SEMI-SUPERVISED MAIZE SEEDLING SEMANTIC SEGMENTATION METHOD BASED ON VISION TRANSFORMER AND CURRICULUM LEARNING
/
基于视觉变换器和课程学习的半监督玉米幼苗识别方法

Zhicheng TANG[1)], Yuxin ZHU[2)], Weiyi FENG[1)], Junke ZHU [*1,3)]
[1)] Shandong University of Technology, College of Agricultural Engineering and Food Science, Department, Zibo / China;
[2)] Shandong Agricultural University, Department of Agricultural and Forestry Economics and Management, Taian / China;
[3)]Qilu Normal University, Jinan/China;
Tel: +8618905338833; E-mail: zhujunke@126.com
Corresponding author: Junke Zhu
DOI: https://doi.org/10.35633/inmateh-78-60

Keywords: *Maize, Deep learning, Convolutional neural network, Semi-supervised learning, Transformer, Smart agriculture*

ABSTRACT
Crop semantic segmentation plays a crucial role in precision agriculture, enabling applications such as growth monitoring, yield prediction, and pest control. However, deep learning methods, such as U-Net, rely heavily on large amounts of labelled data, which are costly and time-consuming to obtain in agricultural settings. To address this limitation, a semi-supervised maize segmentation method based on an improved Vision Transformer within a student-teacher framework is proposed. The model leverages limited labelled data and abundant unlabelled data through consistency training and confidence-based self-training. Experimental results demonstrate that the proposed method achieves a mean Intersection over Union (mIoU) of 0.661, representing a 14.3% improvement over U-Net. These results confirm its effectiveness in reducing annotation costs while achieving superior accuracy in complex farmland environments.

摘要
作物语义分割对于精准农业至关重要，有助于生长监测、产量预测和病虫害控制。然而，像 U-Net 这样的深度学习方法严重依赖大量标记数据，在农业环境中获取这些数据既昂贵又耗时。为此，我们提出了一种基于改进版 Vision Transformer 和师生框架的半监督玉米分割方法。该模型通过一致性训练和基于置信度的自我训练，利用有限的标记数据和大量的未标记数据。实验结果显示其 mIoU 为 0.661，比 U-Net 高 14.3%，表明其在减少注释负担的同时，同时在复杂农田场景中实现了更高的精度。

INTRODUCTION
The rapid development of precision agriculture has become a key driver for enhancing global crop yields, optimising agricultural resource allocation, and alleviating environmental pressures. As a cornerstone of precision agriculture, crop identification technology enables essential functions such as real-time monitoring of crop growth, accurate yield estimation, and early warning of pests and diseases, and is therefore of strategic importance for ensuring food security *(Wu et al., 2025).* In recent years, with the iterative updates of deep learning technologies, significant progress has been made in maize seedling detection and recognition, giving rise to various high-performance improved algorithms. For instance, *Chen et al., (2025),* proposed a real-time detection Transformer combining Convolutional Block Attention Modules (CBAM) and grouped convolutions, effectively enhancing the specificity of feature extraction. *Liu et al., (2025)* achieved lightweight improvements based on the YOLOv8n architecture, realizing efficient positioning of seedlings and weeds in the field. *Tang et al., (2025),* integrated UAV-based multispectral imagery with the CGS-YOLO algorithm, significantly strengthening recognition robustness under complex lighting conditions. These methods have demonstrated excellent detection performance in specific scenarios, providing crucial technical support for maize seedling monitoring *(Fan et al., 2025)*.

However, despite the relative maturity of research on object detection tasks, pixel-level semantic segmentation still faces severe challenges in complex open-field environments. Unlike detection tasks that only require locating bounding boxes, semantic segmentation demands precise classification for every pixel, imposing higher requirements on the model's ability to capture fine-grained features. Firstly, maize plants exhibit significant morphological plasticity at different growth stages; leaf overlap and curling lead to blurred boundaries, greatly increasing the difficulty of pixel-level classification *(Liu et al., 2024; Zhang et al., 2024)*.

Secondly, the field environment is highly unstructured; dynamic changes in lighting conditions, the complexity of soil background textures, and the random distribution of weeds significantly increase the uncertainty of feature extraction, easily leading to model misjudgements *(Yang et al., 2023; Ji et al., 2024)*. More critically, maize seedlings and weeds share high similarity in colour, shape, and texture features. Coupled with the dense canopy occlusion formed as plants grow, the cost of manually annotating high-quality segmentation data is extremely high, and the error rate is difficult to control *(Bai et al., 2024; Lin et al., 2024; Xu et al., 2024).* The coupling effect of these factors makes it extremely difficult to acquire large-scale, high-quality annotated datasets, thereby severely restricting the scalable application and generalization ability of traditional fully supervised deep learning models (such as U-Net and its variants) in actual field scenarios, often resulting in poor performance when facing unseen complex distributions *(Guo et al., 2024; Li et al., 2023).*

To break through the bottleneck of scarce annotated data, Semi-Supervised Learning (SSL) offers a new solution path for agricultural image analysis by synergistically utilizing a small number of labelled samples and a large volume of unlabelled data. Existing studies have generally attempted to introduce strategies such as pseudo-label generation and consistency regularization. However, obvious limitations remain in complex farmland scenarios. They fail to fully utilize the rich structural information contained in massive amounts of unlabelled data, resulting in limited improvements in model generalization. They struggle to effectively capture subtle texture differences between maize and weeds under complex backgrounds, indicating insufficient feature representation capabilities. They have not yet fully leveraged the advantages of self-supervised learning in pre-training or auxiliary tasks, lacking deep mining of the intrinsic distribution laws of the data. Specifically, traditional pseudo-label methods are susceptible to initial model bias, where noise accumulation leads to error propagation. Meanwhile, simple consistency regularization often assumes linear data perturbations, potentially oversimplifying the non-linear variations of the field environment. Furthermore, existing methods based on the YOLO series *(Li et al., 2023)* or lightweight CNNs *(Deng et al., 2022)* mostly focus on supervised training paradigms. They lack an adaptive learning mechanism for the "simple-to-complex" distribution characteristics in unlabelled data, making it difficult to effectively handle the complex transition from clear single-plant samples to densely occluded samples *(Hu et al., 2023; Huang et al., 2023).* This "one-size-fits-all" training approach ignores the heterogeneity of sample difficulty, causing the model to converge slowly or even fall into local optima when processing difficult samples.

