Research Square

Preprints are preliminary reports that have not undergone peer review.
They should not be considered conclusive, used to inform clinical practice,
or referenced by the media as validated information.

# HMTS-YOLO: Detection of Weed Apex and Root/Stem-Base Treatment Regions for Precision Laser Weeding

**Yongbing Chen**
Wenzhou Vocational College of Science and Technology
**Meiling Yuan**
Taizhou University
**Liming Qin**
Taizhou University
**Qian Hu**
Zhejiang Jiangxuan Technology Co., Ltd
**Simon X. Yang**
University of Guelph
**Xiaofeng Ling**
182593414470163.com

Wenzhou Vocational College of Science and Technology

**Article**

**Keywords:** field weed detection, apex, root, morphology awareness, soil-texture suppression, multi-scale feature fusion, YOLO

**Posted Date:** November 3rd, 2026

**DOI:** https://doi.org/10.21203/rs.3.rs-11177082/v1

**License:** This work is licensed under a Creative Commons Attribution 4.0 International License.
Read Full License

**Additional Declarations:** No competing interests reported.

# HMTS-YOLO: Detection of Weed Apex and Root/Stem-Base Treatment Regions for Precision Laser Weeding

Yongbing Chen[a], Meiling Yuan[c], Liming Qin[c], Qian Hu[e], Simon X. Yang[d], Xiaofeng Ling[a,b*]

a Wenzhou Vocational College of Science and Technology, Wenzhou 325006, China

b Wenzhou Key Laboratory of Al Agents for Agriculture, Wenzhou 325006, China

c Research Institute of Motor and Intelligent Control Technology, Taizhou University, Taizhou 318000, China

d Advanced Robotics and Intelligent Systems Laboratory, College of Engineering, University of Guelph, Canada

e Zhejiang Jiangxuan Technology Co., Ltd., Hangzhou 310000, China

* corresponding author e-mail address: 18259341447@163.com

**Abstract:** Under complex field backgrounds, weed apex and root treatment regions are readily affected by soil texture, surface cracks, target-scale variation, and morphological differences, resulting in missed detections, false detections, and bounding-box regression errors, particularly for small-scale and elongated targets. To improve dual-class treatment-region detection for field weeds, a dataset comprising 4,582 images and 13,421 bounding boxes was curated from existing field images and treatment-region annotations, and HMTS-YOLO (Hierarchical Morphology-aware Texture-Suppression YOLO) was developed using YOLOv8s as the baseline. The proposed model introduces STS-C2f, LDM-C2f, and DMA-SPPF at the P2, P3, and P5 feature levels, respectively. These modules improve feature representation across target scales and morphologies through shallow-layer soil-texture suppression, mid-level dual-morphology detail enhancement, and deep multi-scale context modeling. On the validation set, HMTS-YOLO achieved a Precision of 72.736 %, Recall of 68.988 %, mAP50 of 73.129 %, and mAP50-95 of 35.177 %, representing improvements of 3.619, 2.665, 4.341, and 4.514 percentage points, respectively, over the baseline. The model contains 11.572 M parameters and requires 69.157 GFLOPs. Class-wise evaluation showed that the mAP50-95 values for apex and root reached 42.212 % and 28.142 %, respectively, corresponding to gains of 4.969 and 4.059 percentage points over the baseline. These results demonstrate that feature-level mechanisms for texture suppression, morphology modeling, and contextual fusion can effectively improve the overall detection performance of apex and root treatment regions in complex field scenes.

**Keywords:** field weed detection; apex; root; morphology awareness; soil-texture suppression; multi-scale feature fusion; YOLO

## 1 Introduction

Weeds compete with crops for water, nutrients, light, and growing space, causing substantial potential yield losses and increasing production costs[1,2]. With the spread of herbicide-resistant weeds and

increasing environmental and food-safety requirements, long-term reliance on broad-area, uniform herbicide application faces growing sustainability challenges[2,3]. Autonomous agricultural robots integrate machine vision, target recognition, path planning, and site-specific actuation[4–6] to shift weed control from whole-field treatment toward plant-level or localized intervention, providing a technical route for reducing chemical inputs and improving operational selectivity. In laser weeding in particular, treatment effectiveness depends on laser-beam placement, spot size, energy density, and weed growth stage. Consequently, the vision system must not only determine whether a weed is present but also identify candidate treatment regions relevant to downstream operation, such as the apical meristem or the root/stem-base region[7–9].

Machine-vision research has established a relatively mature technical framework encompassing image classification, object detection, semantic segmentation, and instance segmentation. Fully convolutional networks enable pixel-level prediction by combining shallow spatial details with deep semantic information[10]. Faster R-CNN achieves high-accuracy object detection through region proposals[11], whereas YOLO formulates localization and classification as a unified single-stage regression problem[12] to improve detection speed. Mask R-CNN further integrates bounding-box detection with instance-mask prediction[13], providing a methodological basis for target separation in complex scenes. With the development of large-scale datasets such as COCO and standardized evaluation protocols[14], research has increasingly focused on multi-scale feature fusion[15] and the trade-off between detection accuracy and computational efficiency. Representative approaches include EfficientDet, which combines bidirectional feature pyramids with compound scaling[16]; YOLOX, which adopts an anchor-free architecture and decoupled detection head[17]; and YOLOv7 for real-time object detection[18]. In addition, VarifocalNet and Generalized Focal Loss improve the consistency between detection confidence and localization quality[19,20] through localization-quality awareness and continuous bounding-box distribution modeling, respectively, while RT-DETR explores real-time end-to-end detection[21]. Although these methods have advanced general object detection, field applications remain challenging because of small target scales, background-texture similarity, illumination variation, and crop occlusion.

For agricultural weed recognition, previous studies have addressed plant-species classification, crop–weed segmentation, weed object detection, and UAV-based remote-sensing recognition. Dyrmann et al. used convolutional neural networks to classify multiple crop and weed species at early growth stages[22]. Milioto et al. developed a real-time crop–weed semantic-segmentation method[23] for sugar beet fields and demonstrated its feasibility on agricultural robots. Lottes et al. jointly predicted crop and weed classes, plant stem locations, and pixel-level regions[24], thereby providing visual information for plant-specific treatment. Bah et al. investigated inter-row weed detection using UAV imagery and unsupervised annotation[25], reducing the cost of manual labeling in agricultural images. Public resources such as DeepWeeds and Weed25 have further supported comparative studies[26,27] of weed-classification and detection models. Reviews indicate that most existing studies rely on RGB imagery and supervised deep-learning models[28–30], with outputs primarily consisting of weed classes, whole-plant bounding boxes, pixel-level masks, or

field weed density. These outputs can support weed recognition and spraying-zone generation; however, localization of root, stem, or apical growth regions directly relevant to precision laser weeding remains relatively underexplored[6,24], and the center of a whole-plant detection box does not necessarily correspond to the actual laser treatment point.

Weed detection during the early growth stages of soybean provides a representative field scenario. Previous studies have used UAV imagery and convolutional networks to distinguish soybean, soil, grass weeds, and broadleaf weeds[31–34], while other work has applied improved YOLOv7 models to soybean-field weed detection and further evaluated the relationship between detection results and reduced herbicide application. In row crops such as sugar beet, TIA-YOLOv5 improves the detection of small weed targets[35] through data augmentation, Transformer encoding, and multi-scale feature fusion. More recent research on laser weeding has begun to directly target weed stem or root locations[6,9,24], indicating a transition from whole-plant recognition toward localization of executable treatment regions. Keypoint-based detection of weed growth points in maize fields has also shown[36] that occlusion, dense growth, target-scale variation, and illumination changes can substantially affect localization accuracy. Nevertheless, most soybean-field studies still focus on whole-plant weed detection, category recognition, or spraying-area estimation. Unified detection of local regions with distinct morphological characteristics and operational significance—such as weed apices and root/stem-base regions—under complex soil backgrounds remains limited.

To address these issues, this study targets precision laser weeding under complex field backgrounds and curates a dataset for dual-class detection of apex and root treatment regions using existing field images and treatment-region annotations. Here, apex denotes the weed apical meristem or upper growth region, whereas root denotes the visible root or stem-base treatment region in the image and should not be interpreted as an anatomical annotation of the complete underground root system. YOLOv8s is adopted as the baseline. The original C2f at P2/4 is replaced with the proposed STS-C2f, LDM-C2f is added after the P3/8 feature output by the neck, and the original SPPF at P5/32 is replaced with DMA-SPPF. These modifications are designed to address shallow-layer soil-texture interference, mid-level local morphological differences, and deep multi-scale contextual information, respectively. Unified model comparisons, a complete eight-configuration combinatorial ablation study, class-wise analysis, scale- and morphology-grouped statistics, and representative-scene visualizations are conducted to systematically evaluate the effects of feature enhancement at different hierarchical levels on apex and root treatment-region detection, providing a methodological basis for visual perception in precision laser weeding under complex soybean-field conditions.

# 2 Materials and Methods

## 2.1 Data Source and Construction of the Dual-Class Treatment-Region Detection Task

### 2.1.1 Weed Image Source and Dataset Split

The original images used in this study were obtained from the field image dataset described in Yang Xiaozan's master's thesis[9]. The dataset was collected in a soybean field at Qiaosi Farm, Linping District, Hangzhou, China, and originally contained 6,307 images covering soybean seedlings and common field weeds, including *Digitaria sanguinalis*, *Echinochloa crus-galli*, *Chenopodium album*, *Erigeron canadensis*, *Persicaria lapathifolia*, and the source-dataset category “Goosegrass”. Based on the composition of the source dataset, only weed images were retained, and soybean-seedling images were excluded, resulting in 4,582 images. Using the existing treatment-region labels in the source data, the data were filtered, reformatted, and consistently summarized for the dual-class object-detection task in this study. The final dataset contained 13,421 bounding boxes, including 6,094 apex targets and 7,327 root targets. The images preserve natural field characteristics such as soil texture, surface cracks, granular interference, illumination variation, and target-scale differences, thereby reflecting realistic detection conditions in complex field environments.

The dataset was split at the image level into training, validation, and test subsets containing 3,204, 913, and 465 images, respectively, corresponding to 9,432, 2,622, and 1,367 target boxes. The training set was used for model-parameter learning, and the validation set was used for training monitoring, model selection, and standardized performance evaluation across models. The test set was retained as an independent holdout set and was not used for training or model selection. Model comparisons and ablation analyses were conducted on the validation set, whereas only the selected final HMTS-YOLO model was subsequently evaluated on the test set. All comparison and ablation models used the same data split, class definitions, and annotation format to ensure comparability.

The resulting dataset covers a broad range of target scales and pronounced morphological variation. Treatment regions generally occupy a limited proportion of the image, with small- and medium-scale targets accounting for most instances. Apex targets generally exhibit compact local structures, whereas root targets tend to be smaller and more elongated, making them more susceptible to confusion with soil cracks, granular textures, and neighboring plants. Complex backgrounds and uneven distributions across target scales and morphologies further increase detection difficulty. These characteristics enable the evaluation of model performance on small-scale and elongated targets under background-texture interference, while providing a basis for analyzing the shallow texture suppression of STS-C2f, local morphological-detail enhancement of LDM-C2f, and multi-scale context modeling of DMA-SPPF. Dataset statistics are listed in Table 1, whereas Fig. 1 presents representative raw training images of the six weed species; the corresponding n values indicate the numbers of training images for each species.

