Depthwise Separable Network (LDSNet) based on MobileNetV4 to replace the backbone network in YOLOv9. As shown in Fig. 4.

MobileNetV4 creatively proposed the Universal Inverted Bottleneck (UIB) block. The UIB block serves as a configurable component for efficient network design, improving network performance while maintaining the same learning complexity [25].

Throughout the neural network training process in this study, the image data produced substantial redundant feature map data. These data are essentially extracted from individual or groups of images, with their feature points exhibiting high similarity. Handling redundant data not only consumes significant computational resources but also reduces the efficiency of feature learning. Since the similarity between the spike contour area and the background is high, the learning of edge features and central axis features is especially crucial for classifying wheat spike types. Consequently, the network architecture was designed as follows: First, a standard 3×3 convolution layer was used at the input to ensure effective extraction of low-level features from the input images; Next, Extra Depthwise Separable Convolution (EDWConv) was employed from the second to the ninth layer. To improve data transfer efficiency, a Squeeze-and-Excitation (SE) module [28] was incorporated between the Extra Depthwise Convolution (Extra DW) and the Pointwise Convolution (PW) [29] across all EDWConv layers. The module is capable of compressing upper-layer information to normalize the convolutional data, thus strengthening the response of the information channels. Specifically, the second to seventh layers use 3×3 convolution kernels, while the eighth and ninth layers employ 5×5 kernels. To further enhance computational efficiency, the Sigmoid function in the SE module was replaced with the less computationally intensive Hard-Sigmoid function [30] to accelerate regression computations. Moreover, the SPPF module within the backbone network was preserved; it pools feature maps of different sizes, ensuring effective connectivity and facilitating the detection of multi-scale targets.

ELA attention mechanism

The ELA module is a lightweight and efficient local attention mechanism designed to enhance the localization capability of deep convolutional neural networks (CNNs) for target regions. By introducing minimal additional parameters into the network, it significantly improves the model's overall performance [26]. Compared with the commonly used Coordinate Attention (CA) mechanism, ELA overcomes CA's limitations in generalization capability and channel dimensionality reduction, enabling precise positional prediction in spatial dimensions.

Wheat spikes exhibit random horizontal arrangements and vertical height imbalance in imaging. This leads to unstable results during complete contour extraction due to subtle coordinate variations among wheat spikes. ELA employs a strip pooling strategy in spatial dimensions to separately extract horizontal and vertical feature vectors. This design not only captures long-range or offset dependencies among wheat spike samples, but also effectively avoids interference from non-target regions (background) on label prediction (as shown in Fig. 5). Through this approach, ELA generates information-rich positional feature vectors in each direction, providing high-quality feature representations for subsequent attention prediction.

During feature processing, ELA independently processes horizontal and vertical feature vectors, integrates their information through multiplicative operations, and combines spatially deviated wheat spike features in vertical planes, achieving precise localization of target regions of interest. Moreover, considering that overlapping wheat spikes typically manifest as continuous feature vectors in planes, ELA can process these sequential signals using 1D convolution.

Compared with 2D convolution, 1D convolution demonstrates higher computational efficiency and better suitability for processing sequential feature information. Unlike CA's 1×1 convolutional kernels, ELA employs 1D convolutions with kernel sizes of 5 or 7, significantly enhancing feature vector interactions for precise positional embedding, effective edge feature extraction in overlapping regions, and reduced information attenuation.

SIoU Loss Function

The YOLOv9 algorithm employs a compound loss function combining Distributional Focal Loss (DFL) and Complete Intersection over Union (CIoU) loss. The DFL loss function handles targets with ambiguous boundaries or partial occlusion by predicting bounding box probability distributions rather than direct positional coordinates. This approach effectively distinguishes feature similarities/differences among overlapping wheat spikes, reducing excessive coverage or displacement of prediction boxes.

The CIoU loss emphasizes bounding box shape information through correction factors, enhancing robustness to varying sizes/shapes and improving shape accuracy [31]. However, CIoU still shows limitations in handling dense occlusion scenarios like overlapping wheat spikes and leaves.

To optimize regression loss efficiency and improve wheat spike phenotyping contour extraction accuracy, we replace CIoU with Shape Intersection over Union (SIoU) loss while retaining DFL. The SIoU loss function is defined in Eq. (1): Here, IoU denotes the intersection-over-union between predicted and ground-truth boxes, serving as a key metric for positive sample determination. The SIoU loss contains three key components: angle cost Λ, distance cost Δ, and shape cost Ω, defined as:

$$Loss_{SIoU} = 1 - \mathrm{IoU} + \frac{\mathit{\Delta} + \mathit{\Omega}}{2}$$

1

Here, IoU denotes the intersection-over-union between predicted and ground-truth boxes, serving as a key metric for positive sample determination. The SIoU loss contains three key components: angle cost Λ, distance cost Δ, and shape cost Ω, defined as: In these equations, γ represents temporal distance, p_x and p_y are squared ratios of width/height between ground-truth and predicted boxes, θ∈[2, 6] adjusts shape loss emphasis. Figure 6 illustrates schematic parameters in SIoU loss, including centers B (predicted) and BGT (ground-truth), planar distances H and W, angles α and β, center distance σ, and edge distances CH and CW.

$$\wedge = \cos\left(2\left(\arcsin\left(\frac{H}{\sigma}\right)\right) - \frac{\pi}{4}\right)$$

2

$$\mathit{\Delta} = 2 - e^{-\gamma\, p_x} - e^{-\gamma\, p_y}$$

3

$$\mathit{\Omega} = \sum\nolimits_{t=w,h} (1 - e^{-W_t})^{\theta}$$

4

In these equations, γ represents temporal distance, p_x and p_y are squared ratios of width/height between ground-truth and predicted boxes, $\theta \in [2{,}6]$ adjusts shape loss emphasis. Figure 6 illustrates schematic parameters in SIoU loss, including centers B (predicted) and B^{GT}(ground-truth), planar distances H and W, angles α and β, center distance σ, and edge distances C_H and C_W.

By comprehensively considering angular, distance, and shape costs, SIoU optimizes model convergence speed and prediction accuracy while avoiding computational overhead, achieving lightweight wheat spike contour extraction.

Small Target Feature Detection Layer

In agricultural vision detection, spatial overlap between wheat spikes and background or other plants causes significant attenuation of effective edge features. For dataset construction, we employ instance-level copy-paste augmentation with Poisson blending [32], combined with random rotation (± 90°), multi-scale transformation (0.5–1.5×), and random cropping (10–20% regions). Enhanced samples generated via Monte Carlo sampling [33] enable the model to learn robust feature representations under occlusion during training.

Additionally, we augment the original YOLOv9 feature pyramid by adding a small-object detection head at the Backbone's C2 stage (medium-level features: 320×320 input → 160×160 output). This layer provides 4×/16×/64× higher spatial resolution than original detection heads (80×80, 40×40, 20×20 outputs). As calculated by Eq. (5), its 3×3 convolutional kernels achieve 28×28 pixel effective receptive fields, better matching wheat spikes' average 15×15 pixel scale:

$$\mathcal{R} = 2^{(s+1)} \times\ (k-1) + 1$$

5

Multi-layer feature extraction enables shallow high-resolution features (160×160) to capture fine edge textures (avg. 15×15px), while deep features provide categorical semantics. Feature fusion combines shallow local details (e.g., residual edges of partially occluded spikes) with deep global context. This establishes a complete detection framework covering target scales from 5-150 pixels.

