DESIGN OF POTATO SEGMENTATION ALGORITHM FOR STICKY SOIL ENVIRONMENTS
/
面向粘性土壤下马铃薯分割算法设计

Ranbing YANG[1,2], Yihui MIAO[1], Zhiguo PAN*[1], Huan ZHANG[1], Xinlin LI[1], Yue SHI[1], Xuan LUO[1], Hongzhu WU[3], Shuai WANG[1], Tao JIN[1]

[1]College of Mechanical and Electrical Engineering, Qingdao Agricultural University Qingdao / China
[2]College of Mechanical and Electrical Engineering, Hainan University, Haikou / China
[3]Qingdao Hongzhu Agricultural Machinery Co., Ltd., Qingdao / China
Tel: +86 15318715305; E-mail: peter_panzg@163.com
Corresponding author: Zhiguo PAN
DOI: https://doi.org/10.35633/inmateh-78-48

Keywords: *Image Segmentation, Deep Learning, Potato, MTFormer*

ABSTRACT
To address edge blurring, soil-clod interference, and limited feature representation in potato image segmentation under sticky-soil conditions, this study proposes an improved MTFormer segmentation model. The model combines the complementary strengths of convolutional neural networks and Transformers, and further enhances segmentation performance through a multi-stage optimization strategy. For more informative and robust feature learning, a residual CNN-based extractor is introduced to strengthen multi-level feature representations. In addition, an MS-CAM attention mechanism is used to reduce channel redundancy, which helps mitigate adhesion-related target confusion in challenging scenes. Building on these features, the TBFE module promotes cross-channel feature fusion, while a Fourier-based FFCM structure compresses and reconstructs deep features in the frequency domain to improve feature compactness. Experiments on our self-built dataset show that MTFormer achieves an F1 score of 85.19%, an mIoU of 84.62%, and a pixel accuracy of 95.67%. Compared with the baseline model, U-Net, and DeepLabV3+, pixel accuracy increases by 1.75, 0.35, and 1.67 percentage points, respectively. Overall, the proposed approach improves segmentation reliability by strengthening feature representation while limiting unnecessary computation, providing practical support for accurate potato segmentation in sticky-soil environments.

摘要
为解决黏性土壤环境下马铃薯图像分割中存在的边缘模糊、土块干扰及特征表达不足等问题，本文提出一种改进的 MTFormer 分割模型。该模型融合了卷积神经网络与 Transformer 的优势，通过多阶段优化策略提升分割性能，利用残差 CNN 提取器增强特征层次表达，并嵌入 MS-CAM 注意力机制抑制通道冗余，以解决目标粘连与混淆问题，引入 TBFE 模块跨通道特征融合，采用基于傅里叶变换的 FFCM 结构对深层特征进行频域压缩与重构，显著提高了特征的紧凑性。试验表明，MTFormer 在自建数据集上的 F1 分数达到 85.19% ，mIoU 达到 84.62%，像素准确率为 95.67%。相较于原始模型、U-Net 和 DeepLabV3+，像素准确率分别提升 1.75、0.35 和 1.67 个百分点。该方法有效平衡了特征表达与计算冗余，可为黏性土壤环境下的马铃薯精准分割提供可靠技术支撑。

INTRODUCTION
Potatoes rank among the world's most important staple crops, characterized by high yields and strong adaptability, securing a stable position in food security and agricultural supply systems (*Devaux et al., 2020*). Under large-scale production conditions, potatoes are typically harvested using segmented mechanization. After excavation, tubers scatter across the field surface. Their quantity, size, and physical condition (e.g., soil coverage, damage) directly inform yield assessment, sorting operations, and quality control (*Lyu et al., 2015*). Visual segmentation at the harvest site is not merely an image processing task but a critical front-end component in the yield measurement decision chain: segmentation results are used for counting, area/dimension quantification, and subsequent yield estimation. Boundary deviations, misclassifications, and clumping amplify during statistical and regression stages, compromising yield estimation stability. Concurrently, industrial deployment imposes constraints on model parameter scale and inference speed, requiring algorithms to achieve an engineering-viable balance between accuracy and computational overhead. Harvest site imagery under sticky soil conditions exhibits pronounced unstructured characteristics. Potato tubers and soil clumps exhibit similar appearances with weak local boundaries; occlusions, adhesions, and scale

variations are prevalent; shadows and reflections introduce additional noise. Such scenarios impose two core requirements on segmentation models: sufficient local detail representation to reliably recover contours and separate adjacent instances, coupled with global context modeling and noise suppression capabilities to prevent background textures from dominating feature responses and causing misclassifications.