Table 1. Dataset split and sample statistics

| Split | Images | Apex annotations | Root annotations | Total boxes |
|---|---|---|---|---|
| Training | 3204 | 4265 | 5167 | 9432 |
| Validation | 913 | 1206 | 1416 | 2622 |
| Test | 465 | 623 | 744 | 1367 |
| Total | 4582 | 6094 | 7327 | 13421 |

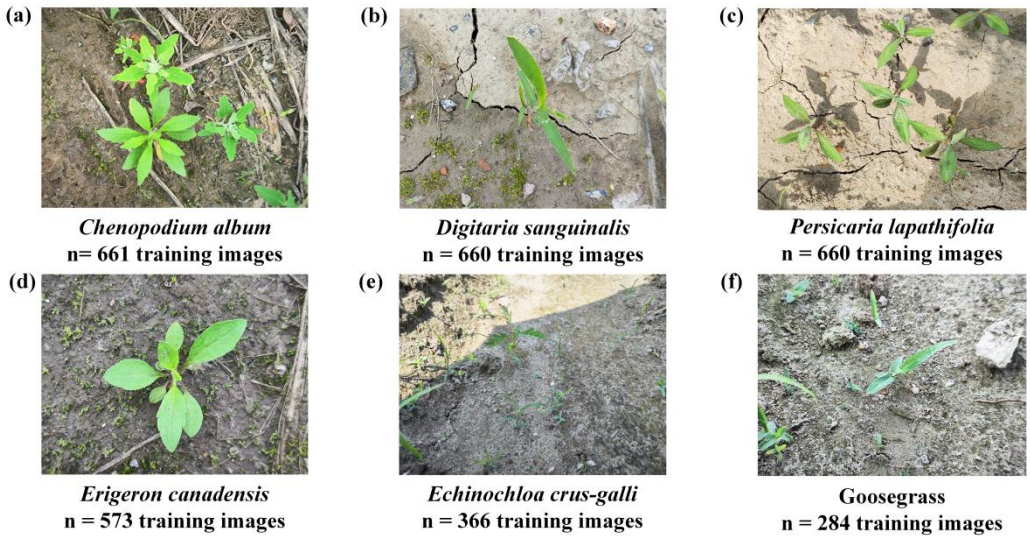



Fig. 1. Representative raw training images of different weed species

**2.1.2 Definition of Dual-Class Treatment Regions and Analysis of Target Morphology**

This study follows the weed treatment-region label definitions established during construction of the source dataset[9] in the master's thesis of Yang Xiaozan, without introducing additional whole-plant weed annotations. The source data define two target classes related to laser weeding: weed_apical_meristem (apex), representing the weed apical meristem or upper growth region, and weed_root (root), representing the visible root or stem-base treatment region. It should be emphasized that root in this study refers only to the root or stem-base region visible in the image and is not equivalent to an anatomical annotation of the complete underground root system.

According to the original labeling protocol, apex and root correspond to local weed treatment regions rather than the entire plant extent. The annotations cover only visible local regions and exclude irrelevant background areas, such as soil, cracks, granules, and other plants, as far as possible. No speculative extension is made to invisible underground roots or occluded regions. For the current object-detection task, the existing treatment-region labels were uniformly converted to YOLO-format class and bounding-box annotations, and multiple treatment regions within the same image were retained as separate instances. The

small, medium, and large target groups and the compact and elongated morphology groups used in subsequent experiments were computed from the area and aspect ratio of these bounding boxes.

This target definition allows model outputs to correspond to potential laser treatment regions rather than merely indicating the approximate location of an entire weed, thereby providing a visual basis for subsequent candidate-operation decisions based on detected boxes. The dual-class annotations and representative morphologies are illustrated in Fig. 2.

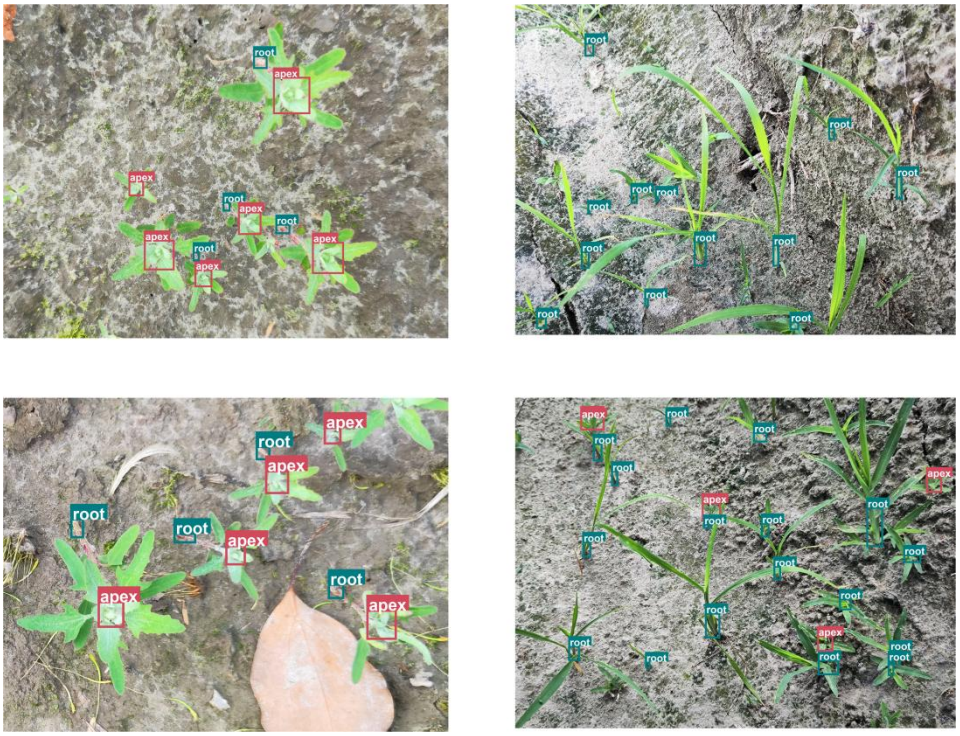



Fig. 2. Illustration of dual-class target annotations and morphological characteristics

## 2.2 HMTS-YOLO Architecture and Core Module Design

### 2.2.1 Overall Architecture of HMTS-YOLO

HMTS-YOLO, namely the proposed Hierarchical Morphology-aware Texture-Suppression YOLO model, uses YOLOv8s as the baseline[37] while retaining the overall Backbone – Neck – Detect framework and the original detection loss. Input images are resized with aspect-ratio preservation and padding to 960 × 960 pixels. The backbone then generates four feature levels, P2/4, P3/8, P4/16, and P5/32, with spatial resolutions of 240 × 240, 120 × 120, 60 × 60, and 30 × 30 and channel dimensions of 64, 128, 256, and 512, respectively. Targeted enhancement modules are configured at different levels: the native C2f at P2/4 is replaced by STS-C2f, DMA-SPPF is used at P5/32, and LDM-C2f is inserted after the P3/8 detection feature output by the neck.

$$P_i \in \mathbb{R}^{C_i \times \frac{H}{2^i} \times \frac{W}{2^i}}, \quad i \in \{2,3,4,5\} \tag{1}$$

Here, $P_i$ denotes the feature map at stride $2^i$; $C_i$ is its channel dimension, and H and W denote the

input image height and width, respectively. The neck first progressively upsamples P5 and concatenates it with the P4 and P3 features from the backbone, followed by C2f blocks for top-down semantic fusion. Downsampling and lateral connections are then used to restore bottom-up multi-scale fusion.

$$\hat{Y} = \mathrm{Detect}\left(P_3^*, P_4^*, P_5^*\right) \tag{2}$$

The Detect module ultimately receives the enhanced P3, P4, and P5 features and outputs the predicted detection result shown in Eq. (2), including bounding boxes and class confidences for the apex and root classes. Here, the hat symbol denotes a predicted output, whereas the superscript * denotes feature maps processed by the neck and the corresponding enhancement module. Thus, HMTS-YOLO modifies feature representation and cross-level information fusion without changing the target classes, detection-head type, or task definition. The overall architecture and module insertion positions are shown in Fig. 3.

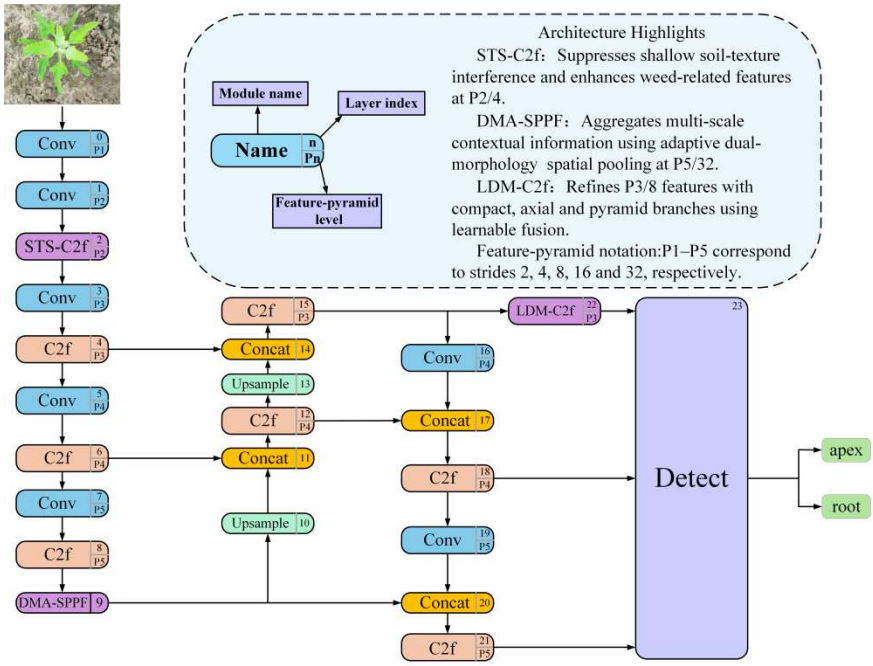



Fig. 3. Overall architecture of HMTS-YOLO and locations of the proposed modules

### 2.2.2 DMA-SPPF: Dual-Morphology Adaptive Spatial Pyramid Pooling-Fast

At the P5/32 stage of the backbone, DMA-SPPF replaces the original SPPF implementation in YOLOv8[37] and retains the spatial-pyramid pooling principle of SPP[38], with the aim of enlarging the effective receptive field at low spatial resolution and a high semantic level. Let the input feature be X. A 1 × 1 convolution is first applied to obtain the compressed feature $X_r$. At the P5 stage of HMTS-YOLO, both the input and output channel dimensions are 512, and the hidden channel dimension is 256.

The compressed feature $X_r$ is fed into three complementary branches. The context branch applies 5 × 5 max pooling with stride 1 and padding three consecutive times to capture local context over progressively larger ranges; the original compressed feature is also retained for concatenation.

$$P_1 = \mathrm{MaxPool}_{5\times5}(X_r), \quad P_2 = \mathrm{MaxPool}_{5\times5}(P_1), \quad P_3 = \mathrm{MaxPool}_{5\times5}(P_2) \tag{3}$$

The three pooled features $P_1$, $P_2$ and $P_3$ are concatenated along the channel dimension and projected through a 1 × 1 convolution to obtain the contextual feature. The compact-morphology branch uses 3 × 3 and 5 × 5 depthwise convolutions to extract local-neighborhood responses, followed by a 1 × 1 convolution for channel compression. The directional branch uses 1 × 7 and 7 × 1 depthwise convolutions to model horizontal and vertical structures, respectively.