Evaluation indicators

This study employs a comprehensive model evaluation framework covering two dimensions: detection accuracy and computational efficiency. Detection accuracy is measured by Precision (P), Recall (R), and mean Average Precision (mAP). The relevant formulas are: Here, P represents the proportion of true positive samples among all predicted positive samples, reflecting detection reliability; R indicates the proportion of correctly predicted positive samples among all actual positives, demonstrating model coverage capability; mAP is the arithmetic mean of Average Precision (AP) across all categories, evaluating overall performance in multi-class detection tasks. AP is defined in Eq. (9):

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

6

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

7

$$mAP = \frac{1}{|Q_R|} \sum {}_{q \in Q_R} AP\left(q\right)$$

8

Here, P represents the proportion of true positive samples among all predicted positive samples, reflecting detection reliability; R indicates the proportion of correctly predicted positive samples among all actual positives, demonstrating model coverage capability; mAP is the arithmetic mean of Average Precision (AP) across all categories, evaluating overall performance in multi-class detection tasks. AP is defined in Eq. (9):

$$AP = \int_0^1 P\left(R\right) dR$$

9

Where TP (True Positive) counts correctly predicted positives, FP (False Positive) counts incorrectly predicted positives, and FN (False Negative) counts positives incorrectly predicted as negatives. Q_R denotes the target category set, q represents a detection category, and AP(q) is the average precision for category q.

To intuitively evaluate model prediction performance, the coefficient of determination (R^2) is introduced to quantify the proportion of data variability explained by the model. Mean Squared Error (MSE) and Root Mean Squared Error (RMSE) are adopted, where MSE represents the average of squared differences between predicted and true values, and RMSE is the square root of MSE.

$$p_i = \frac{|x_i - y_i|}{x_i} \times \ 100\backslash\%.$$

10

$$R^2 = 1 - \frac{\sum_{i=1}^{n}(x_i - y_i)^2}{\sum_{i=1}^{n}(x_i - \bar{x})^2}.$$

11

$$RMSE = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(x_i - y_i)^2}.$$

12

where P_i is a relative error for a single sample; x_i is the true number of the spike; y_i is the predicted number of the spike; $\bar{x}$ is the average of the true values of the number of the spike; n is the number of samples per variety.

Computational efficiency is evaluated using Floating-point Operations (FLOPs), parameter count (Params), and Frames Per Second (FPS) as primary metrics. FLOPs and Params assess model lightweighting from computational resource perspectives, while FPS directly reflects real-time performance in practical deployment.

where P_i is a relative error for a single sample; x_i is the true number of grains on the spike; y_i is the predicted number of grains on the spike; $\bar{x}$ is the average of the true values of the number of grains on the spike; n is the number of samples per variety.

Performance analysis

Experimental Setup

All experiments in this study were conducted under a unified hardware and software environment configured as follows: Intel® Xeon(R) Silver4114 CPU @2.20GHz, NVIDIA Quadro T400 GPU (16GB VRAM), 64GB RAM; with software environment built on Ubuntu 20.04 LTS OS, Python 3.11, and PyTorch 2.0.0 deep learning framework.

During the initial training phase, random initialization of network weights results in weak feature extraction capability, leading to significant performance degradation [34]. To address this issue, we employ a progressive training strategy with dynamic hyperparameter adjustment to optimize the training process, ensuring effective loss function convergence while preventing overfitting. Specific parameter settings include: Cosine annealing learning rate scheduler [35] with initial learning rate (lr_0) = 0.01, final learning rate (lr_f) = 0.05, following Eq. (13):

$$lr = lr_f + \frac{1}{2}(lr_0 - lr_f)(1 + \cos(\frac{T_{cur}}{T_{max}}\pi\ ))$$

13

where T_{cur} denotes current training epoch and T_{max} represents total epochs.

Confidence threshold (conf_threshold) = 0.5 for filtering low-confidence predictions;IoU threshold = 0.45 for Non-Maximum Suppression (NMS);Stochastic Gradient Descent (SGD) optimizer with momentum = 0.937 and weight decay = 0.0005.

Ablation Study

To validate the effectiveness of the YOLOv9-LDS algorithm, we conducted systematic ablation experiments under identical hardware, software environments, and datasets. All comparative models used identical training parameters throughout the experiments to ensure fairness. To mitigate randomness effects, each model configuration underwent 10 independent trials with arithmetic mean as final result (see Table 2).

Table 2
Results of ablation experiment

Serial Number	LDSNet	ELA	SIoU	Feature Detection Layer	P/%	R/%	mAP/%	GFLOPs/G	Params/M	FPS
1	✗	✗	✗	✗	80.7	79.3	87.1	102.1	3.56	101
2	√	✗	✗	✗	81.1	82.1	85.1	**55.8**	1.85	**175**
3	√	√	✗	✗	80.9	79.9	86.9	68.9	1.91	121
4	√	✗	√	✗	79.9	80.2	85.9	65.7	1.99	115
5	√	✗	✗	√	82.2	81.1	87.3	72.6	**1.80**	117
6	√	√	√	✗	83.4	80.7	89.5	64.3	2.24	114
7	√	√	✗	√	83.1	82.1	88.4	68.9	2.13	124
8	√	✗	√	√	82.4	81.3	90.1	70.1	2.36	120
9	√	√	√	√	**83.9**	**84.1**	**92.4**	78.9	2.38	133

With LDSNet backbone, compared to baseline model (No.1), mAP decreased 2.0% while computational load (GFLOPs) reduced 45.3%, parameters (Params) decreased 48.0%, and FPS increased 73.3%, demonstrating LDSNet's significant computational/parametric efficiency at the cost of partial feature extraction capability. Adding ELA module to LDSNet improved mAP by 1.8% with 23.5% computational increase and 30.9% FPS reduction, proving ELA enhances feature representation while introducing computational overhead. Adopting SIoU loss improved mAP by 0.8% with 9.6% parameter reduction, indicating SIoU optimizes bounding box regression and accelerates convergence. Incorporating small-target feature layers boosted mAP by 2.2% with 2.7% parameter decrease, demonstrating enhanced small-object detection capability.

Pairwise module combinations (No.6–8) achieved mAP improvements (88.4%-90.1%) with 15.2%-25.6% computational increases while maintaining 114–124 FPS. Multi-module combinations further enhanced detection accuracy at increased computational complexity. The full-module configuration achieved optimal performance: 5.3% mAP improvement with 22.7% GFLOPs reduction, 33.1% parameter decrease, and 31.7% FPS increase. This demonstrates the full-module design optimally balances accuracy/efficiency, achieving 5.3% mAP gain with 22.7% computation and 33.1% parameter reductions, validating algorithm effectiveness. The enhanced model demonstrated superior performance in complex farmland scenarios, effectively extracting wheat spike contours and performing phenotyping classification.