Existing studies generally have the following limitations: they fail to fully utilize unlabelled data, have difficulty in effectively capturing subtle features in complex farmland environments, and do not fully leverage the advantages of self-supervised learning. To address these issues, this paper proposes a dual-stream complementary learning structure based on an improved Vision Transformer, which enhances the consistency and stability of feature extraction and pseudo-label generation through a student-teacher model framework. The student network learns supervised information from labelled data and processes unlabelled data using pseudo-labels generated by the teacher network with the help of consistency constraints. The teacher network updates its parameters through exponential moving average (EMA) to provide stable target predictions. The proposed method not only performs excellently in the maize segmentation task but also achieves a mean intersection over union ($mIOU$) of 0.661, which is 14.3% higher than that of U-Net and outperforms other semi-supervised methods. This validates its effectiveness under limited labelled data and provides new theoretical and practical references for the application of semi-supervised learning in agricultural image analysis.

MATERIALS AND METHODS

Overview of the study area

The study area is located in a typical farmland in Zibo, Shandong Province, China (36°48′ N, 118°02′ E). As shown in Fig. 1, the experimental site features a consistent soil background and uniform maize seedling distribution. The multispectral images were captured using a UAV platform at a flight altitude of 30 m, yielding a ground sampling distance (GSD) of 1.6 cm/pixel. A total of 1,613 raw images were initially collected. After rigorous quality control to remove blurred or overexposed samples, all 1,613 high-quality images were retained for dataset construction. From this pool, 242 independent images (15%) were strictly reserved as the test set to ensure unbiased evaluation, while the remaining images were split into training and validation sets. All images were pre-processed and resized to a standard resolution of 512 × 512 pixels

for model training and inference. The annotation was performed by professional agronomists using LabelMe software to generate pixel-level masks, categorizing each pixel into two classes: maize seedling and background.

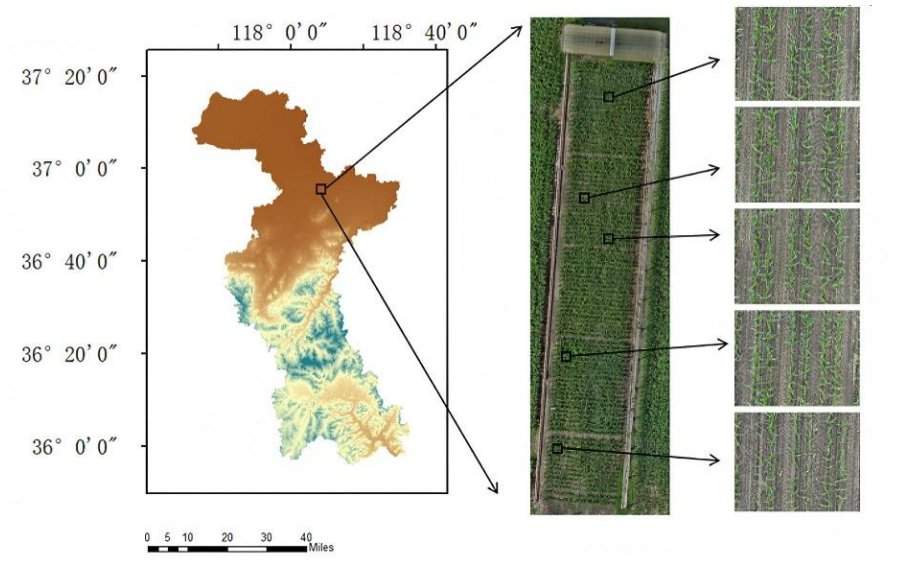

Fig. 1 - Overview map of the study area

UAV Parameters

In this experiment, visible light and multispectral images of maize at the jointing stage were acquired using the DJI Mavic 3 Multispectral (M3M) on July 5, 2024. The resolution of the visible light camera was 5280 × 3956 pixels, as shown in Figure 2. Table 1 details the UAV parameters. Data acquisition was conducted at 10:00 local time under clear weather conditions to ensure optimal illumination. The UAV flew at a constant altitude of 30 m with a flight speed of 2 m/s. The camera captured images at equal intervals along 6 automatically planned flight routes. Upon completion, the UAV automatically returned and landed, generating a total of 1,613 original images.

Fig. 2 - Unmanned aerial vehicles (UAV) used in the experiment

Table 1

Parameters of the experimental drones

Specification	Technical parameters	
	Visible light camera	**Multispectral camera**
Image sensor [effective pixels]	20 million	5 million
Lens	visual angle: 84° focal length: 24 [mm] f/2.8–f/11 1 m to infinity	visual angle: 73.91° focal length: 25 [mm] f/2.0 Fixed focus
Shutter speed [s]	8–1/8000	1/30~1/12800
Photo size	5280×3956	2592×1944

Specification	Technical parameters	
	Visible light camera	Multispectral camera
Photo capture mode	time-lapse shooting	time-lapse shooting
Photo format	JPEG/DNG (RAW)	TIFF
Camera band [nm]	visible light	(G): 560±15 (R): 650±15 [nm] (RE): 730±15 [nm] (NIR): 860±30 [nm]

Dataset construction

From the 1,613 original images, the dataset was split using a stratified random protocol: 70% (1,129 images) for training, 15% (242 images) for validation, and 15% (242 images) strictly reserved for the test set. To simulate a semi-supervised learning scenario, the training set was further divided: 50% (564 images) were manually annotated by professional agronomists to form the labelled subset, while the remaining 50% (565 images) served as the unlabelled subset. All images were pre-processed and resized to a standard resolution of 512 × 512 pixels. Annotations were generated using LabelMe software, creating pixel-level masks for two classes: maize seedling and background. Selecting maize images at the jointing stage is critical for this research, not only due to its importance in agricultural production but also because maize plants at this stage are highly sensitive to external environmental changes (such as drought, flood, pests, and diseases). Any adverse factors may lead to poor plant development, thereby affecting the final yield. Precise monitoring of maize plants at the jointing stage through high-resolution images can timely detect potential problems and facilitate corresponding measures, thus improving crop yield and quality. In addition, the jointing stage represents one of the most significant periods of morphological changes in the maize growth process. Analysis of maize images at this stage provides a deeper understanding of plant growth patterns and their responses to environmental changes, thereby offering a scientific basis for further optimisation of agricultural production.