The gating network first applies global average pooling to the concatenated features from the three branches and then uses a 1 × 1 convolution, SiLU, and a second 1 × 1 convolution to generate channel-wise gating logits for each branch logit $z_{j,u}$.

Softmax normalization is subsequently applied along the branch dimension:

$$\alpha_{j,u} = \frac{\exp(z_{j,u})}{\sum_{q\in\{c,k,d\}} \exp(z_{q,u})}, \quad j \in \{c,k,d\} \tag{4}$$

$$F = \sum_{j\in\{c,k,d\}} \alpha_j \odot B_j \tag{5}$$

Let the context, compact, and directional features be denoted by $B_c$, $B_k$ and $B_d$, respectively. These features are fused using channel-adaptive weights to obtain F. F and $X_r$ are concatenated along the channel dimension, mapped back to the output channel dimension using a 1 × 1 convolution, and combined through a residual connection to produce the module output.

$$Y = X + \gamma \cdot \mathrm{Conv}_{1\times1}\left([X_r \| F]\right), \quad \gamma_0 = 0.5 \tag{6}$$

The residual scale $\gamma$ is learnable and initialized to 0.5. The gating bottleneck width is calculated as $max\left(8, h/16\right)$, where h is the hidden channel dimension. At P5, the gating network therefore uses a 768→16→768 channel transformation. The overall structure of DMA-SPPF is shown in Fig. 4.

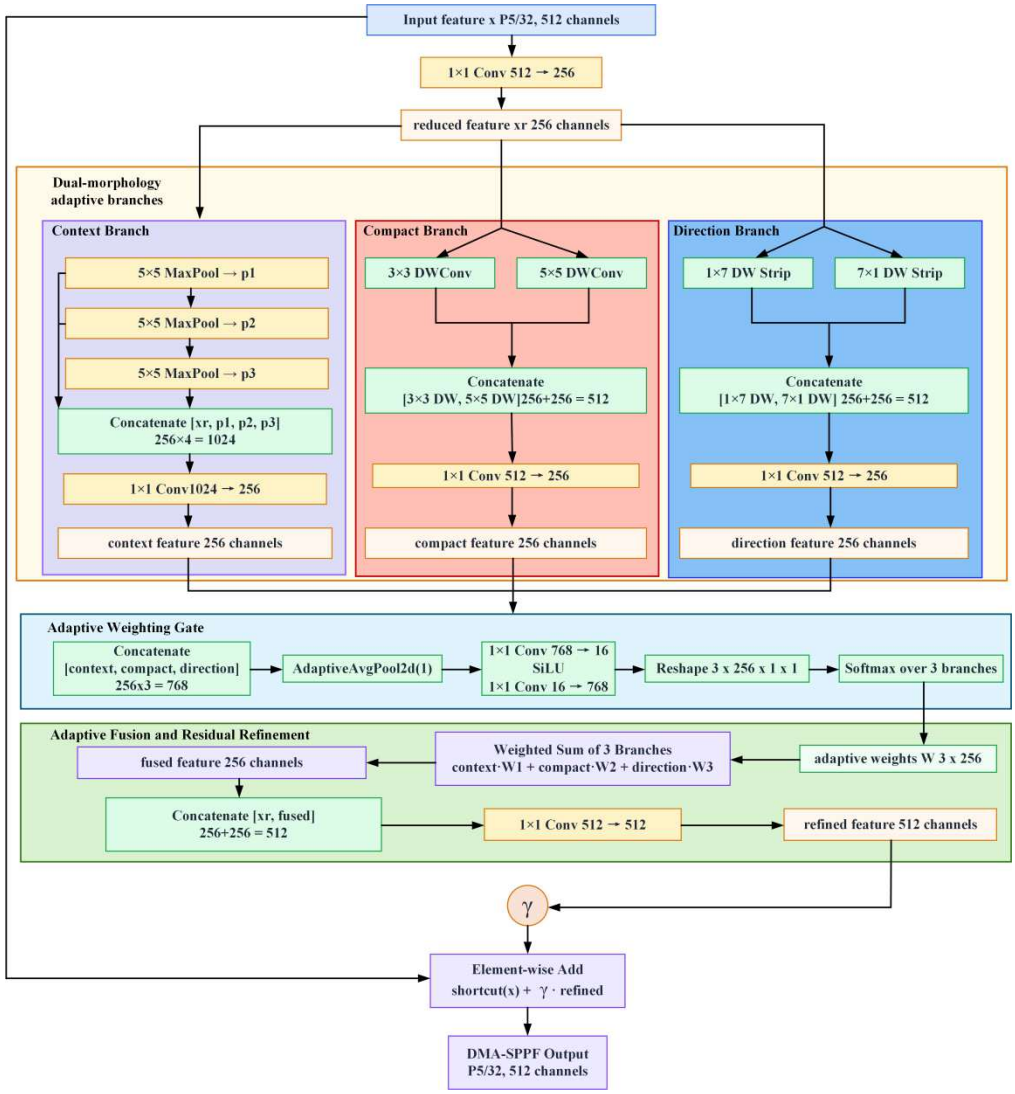



Fig. 4. Architecture of the DMA-SPPF module

### 2.2.3 LDM-C2f: High-Resolution Lightweight Dual-Morphology Detail Enhancement

LDM-C2f is inserted after the high-resolution P3/8 detection feature output by the neck to compensate for local information loss in small-scale and elongated treatment regions. The module first performs a native C2f transformation to obtain the main-path feature B; both this main path and its internal Bottleneck structure retain the original YOLOv8 formulation[37]. A low-dimensional morphology branch is then constructed from B. The compact branch uses depthwise convolution to extract local responses[39], the axial branch uses grouped strip convolutions, and the local-pyramid branch uses max pooling at different scales.

$$B = \mathrm{C2f}(X), \quad R = \mathrm{Conv}_{1\times1}(B) \tag{7}$$

where $C_{out}$ denotes the output-channel dimension of the native C2f path. The low-dimensional branch ratio is $r = 0.25$, and the detail-channel number is calculated as:

$$C_r = \max\left(16, 8 \cdot \text{round}\left(\frac{rC_{out}}{8}\right)\right) \tag{8}$$

At P3, $C_{out} = 128$ resulting in $C_r = 32$ detail channels. Using R as the shared input, the detail branch constructs compact, axial, and local-pyramid morphological features. The compact branch uses a $3\times 3$ depthwise convolution. The axial branch uses grouped $1\times 5$ and $5\times 1$ strip convolutions, whereas the local-pyramid branch uses $3\times 3$ and $5\times 5$ max-pooling operations.

The three detail branches avoid complex spatially varying attention. Instead, three learnable global scalar logits $\beta_c, \beta_a, \beta_p$ are used for branch selection. Their normalized weights are calculated by:

$$w_j = \frac{\exp(\beta_j)}{\sum_{q\in\{c,a,p\}} \exp(\beta_q)}, \quad j \in \{c, a, p\} \tag{9}$$

where $\beta_c$ , $\beta_a$ , and $\beta_p$ are the learnable logits for the compact, axial, and local-pyramid branches, respectively. All three logits are initialized to zero, resulting in equal initial weights of 1/3.

$$F = w_c F_c + w_a F_a + w_p F_p \tag{10}$$

The fused feature is mapped back to 128 output channels through a 1 × 1 convolution without activation and injected into the native C2f main path using a learnable residual scale $\lambda$ . The scale $\lambda$ is initialized to 0.05 to limit the initial perturbation of pretrained features by the newly introduced branch. The overall structure of LDM-C2f is shown in Fig. 5.

$$Q = \text{Conv}_{1\times 1}(F), \quad Y = B + \lambda Q, \quad \lambda_0 = 0.05 \tag{11}$$

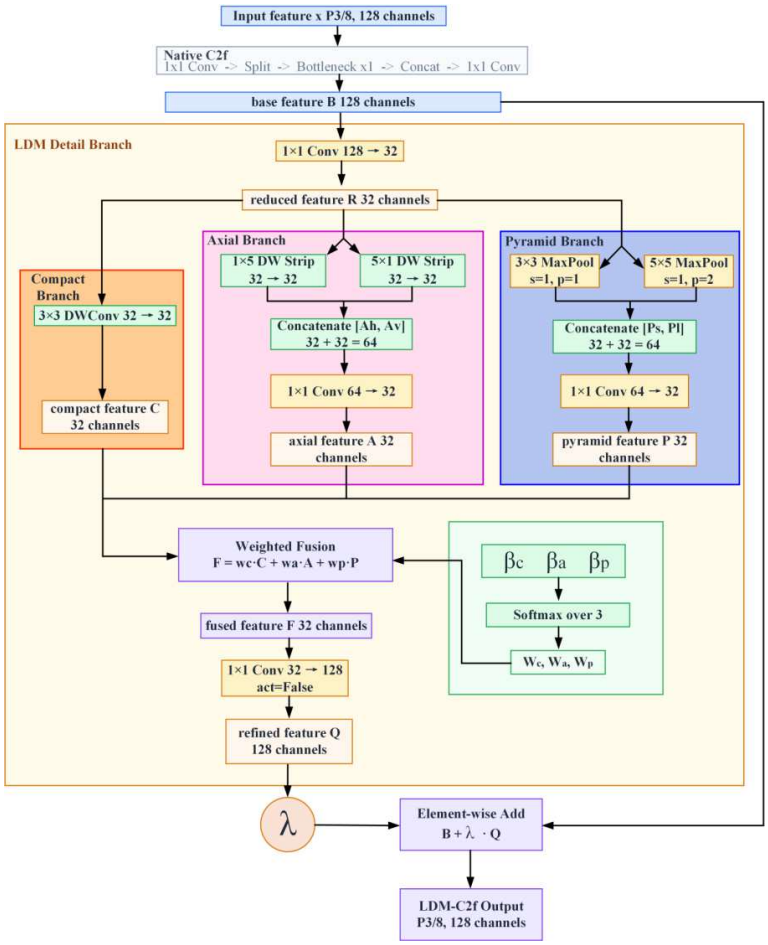



Fig. 5. Architecture of the LDM-C2f module

### 2.2.4 STS-C2f: Shallow Soil-Texture Suppression

At the P2/4 stage of the backbone, the native C2f is replaced with the proposed STS-C2f module to suppress interference from soil texture, debris, and illumination variation contained in high-resolution shallow features. The module retains the native C2f main path to obtain a base feature B. Let $R = \mathrm{Conv}_{1\times1}(B)$ denote the reduced feature. At P2, the native output channel dimension is 64, corresponding to 16 channels in the low-dimensional branch.

$$L = \mathrm{AvgPool}_{3\times3}(R), \quad H = R - L \tag{12}$$

The average-pooled feature L is used as an estimate of local low-frequency structure, and the difference H between R and L is treated as a relative high-frequency response. The low-frequency feature L is then processed by a 3 × 3 depthwise convolution to obtain a coherent structural feature C, thereby further modeling spatially continuous structural responses.

$$G = \mathrm{Sigmoid}\left(\mathrm{Conv}_{1\times1}\left(\mathrm{SiLU}\left(\mathrm{Conv}_{1\times1}\left(\mathrm{Concat}\left(|H|, C\right)\right)\right)\right)\right) \tag{13}$$

The gating network takes the channel-wise concatenation of |H| and C as input and applies two 1 × 1

convolutions with SiLU and Sigmoid activations to generate the gating map G. In this formulation, G acts as a retention gate for the high-frequency residual, whereas $1-\mathrm{G}$ determines the magnitude of the high-frequency component subtracted from the coherent low-frequency feature. The gated correction feature is defined as:

$$E = C \cdot (1-G) \square\ H \tag{14}$$

When G is small, a larger portion of the high-frequency residual is subtracted, resulting in stronger suppression of high-frequency texture. When G is large, the subtraction is weakened and more high-frequency information is retained. Because E is used as a residual correction term rather than a direct reconstruction of R, it is subsequently projected and added to the native C2f feature through the learnable residual scale $\lambda$ :

$$Q = \mathrm{Conv}_{1\times1}(E), \quad Y = B + \lambda Q, \quad \lambda_0 = 0.05 \tag{15}$$

The correction feature E is restored to 64 channels through a 1 $\times$ 1 convolution without activation and is then reinjected into the native C2f main path. P2 is used only for feature extraction and does not introduce an independent detection head; detection is still performed using P3, P4, and P5. The overall structure of STS-C2f is shown in Fig. 6.