Comparative Experiment on Segmentation Accuracy Across Models

To validate the effectiveness of the proposed YOLOv9-LDS algorithm on wheat spike phenotyping datasets in complex farmland scenarios, we conducted comparative experiments with mainstream single-stage and two-stage object detection models. The selected models include single-stage detectors (YOLOv8, YOLOv9, YOLOv10, Edge-YOLO, Gold-YOLO) and the two-stage model Faster R-CNN. All experiments were conducted under identical hardware/software configurations with consistent training parameters. To mitigate randomness, each model underwent 10 independent trials with arithmetic mean as final result (Table 3).

Table 3
Results of comparative experimental

Models	P/%	R/%	mAP/%
Faster R-CNN	70.1	78.3	74.1
YOLOv8	80.5	82.6	88.3
YOLOv9	82.7	83.8	88.6
YOLOv10	79.5	80.2	87.4
Edge-YOLO	75.6	81.4	85.6
Gold-YOLO	81.4	78.6	82.4
OUR	**83.9**	**84.1**	**92.4**

Results show Faster R-CNN achieved 74.1% mAP, significantly lower than single-stage models, indicating limited small-object detection capability in complex farmland scenarios. This is primarily due to interference in two-stage models' feature extraction and region proposal mechanisms under complex backgrounds, degrading small-object detection performance.

YOLOv8, YOLOv9, and YOLOv10 achieved comparable mAP scores of 88.3%, 88.6%, and 87.4% respectively. Edge-YOLO and Gold-YOLO showed slightly lower mAP (85.6% and 82.4%) compared to the YOLO series. The YOLO series enhances detection capability in complex scenarios through multi-scale feature fusion and loss function optimization, while Edge-YOLO/Gold-YOLO prioritize lightweight design at some accuracy cost.

Our model achieved superior precision (83.9%) and recall (84.1%) over counterparts, demonstrating significant advantages in reducing false positives and negatives. With 92.4% mAP, our model outperformed the best baseline (YOLOv9, 88.6%) by 3.8 percentage points, confirming its superiority in complex farmland scenarios. This demonstrates the model's excellent performance in complex farmland environments, effectively suppressing background interference to accurately extract wheat spike contours and perform phenotyping classification.

Contour Extraction Performance Across Growth Stages

Morphological and physiological changes during wheat growth stages significantly impact model performance. Neural networks require adjustments based on growth stage characteristics to enhance model robustness.

To evaluate the influence of growth-stage traits on wheat spike contour extraction, 150 field images from complex agricultural scenarios over a two-year period were selected for testing. Detailed extraction results are presented in Fig. 7.

Figure 7 demonstrates comparative contour extraction across growth stages and models. The main spike contours were successfully extracted across all stages with acceptable variation ranges. All models accurately captured main spike structures, but the improved model showed superior sensitivity to edge tone variations and better resistance to feature drift. Our model demonstrated significant improvements in overlapping region predictions with higher accuracy compared to counterparts, effectively filtering incomplete spikes during extraction. The algorithm effectively meets practical requirements for detecting spikelet surfaces and extracting contours in complex field environments.

Spike Recognition Performance Across Growth Stages

To ensure simultaneous spike contour extraction and phenotyping capabilities, the enhanced model was evaluated for spike detection performance across growth stages. 150 field images (50 per early/mid/late grain filling stages) from two-year complex agricultural scenarios were selected. Detection target statistics and results are shown in Fig. 8, with visual demonstrations in Fig. 9.

Figure 8 reveals relatively dispersed predictions in early grain filling stage, primarily due to incomplete spike contours, though basic phenotyping features suffice for majority detections. Mid/late stages exhibit well-defined contours with stable spike-background contrast, distinct edge differentiation, and near-perfect alignment with ground truth due to enhanced feature discriminability. The model demonstrates robust spike extraction and classification in complex environments, providing reliable scientific data for wheat yield estimation. This establishes critical prerequisites for automated yield measurement systems.

Discussion

With the acceleration of agricultural intelligence, deep learning-based wheat phenotyping detection has emerged as a critical approach for enhancing yield prediction accuracy [18]. However, significant phenotypic variations across growth stages and severe background interference in complex field scenarios persist, while conventional spike detection methods often rely on costly equipment (e.g., LiDAR) for 3D point cloud acquisition and manual geometric feature engineering, limiting algorithm generalization.

Liu et al. [36] proposed LiDAR combined with Kernel Prediction CNN (KP-CNN) achieving 84.62% recognition accuracy, yet its dependence on specialized hardware and manual annotations hinders large-scale field deployment. Similarly, international UAV-based high-throughput phenotyping analyses capture multidimensional data but exhibit light sensitivity and high computational complexity, struggling to meet real-time requirements.

The proposed YOLOv9-LDS algorithm reduces computational load by 45.3% (GFLOPs = 55.8) via LDSNet backbone and channel re-parameterization, while enhancing edge features through ELA local attention to achieve 92.4% mAP@0.5 - a 5.3% improvement over baseline (YOLOv9-C). This outperforms LiDAR solutions (84.62% segmentation accuracy) without expensive hardware dependencies, significantly enhancing practicality. The SIoU loss function with dynamic focus factors alleviates dense spikelet feature loss, improving small-object recall by 19.4%.Poisson blending and random cropping augmentation strategies enhance robustness against complex backgrounds (e.g., spike-leaf overlaps), reducing false positives by 27.6%.

Despite excellent performance in standard conditions, limitations include: training data concentration in Henan's clear/cloudy weather, with reduced adaptability to rain/fog/glare (error rate increases ~ 15%).Future work requires cross-climatic datasets and multispectral fusion (NIR/thermal imaging) to enhance feature representation; Though achieving 133 FPS, parameter reduction (current 2.38M) is needed for robotic arm integration. TensorRT quantization and adaptive confidence strategies (dynamic threshold 0.4→0.6 during grain filling) could enable embedded deployment with < 50 ms latency; Current research neglects growth-stage temporal features. Future designs could incorporate spatiotemporal fusion networks with Spike Morphology Development Index (SMDI) for full-growth-stage phenotyping prediction.

Through algorithmic innovation and systematic engineering design, this study provides an efficient solution for spike phenotyping in complex farmland. It establishes a technical paradigm for translating crop phenomics from laboratory research to field engineering applications. Future multimodal data fusion and embedded deployment optimization may drive transformative development in smart wheat breeding and precision agriculture.

Conclusion

This study proposes an enhanced YOLOv9-LDS algorithm to address the challenges of wheat spike contour extraction and phenotyping classification in complex farmland scenarios. A two-year field experiment (covering three varieties: Xinmai 26, Kexing 3302, Zhengmai 136) established a multi-source dataset of 6,572 images, enhanced through Poisson blending augmentation, random cropping, and multi-scale transformation strategies to improve data diversity. Replacing the original YOLOv9-C backbone with a lightweight depthwise separable network (LDSNet) based on MobileNetV4 architecture significantly improved computational efficiency (45.3% GFLOPs reduction to 55.8G, 73.3% FPS increase to 175 fps), while the ELA local attention module enhanced edge feature perception and SIoU loss function improved spatial localization accuracy (18.6% reduction in bounding box regression error). Ablation studies demonstrated that the full-module configuration achieved 83.9% precision, 84.1% recall, and 92.4% mAP@0.5 on test sets - improvements of 3.2%, 4.8%, and 5.3% over baseline (YOLOv9-C) respectively, with 33.1% parameter reduction to 2.38M, confirming algorithm robustness in complex backgrounds.