With the evolution of computer vision technology, crop detection has shifted from traditional machine learning to a deep learning paradigm *(ElMasry et al., 2012; Razmjooy et al., 2012; Oppenheim et al., 2019)*. In potato object detection research, significant progress has been achieved. *Ma et al. (2016),* and *Xi et al. (2020),* achieved high-precision detection of lesions and bud eyes using *SSD (Liu et al., 2016)* and Faster R-CNN *(Ren et al., 2016)*, respectively; *Zhang et al., (2022), Geng et al. (2024)*, and *Liao et al*., *(2025)*. achieved breakthroughs in lightweight and real-time performance by improving YOLO *(Redmon et al., 2016)* series models. Building upon this foundation, Mask *R-CNN (He et al., 2017)* extends the segmentation framework to instance-level mask prediction, enabling precise contour modeling. However, segmentation boxes fail to meet the demands of detailed morphological analysis, making segmentation networks like DeepLab *(Chen et al., 2017)* and U-Net *(Ronneberger et al., 2015)* the mainstream choice. While *CNNs* excel in segmentation tasks, their convolutional operations' “local receptive fields” struggle to capture global contextual information. Particularly in sticky soil environments, potato and soil texture features overlap significantly. *CNNs*, unable to establish long-range pixel dependencies, often produce blurred segmentation edges or even miss segmentations. In contrast, self-attention-based Transformers *(Dosovitskiy et al., 2020)* (e.g., SegFormer *(Xie et al., 2020)* possess robust global modeling capabilities, better handling background interference. However, lacking inductive bias, *SegFormer* often overlooks minute textures and edge details in sticky soil backgrounds during shallow feature extraction. To address this, this study proposes an enhanced *MTFormer* model. This model integrates a *CNN* residual structure to strengthen shallow geometric features, and combines the *MS-CAM (Chen et al., 2025)* attention mechanism, a dual-branch feature extraction module (*TBFE*) *(Dai et al., 2021)*, and a Fourier convolution mixer (*FFCM*)*(Gao et al., 2024)* to solve the challenge of precisely separating potatoes in a sticky soil environment. This paper conducts comparative experiments and ablation studies on a self-built potato harvesting field dataset. It not only validates the segmentation accuracy of the model but also evaluates its engineering deployment feasibility from the perspectives of parameter scale, floating-point operations, and inference speed. This provides a viable segmentation solution and technical reference for agricultural machinery yield estimation systems.

MATERIALS AND METHODS

Image acquisition and dataset creation

Image acquisition

Because no publicly available dataset contains potato images captured in sticky-soil environments, this study constructed a dedicated dataset to support segmentation under diverse and challenging field conditions. Four farm-grown varieties—Fertile Soil No. 1, McKen, Snow Valley Red, and Lucinda— were selected. Data were collected at multiple sites in Inner Mongolia (Malianqu Township, Jining District, Ulanqab City; Meiguiying Town, Qianqi Banner, Chahar Right Front League, Ulanqab City; Xiao Sanjing Village, Qianqi Banner, Chahar Right Front Banner, Ulanqab City; and Gaojiadi Village, Qianqi Banner, Chahar Right Front Banner, Ulanqab City) and in Hebei Province (Zhangbei Town, Zhangbei County, Zhangjiakou City; and Dahe Town, Zhangbei County, Zhangjiakou City). Image acquisition was conducted from October 1 to 7, 2024, using multiple sensing devices (e.g., smartphones and industrial cameras), and was synchronized with the harvesting process to faithfully capture real field complexity. As shown in Figure 1, the dataset includes images recorded at different harvesting stages across the four varieties.