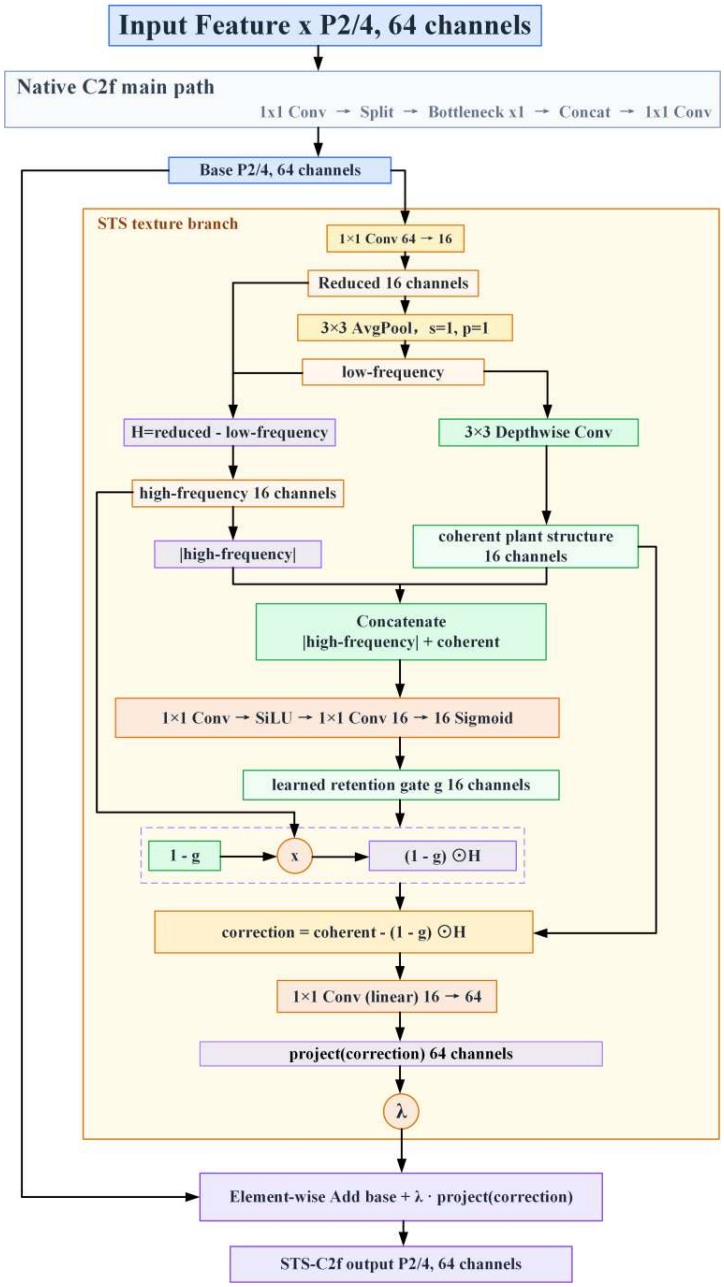



Fig. 6. Architecture of the STS-C2f module

## 2.3 Experimental Settings and Comparison Protocols

### 2.3.1 Experimental Environment and Training Configuration

All comparison and ablation models used the same dataset split, input size, and training protocol. Models were trained for a maximum of 200 epochs with an early-stopping patience of 40, a batch size of 4, and an input resolution of 960 × 960. AdamW was used as the optimizer with an initial learning rate of 0.001, a final learning-rate factor of 0.01, and cosine learning-rate decay. The weight-decay coefficient was 0.0005. The warm-up stage lasted 3 epochs, with a warm-up momentum of 0.8 and an initial warm-up bias learning rate of 0.1. Four data-loader workers were used, the random seed was set to 42, and deterministic

computation and automatic mixed precision were enabled. Each comparison model used its corresponding pretrained weights, whereas HMTS-YOLO and all of its ablation variants were initialized from YOLOv8s pretrained weights. The main model comparisons were conducted using Ultralytics 8.3.221. The detailed experimental environment and training parameters are listed in Table 2. The principal detection augmentations included Mosaic with a probability of 0.5, while Mosaic was disabled during the final 15 epochs. MixUp, CutMix, and Copy-Paste were disabled.

Table 2. Experimental environment and model-training parameters

| Item | Setting | Item | Setting |
|---|---|---|---|
| GPU | NVIDIA GeForce RTX 3090 24 GiB | Python | 3.9.25 |
| PyTorch | 2.5.1 + cu121 | CUDA | 12.1 |
| Ultralytics | 8.3.221 | Input size | 960× 960 |
| Batch size | 4 | Maximum epochs | 200 |
| Early stopping | Patience = 40 | Initial learning rate | 0.001 |
| Final LR factor | 0.01 | Weight decay | 0.0005 |
| Warm-up epochs | 3 | Warm-up momentum | 0.8 |
| Warm-up bias LR | 0.1 | Data-loader workers | 4 |
| Random seed | 42 | Validation confidence threshold | 0.001 |
| NMS IoU threshold | 0.70 | Maximum detections | 300 |
| Complexity-analysis input | 960 × 960 | Mosaic probability | 0.5 |

**2.3.2 Comparison Models and Ablation Design**

The original YOLOv8s[37] was used as the baseline. YOLOv5su[40], YOLOv9s[41], YOLOv10s[42], YOLO11s[43], and HMTS-YOLO were evaluated under the same dataset split and training protocol. To ensure consistent comparison, all models used identical class definitions, input sizes, and training settings and were evaluated on the same validation set. Model performance was determined using the best.pt checkpoint corresponding to the highest validation-set mAP50-95 during training. The three core modules were evaluated using a full eight-configuration combinatorial ablation comprising the baseline, three single-module variants, three dual-module variants, and the complete HMTS-YOLO. Precision, Recall, mAP50, and mAP50-95 were recorded for all models. Table 4 reports the eight ablation configurations and their detection metrics, while Tables 5–7 further analyze four progressive configurations in terms of class, scale, and target morphology. Apart from architectural differences, all models were initialized with their corresponding pretrained weights; HMTS-YOLO and its ablation variants used YOLOv8s pretrained weights.

**2.3.3 Real-World System Demonstration and Coordinate-Output Protocol**

To demonstrate the practical workflow from HMTS-YOLO detection to galvanometer control, a

hardware demonstration platform was established. The platform integrated a dual-camera acquisition module, a computer, a galvanometer control board, a laser module, a cooling unit, and a specimen-positioning platform. It used a rated 80-W CO2 laser with a central wavelength of 10.6 μm, a threefold beam-expansion and collimation system, a coaxial 650-nm red alignment beam, an RC2808 galvanometer scanner (Sino-Galvo), and an f-theta lens with an effective scanning field of 300 mm × 300 mm. The complete platform is shown in Fig. 7.

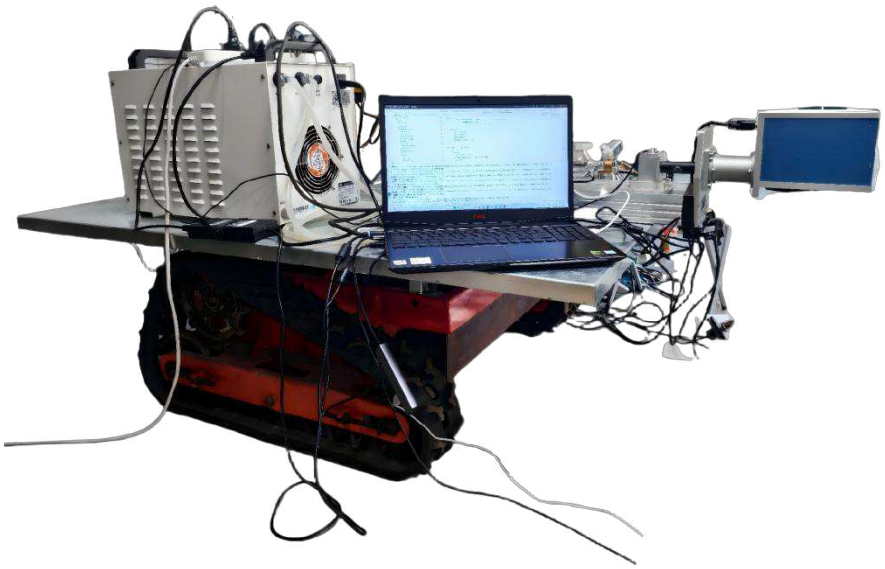


Fig. 7. Real-world demonstration platform of the HMTS-YOLO-based precision laser system

For each camera image, the model detected apex and root treatment regions, and the bounding-box centers were used as candidate treatment points. Corresponding detections from the two cameras were matched to calculate the target position in three-dimensional space and were then converted into $G_x$ and $G_y$ galvanometer coordinates using the calibrated mapping model. During the demonstration runs, MOVE commands were used with laser emission disabled, whereas automatic continuous firing was not enabled.

## 2.4 Evaluation Metrics and Grouped Analysis

### 2.4.1 Detection Performance Metrics

Bounding-box regression quality was measured using Intersection over Union (IoU). Let Bp denote a predicted box and Bg the corresponding ground-truth box:

$$IoU(B_p, B_g) = \frac{|B_p \cap B_g|}{|B_p \cup B_g|} \tag{16}$$

At a given IoU threshold $\tau$, predictions were ranked in descending order of confidence, from which true positives (TP), false positives (FP), and false negatives (FN) were determined. Precision and Recall were defined as follows:

$$P = \frac{TP}{TP + FP}, \quad R = \frac{TP}{TP + FN} \tag{17}$$

Average Precision (AP) was calculated from the precision–recall curve. For grouped diagnostic analysis, a 101-point interpolation scheme was used[14]:

$$AP_{\tau,c} = \frac{1}{101} \sum_{k=0}^{100} \max_{\rho \geq k/100} p_{\tau,c}(\rho) \tag{18}$$

where $\tau$ denotes the IoU threshold, c is the class index, and $p(\rho)$ is the interpolated precision at recall $\rho$ for the specified IoU threshold.