Model Architecture and Methodology

In the task of agricultural maize field image segmentation, it is extremely difficult to obtain a large amount of labelled data due to the high annotation cost. Semi-supervised learning utilizes a small amount of labelled data and a large amount of unlabelled data to improve the performance of the model.

The semi-supervised learning framework for image segmentation of agricultural maize fields mainly consists of the following parts: data input layer, data augmentation module, student model, teacher model, and loss function. These components work together to ensure the accuracy of the model on labelled data and the consistency on unlabelled data. The model in this paper introduces an adaptive consistency mechanism strategy. By dynamically adjusting the consistency constraint strength of different regions and enhancing the intra-class feature similarity, it realizes the efficient utilization of unlabelled data. In addition, an uncertainty-guided pseudo-label generation strategy is designed to screen high-quality pseudo-labels according to the prediction confidence of the teacher model, reducing error accumulation.

First, the data input layer receives the original maize field images and passes them to the data augmentation module. This module uses various transformation techniques (such as geometric transformation, colour transformation, and noise addition) to increase the diversity and robustness of the training data. Subsequently, the augmented data is input into the student model and the teacher model for processing respectively. The student model is optimized by minimizing the supervised label loss and the consistency loss, while the teacher model updates its parameters through the exponential moving average (EMA) to provide stable target predictions. Among them, the supervised label loss uses the Cross-Entropy Loss function, which is defined as:

$$L_{CE} = -\sum_{C=1}^{C} \mathrm{y}_{o,c} \log(p_{o,}c) \quad (1)$$

$y_{o,c}$ is the true label (one-hot encoding) of the labelled data, p_o,c is the predicted probability of the student network that the sample o belongs to category c, and C is the total number of categories (e.g., maize plants, background, etc.). This loss forces the student network to learn accurate semantic segmentation features from limited labelled data, ensuring the semantic segmentation accuracy of the model on known categories.

Ultimately, the loss function combines the supervised loss and the consistency loss to ensure the accuracy and consistency of the model.

The final total loss function is:

$$L_{total} = \alpha \cdot L_{CE}(student\_labeled) + \beta \cdot L_{cons}^{adaptive}(student\_unlabeled, teacher\_pseudo\_labels) \quad (2)$$

where: α and β are coefficients for balancing the two losses (e.g., α = 1, β = 0.1). If there is a small amount of labelled data, β can be appropriately increased to fully utilize the unlabelled data; if the quality of the unlabelled data is low, β should be decreased to avoid noise interference.

The model architecture proposed in this paper is shown in Figure 3. The overall model adopts a two-stream architecture, consisting of two branches: the student network and the teacher network. The student network is responsible for learning supervised information from labelled data and obtaining pseudo-label information of unlabelled data from the teacher network through consistency constraints. The teacher network updates its parameters using the exponential moving average method to provide stable target predictions for the student network. In the feature extraction stage, a local attention module and a global context module are designed. To further improve the model's performance in maize plant segmentation, a self-training strategy based on curriculum learning is designed. The $mIOU$ is used as a criterion to rank and filter unlabelled data, and then the model performance is gradually optimized through retraining. This strategy screens high-quality unlabelled data through reliability ranking and gradually integrates them into iterative training, enhancing the model's adaptability to complex data distributions and improving its application effect in agricultural scenarios.

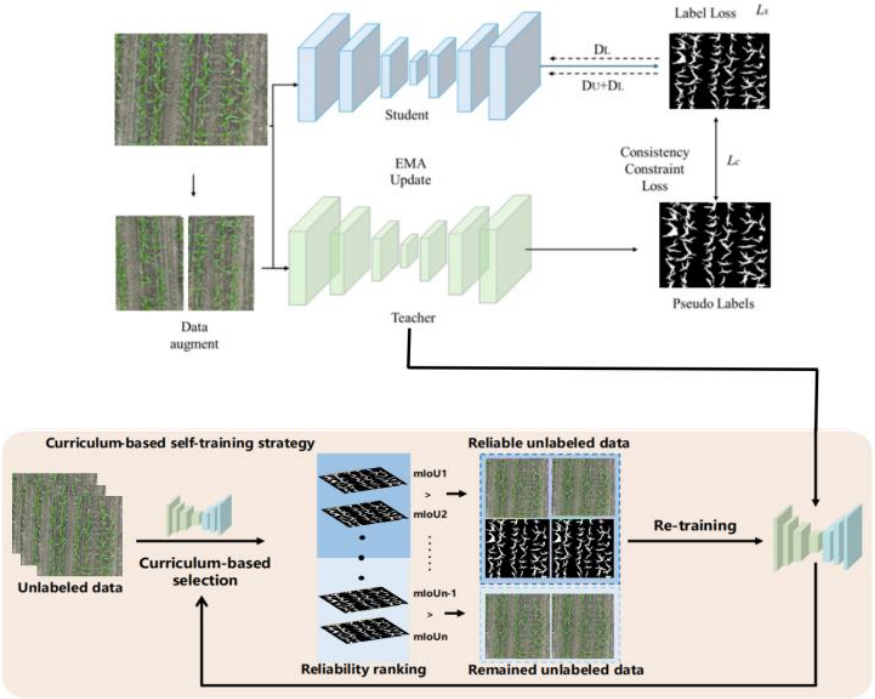

Fig. 3 - Structural diagram of the model

Data Augmentation

Maize field images consist of two main categories: background and maize crops. To increase the diversity of the training data and improve model robustness, various data augmentation techniques were employed. These include random cropping, rotation, and flipping to simulate maize distribution from different

perspectives; adjustments in brightness, contrast, saturation, and hue to mimic varying lighting conditions and weather scenarios; and the addition of Gaussian noise or salt-and-pepper noise to simulate sensor noise and transmission errors.

These augmentation techniques not only enhance the model's generalization ability but also provide a basis for consistency constraints. The core idea of consistency constraints is that the teacher model and the student model should produce similar prediction results under different augmented views.