This study provides a high-precision, lightweight technical solution for automated wheat yield estimation, with its modular improvement framework being extensible to other crop phenotyping analyses.

Declarations

Ethics approval and consent to participate

Not applicable.

Consent for publication

Not applicable.

Competing interests

The authors declare no competing interests.

Author details

[1] College of Information and Management Science, Henan Agricultural University, Zhengzhou 450002, China;

[2] Agricultural College, Henan Agricultural University, Zhengzhou 450002, China

Funding

Major Science and Technology Project of Henan Province, China. (241100110300)

Author Contribution

Xin Xu and Haiyang Zhang wrote the manuscript. Haiyang Zhang and Jiangchuan Lu performed the in-field imaging. Juanjuan Zhang, Jibo Yue, Yuanyuan Fu, and Xinming Ma supervised wheat field experiments and provided biological expertise. Xin Xu, Haiyang Zhang, and Ziyi Guo designed the research. Haiyang Zhang built and tested the deep learning models. All authors read and approved the final manuscript.

Acknowledgments

The authors thank the graduate students at the College of Information and Management Science and the Ministry of Agriculture at Henan Agricultural University for their continued support of our research.

Availability of data and materials

The datasets used for the analysis are available from the corresponding author upon reasonable request.

References

1. Guarin JR, Martre P, Ewert F, Webber H, Dueri S, Calderini D, et al. Evidence for increasing global wheat yield potential. Environ Res Lett. 2022;17:124045.
2. Hu W-J, Fan J, Du Y-X, Li B-S, Xiong N, Bekkering E. MDFC–ResNet: An agricultural IoT system to accurately recognize crop diseases. IEEE Access. 2020;8:115287–98.
3. Tanabe R, Matsui T, Tanaka TST. Winter wheat yield prediction using convolutional neural networks and UAV-based multispectral imagery. Field Crops Res. 2023;291:108786.
4. Carlier A, Dandrifosse S, Dumont B, Mercatoris B. Wheat ear segmentation based on a multisensor system and superpixel classification. Plant Phenomics [Internet]. 2022 [cited 2024 Oct 31];2022. Available from: https://spj.science.org/doi/full/10.34133/2022/9841985
5. Xu X, Geng Q, Gao F, Xiong D, Qiao H, Ma X. Segmentation and counting of wheat spike grains based on deep learning and textural feature. Plant Methods. 2023;19:77.
6. Geng Q, Zhang H, Gao M, Qiao H, Xu X, Ma X. A rapid, low-cost wheat spike grain segmentation and counting system based on deep learning and image processing. Eur J Agron. 2024;156:127158.
7. Zou Z, Chen K, Shi Z, Guo Y, Ye J. Object detection in 20 years: A survey. Proc IEEE. 2023;111:257–76.
8. Dandrifosse S, Ennadifi E, Carlier A, Gosselin B, Dumont B, Mercatoris B. Deep learning for wheat ear segmentation and ear density measurement: From heading to maturity. Comput Electron Agric. 2022;199:107161.
9. Wang Y, Qin Y, Cui J. Occlusion robust wheat ear counting algorithm based on deep learning. Front Plant Sci [Internet]. 2021 [cited 2025 Feb 28];12. Available from: https://www.frontiersin.org/journals/plant-science/articles/10.3389/fpls.2021.645899/full
10. Wang D, Zhang D, Yang G, Xu B, Luo Y, Yang X, SSRNet. In-field counting wheat ears using multi-stage convolutional neural network. IEEE Trans Geosci Remote Sens. 2022;60:1–11.
11. Li Z, Zhu Y, Sui S, Zhao Y, Liu P, Li X. Real-time detection and counting of wheat ears based on improved YOLOv7. Comput Electron Agric. 2024;218:108670.
12. Ren S, He K, Girshick R, Sun J, Faster R-CNN. Towards real-time object detection with region proposal networks. IEEE Trans Pattern Anal Mach Intell. 2017;39:1137–49.
13. He K, Gkioxari G, Dollár P, Girshick R, Mask. R-CNN [Internet]. arXiv; 2018 [cited 2024 Jul 18]. Available from: http://arxiv.org/abs/1703.06870
14. Tian Y, Yang G, Wang Z, Li E, Liang Z. Instance segmentation of apple flowers using the improved mask R–CNN model. Biosyst Eng. 2020;193:264–78.
15. Liu Z, Jin S, Liu X, Yang Q, Li Q, Zang J, et al. Extraction of wheat spike phenotypes from field-collected lidar data and exploration of their relationships with wheat yield. IEEE Trans Geosci Remote Sens. 2023;61:1–13.
16. Maji AK, Marwaha S, Kumar S, Arora A, Chinnusamy V, Islam S, SlypNet. Spikelet-based yield prediction of wheat using advanced plant phenotyping and computer vision techniques. Front Plant Sci [Internet]. 2022 [cited 2025 Feb 28];13. Available from: https://www.frontiersin.org/journals/plant-science/articles/10.3389/fpls.2022.889853/full
17. Wen C, Wu J, Chen H, Su H, Chen X, Li Z et al. Wheat spike detection and counting in the field based on SpikeRetinaNet. Front Plant Sci [Internet]. 2022 [cited 2025 Feb 28];13. Available from: https://www.frontiersin.org/journals/plant-science/articles/10.3389/fpls.2022.821717/full
18. Wang S, Zhao J, Cai Y, Li Y, Qi X, Qiu X, et al. A method for small-sized wheat seedlings detection: From annotation mode to model construction. Plant Methods. 2024;20:15.
19. Farjon G, Edan Y. AgroCounters—a repository for counting objects in images in the agricultural domain by using deep-learning algorithms: Framework and evaluation. Comput Electron Agric. 2024;222:108988.
20. Fernandez-Gallego JA, Lootens P, Borra-Serrano I, Derycke V, Haesaert G, Roldán-Ruiz I, et al. Automatic wheat ear counting using machine learning based on RGB UAV imagery. Plant J. 2020;103:1603–13.