Main-model evaluation was performed using the Ultralytics validation interface on the validation set with a 960 × 960 input size, a confidence threshold of 0.001, an NMS IoU threshold of 0.70, and a maximum of 300 detections per image. Reported validation performance was obtained from the best.pt checkpoint achieving the highest validation-set mAP50-95.

mAP50 denotes the mean AP across classes at IoU = 0.50, whereas mAP50-95 denotes the mean mAP over 10 IoU thresholds from 0.50 to 0.95 in increments of 0.05.

**2.4.2 Class-, Scale-, and Morphology-Specific Statistical Analysis**

To analyze detection performance across different target scales and morphologies, normalized YOLO annotation boxes were first restored to the original image coordinates and then converted to pixel-area values according to the 960-pixel input scale used by the implementation:

$$w = b_w W, \quad h = b_h H, \quad s = \min\left(\frac{960}{W}, \frac{960}{H}\right), \quad A = (sw)(sh) = s^2 wh \tag{19}$$

For diagnostic purposes, COCO-style area thresholds were applied after mapping the bounding boxes to the 960-pixel model-input scale.

$$small: A < 32^2; \quad medium: 32^2 \leq A < 96^2; \quad large: A \geq 96^2 \tag{20}$$

where W and H are the original image width and height; bw and bh are the normalized bounding-box width and height; w and h are the corresponding box dimensions restored to the original image; s is the scaling factor; and A is the target-box area mapped to the 960 × 960 input scale. Target morphology was grouped according to bounding-box elongation:

$$e = \max\left(\frac{w}{h}, \frac{h}{w}\right) \tag{21}$$

$$compact: e < e_0; \quad elongated: e \geq e_0 \tag{22}$$

where $e_0 = 2$ is the predefined elongation threshold used for morphology-based diagnostic grouping.

For a given scale or morphology group, ground-truth boxes of the same class were first divided into in-group GT and out-of-group GT. When calculating AP for that group, all predictions of the corresponding class in the validation images were retained and ranked by confidence. Predicted boxes were matched one-to-one to unmatched in-group GT boxes of the same class, with each in-group GT allowed to form at most one valid match. A prediction successfully matched to an in-group GT at the current IoU threshold was counted as a TP. If it failed to match an in-group GT but its maximum IoU with an out-of-group GT of the same class reached the current threshold, it was marked as ignore and excluded from both TP and FP counts. Predictions satisfying neither condition were counted as FP. After evaluation, unmatched in-group GT boxes were counted as FN.

For each group g, class c, and IoU threshold $\tau$, the resulting average precision is denoted by $AP_{g,c,\tau}$. For mAP50, $T = \{0.50\}$; for mAP50-95, $T = \{0.50, 0.55, ..., 0.95\}$. The grouped mAP was calculated as:

$$\mathrm{mAP}_g = \frac{1}{|C_g|}\sum_{c \in C_g}\left(\frac{1}{|T|}\sum_{\tau \in T} AP_{g,c,\tau}\right) \tag{23}$$

where $C_g$ is the set of valid classes containing ground-truth samples in group g; classes without ground-truth samples in that group were excluded from the macro-average.

### 2.4.3 Model Complexity and Theoretical Computational Cost

Model complexity was quantified using the number of parameters and theoretical floating-point operations. The total number of network parameters was obtained using the model-analysis tool:

$$N_{\text{param}} = |\theta| \tag{24}$$

$$\mathrm{GFLOPs} = \frac{\mathrm{FLOPs}(960 \times 960)}{10^9} \tag{25}$$

where $\theta$ denotes the set of all trainable parameters in the network and $|\theta|$ is the total parameter count. GFLOPs were calculated by the model-analysis tool at an input size of 960 × 960 and represent the billions of floating-point operations required for one forward pass of a single 960 × 960 image; no proportional conversion across input resolutions was applied.

# 3 Results and Analysis

## 3.1 Performance Comparison between HMTS-YOLO and Other Detection Models

### 3.1.1 Comparison of Detection Performance and Model Complexity

To compare HMTS-YOLO with different YOLO-series detectors under a unified training and evaluation protocol, YOLOv5su, YOLOv8s, YOLOv9s, YOLOv10s, and YOLO11s were selected as comparison models. The results are presented in Table 3.

Table 3. Comparison of detection performance and model complexity across different models

| Model | Precision / % | Recall / % | mAP50 / % | mAP50-95 / % | Parameters / M | GFLOPs |
|---|---|---|---|---|---|---|
| YOLOv5su | 66.344 | 66.076 | 65.945 | 26.832 | 9.123 | 54.102 |
| YOLOv8s | 69.117 | 66.323 | 68.788 | 30.663 | 11.136 | 64.461 |
| YOLOv9s | 70.203 | 66.374 | 70.857 | 32.951 | 7.288 | 61.622 |
| YOLOv10s | 63.793 | 63.980 | 66.275 | 31.887 | 8.068 | 55.742 |
| YOLO11s | 68.657 | 66.766 | 69.752 | 33.127 | 9.429 | 48.489 |
| HMTS-YOLO | 72.736 | 68.988 | 73.129 | 35.177 | 11.572 | 69.157 |

On the validation set, HMTS-YOLO achieved a Precision of 72.736 %, Recall of 68.988 %, mAP50 of 73.129 %, and mAP50-95 of 35.177 %, all of which were the highest among the six compared models.

Compared with the baseline, HMTS-YOLO improved Precision, Recall, mAP50, and mAP50-95 by 3.619, 2.665, 4.341, and 4.514 percentage points, respectively. It also exceeded YOLO11s, the second-best model in mAP50-95, by 2.050 percentage points. The fact that HMTS-YOLO maintained the best result under the more stringent mAP50-95 metric indicates that its performance gain was not limited to a relatively permissive IoU threshold.

Relative comparisons further show that HMTS-YOLO exceeded YOLOv9s by 2.533, 2.614, 2.272, and 2.226 percentage points in Precision, Recall, mAP50, and mAP50-95, respectively, and exceeded YOLO11s by 4.079, 2.222, 3.377, and 2.050 percentage points. Thus, the advantage of HMTS-YOLO does not arise solely from comparison with a weaker baseline; it also remains superior to the relatively high-performing YOLOv9s and YOLO11s under the present experimental setting.

After model selection based on the validation set, the final HMTS-YOLO model was evaluated on the independent test set containing 465 images and 1,367 annotated boxes. It achieved a Precision of 70.703 %, Recall of 69.644 %, mAP50 of 72.253 %, and mAP50-95 of 35.274 %.

### 3.1.2 Relationship between Model Complexity and Detection Accuracy

To further examine the trade-off between improved detection accuracy and theoretical computational cost, an accuracy–complexity plot was constructed using the GFLOPs and mAP50-95 values in Table 3, as shown in Fig. 8.

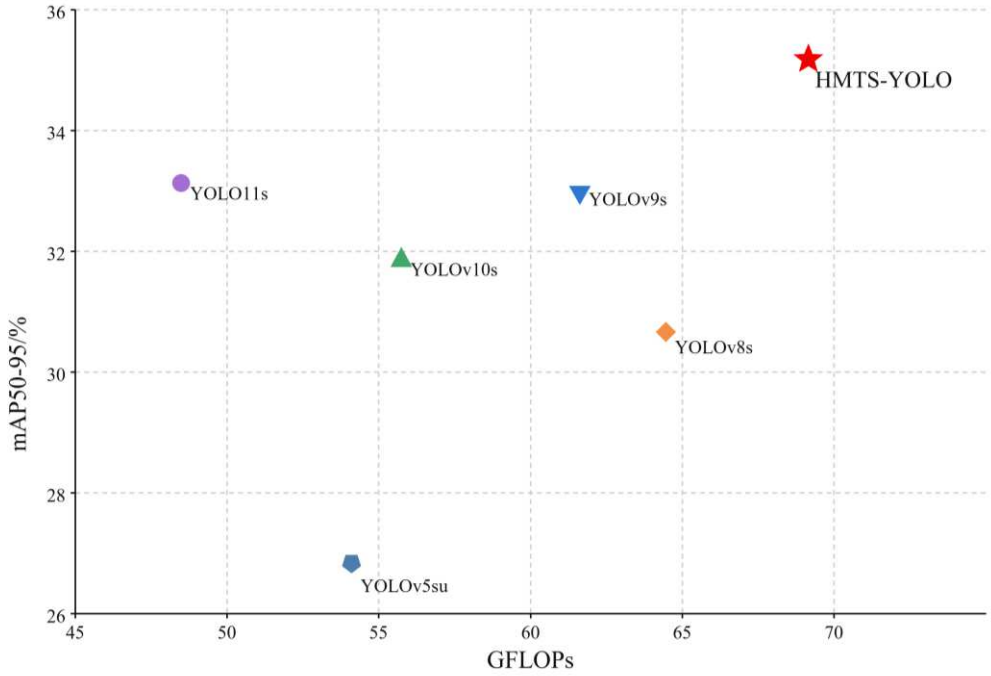



Fig. 8. Relationship between detection accuracy and model complexity

According to Table 3, HMTS-YOLO achieved an mAP50-95 of 35.177 % at 69.157 GFLOPs, outperforming YOLOv9s (32.951 %), YOLO11s (33.127 %), and the baseline (30.663 %). Relative to the baseline, computational cost increased by 4.696 GFLOPs while mAP50-95 increased by 4.514 percentage points.

In terms of parameter count, HMTS-YOLO contains 11.572 M parameters, only 0.436 M more than the baseline, although it has the largest parameter count among the models in Table 3. This indicates that the increase in parameter scale is relatively limited and that the principal complexity difference is reflected in theoretical computation.

The points in Fig. 8 indicate that theoretical computational cost and detection accuracy do not exhibit a simple linear relationship. YOLO11s requires only 48.489 GFLOPs, substantially less than the 69.157 GFLOPs of HMTS-YOLO, but its mAP50-95 is 2.050 percentage points lower; the difference in computational cost between the two models is 20.668 GFLOPs. Therefore, HMTS-YOLO is not the least computationally demanding model but lies toward the high-accuracy end of the comparison. Because actual inference latency and throughput were not measured in this study, Fig. 8 should be interpreted primarily as a comparison between theoretical computational cost and detection accuracy.

## 3.2 Complete Combinatorial Ablation and Training-Behavior Analysis of the Core Modules

### 3.2.1 Complete Combinatorial Ablation of the Three Core Modules

To further distinguish the individual contributions of DMA-SPPF, LDM-C2f, and STS-C2f across different module configurations, a full eight-configuration combinatorial ablation was conducted. The results

are shown in Table 4.