The calculation formula for the consistency loss is as follows:

$$L_{cons}^{adaptive} = \sum_{i} w_i \cdot d\left(f_{student}(x_i), f_{teacher}(x_i)\right) \quad (3)$$

where: w_i is the adaptive weight based on prediction entropy (the weight is small in high-uncertainty regions and large in low-uncertainty regions), and d is the distance metric function (such as KL divergence or MSE). This loss enhances the model's generalization ability by utilizing unlabelled data through constraining the student network to output consistent prediction results for different augmented versions (e.g., rotated or cropped images) of the same input.

Student and Teacher Models

In the image segmentation task of agricultural maize fields, a Vision Transformer (ViT)-based visual encoder was employed for feature extraction. Different from traditional convolutional neural networks (such as Faster R-CNN and ResNet), ViT processes image information through the Transformer self-attention mechanism, which can effectively capture the long-range dependencies between different parts of the image, thereby extracting deep-level image feature representations. The Vision Transformer (ViT) backbone utilized a patch size of 16×16 pixels. The curriculum learning strategy ranked unlabelled samples based on prediction confidence (negative entropy) generated by the teacher model. A dynamic threshold τ τ was applied to filter pseudo-labels. The unsupervised loss weight β β followed a linear growth schedule, increasing from 0 to its maximum value over the first 50 epochs (warm-up period) to prioritize reliable supervised signals before incorporating complex unlabelled data.

Given an input image $X_v \in \mathbb{R}^{H\times W\times C}$ with a height of H pixels, a length of W pixels, and a channel number of C, the ViT visual encoder extracts image features mainly through image embedding and multi - layer feature extraction steps. The image is processed into N P×P image patches, where the length and width of each image patch are both P pixels. Each image patch is linearly projected using a learnable embedding matrix V to obtain a sequence of patch embedding vectors, and a learnable [CLS] token is prepended to the sequence to capture global image features. Finally, positional encoding is added to the sequence of image embedding vectors to provide the positional information of the image patches, resulting in the input sequence of the image. The calculation process is shown in Eq. 4:

$$z_0 = (v_{cls}; v_p^1 V; v_p^2 V \ldots; v_p^n V) + V_{pos} \quad (4)$$

The input sequence of the image undergoes feature extraction through multiple layers of Transformer encoders. Each layer of the encoder consists of two core components: a multi - head self - attention layer and a multi - layer perceptron. The calculation process is as follows:

$$z'_1 = MSA(LN(z_{l-1})) + z_{l-1}, \quad I = 1, \ldots\ldots, L \quad (5)$$

$$z'_1 = MLP\left(LN\left(z'_l\right)\right) + z'_l, I = 1, \ldots, L \quad (6)$$

ViT can effectively extract deep feature representations of images, which not only improves the accuracy of the model on labelled data but also maintains a high level of consistency and stability on unlabelled data. This method is particularly suitable for the image segmentation task of agricultural maize fields, enabling more accurate identification and semantic segmentation of maize crops and their background areas.

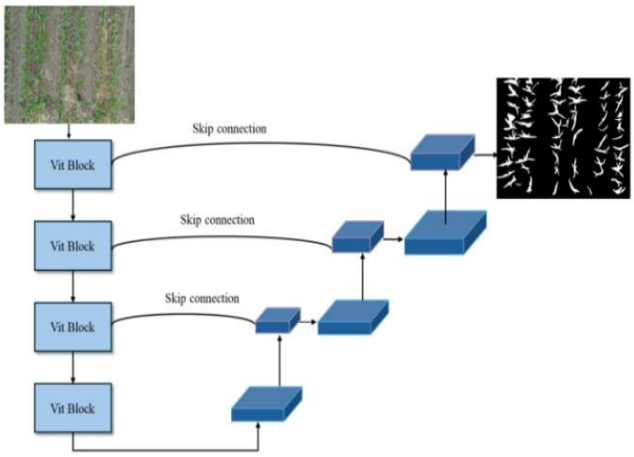

Fig. 4 - Structure diagram of the model

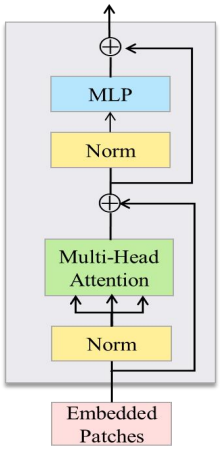

Fig. 5 - Schematic diagram of the Transformer module

Evaluation Indicators

Multiple evaluation metrics were employed, including the mean Intersection over Union (*mIOU*), the mean *F1* score (*mF1*), the mean precision (*m*P*recision*), and the mean recall (*mRecall*). The *mIOU* metric measures the degree of overlap between the predicted results and the ground truth labels, and its calculation formula is:

$$mIOU = \frac{1}{N}\sum_{i-1}^{N}\frac{TP_i}{TP_i + FP_i + FN_i} \tag{7}$$

where N is the total number of categories, and TP_i, FP_i, and FN_i represent the numbers of true positive, false positive, and false negative pixels of the i-th category, respectively.

mPrecision and *mRecall* represent the average values of precision and recall for all categories, respectively.

$$m\,\mathrm{Pr}\,ecision = \frac{1}{N}\sum_{i\text{-}1}^{N}\frac{TP_i}{TP_i + FP_i} \tag{8}$$

$$\mathrm{mRecall} = \frac{1}{N}\sum_{i-1}^{N}\frac{\mathrm{TP}_i + \mathrm{TP}_i}{\mathrm{TP}_i} \tag{9}$$

where *mF1* is the harmonic mean of precision and recall, which is used for comprehensively evaluating the performance of the model.

$$\mathrm{mFl} = \frac{1}{N}\sum_{i-1}^{N}\frac{2\cdot \mathrm{Precision}_i \cdot \mathrm{Recall}_i}{\mathrm{Precision}_i + \mathrm{Recall}_i} \tag{10}$$

RESULTS AND DISCUSSIONS

All experiments were implemented using Python and PyTorch on a workstation equipped with an NVIDIA RTX 3090 GPU. To ensure a fair and rigorous comparison, all baseline methods (Supervised, Mean Teacher, CPS, and UniMatch) were re-implemented using the identical Vision Transformer (ViT) backbone

and shared hyperparameters: input resolution of 512×512, patch size of 16×16, Adam optimizer (learning rate 0.001, weight decay 1e-4), batch size of 8, and 100 training epochs. Each experiment was repeated 3 times to ensure reliability, with average values reported as final results. This consistent protocol eliminates performance discrepancies arising from architectural differences or tuning biases, ensuring that observed improvements are attributable solely to the proposed semi-supervised strategy and curriculum learning mechanism.