Fig. 1 – Image Acquisition Environment

Dataset Construction

To ensure consistent inputs across different network models, all images were resized to a unified resolution, which reduces computational cost and speeds up training. Potatoes in each image were annotated using the LabelMe tool, and, as illustrated in Figure 2, instance-level segmentation masks were generated for tubers in sticky-soil scenes (highlighted in green).

As illustrated in Figure 2, the dataset collected in this study covers a wide range of field conditions, including substantial variation in background soils (e.g., dry, light-colored clay and wet, dark-colored clay) and non-uniform illumination across different times of day, which introduces pronounced shadows and specular reflections in some images. To capture meaningful variation in target appearance, tubers with different sizes, shapes, skin colors, and degrees of soil coverage were recorded from multiple viewpoints, and many images contain notable occlusion and overlap. Polygon-based contour labeling was adopted to provide accurate object boundaries, supporting both segmentation and segmentation tasks in clay-soil scenes.

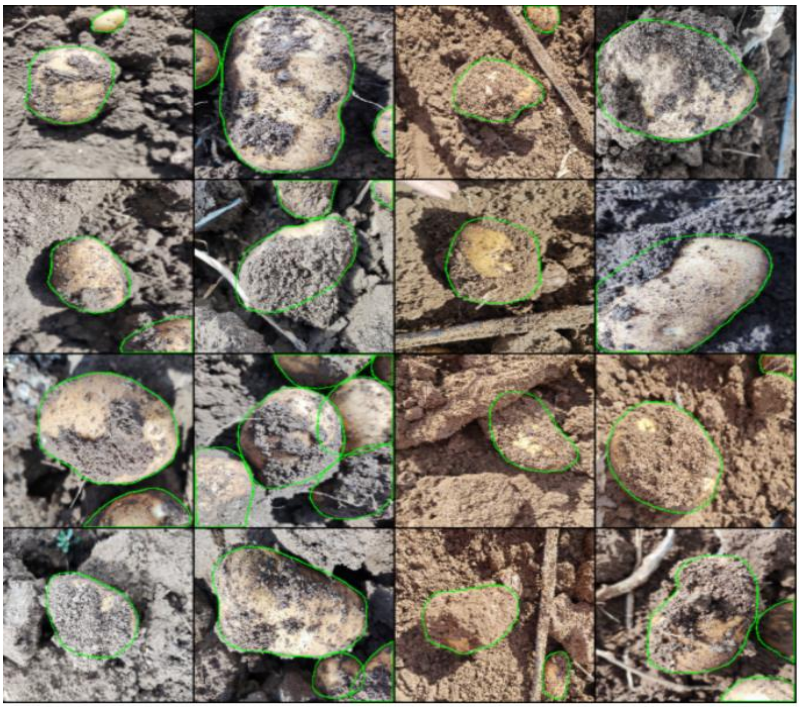

Fig. 2 –Potato samples in clay soil

Image Preprocessing

Potatoes in complex field scenes were annotated using the open-source *LabelMe* tool, and all labels were saved in a standardized *JSON* format. To improve model generalization and robustness while mitigating overfitting, several data augmentation techniques were applied, including horizontal and vertical flipping, brightness adjustment, motion blur, and contrast enhancement. After augmentation, the dataset contained 1,245 images, which were randomly shuffled and split into a training set of 996 images and a test set of 249 images using an 8:2 ratio.

Improved method for potato segmentation in sticky soil

To address segmentation in cohesive-soil scenes—where tubers are visually similar to soil clods, texture cues are weak, and targets are frequently partially buried—this study improves SegFormer by strengthening multi-scale feature representation and global context modeling. The proposed design begins with a residual *CNN* front-end to enhance low-level discriminative features, and integrates an MS-CAM attention mechanism to reduce channel redundancy and alleviate feature ambiguity under heavy background noise. Next, the *TBFE* module promotes richer cross-channel interactions, compensating for the limited ability of standard 2D convolutions to capture complex inter-channel relationships. Finally, a Fourier-based *FFCM* structure performs frequency-domain reconstruction to obtain more compact and informative deep features. Overall, the enhanced network retains the Transformer's global modeling benefits while improving semantic representation and inference robustness in complex field environments. The resulting model, termed *MTFormer*, is shown in Figure 3.