Table 4. Complete combinatorial ablation of the three core modules

| Model | DMA-SPPF | LDM-C2f | STS-C2f | Precision / % | Recall / % | mAP50 / % | mAP50-95 / % |
|---|---|---|---|---|---|---|---|
| Baseline | × | × | × | 69.117 | 66.323 | 68.788 | 30.663 |
| + DMA-SPPF | √ | × | × | 70.682 | 69.240 | 72.467 | 33.647 |
| + LDM-C2f | × | √ | × | 69.665 | 66.981 | 69.831 | 30.655 |
| + STS-C2f | × | × | √ | 69.125 | 66.403 | 68.003 | 29.891 |
| + DMA-SPPF + LDM-C2f | √ | √ | × | 70.915 | 68.039 | 71.272 | 34.208 |
| + DMA-SPPF + STS-C2f | √ | × | √ | 70.714 | 66.874 | 70.800 | 33.937 |
| + LDM-C2f + STS-C2f | × | √ | √ | 70.185 | 65.645 | 68.139 | 29.210 |
| HMTS-YOLO | √ | √ | √ | 72.736 | 68.988 | 73.129 | 35.177 |

Using mAP50-95 as the primary evaluation metric, DMA-SPPF produced the clearest gain among the single-module variants, increasing performance from 30.663 % for the baseline to 33.647 %, a gain of 2.984 percentage points. Neither LDM-C2f nor STS-C2f alone exceeded the baseline. Among the dual-module variants, DMA-SPPF + LDM-C2f achieved the highest result of 34.208 %, 3.545 percentage points above the baseline. With all three modules combined, HMTS-YOLO further reached 35.177 %, a 4.514-percentage-point improvement and the best result among all eight configurations.

The observed performance changes across configurations do not follow a simple additive pattern in the present experiment. DMA-SPPF provides a clear baseline improvement, whereas LDM-C2f and STS-C2f do not yield an overall gain when used individually. However, performance improves further when they are combined with DMA-SPPF, suggesting a configuration-dependent complementary effect among enhancement mechanisms operating at different feature hierarchies. In addition, the complete HMTS-YOLO achieved the highest Precision, mAP50, and mAP50-95 among all eight configurations, indicating superior overall detection performance for the three-module design.

### 3.2.2 Performance Evolution of the Baseline and HMTS-YOLO

To examine how the architectural modifications affect the training process and the evolution of validation performance, the mAP50 and mAP50-95 curves of the baseline and HMTS-YOLO during training were compared, as shown in Fig. 9.

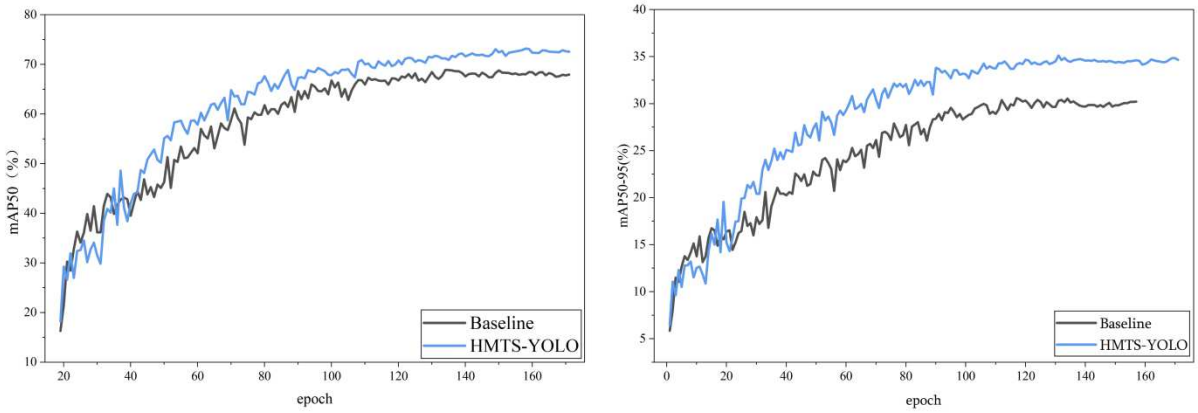



Fig. 9. Validation mAP curves of the baseline and HMTS-YOLO during training

Near the end of training, the mAP50 and mAP50-95 values of HMTS-YOLO were approximately 73.2 %

and 34.7 %, respectively, whereas those of the baseline were approximately 68.5 % and 30.2 %. The two models therefore maintained the same relative ranking as in the final evaluation. Across the full training process, both models exhibited some fluctuations during the early epochs and gradually stabilized, while HMTS-YOLO generally remained above the baseline in both metrics during the middle and late stages of training.

These results indicate that the performance advantage of HMTS-YOLO over the baseline is not confined to a single training epoch but is sustained through the middle and later stages of training. It should be noted that Fig. 9 shows dynamic validation results at individual epochs, whereas Table 4 reports the standardized evaluation obtained from the best checkpoint saved during training. Consequently, the final points of the curves do not exactly match the values in Table 4, and the final quantitative comparison between models should be based on Table 4.

## 3.3 Class-Wise Detection Performance of Four Progressive Configurations

To determine whether the complete model improves both apex and root treatment-region detection, class-wise results were further calculated for four progressive configurations, as shown in Table 5.

Table 5. Class-wise detection performance of four progressive configurations

| Model | Class | Precision / % | Recall / % | mAP50 / % | mAP50-95 / % |
|---|---|---|---|---|---|
| Baseline | apex | 70.445 | 65.615 | 70.051 | 37.243 |
| | root | 67.790 | 67.032 | 67.526 | 24.083 |
| + DMA-SPPF | apex | 71.089 | 68.988 | 73.438 | 40.514 |
| | root | 70.275 | 69.492 | 71.496 | 26.781 |
| + DMA-SPPF + LDM-C2f | apex | 72.028 | 68.823 | 73.492 | 41.811 |
| | root | 69.802 | 67.256 | 69.052 | 26.606 |
| HMTS-YOLO | apex | 73.080 | 69.403 | 74.336 | 42.212 |
| | root | 72.393 | 68.573 | 71.923 | 28.142 |

HMTS-YOLO achieved mAP50-95 values of 42.212 % for apex and 28.142 % for root, representing gains of 4.969 and 4.059 percentage points over the baseline, respectively. For mAP50-95, both classes achieved their highest values with the complete model. However, root Recall was highest for +DMA-SPPF (69.492 %), slightly above HMTS-YOLO (68.573 %), indicating that the complete model’s advantage was not uniform across all evaluation metrics. Across the progressive configurations, apex mAP50-95 increased consistently from 37.243 % to 42.212 %, whereas root performance showed a slight fluctuation after the addition of LDM-C2f before ultimately increasing to 28.142 %, indicating that the two target classes do not respond identically to different module configurations.

In the final results, root mAP50-95 remained 14.070 percentage points lower than apex, indicating that root is still the more difficult class under stringent IoU criteria. Comparison of the other metrics shows a larger Recall improvement for apex and a larger Precision improvement for root, suggesting that the

complete model improves the two classes in somewhat different ways. Table 5 is intended primarily to characterize class-level response patterns, whereas the combined effects of the three modules are assessed from the complete ablation results in Table 4.

## 3.4 Detection Performance across Target Scales and Morphologies

### 3.4.1 Detection Performance across Target Scales

To analyze how model gains are distributed across target scales, validation targets were divided into small, medium, and large groups, and the four progressive configurations were evaluated separately for each group. The results are presented in Table 6.

Table 6. mAP comparison of four progressive configurations across target scales

| Group | Annotations | Baseline | | + DMA-SPPF | | + DMA-SPPF + LDM-C2f | | HMTS-YOLO | |
|---|---|---|---|---|---|---|---|---|---|
| | | mAP50 / % | mAP50-95 / % | mAP50 / % | mAP50-95 / % | mAP50 / % | mAP50-95 / % | mAP50 / % | mAP50-95 / % |
| Small targets | 849 | 24.380 | 6.878 | 28.645 | 8.642 | 26.499 | 8.119 | 28.277 | 8.797 |
| Medium targets | 1499 | 67.777 | 28.068 | 70.899 | 31.147 | 70.364 | 30.938 | 71.723 | 32.214 |
| Large targets | 274 | 28.689 | 16.209 | 31.356 | 19.098 | 33.162 | 21.064 | 32.095 | 20.194 |

For small, medium, and large targets, HMTS-YOLO achieved mAP50-95 values of 8.797 %, 32.214 %, and 20.194 %, respectively, representing gains of 1.919, 4.146, and 3.985 percentage points over the baseline. Thus, the complete model improved performance across all three scale groups. Medium targets achieved the highest absolute performance, whereas small targets remained the most difficult, indicating pronounced differences among target-scale groups.

The best-performing configuration was not identical across all scales. For small targets, the mAP50 of DMA-SPPF alone was slightly higher than that of the complete model, whereas for large targets both metrics of DMA-SPPF + LDM-C2f exceeded those of HMTS-YOLO. The complete model performed best for medium targets. Therefore, the principal advantage of HMTS-YOLO lies in overall performance and cross-scale robustness rather than absolute dominance for every scale and metric. These results show that HMTS-YOLO provides the strongest overall cross-scale performance, although the best configuration varies with target scale and evaluation metric.

### 3.4.2 Detection Performance for Compact and Elongated Targets

To further examine the effect of target morphology on detection performance, targets were divided into compact and elongated groups according to bounding-box elongation, and the four progressive configurations were compared. The results are shown in Table 7.

Table 7. Comparison of mAP50 and mAP50-95 across target morphologies for four progressive configurations

| Group | Annotations | Baseline | | + DMA-SPPF | | + DMA-SPPF + LDM-C2f | | HMTS-YOLO | |
|---|---|---|---|---|---|---|---|---|---|
| | | mAP50 / % | mAP50-95 / % | mAP50 / % | mAP50-95 / % | mAP50 / % | mAP50-95 / % | mAP50 / % | mAP50-95 / % |
| Compact targets | 2150 | 64.936 | 28.784 | 68.618 | 31.926 | 68.140 | 32.143 | 69.294 | 32.934 |
| Elongated targets | 472 | 52.711 | 14.593 | 59.196 | 18.344 | 54.280 | 17.311 | 57.942 | 18.932 |

Table 7 contains 2,150 compact targets and 472 elongated targets. For compact targets, HMTS-YOLO achieved an mAP50 of 69.294 % and an mAP50-95 of 32.934 %; for elongated targets, the corresponding values were 57.942 % and 18.932 %. Relative to the baseline, mAP50-95 improved by 4.150 and 4.339 percentage points for compact and elongated targets, respectively. The complete model achieved the highest values for both metrics on compact targets. For elongated targets, HMTS-YOLO achieved the highest mAP50-95, whereas the highest mAP50 (59.196 %) was obtained by the +DMA-SPPF configuration.

In absolute terms, mAP50-95 for elongated targets was 14.002 percentage points lower than that for compact targets, indicating substantially poorer overall detection performance for elongated structures under the current dataset and evaluation settings. Meanwhile, HMTS-YOLO improved mAP50 for elongated targets by 5.231 percentage points, exceeding the 4.358-percentage-point gain for compact targets; however, this relative improvement does not imply that the intrinsic detection difficulty of elongated targets has been eliminated.

Model rankings differed across evaluation metrics. For elongated targets, +DMA-SPPF achieved an mAP50 of 59.196 %, higher than the 57.942 % of HMTS-YOLO. Under the more stringent mAP50-95 metric, however, HMTS-YOLO reached 18.932 %, exceeding the 18.344 % of +DMA-SPPF. Thus, the advantage of HMTS-YOLO is more evident under the stricter mAP50-95 metric.