21. Gu Y, Wang Y, Wu Y, Warner TA, Guo T, Ai H, et al. Novel 3D photosynthetic traits derived from the fusion of UAV LiDAR point cloud and multispectral imagery in wheat. Remote Sens Environ. 2024;311:114244.
22. Misra T, Arora A, Marwaha S, Jha RR, Ray M, Jain R, et al. Web-SpikeSegNet: Deep learning framework for recognition and counting of spikes from visual images of wheat plants. IEEE Access. 2021;9:76235–47.
23. Hu G, Qian L, Liang D, Wan M. Self-adversarial training and attention for multi-task wheat phenotyping. Appl Eng Agric. 2019;35:1009–14.
24. Li J, Feng Y, Shao Y, Liu F. IDP-YOLOV9: Improvement of object detection model in severe weather scenarios from drone perspective. Appl Sci. 2024;14:5277.
25. Qin D, Leichner C, Delakis M, Fornoni M, Luo S, Yang F et al. MobileNetV4 -- universal models for the mobile ecosystem [Internet]. arXiv; 2024 [cited 2024 Oct 14]. Available from: http://arxiv.org/abs/2404.10518
26. Xu W, Wan Y. ELA: Efficient local attention for deep convolutional neural networks [Internet]. arXiv; 2024 [cited 2024 Jul 21]. Available from: http://arxiv.org/abs/2403.01123
27. Gevorgyan Z. SIoU loss: More powerful learning for bounding box regression [Internet]. arXiv; 2022 [cited 2024 Oct 15]. Available from: http://arxiv.org/abs/2205.12740
28. Hu J, Shen L, Sun G. Squeeze-and-excitation networks. 2018 [cited 2024 Nov 3]. pp. 7132–41. Available from: https://openaccess.thecvf.com/content_cvpr_2018/html/Hu_Squeeze-and-Excitation_Networks_CVPR_2018_paper.html
29. Hua B-S, Tran M-K, Yeung S-K. Pointwise convolutional neural networks. 2018 [cited 2024 Nov 3]. pp. 984–93. Available from: https://openaccess.thecvf.com/content_cvpr_2018/html/Hua_Pointwise_Convolutional_Neural_CVPR_2018_paper.html
30. Louizos C, Welling M, Kingma DP. Learning sparse neural networks through l_0 regularization [Internet]. arXiv; 2018 [cited 2024 Nov 3]. Available from: http://arxiv.org/abs/1712.01312
31. Wang X, Song J, ICIoU. Improved loss based on complete intersection over union for bounding box regression. IEEE Access. 2021;9:105686–95.
32. Ghiasi G, Cui Y, Srinivas A, Qian R, Lin T-Y, Cubuk ED et al. Simple copy-paste is a strong data augmentation method for instance segmentation. 2021 [cited 2025 Feb 21]. pp. 2918–28. Available from: https://openaccess.thecvf.com/content/CVPR2021/html/Ghiasi_Simple_Copy-Paste_Is_a_Strong_Data_Augmentation_Method_for_Instance_CVPR_2021_paper.html?ref=https://githubhelp.com
33. Harrison RL. Introduction to monte carlo simulation. AIP Conf Proc. 2010;1204:17–21.
34. Khaki S, Safaei N, Pham H, Wang L, WheatNet:. A lightweight convolutional neural network for high-throughput image-based wheat head detection and counting. Neurocomputing. 2022;489:78–89.
35. Wang J, He L, Zhou X. Optimizing inception-V3 for brain tumor classification using hybrid precision training and cosine annealing learning rate. 2024 7th Int Conf Adv Algorithms Control Eng ICAACE [Internet]. 2024 [cited 2025 Feb 21]. pp. 528–32. Available from: https://ieeexplore.ieee.org/abstract/document/10548577
36. Liu Z, Jin S, Liu X, Yang Q, Li Q, Zang J, et al. Extraction of wheat spike phenotypes from field-collected lidar data and exploration of their relationships with wheat yield. IEEE Trans Geosci Remote Sens. 2023;61:1–13.