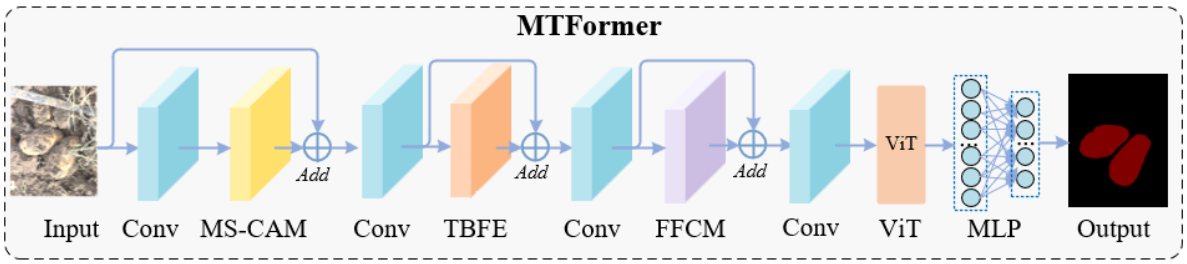

Fig. 3 – The structure of MTFormer Model

Residual structure-based CNN feature extractor

Although *SegFormer* employs *Transformer* encoders to model long-range semantic dependencies, it does not explicitly encode local inductive priors, which limits its ability to capture local spatial structures—such as object boundaries and fine textures—particularly in sticky-soil scenes. To strengthen early-stage local feature perception and improve training stability, this study adds a residual *CNN* feature extractor at the front end of the encoder, as illustrated in Figure 4. Following the residual learning concept proposed by He et al., the module uses an identity shortcut to encourage feature reuse and facilitate gradient flow. Concretely, the input is processed along two paths: the main branch applies convolutional filtering to extract local texture cues and uses batch normalization (BN) to regulate feature distributions, while the shortcut branch directly forwards the input as an identity mapping. The two branches are then fused by element-wise addition. This design preserves low-level geometric information while providing a smoother backpropagation path, improving the stability and convergence efficiency of early feature extraction.

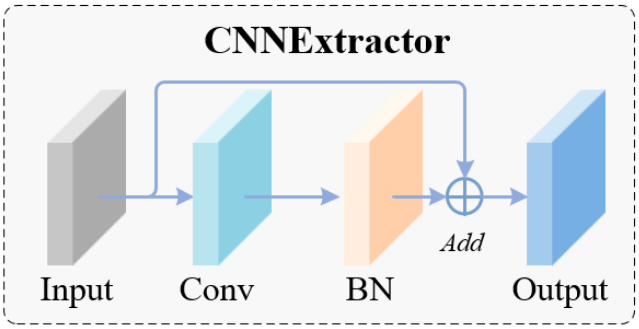

Fig. 4 – Block Diagram of Residual-Based CNN Feature Extractor Architecture

MS-CAM

In sticky-soil scenes, potato segmentation is challenged by large scale variation, severe occlusion, and the extremely small pixel footprint of many targets. Conventional channel-attention mechanisms typically model channel relationships using only global average pooling (*GAP*), which can be dominated by large background regions and consequently weaken the response to small or partially occluded tubers. To address this limitation, a multi-branch channel attention module (*MS-CAM*; Figure 5) is introduced in the encoder to enhance feature discrimination through joint global–local channel modeling. Given an input feature map ($X \in \mathbb{R}^{\{C\times H\times W}$, *MS-CAM* performs channel modeling through two complementary branches. The local branch *avoids* the downsampling effect of pooling and instead constructs a channel bottleneck directly from the original feature map using pointwise convolutions, as defined by:

$$L(X) = \mathcal{B}(PW_2(\delta(\mathcal{B}(PW_1(X))))) \quad (1)$$

This branch preserves the feature map's spatial resolution ($H \times W$) while refining channel relationships, thereby avoiding extra spatial information loss and helping retain discriminative cues at object boundaries and fine-detail regions. In contrast, the global branch first applies *GAP* to aggregate (X) over the spatial dimensions, producing a global channel-statistics vector. It then captures long-range inter-channel dependencies using a pointwise-convolution bottleneck similar to the local branch, yielding a scene-level channel prior ($G(X)$). For fusion, the local output ($L(X)$) and the global output ($G(X)$) are combined through channel-wise broadcast addition, and a Sigmoid function produces channel attention weights to adaptively re-weight the input features. This operation can be expressed as:

$$\mathrm{X}' = \mathrm{X} \otimes \sigma(\mathrm{L}(\mathrm{X}) \oplus \mathrm{G}(\mathrm{X})) \quad (2)$$

Through the complementary modeling of local inter-channel interactions ($L(X)$) and the global semantic prior ($G(X)$), the model simultaneously captures fine-grained geometric details and scene-level target cues. This joint attention suppresses redundant channel activations that closely resemble background textures in sticky-soil scenes, thereby improving segmentation robustness for small targets and partially buried potatoes. Where, $PW_1(\cdot)$ and $\mathrm{PW}_2(\cdot)$ denote 1×1 pointwise convolutions used for channel reduction and channel expansion, respectively, to capture local inter-channel interactions; $\mathcal{B}(\cdot)$ represents Batch Normalization for stabilizing feature distributions; and $\delta(\cdot)$ denotes the $ReLU$ activation function, which introduces nonlinearity into the mapping.$\oplus$ denotes Broadcasting Addition, $\otimes$ denotes Element-wise Multiplication, $\sigma(\cdot)$ denotes the $Sigmoid$ activation function, which normalizes channel weights to the range [0,1].