## 3.5 Field-Scene Detection Visualization and Real-World System Demonstration

### 3.5.1 Detection Visualization in Field Scenes

To qualitatively compare the predictions of the Baseline and HMTS-YOLO with the corresponding ground-truth annotations under complex soil textures, surface cracks, target-scale variation, and different treatment-region morphologies, representative field scenes were selected, as shown in Fig. 10. The top, middle, and bottom rows show the ground-truth annotations, Baseline predictions, and HMTS-YOLO predictions, respectively. Red and teal bounding boxes indicate apex and root, respectively.

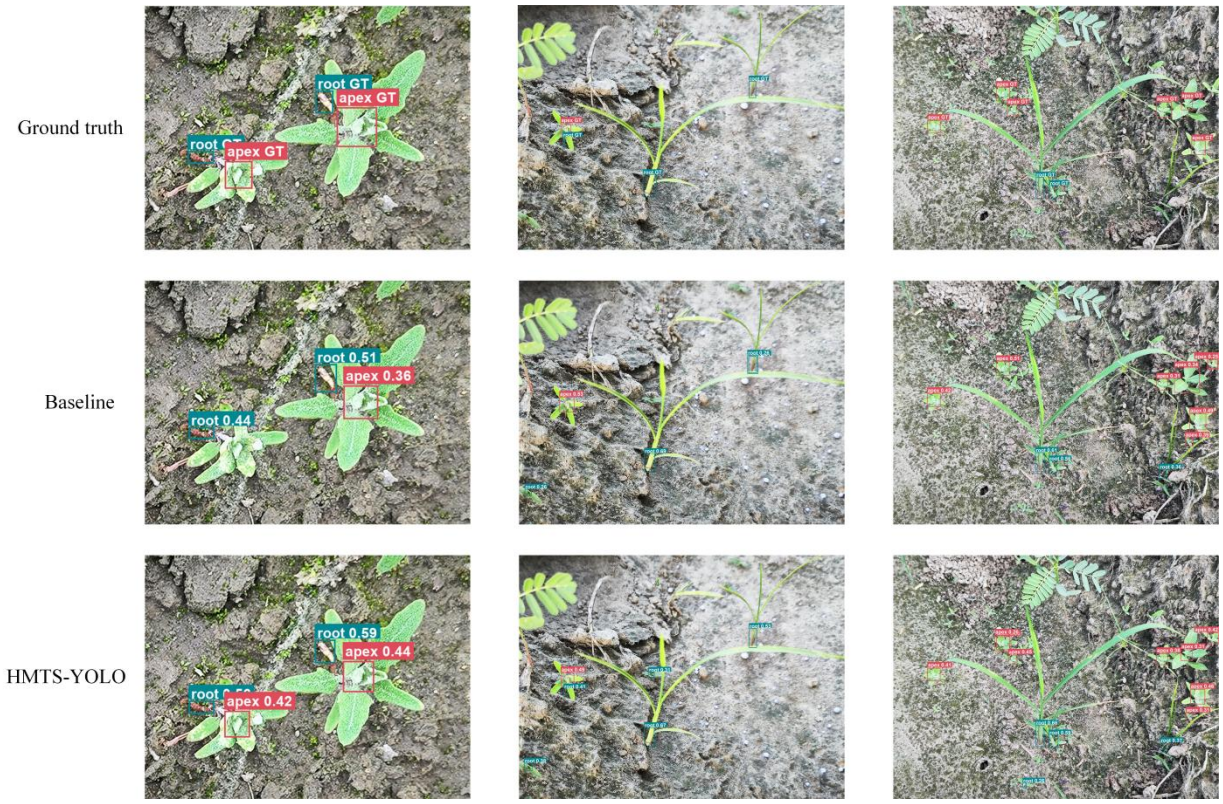



Fig. 10. Qualitative comparison of ground-truth annotations, Baseline predictions, and HMTS-YOLO predictions in representative field scenes.

The comparison shows differences between the two models in the number of predicted apex and root regions, bounding-box locations, and confidence scores under complex backgrounds. Root regions are generally smaller and more elongated and are therefore more susceptible to interference from soil cracks, granular textures, and neighboring plants. The quantitative performance of the two models is evaluated using the metrics reported in Table 4.

### 3.5.2 Real-World Dual-Camera Detection and Galvanometer-Coordinate Output

During the preliminary real-world system demonstration, HMTS-YOLO was integrated into the dual-camera acquisition and processing workflow. The interface displayed apex and root detection boxes, confidence scores, bounding-box centers, stereo-matching results, three-dimensional coordinates, and galvanometer coordinates. As shown in Fig. 11, the system generated multiple candidate targets under soil background, plant occlusion, and varying target numbers.

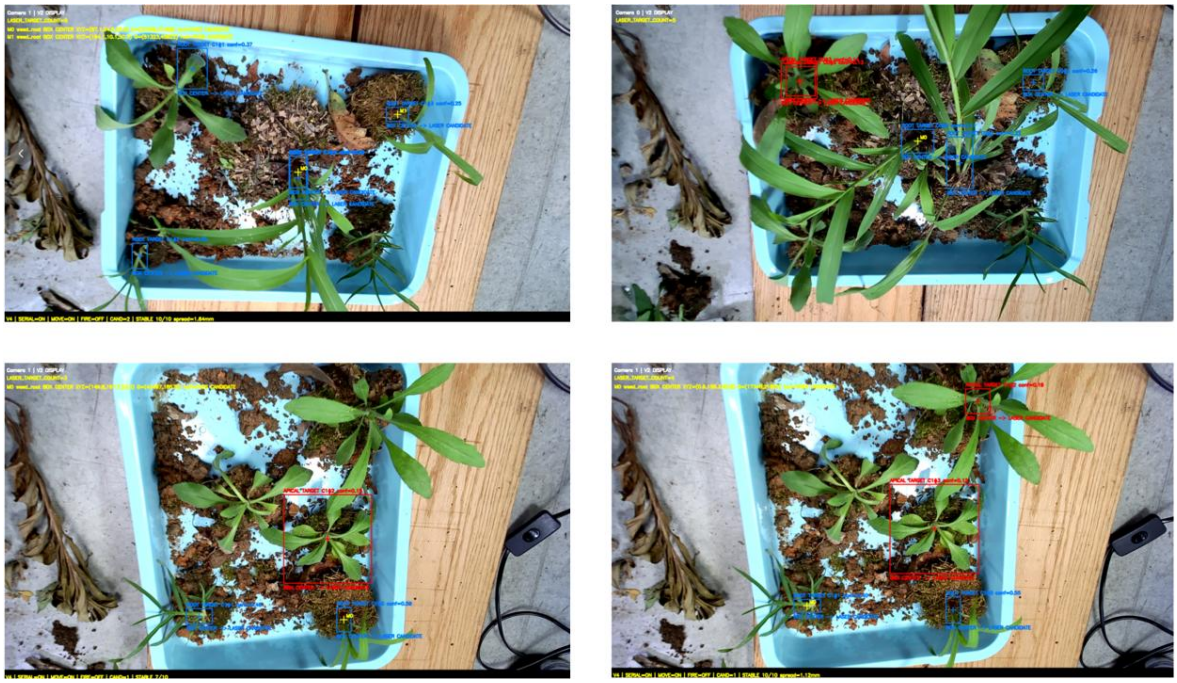


Fig. 11. Real-world dual-camera detection, stereo localization, and galvanometer-coordinate output

The reported runs were conducted with MOVE enabled and FIRE disabled. The results demonstrate the integrated process from dual-camera detection and stereo localization to three-dimensional coordinate estimation and galvanometer-coordinate generation. Because independent laser-spot measurements were not available, aiming accuracy and laser-firing success rate were not quantitatively evaluated.

# 4 Discussion

DMA-SPPF provides the clearest individual gain: mAP50-95 increases from 30.663 % to 33.647 %, rises to 34.208 % when LDM-C2f is added, and reaches 35.177 % in the complete model. This result agrees with the role of feature pyramids and multi-scale aggregation in object detection[15,16]. LDM-C2f and STS-C2f do not improve the overall score alone, and their joint configuration remains below the baseline. Their effects become favorable when combined with DMA-SPPF, indicating configuration-dependent complementarity rather than simple additive gains. This pattern suggests that, in the current architecture, the benefits of LDM-C2f and STS-C2f emerge more clearly when combined with DMA-SPPF.

Root, small, and elongated targets remain the main difficulties. HMTS-YOLO achieves 42.212 % mAP50-95 for apex and 28.142 % for root. The values for small, medium, and large targets are 8.797 %, 32.214 %, and 20.194 %, respectively, while compact and elongated targets reach 32.934 % and 18.932 %. Lower performance for root and elongated targets is plausibly related to smaller visible regions, weaker boundaries, and greater similarity to soil texture, as reported in agricultural weed datasets and detection studies[26,27,29,30]. The gap between compact and elongated groups further indicates that object geometry affects the reliability of local feature responses.

Detecting apex and root as local treatment regions may support site-specific weed management and

laser-based operation. The P2-to-P5 placement of STS-C2f, LDM-C2f, and DMA-SPPF follows a local-to-context hierarchy related to feature-pyramid and multi-scale detector designs. This helps explain the complete model's broad improvement across target groups, although the best configuration varies by subgroup and metric. With 11.572 M parameters and 69.157 GFLOPs, HMTS-YOLO remains an accuracy-oriented design.

The remaining errors were concentrated in root, small, and elongated targets, indicating sensitivity to target visibility, boundary quality, and soil-background similarity. Performance should therefore be interpreted within the current dataset and evaluation protocol; broader field conditions and quantitative laser-spot measurements remain topics for future work.

# 5 Conclusions

This study developed HMTS-YOLO to address the susceptibility of dual-class weed apex and root targets to soil texture, scale variation, and morphological differences in complex field scenes. Based on YOLOv8s, the model introduces STS-C2f, LDM-C2f, and DMA-SPPF at the P2, P3, and P5 feature levels, respectively, to enhance feature representation across target scales and morphologies through shallow texture suppression, mid-level morphology modeling, and deep multi-scale contextual aggregation.

Model comparisons showed that HMTS-YOLO achieved a Precision of 72.736 %, Recall of 68.988 %, mAP50 of 73.129 %, and mAP50-95 of 35.177 % on the validation set, all higher than the baseline; mAP50-95 increased by 4.514 percentage points. The model contains 11.572 M parameters and requires 69.157 GFLOPs, indicating that the proposed improvement in detection performance is accompanied by additional computational cost and should currently be interpreted primarily as an accuracy-oriented architectural enhancement.

The complete eight-configuration ablation study showed that DMA-SPPF alone improved mAP50-95 by 2.984 percentage points, DMA-SPPF combined with LDM-C2f improved it by 3.545 percentage points, and the complete HMTS-YOLO achieved the largest gain of 4.514 percentage points and the best overall result.

In summary, HMTS-YOLO improves apex and root detection in complex field scenes and provides a basis for future precision treatment.

**CRediT authorship contribution statement**

**Yongbing Chen:** Writing-original draft, Methodology, Investigation, Data curation, Funding acquisition. **Meiling Yuan:** Formal analysis, Visualization, Software, Writing-review & editing, Methodology. **Liming Qin:** Writing-original draft, Software, Writing-review & editing, Formal analysis, Supervision. **Qian Hu:** Formal analysis, Validation, Software. **Simon X. Yang:** Methodology, Formal analysis, Resources. **Xiaofeng Ling:** Supervision, Resources, Project administration, Conceptualization.