Analysis of experimental results

As shown in the figure, the training and validation loss curves of supervised learning and semi-supervised learning under different numbers of training epochs were compared. It can be clearly observed from the figure that under the same number of training epochs, the validation loss of the semi-supervised learning method is significantly lower than that of the pure supervised learning method. Since semi-supervised learning fully utilizes the rich information in the unlabelled data, its validation loss shows a steady downward trend and finally converges to a relatively low level. In contrast, due to the smaller amount of data in supervised learning, the validation loss curve shows an obvious upward trend after about 60 epochs, indicating that the model has overfitted the limited labelled data.

In addition, the training loss and validation loss curves of the semi-supervised learning method exhibit better consistency, with a smaller gap between the two curves, indicating that the model has better generalization ability. This result fully demonstrates that the proposed semi-supervised learning method can effectively alleviate the overfitting problem caused by insufficient labelled data and improve the model's performance on unseen data.

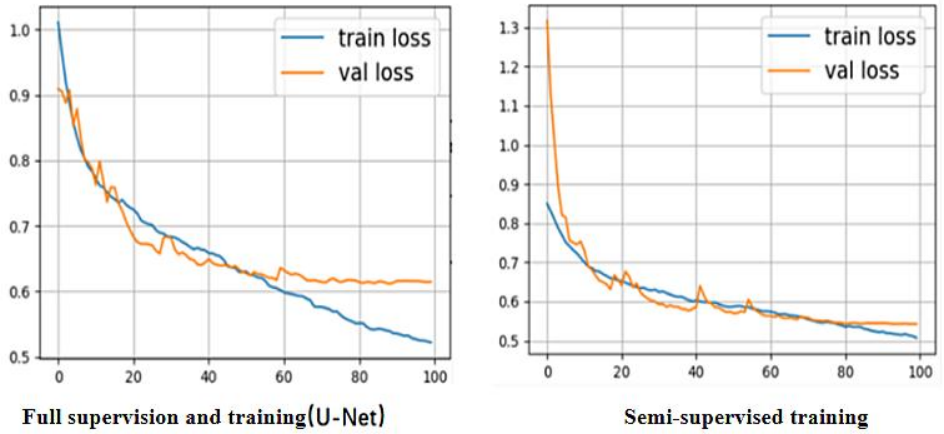

Fig. 6 - Comparison diagram of loss functions

Visualization of Uncertainty

This uncertainty visualization graph compares the performance of different methods in the image segmentation task, including "Our method", supervised learning, Mean Teacher, Cps, and UniMatch. The results show that "Our method" generally exhibits the lowest uncertainty and is closest to the ground truth labels, especially in the detailed parts, indicating its superiority in terms of reliability and accuracy. In contrast, supervised learning and Cps show relatively high uncertainty in certain regions, which may affect their performance in complex scenarios. Mean Teacher maintains good overall consistency but shows a slight deficiency in detail handling. Although UniMatch has a relatively uniform uncertainty distribution, the high uncertainty in specific regions suggests that there is still room for improvement in its precision. "Our method" stands out in these comparisons with its lower uncertainty and higher accuracy, providing a better choice for practical applications.

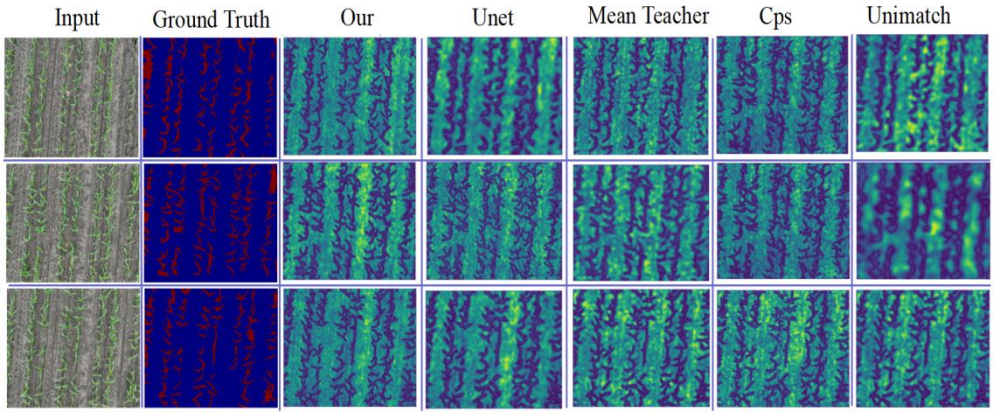

Fig. 7 - Visualization of uncertainty map

Comparative analysis of different methods

The curriculum learning (CL) strategy is integrated to guide the model from simple to complex samples. The ranking of sample difficulty is determined by the prediction confidence scores (negative entropy) generated by the Teacher model. Specifically, pixels with higher confidence are prioritized as 'easy' samples for early-stage training. The learning criterion is defined by a dynamic threshold τ, which filters the pseudo-labels produced by the Teacher model. To control the learning pace, a linear growth schedule is applied to the unsupervised loss weight β: during the first 50 epochs (the warm-up period), β increases from 0 to its maximum value, allowing the model to focus on reliable supervised signals before gradually incorporating more complex unlabelled information.