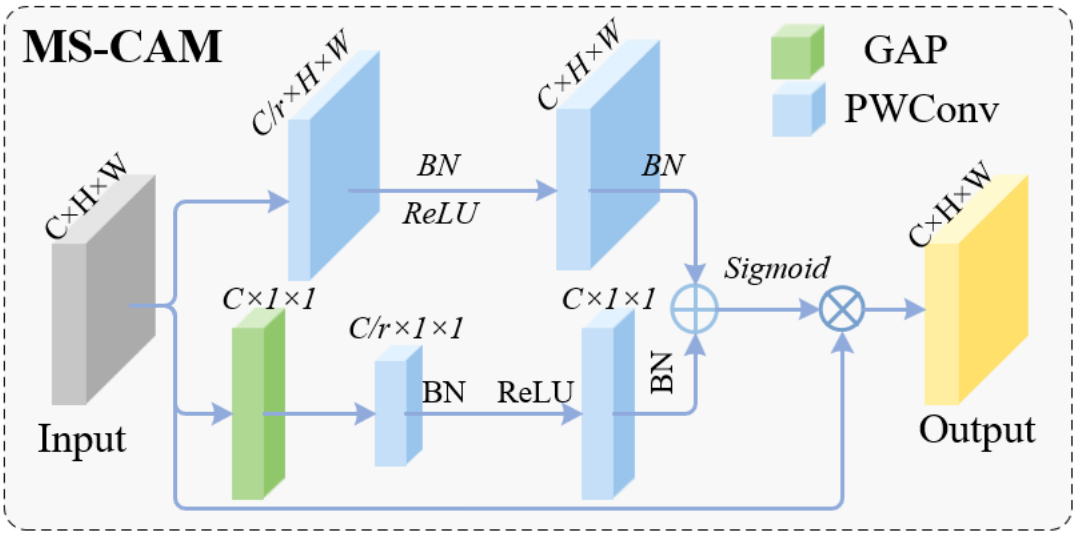

Fig. 5 –MS-CAM Structural Diagrams

TBFE module

In sticky-soil conditions, potato surfaces often closely resemble surrounding soil clods in both texture and color, requiring strong local modeling to delineate boundaries accurately. To overcome the limited decoupling and channel modeling capacity of a single convolution operator, this study introduces a Two-Branch Feature Extraction ($TBFE$) module (Figure 6). $TBFE$ jointly models inter-channel dependencies and spatial geometry by combining a 3D-convolution branch with a 2D-convolution branch operating in parallel. After an initial (1×1) pointwise convolution for channel compression and information integration, the input features are transformed into an intermediate representation: $F \in \mathbb{R}^{C' \times H \times W}$In the 3D-convolution branch, the channel dimension is reshaped into the depth dimension of a volumetric representation, allowing channel-wise filtering with a $(3 \times 1 \times 1)$ (depth $(\times)$ height $(\times)$ width) kernel. Unlike standard (1×1) convolutions, which perform dense mixing across all channels, this design emphasizes nonlinear correlations within local channel groups. As a result, it reduces parameter cost while strengthening the modeling of localized inter-channel dependencies. The computation is given by:

$$\mathrm{F}_{3\mathrm{d}} = \Phi(\mathrm{K}_{3\mathrm{d}} * \mathrm{F} + \mathrm{b}_{3\mathrm{d}}) \tag{3}$$

In the 2D-convolution branch, a (3×3) kernel is applied to perform spatial filtering on the feature map (F), capturing local texture cues and boundary information. Zero padding is used to preserve the spatial resolution so that the output retains the same height and width as the input. This operation can be written as:

$$\mathrm{F}_{2\mathrm{d}} = \Phi(\mathrm{F} * \mathrm{w}_{2\mathrm{d}} + \mathrm{b}_{2\mathrm{d}}) \tag{4}$$

Finally, the outputs of the two branches are concatenated along the channel dimension, fusing channel-dependent cues with spatial structural information to form an enhanced feature representation:

$$\mathrm{M}_{\mathrm{out}} = \mathrm{Concat}(\mathrm{F}_{3\mathrm{d}}, \mathrm{F}_{2\mathrm{d}}) \tag{5}$$

Where, (C') denotes the number of channels after compression. In the 3D branch, $(*)$ denotes the 3D convolution operation, where $(\mathrm{K}_{3\mathrm{d}})$ is a 3D convolution kernel with spatial size (1×1) that acts primarily along the channel (depth) dimension; $(\mathrm{b}_{3\mathrm{d}})$ is the corresponding bias term; and $(\Phi(\cdot))$ is a nonlinear activation function (e.g., *ReLU*). In the 2D branch, () denotes the 2D convolution operation, where $(\mathrm{w}_{2\mathrm{d}})$ and $(\mathrm{b}_{2\mathrm{d}})$ are the 2D convolution weights and bias, respectively.