Figures

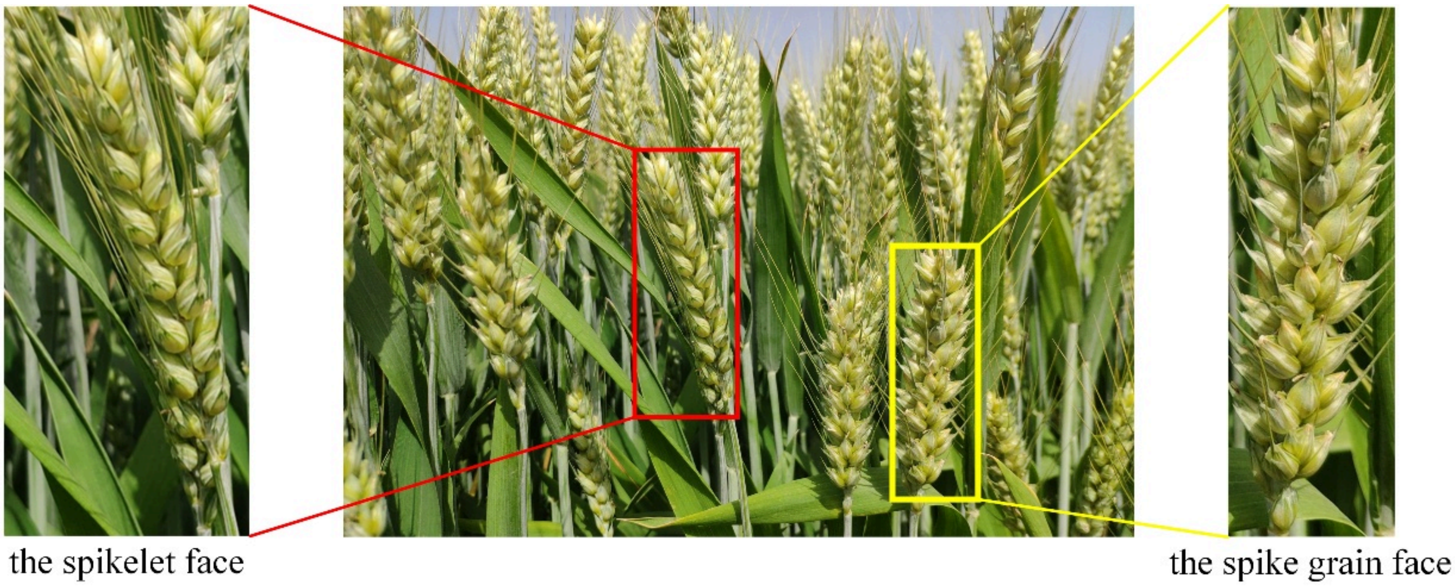

Figure 1

Example diagram of the spikelet face and the spike grain face

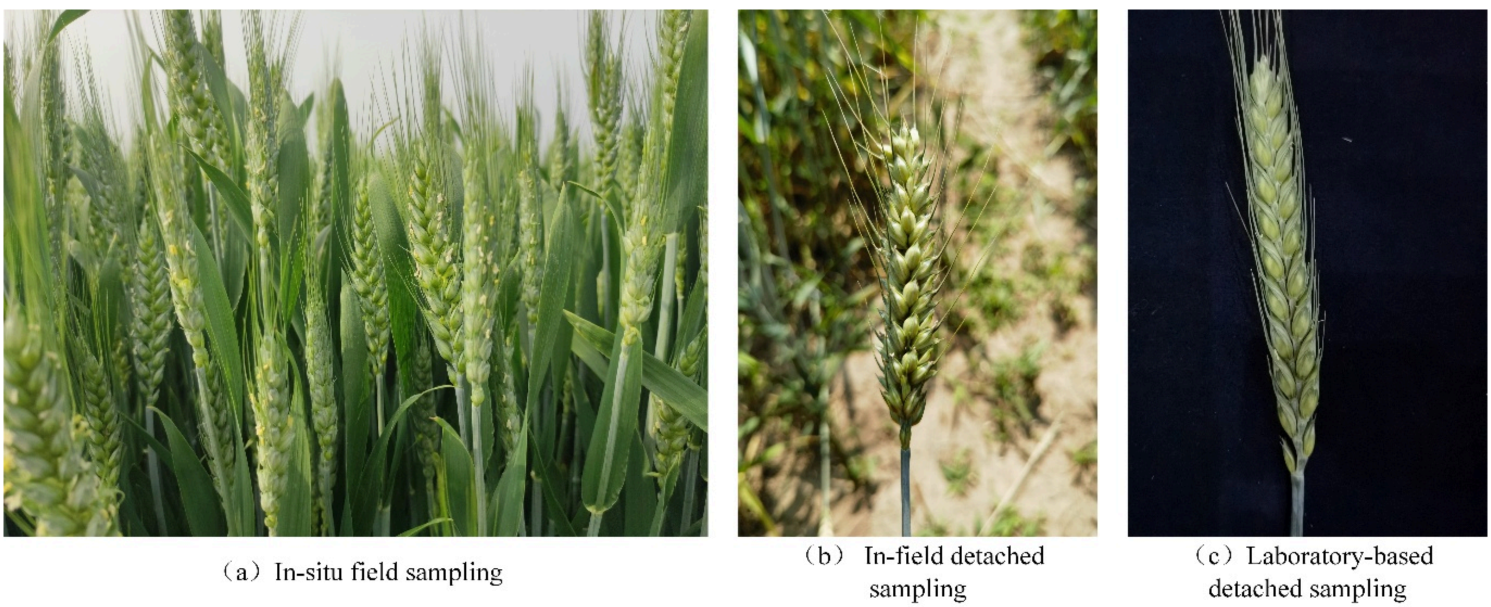

Figure 2

Examples of raw data

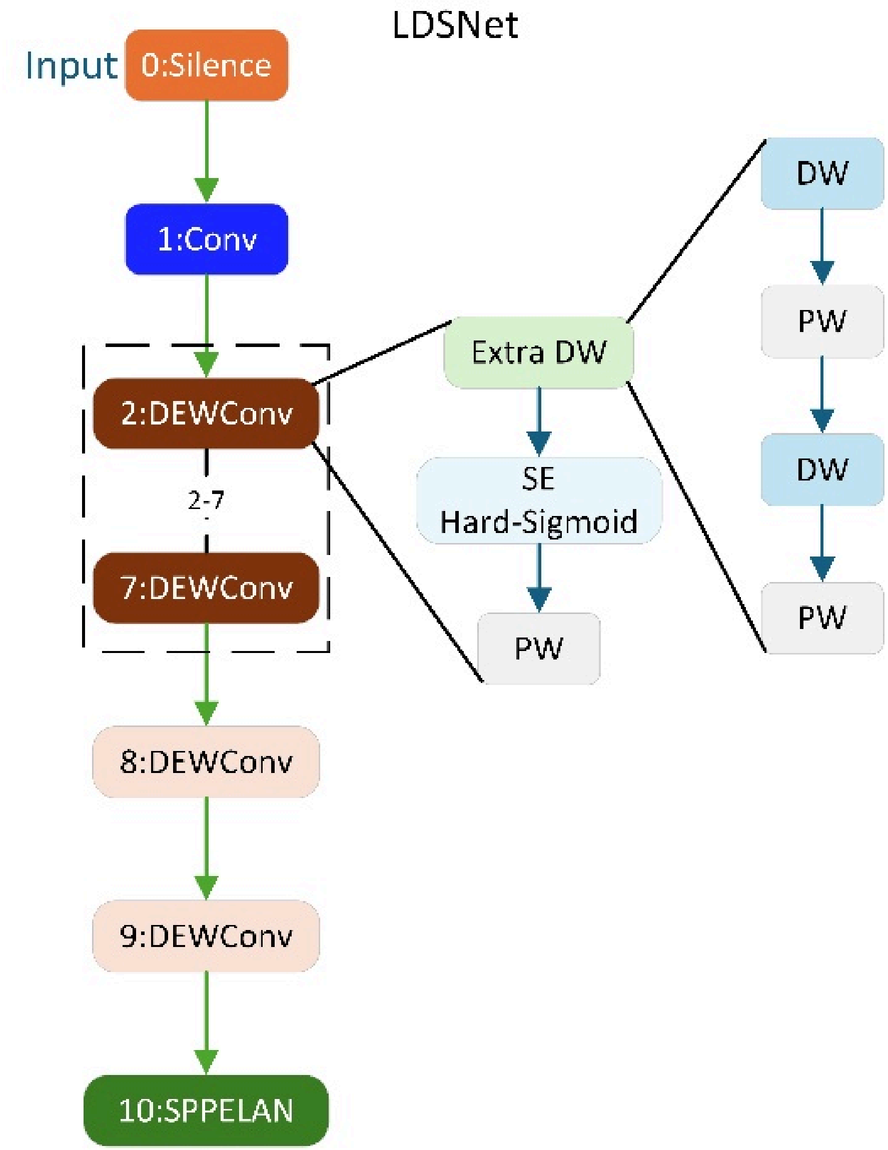

Figure 3

YOLOv9-LDS Network Architecture Diagram