**Declaration of competing interest**

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

**Funding**

This work was supported by the Wenzhou Key Laboratory of Embodied Intelligent Agents for Agriculture (Grant no. AIAA2504); Taizhou Science and Technology Planning Project (grant no.: 24nya06); Taizhou Science and Technology Planning Project (grant no.: 25nyb38); National College Students Innovation and Entrepreneurship Training Program (grant no. 2024103500024).

**Data availability**

Data will be made available on request.

**Acknowledgments**

Thanks to all the authors cited in this article and the referees for their helpful comments and suggestions.

# References

[1] Oerke E-C. Crop losses to pests. J Agric Sci 2006;144:31–43. https://doi.org/10.1017/S0021859605005708.

[2] Heap I. Global perspective of herbicide-resistant weeds. Pest Management Science 2014;70:1306–15. https://doi.org/10.1002/ps.3696.

[3] Beckie HJ. Herbicide-Resistant Weeds: Management Tactics and Practices. Weed Technol 2006;20:793–814. https://doi.org/10.1614/WT-05-084R1.1.

[4] Gonzalez-de-Santos P, Ribeiro A, Fernandez-Quintanilla C, Lopez-Granados F, Brandstoetter M, Tomic S, et al. Fleets of robots for environmentally-safe pest control in agriculture. Precision Agric 2017;18:574–614. https://doi.org/10.1007/s11119-016-9476-3.

[5] Xu Y, Bai Y, Fu D, Cong X, Jing H, Liu Z, et al. Multi-species weed detection and variable spraying system for farmland based on W-YOLOv5. Crop Protection 2024;182:106720. https://doi.org/10.1016/j.cropro.2024.106720.

[6] Liu D, Li J, Su H, Bei Cui, Wang Z, Yuan Q, et al. Towards Efficient and Intelligent Laser Weeding: Method and Dataset for Weed Stem Detection. Proceedings of the AAAI Conference on Artificial Intelligence, vol. 39, 2025, p. 28204–12. https://doi.org/10.1609/aaai.v39i27.35040.

[7] Mathiassen SK, Bak T, Christensen S, Kudsk P. The Effect of Laser Treatment as a Weed Control Method. Biosystems Engineering 2006;95:497–505. https://doi.org/10.1016/j.biosystemseng.2006.08.010.

[8] Andreasen C, Scholle K, Saberi M. Laser Weeding With Small Autonomous Vehicles: Friends or Foes? Front Agron 2022;4:841086. https://doi.org/10.3389/fagro.2022.841086.

[9] Yang X. Design and Experiment of Laser Weeding Robot. Master's thesis. Zhejiang University of Science and Technology, 2025.

[10] Long J, Shelhamer E, Darrell T. Fully convolutional networks for semantic segmentation. 2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Boston, MA, USA: IEEE; 2015, p. 3431–40. https://doi.org/10.1109/CVPR.2015.7298965.

[11] Ren S, He K, Girshick R, Sun J. Faster R-CNN: Towards Real-Time Object Detection with Region Proposal Networks. Advances in Neural Information Processing Systems, vol. 28, 2015, p. 91–9.

[12] Redmon J, Divvala S, Girshick R, Farhadi A. You Only Look Once: Unified, Real-Time Object Detection. 2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Las Vegas, NV, USA: IEEE; 2016, p. 779–88. https://doi.org/10.1109/CVPR.2016.91.
[13] He K, Gkioxari G, Dollár P, Girshick R. Mask R-CNN. 2017 IEEE International Conference on Computer Vision (ICCV), Venice, Italy: IEEE; 2017, p. 2961–9. https://doi.org/10.1109/ICCV.2017.322.
[14] Lin T-Y, Maire M, Belongie S, Hays J, Perona P, Ramanan D, et al. Microsoft COCO: Common Objects in Context. In: Fleet D, Pajdla T, Schiele B, Tuytelaars T, editors. Computer Vision – ECCV 2014, vol. 8693, Cham: Springer International Publishing; 2014, p. 740–55. https://doi.org/10.1007/978-3-319-10602-1_48.
[15] Lin T-Y, Dollar P, Girshick R, He K, Hariharan B, Belongie S. Feature Pyramid Networks for Object Detection. 2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Honolulu, HI: IEEE; 2017, p. 936–44. https://doi.org/10.1109/CVPR.2017.106.
[16] Tan M, Pang R, Le QV. EfficientDet: Scalable and Efficient Object Detection. Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2020, p. 10781–90. https://doi.org/10.1109/CVPR42600.2020.01079.
[17] Ge Z, Liu S, Wang F, Li Z, Sun J. YOLOX: Exceeding YOLO Series in 2021 2021. https://doi.org/10.48550/ARXIV.2107.08430.
[18] Wang C-Y, Bochkovskiy A, Liao H-YM. YOLOv7: Trainable bag-of-freebies sets new state-of-the-art for real-time object detectors. Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2023, p. 7464–75. https://doi.org/10.1109/CVPR52729.2023.00721.
[19] Zhang H, Wang Y, Dayoub F, Sunderhauf N. VarifocalNet: An IoU-aware Dense Object Detector. 2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), Nashville, TN, USA: IEEE; 2021, p. 8514–23. https://doi.org/10.1109/CVPR46437.2021.00841.
[20] Li X, Wang W, Wu L, Chen S, Hu X, Li J, et al. Generalized Focal Loss: Learning Qualified and Distributed Bounding Boxes for Dense Object Detection. Advances in Neural Information Processing Systems, vol. 33, 2020, p. 21075–85.
[21] Zhao Y, Lv W, Xu S, Wei J, Wang G, Dang Q, et al. DETRs Beat YOLOs on Real-time Object Detection. Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2024, p. 16965–74. https://doi.org/10.1109/CVPR52733.2024.01605.
[22] Dyrmann M, Karstoft H, Midtiby HS. Plant species classification using deep convolutional neural network. Biosystems Engineering 2016;151:72–80. https://doi.org/10.1016/j.biosystemseng.2016.08.024.
[23] Milioto A, Lottes P, Stachniss C. Real-Time Semantic Segmentation of Crop and Weed for Precision Agriculture Robots Leveraging Background Knowledge in CNNs. 2018 IEEE International Conference on Robotics and Automation (ICRA), Brisbane, QLD, Australia: IEEE; 2018, p. 2229–35. https://doi.org/10.1109/ICRA.2018.8460962.
[24] Lottes P, Behley J, Chebrolu N, Milioto A, Stachniss C. Joint Stem Detection and Crop-Weed Classification for Plant-Specific Treatment in Precision Farming. 2018 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS), Madrid: IEEE; 2018, p. 8233–8. https://doi.org/10.1109/IROS.2018.8593678.
[25] Bah MD, Hafiane A, Canals R. Deep Learning with Unsupervised Data Labeling for Weed Detection in Line Crops in UAV Images. Remote Sensing 2018;10:1690. https://doi.org/10.3390/rs10111690.
[26] Olsen A, Konovalov DA, Philippa B, Ridd P, Wood JC, Johns J, et al. DeepWeeds: A Multiclass Weed Species Image Dataset for Deep Learning. Sci Rep 2019;9:2058. https://doi.org/10.1038/s41598-018-38343-3.
[27] Wang P, Tang Y, Luo F, Wang L, Li C, Niu Q, et al. Weed25: A deep learning dataset for weed identification. Front Plant Sci 2022;13:1053329. https://doi.org/10.3389/fpls.2022.1053329.
[28] Hasan ASMM, Sohel F, Diepeveen D, Laga H, Jones MGK. A Survey of Deep Learning Techniques for Weed Detection from Images. Computers and Electronics in Agriculture 2021;184:106067. https://doi.org/10.1016/j.compag.2021.106067.
[29] Hu W, Wane SO, Zhu J, Li D, Zhang Q, Bie X, et al. Review of deep learning-based weed identification in crop fields. International Journal of Agricultural and Biological Engineering 2023;16:1–10.

https://doi.org/10.25165/j.ijabe.20231604.8364.
[30] Murad NY, Mahmood T, Forkan ARM, Morshed A, Jayaraman PP, Siddiqui MS. Weed Detection Using Deep Learning: A Systematic Literature Review. Sensors 2023;23:3670. https://doi.org/10.3390/s23073670.
[31] Yu H, Che M, Yu H, Ma Y. Research on weed identification in soybean fields based on the lightweight segmentation model DCSAnet. Front Plant Sci 2023;14:1268218. https://doi.org/10.3389/fpls.2023.1268218.
[32] Li J, Zhang W, Zhou H, Yu C, Li Q. Weed detection in soybean fields using improved YOLOv7 and evaluating herbicide reduction efficacy. Front Plant Sci 2024;14:1284338. https://doi.org/10.3389/fpls.2023.1284338.
[33] Dos Santos Ferreira A, Matte Freitas D, Gonçalves Da Silva G, Pistori H, Theophilo Folhes M. Weed detection in soybean crops using ConvNets. Computers and Electronics in Agriculture 2017;143:314–24. https://doi.org/10.1016/j.compag.2017.10.027.
[34] Wang S, Wang S, Zhang H, Wen C. Soybean field weed recognition based on light sum-product networks and UAV remote sensing images. Transactions of the Chinese Society of Agricultural Engineering 2019;35:81–9.
[35] Wang A, Peng T, Cao H, Xu Y, Wei X, Cui B. TIA-YOLOv5: An improved YOLOv5 network for real-time detection of crop and weed in the field. Front Plant Sci 2022;13:1091655. https://doi.org/10.3389/fpls.2022.1091655.
[36] Liu M, Xu X, Tian T, Shang M, Song Z, Tian F, et al. A keypoint-based method for detecting weed growth points in corn field environments. Plant Phenomics 2025;7:100072. https://doi.org/10.1016/j.plaphe.2025.100072.
[37] Jocher G, Chaurasia A, Qiu J. Ultralytics YOLOv8. Ultralytics Docs 2023. https://docs.ultralytics.com/models/yolov8/.
[38] He K, Zhang X, Ren S, Sun J. Spatial Pyramid Pooling in Deep Convolutional Networks for Visual Recognition. IEEE Trans Pattern Anal Mach Intell 2015;37:1904–16. https://doi.org/10.1109/TPAMI.2015.2389824.
[39] Howard AG, Zhu M, Chen B, Kalenichenko D, Wang W, Weyand T, et al. MobileNets: Efficient Convolutional Neural Networks for Mobile Vision Applications 2017. https://doi.org/10.48550/arXiv.1704.04861.
[40] Ultralytics. Ultralytics YOLOv5. Ultralytics n.d. https://docs.ultralytics.com/models/yolov5/ (accessed August 20, 2026).
[41] Wang C-Y, Yeh I-H, Liao H-YM. YOLOv9: Learning What You Want to Learn Using Programmable Gradient Information. Computer Vision–ECCV 2024, vol. 15089, Springer; 2024, p. 1–21. https://doi.org/10.1007/978-3-031-72751-1_1.
[42] Wang A, Chen H, Liu L, Chen K, Lin Z, Han J, et al. YOLOv10: Real-Time End-to-End Object Detection. Advances in Neural Information Processing Systems, vol. 37, 2024. https://doi.org/10.52202/079017-3429.
[43] Ultralytics. Ultralytics YOLO11 2024. https://docs.ultralytics.com/models/yolo11/.