As shown in Table 2, a comprehensive comparison of test set metrics across different semi-supervised learning methods was conducted. To ensure statistical reliability, all results are reported as Mean ± Standard Deviation (SD) based on three independent experimental runs with distinct random seeds. The data demonstrates that the proposed method achieves superior performance across multiple evaluation metrics, highlighting its significant advantage in leveraging unlabelled data. Compared with the traditional supervised baseline, our method improves the $mIoU$ by 14.3% (from 0.518 to 0.661) and the mF1-score by 5.9% (from 0.715 to 0.774). A paired two-sample t-test confirms that these improvements are statistically significant ($p < 0.05$). Given the limited number of independent runs (n=3), the Standard Deviation (SD) is reported as the primary measure of dispersion, which, combined with the significant p-values, robustly demonstrates the stability and reliability of the proposed method without the potential over-interpretation of wide confidence intervals derived from small sample sizes. Furthermore, when benchmarked against other state-of-the-art semi-supervised strategies—specifically Mean Teacher, CPS, and UniMatch —the proposed method exhibits both higher accuracy and greater stability (indicated by consistently lower SD values).

Further analysis reveals that our method holds a distinct advantage, particularly in the $mIoU$ metric, outperforming the second-best method (UniMatch) by an absolute margin of 2.8% (0.661 − 0.633 = 0.028). This gain indicates an enhanced capability to delineate category boundaries and resolve complex regions (e.g., heavy weed occlusion), thereby significantly reducing mis-segmentation errors. In terms of mPrecision, the proposed method also achieves the peak result of 0.764 ± 0.006, suggesting a markedly lower false-positive rate and superior specificity in identifying true maize seedling regions compared to UniMatch (0.737 ± 0.005) and CPS (0.732 ± 0.006).

From a technical perspective, this superiority stems from the synergistic effect of the Vision Transformer (ViT) encoder and the Local-Global Feature Fusion (LGF) module. Unlike Convolutional Neural Networks (CNNs), which are constrained by limited local receptive fields, the ViT's self-attention mechanism captures long-range spatial dependencies. This enables the model to maintain the structural continuity of maize rows even under severe weed interference. Concurrently, the proposed Curriculum Learning strategy effectively mitigates confirmation bias—a common pitfall in semi-supervised learning—by preventing the model from assimilating low-quality pseudo-labels during the critical early iterations, thus ensuring a more robust convergence.

Table 2

Comparison of indicators for different semi-supervised learning methods

model	mIOU	mF1	mPrecision	mRecall

model	mIOU	mF1	mPrecision	mRecall
Supervised learning	0.518±0.008	0.715±0.006	0.739±0.007	0.736±0.009
Mean Teacher	0.513±0.007	0.690±0.008	0.669±0.010	0.716±0.007
CPS	0.621±0.005	0.750±0.004	0.732±0.006	0.773±0.005
UniMatch	0.633±0.004	0.760±0.003	0.737±0.005	0.788±0.004
Proposed method	**0.661±0.003**	**0.774±0.002**	**0.764±0.006**	**0.795±0.003**

To rigorously verify the effectiveness of the proposed method, a systematic ablation study was conducted based on the Mean Teacher (MT) framework, progressively integrating the improved Vision Transformer (ViT) backbone, Curriculum Self-Training (CST) strategy, and Local-Global Feature Fusion (LGF) module. All results are reported as mean ± standard deviation (SD) over three independent runs, and statistical significance was assessed using a paired two-sample t-test ($p < 0.05$). As shown in Table 3, the baseline MT model achieved an $mIoU$ of 0.520 ± [0.004]. Incorporating the improved ViT backbone increased the $mIOU$ to 0.545 ± [0.003] ($p < 0.05$). This 2.5% improvement, coupled with a reduction in performance variance, indicates that the enhanced self-attention mechanism effectively captures long-range dependencies, providing superior semantic representation for distinguishing maize plants from complex soil backgrounds compared to standard convolutional structures. Further introducing the CST strategy boosted the $mIoU$ to 0.570 ± [0.002] ($p < 0.05$). Notably, the decrease in standard deviation suggests that the curriculum learning mechanism stabilizes the training process by filtering out noisy pseudo-labels in early epochs, thereby enhancing model robustness against label noise. Finally, integrating the LGF module yielded the most substantial gain, elevating the $mIoU$ to 0.661 ± [0.003] ($p < 0.05$), which represents a 9.1% absolute improvement over the CST-only variant. This significant jump demonstrates the critical synergy of the module: by explicitly fusing local texture details with global contextual information, the model successfully resolves boundary ambiguities in overlapping maize canopies—a scenario where previous configurations struggled.

Table 3

Ablation Experiments

MT	ViT	CST	LGF	mIOU	F1	Precision	Recall
✓				0.520 ± 0.004	0.715 ± 0.003	0.739 ± 0.005	0.736 ± 0.004
✓	✓			0.545 ± 0.003	0.725 ± 0.002	0.745 ± 0.003	0.742 ± 0.003
✓	✓	✓		0.570 ± 0.002	0.740 ± 0.003	0.750 ± 0.002	0.748 ± 0.002
✓	✓	✓	✓	0.661 ± 0.003	0.774 ± 0.002	0.764 ± 0.003	0.795 ± 0.002

As shown in Fig. 8, each method exhibits distinct advantages. "Our method" performs well in detail capture and boundary segmentation, and can locate the target area relatively accurately. In contrast, although the supervised learning method can also effectively identify most targets, it may be slightly insufficient in some complex scenarios. Methods such as Mean Teacher and CPS offer good stability and consistency, but may be less accurate than other methods. UniMatch also demonstrates its practicality under specific conditions, but its overall performance is slightly inferior.