Figure 4

LDSNet Network Architecture Diagram

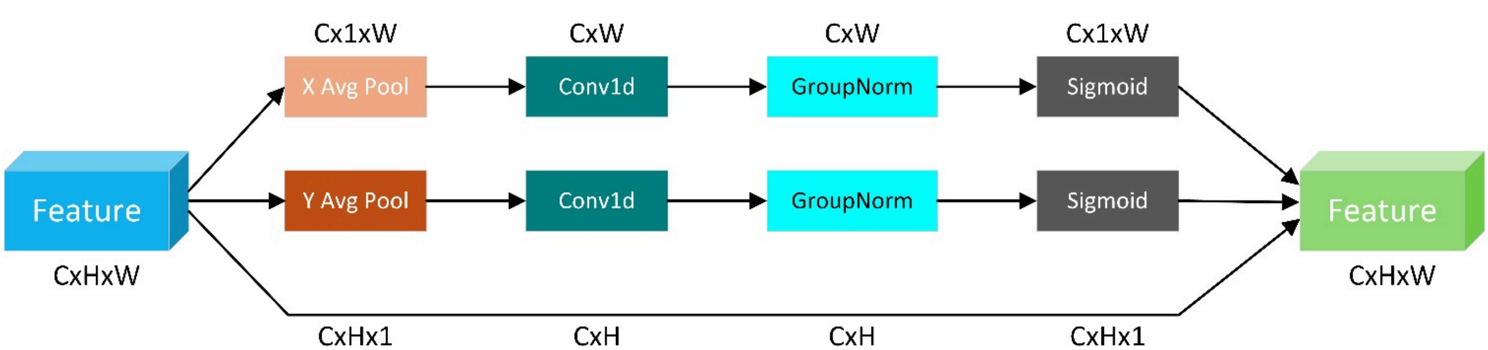

Figure 5

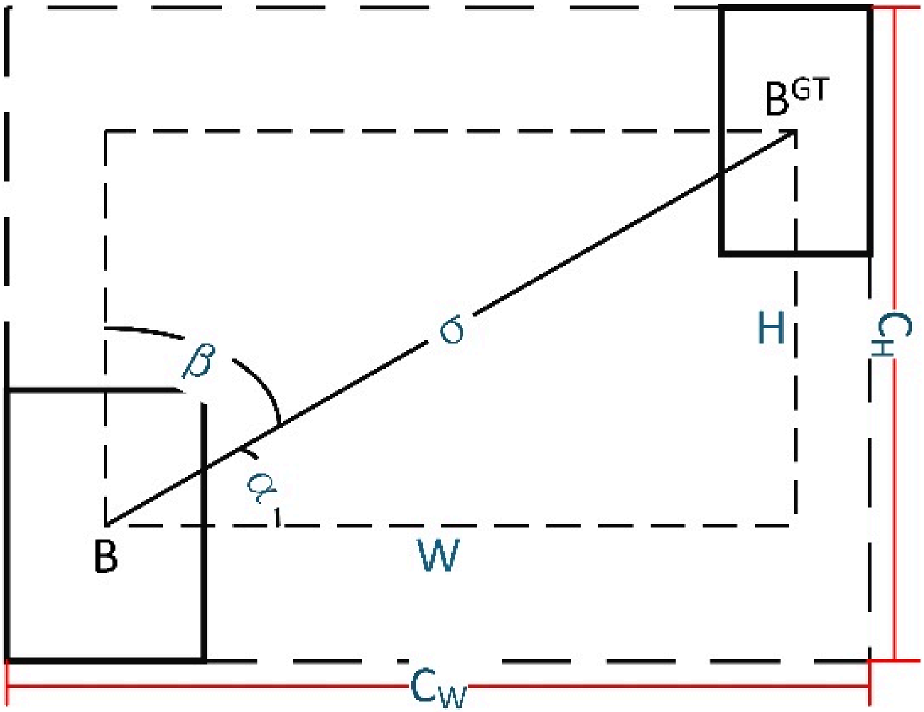

Figure 6

SIoU Loss Function Parameter Relationship Example

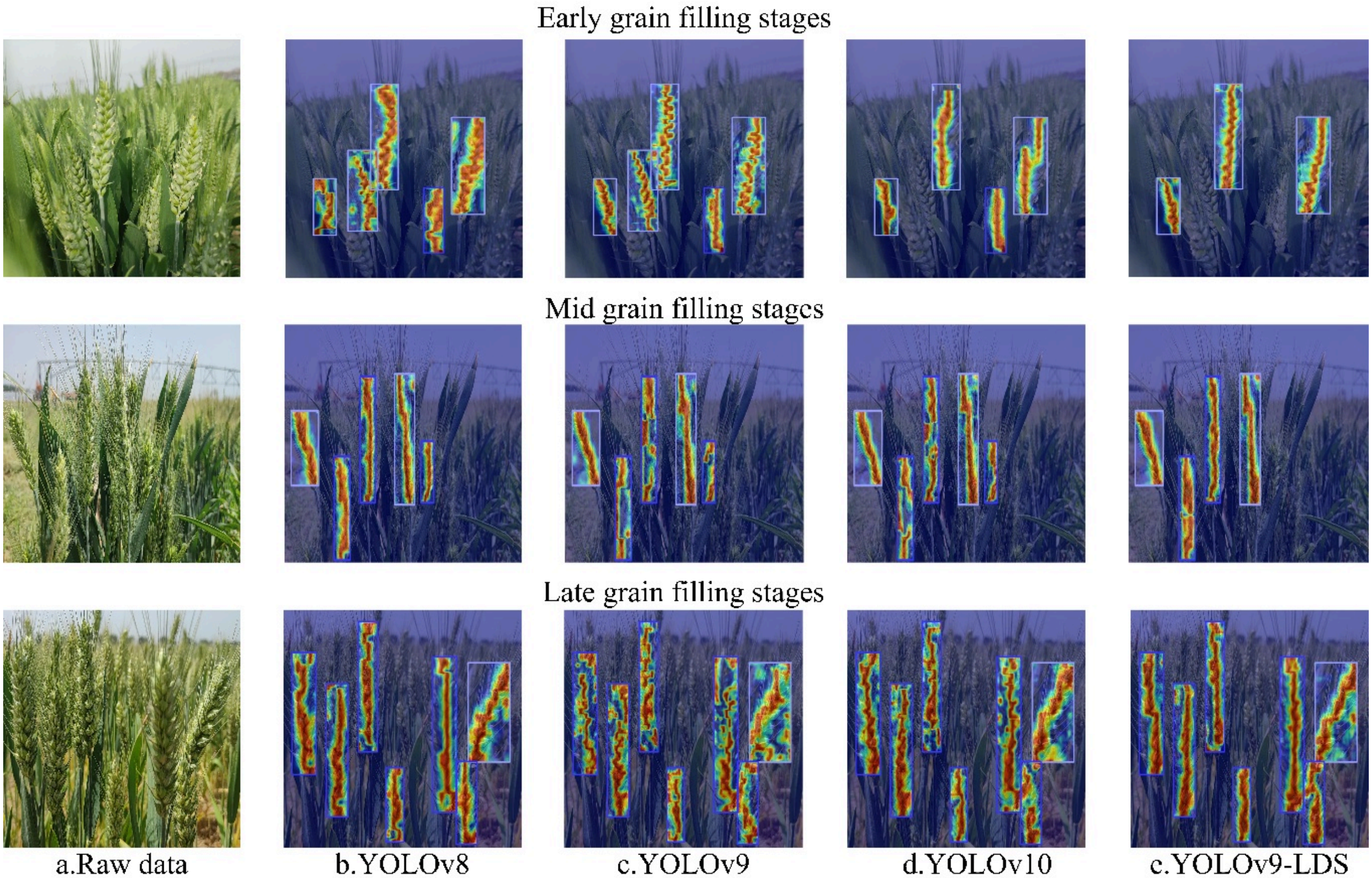

Figure 7