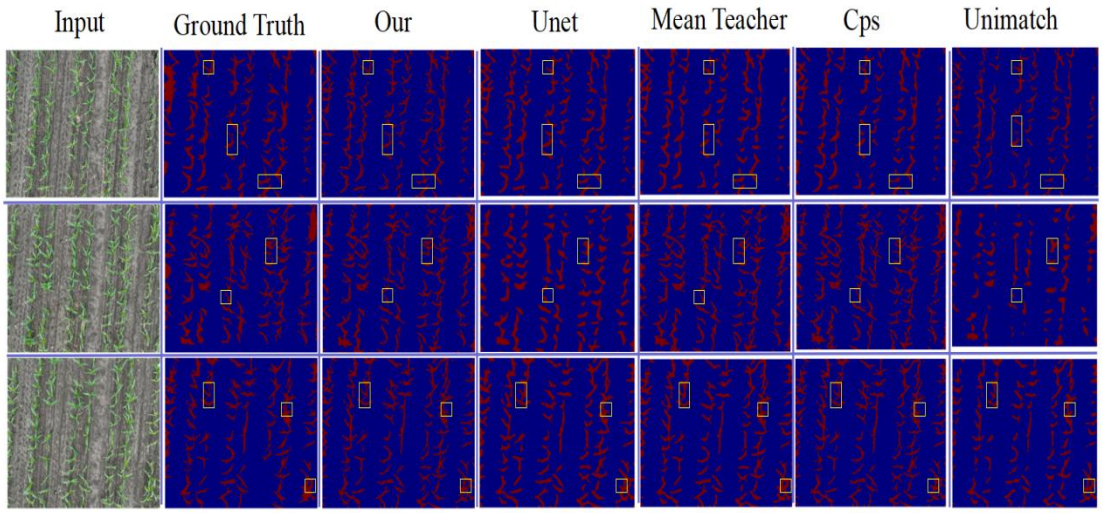

Fig. 8 - Comparison diagram of model performance

CONCLUSIONS

This paper proposes a semi-supervised maize semantic segmentation method based on local-global feature fusion. This method adopts a two-stream architecture that combines a student network and a teacher network, and extracts features through an improved Vision Transformer (ViT) visual encoder. Specifically, a local attention module and a global context module were designed to capture fine-grained features of maize plants and the overall spatial distribution of farmland, respectively. Multi-scale information was then effectively integrated through a feature fusion network. To further improve model performance, a consistency loss was introduced, and a curriculum learning-based self-training strategy (CST) was proposed to enhance the utilisation efficiency of unlabelled data. The experimental results, reported as Mean ± Standard Deviation (SD) based on three independent runs to ensure statistical reliability, show that this method performs excellently on multiple evaluation metrics. The $mIOU$ reaches 0.661±0.005, which is 14.3% higher than that of the traditional U-Net method ($p<0.05$ via T-test), and the mF1 is improved to 0.774, significantly outperforming other semi-supervised methods. These results verify the effectiveness of the proposed method in handling unlabelled data and its adaptability to complex farmland scenarios. From a technical perspective, this superiority stems from the Vision Transformer (ViT) encoder and the Local-Global Feature Fusion (LGF) module. Unlike CNNs that have limited receptive fields, the ViT captures long-range spatial dependencies, enabling the model to maintain the structural continuity of maize rows even under weed interference. The curriculum strategy effectively mitigates the confirmation bias by preventing the model from learning from low-quality pseudo-labels during early iterations.

Ablation experiments confirm the effectiveness and necessity of each component. These findings not only provide a new theoretical basis for the development of precision agriculture technology but also lay the foundation for the practical application of semi-supervised learning in agricultural image analysis. Future research will further explore the feature extraction strategies for different crops and their growth stages to promote the improvement of agricultural production efficiency and sustainability.

ACKNOWLEDGEMENT

Funding Declaration

This research was supported by the National Natural Science Foundation of China (32301716), National Natural Science Foundation of China (32302410), Ningxia key research and development plan (2023BCF01051), National Key R&D Program of China (2023YFD2000200). We would also like to thank the anonymous reviewers who provided helpful comments for the manuscript improvement.

Data Availability Statement

The datasets used and/or analysed during the current study are available from the corresponding author on reasonable request.

Author Contributions

Zhicheng Tang; Writing – original draft, Visualization, Validation, Software, Investigation, Formal analysis, Data curation, Conceptualization. Yuxin Zhu; Supervision, Methodology, Concept–aviation. Weiyi Feng; Writing – review & editing, Supervision, Project administration, Funding acquisition. Junke Zhu; Writing – review & editing.

Conflicts of 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.

REFERENCES

[1] Bai, N. (2024). *Research on seedling and weed recognition method in maize seedling field based on SRDS_Unet* (in Chinese) [Master's thesis, Heilongjiang Bayi Agricultural University].

[2] Bai, Y. (2022). *Research on remote sensing monitoring of maize seedling condition based on UAV multi-source images* (in Chinese) [Master's thesis, Chinese Academy of Agricultural Sciences].

[3] Bernklau, E.J., Hibbard, B.E., Dick, D.L., Rithner, C.D., Bjostad, L.B. (2015). Monogalactosyldiacylglycerols as host recognition cues for western corn rootworm larvae (Coleoptera: Chrysomelidae). *Journal of Economic Entomology*, Vol. 108, pp. 539-548, United Kingdom.

[4] Chen, Y., Wu, Z., Zhao, B., Fan, C., Shi, S. (2020). Weed and corn seedling detection in field based on multi feature fusion and support vector machine. *Sensors*, Vol. 21, pp. 212, Switzerland.

[5] Diao, Z., Yan, J., He, Z., Zhao, S., Guo, P. (2022). corn seedling recognition algorithm based on hyperspectral image and lightweight-3D-CNN. *Computers and Electronics in Agriculture*, Vol. 201, pp107343, Netherlands.

[6] Fan, X.X. (2025). *Research on maize seedling and weed recognition model based on improved YOLOv7 (基于改进 YOLOv7 的玉米幼苗与杂草识别模型研究)* [Master's thesis, Heilongjiang Bayi Agricultural University].

[7] Guo, X., Ge, Y., Liu, F., Yang, J. (2023). Identification of maize and wheat seedlings and weeds based on deep learning. *Frontiers in Earth Science*, Vol. 11, pp. 1146558, Switzerland.

[8] Guo, X. Q. (2024). *Research and application of recognition model for maize and wheat seedlings and field weeds based on deep learning (基于深度学习的玉米小麦幼苗及田间杂草识别模型研究与应用)* [Master's thesis, Hebei North University].

[9] Hu, M. C. (2024). Experiment and research on root-zone fertilization robot based on stem recognition (基于茎秆识别的根区施肥机器人实验与研究) [Master's thesis, Tianjin University of Technology].

[10] Hu, W.Z., Wang, B.J., Geng, L.J., Lan, Y.B., Li, W.H., Li, D.S. (2023). Maize seedling detection based on Cascade R-CNN (基于 Cascade R-CNN 的玉米幼苗检测). *Journal of Agricultural Mechanization Research*, Vol. 45, pp. 26-31, Heilongjiang/China.