Comparison of Contour Extraction Effects at Different Periods

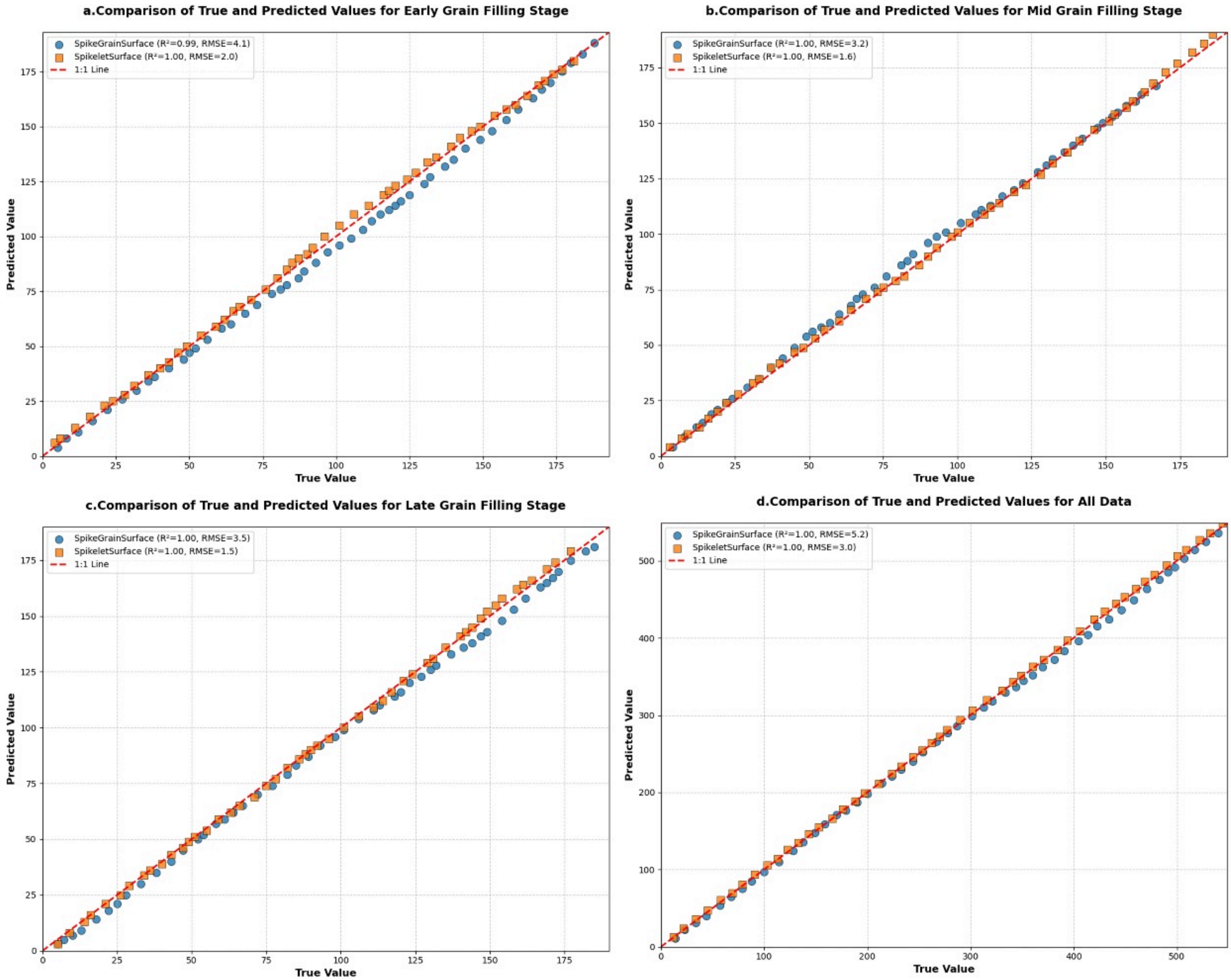

Figure 8

Analysis of Spike Counting Results Across Growth Stages

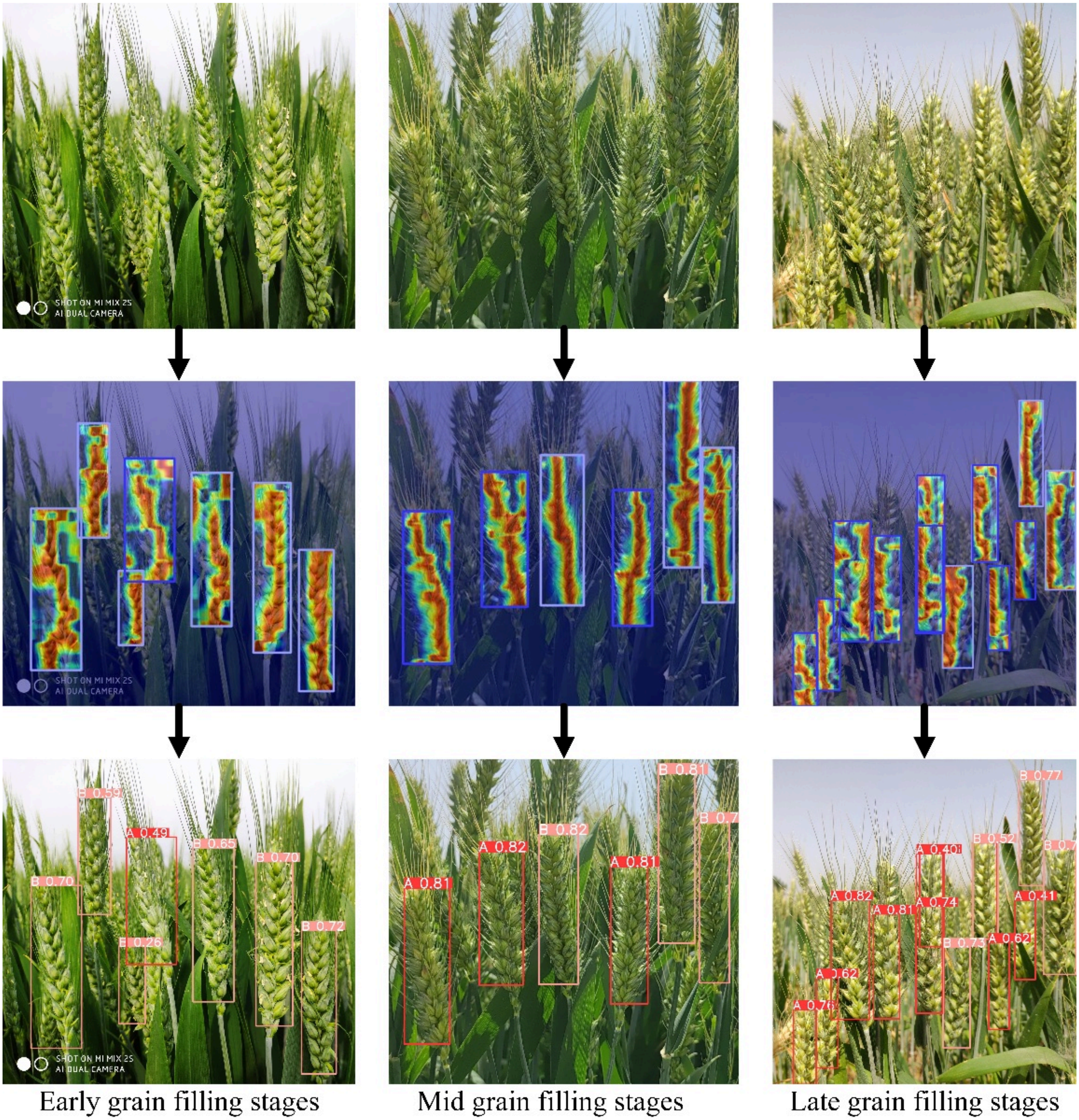

Figure 9

Effect diagrams of detection at different periods