[11] Huang, S.K. (2023). *Research on image recognition of maize growth period based on UAV (基于无人机的玉米生育期图像识别研究)* [Master's thesis, Jilin Agricultural University].

[12] Ji, W.L., Liu, Z., Xing, H.H. (2024). Lightweight method for farmland weed recognition based on YOLO v5 (基于 YOLO v5 的农田杂草轻量化识别方法). *Transactions of the Chinese Society of Agricultural Machinery*, Vol. 55, pp. 212-222, Beijing/China.

[13] Jiang, T.T. (2025). *Research on maize seedling monitoring based on RGB images and deep learning algorithms (基于 RGB 图像和深度学习算法的玉米幼苗监测研究)* [Master's thesis, Anhui University of Science and Technology].

[14] Li, D. S. (2023). *Research and application of maize seedling and weed detection algorithm based on deep learning (基于深度学习的玉米幼苗与杂草检测算法研究与应用)* [Master's thesis, Shandong University of Technology].

[15] Li, X.F., Shi, C.H., Song, M.G., Gan, Z.H., Qiao, Z.R., Meng, Q.K. (2023). Maize seedling and weed recognition and detection based on Yolov4 model (基于 Yolov4 模型的玉米幼苗与杂草识别检测). *Tropical Agricultural Engineering*, Vol. 47, pp. 1-6, Hainan/China.

[16] Lin, Y. T. (2024). *Research on weed segmentation model in maize seedling stage based on deep learning (基于深度学习的玉米苗期杂草分割模型研究)* [Master's thesis, Shenyang Agricultural University].

[17] Liu, B. J., Zhou, Y. N., Zhou, X. H., Ding, L., Li, H., Wang, W. Z. (2024). Research on weed recognition model in maize field based on deep learning (基于深度学习的玉米田杂草识别模型研究). *Journal of Henan Agricultural University*, Vol. 58, pp. 279-286, Henan/China.

[18] Liu, S., Yin, D., Feng, H., Li, Z., Xu, X., Shi, L., Jin, X. (2022). Estimating maize seedling number with UAV RGB images and advanced image processing methods. *Precision Agriculture*, Vol. 23, pp. 1604-1632, Netherlands.

[19] Liu, W. (2025). *Research on identification and localization of maize seedlings and weeds in field based on YOLOv8n (基于 YOLOv8n 的田间玉米幼苗与杂草识别定位研究)* [Master's thesis, Inner Mongolia Agricultural University].

[20] Liu Z. (2023). *Research on weed identification methods in farmland (in Chinese)* [Master's thesis, Xi'an University of Science and Technology].

[21] Lyu, M. Y. (2025). *Research on field maize weed detection method based on visual attention mechanism (基于视觉注意力机制的田间玉米杂草检测方法研究)* [Master's thesis, Jilin University].

[22] Ma, Z. X. (2023). *Research and application of main weed detection in maize seedling field based on YOLOX (基于 YOLOX 的玉米苗期主要杂草检测研究与应用)* [Master's thesis, Shandong Agricultural University].

[23] Tang, B., Zhou, J., Pan, Y., Qu, X., Cui, Y., Liu, C., Gu, X. (2025). Recognition of maize seedling under weed disturbance using improved YOLOv5 algorithm. *Measurement*, Vol. 242, pp. 115938, Netherlands.

[24] Tang, B., Zhou, J., Zhao, C., Pan, Y., Lu, Y., Liu, C., Gu, X. (2025). Using UAV-based multispectral images and CGS-YOLO algorithm to distinguish maize seeding from weed. *Artificial Intelligence in Agriculture*, Vol. 15, pp. 162-181, Netherlands.

[25] Wang, B. J. (2022). *Research on maize seedling and weed identification based on deep learning (基于深度学习的玉米幼苗与杂草识别研究)* [Master's thesis, Shandong University of Technology].

[26] Wang, S. W. (2024). *Research on maize seedling counting method based on UAV optical images (基于无人机光学图像的玉米幼苗计数方法研究)* [Master's thesis, Shandong Agricultural University].

[27] Wang, Y. F., Chao, Q., Li, Z., Lu, T. C., Zheng, H. Y., Zhao, C. F., Wang, B. C. (2019). Large-scale identification and time-course quantification of ubiquitylation events during maize seedling de-etiolation. *Genomics, Proteomics & Bioinformatics*, Vol. 17, pp. 603-622, China.

[28] Wu, Y., Yuan, S., Tang, Y., Tang, L. (2025). Application of real-time detection transformer based on convolutional block attention module and grouped convolution in maize seedling. *Frontiers in Plant Science*, Vol. 16, pp. 1672746, Switzerland.

[29] Xu, S. (2024). *Research on maize seedling and weed recognition based on convolutional neural network (基于卷积神经网络的玉米幼苗与杂草识别研究)* [Master's thesis, Jiangxi Agricultural University].

[30] Yang, X. Y., Li, P. J. (2023). *Extraction of maize seedling distribution information from UAV images using spectral features, morphological features and Hough transform (利用光谱特征、形态特征和 Hough 变换从无人机图像中提取玉米幼苗分布信息). Acta Scientiarum Naturalium Universitatis Pekinensis*, Vol. 59, pp. 843-853, Beijing/China.

[31] Zhang, T. Y. (2024). *Research on maize seedling detection based on optimized YOLOv5s algorithm (基于优化 YOLOv5s 算法的玉米幼苗检测研究)* [Master's thesis, Heilongjiang Bayi Agricultural University].

[32] Zhang, X. L. (2022). *Research on evaluation method of maize emergence quality based on UAV images (基于无人机图像的玉米出苗质量评价方法研究)* [Master's thesis, Shihezi University].

[33] Zhao, Y. P. (2021). *Research on weed recognition in corn fields based on convolutional neural network (基于卷积神经网络的玉米田杂草识别研究)* [Master's thesis, Shanxi Agricultural University].

[34] Zhi, Z., Li, Y., Liu, G., Ou, Q. (2025). Identification and detection of label-free polystyrene microplastics in maize seedlings by Raman spectroscopy. *Science of The Total Environment*, Vol. 958, pp. 178093, Netherlands.