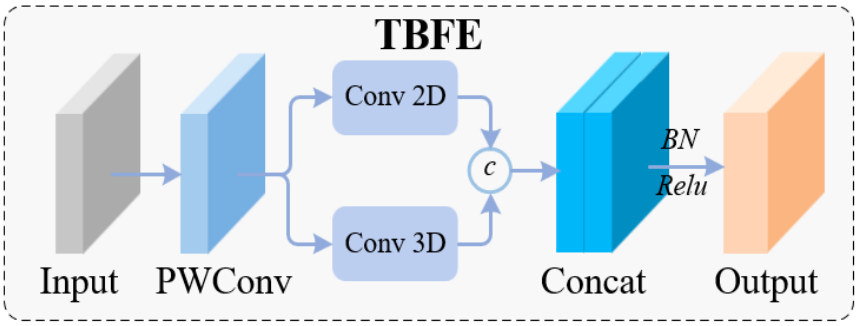

Fig. 6 –TBFE Module Structure Diagrame

FFCM module

To mitigate unstructured texture interference and boundary ambiguity in potato tuber segmentation under cohesive-soil conditions, this study introduces a Fourier Feature Coupling Module (*FFCM*) to strengthen feature discrimination in complex backgrounds. *FFCM* adopts a cascaded spatial–frequency hybrid design: it first refines local spatial cues and then leverages frequency-domain global information to suppress components that are strongly correlated with background textures. In the spatial modeling stage, the input features are first projected through pointwise convolution (*PConv*) to facilitate inter-channel mixing. The resulting features are then processed by two parallel depthwise convolution branches (*DWConv*) with 3×3 and 5×5 kernels to extract local patterns under different receptive fields. Each branch applies the *GeLU* activation to enhance nonlinear representation, and the branch outputs are concatenated along the channel dimension to form a multi-scale spatial feature representation. In the frequency-domain modulation stage, the fused spatial features are transformed via the Fast Fourier Transform (*FFT*) into spectral features with real and imaginary components. These components are concatenated along the channel dimension and then modulated through channel-wise interactions implemented by pointwise convolution, followed by normalization and nonlinear activation to stabilize training and amplify informative spectral responses. The refined spectral features are subsequently mapped back to the spatial domain using the inverse *FFT* (*IFFT*), producing features enriched with global context. Finally, *FFCM* fuses the frequency-enhanced tuber features with the original spatial features using a residual connection via element-wise addition, and a pointwise convolution performs channel compression and feature reconstruction. By jointly exploiting local texture cues and global frequency-domain priors, this spatial–frequency coupling strategy improves robustness to background interference and enhances the separability of potato tuber features in cohesive-soil scenes.

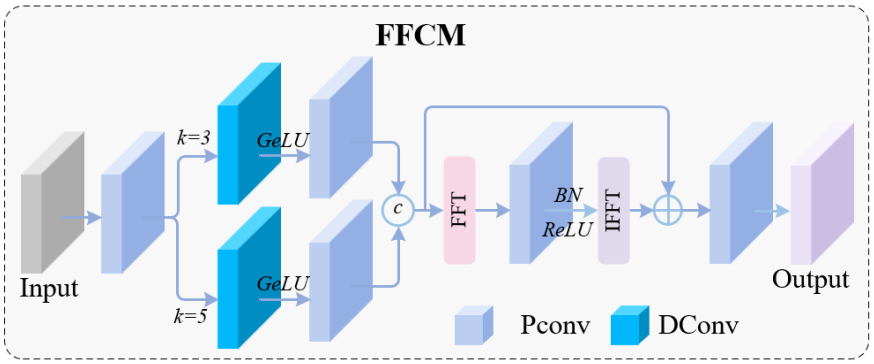

Fig. 7 –FFCM Module Structure Diagram

Segmentation-Based Potato Counting Method

To translate the pixel-level segmentation capabilities of the MTFormer model into an intuitive assessment metric for potato yield, this study constructs a simplified potato counting framework based on semantic segmentation networks. The specific workflow is as follows. First, the binary segmentation mask obtained through MTFormer model inference is used as input. Considering potential minor clumping or segmentation noise in sticky soil environments, morphological opening operations are applied to smooth the segmentation results. This eliminates interference from isolated soil clumps that are too small and breaks minor edge adhesions. Next, a connected component labeling algorithm is applied to the processed binary image to extract connected regions.

By setting a pixel area threshold, non-target interference items smaller than this threshold are filtered out. The total number of independent connected regions retained is counted, representing the model-predicted potato count. This count value is directly used for subsequent yield estimation and manual counting comparison.

RESULTS AND DISCUSSION

Test environment and parameter configuration

All experiments and model training were performed on the same workstation running Windows 11, equipped with a 16-core AMD R9-7940HX CPU, 32 GB RAM, and an NVIDIA GeForce RTX 4070 GPU with 8 GB VRAM. The implementation was based on Python 3.10, using PyTorch 2.0.1 with CUDA 12.1 and cuDNN 8.9, while all other settings followed the official defaults.

Evaluation indicators

To comprehensively and consistently evaluate the proposed method, Precision, Recall, F1 score, mean Pixel Accuracy (mPA), and mean Intersection over Union (mIoU) were adopted as quantitative metrics for potato tuber segmentation performance. The definitions of these metrics are given as follows:

$$P = \frac{TP}{TP+FP} \tag{6}$$

$$\mathrm{R} = \frac{\mathrm{TP}}{\mathrm{TP+FN}} \tag{7}$$

$$AP_i = \int_0^1 P\,(R)\,dR \quad or \quad AP_i = \sum_{k=1}^{n} (R_k - R_{k-1})P_k \tag{8}$$

$$mAP = \frac{1}{C}\sum_{i=1}^{C} AP_i = \frac{1}{C}\sum_{i=1}^{C}\int_0^1 P_i\,(R)\,dR \tag{9}$$

$$F1 = 2 \times \frac{P \times R}{P + R} \tag{10}$$

$$mIoU = \frac{1}{C}\sum_{i=1}^{N} \frac{TP_i}{TP_i + FP_i + FN_i} \tag{11}$$

where:

TP, FP, and FN denote the numbers of pixel-wise true positives, false positives, and false negatives, respectively. For mAP, AP_i represents the Average Precision for class i, calculated as the area under the Precision-Recall curve $(P(R))$ obtained by varying the confidence threshold. C is the number of classes (for binary tuber–background segmentation, $C = 2$). mIoU measures the overlap between predicted masks and ground-truth annotations averaged over all classes.

Analysis of attention mechanism effectiveness

Due to the complex nature of clay soil environments, segmentation accuracy may decline under environmental influences. To mitigate the impact of complex backgrounds on model segmentation performance, this study incorporates various attention mechanisms into the backbone architecture for comparative analysis. Table 1 demonstrates the segmentation *performance* of the original RC_Block model when incorporating six attention mechanisms—*CA, CBAM, ECA, FF CM, GAM,* and *MS-CAM*—for potatoes in cohesive soil. The results indicate that incorporating the *CA, CBAM, ECA, FFCM, GAM*, and *MS-CAM* attention mechanisms yields varying effects on model performance. Among them, the MS-CAM module demonstrated the best gain across all metrics. Compared to the original model, it improved *mIoU*, *Accuracy*, and mAP by 2.23, 0.61, and 2.34 percentage points respectively, while also increasing the F1 score by 2.66 percentage points to 83.90%. In contrast, other attention mechanisms (such as *CA* and *ECA*) showed fluctuations in some metrics but failed to achieve the same level of overall improvement as *MS-CAM*. In summary, the *MS-CAM* mechanism more effectively focuses on potato target areas while suppressing background noise, demonstrating higher practicality and segmentation efficiency. It is the preferred choice for potato appearance segmentation tasks in sticky soil environments.

Table 1

Comparison of the results of different attention mechanisms

Attention mechanism	F1%	mIoU%	Accuracy%	mAP%
——	81.24	81.18	94.72	86.78
CA	80.60	80.59	94.44	86.99
CBAM	80.60	80.61	94.54	86.99
ECA	80.70	80.73	94.53	86.42
FFCM	80.90	80.82	94.49	87.37
GAM	81.80	81.76	94.89	87.28
MS-CAM	83.90	83.41	95.33	89.12

Ablation experiment

To validate the impact of the introduced residual-based CNN feature extractor (denoted as RC_Block in Table 2), multi-scale channel attention module (denoted as MS-CAM in Table 2), dual-branch feature extraction module (denoted as TBFE in Table 2), and fusion Fourier convolution mixer (denoted as FFCM in Table 2) on improving segmentation accuracy and computational complexity of the SegFormer model, this paper conducted ablation experiments, with results shown in Table 2. Compared to the original SegFormer, introducing the RC_Block feature extractor enables the model to better capture shallow texture information, improving F1 and mIoU by 2.22 and 1.93 percentage points, respectively. Notably, this module optimized the front-end feature extraction architecture, reducing model parameters from 3.71 M to 3.47 M and significantly decreasing FLOPs from 13.53 G to 10.98 G, achieving a win-win scenario of improved accuracy and lightweight performance. Following the integration of the MS-CAM attention mechanism, F1 and mIoU further increased by 2.59 and 2.23 percentage points respectively compared to the previous combination, while MPA improved by 2.34 percentage points. At this stage, the number of parameters and FLOPs only slightly increased to 3.57 million and 11.02 G, respectively, indicating that multi-scale attention effectively enhances feature discriminability at minimal computational cost. Further incorporating the TBFE module improved F1 and mIoU by 0.84 and 0.79 percentage points, respectively, while maintaining low computational complexity (3.65 M parameters, 11.49 G FLOPs). Finally, incorporating the FFCM frequency domain module to suppress background noise yields optimal performance across all metrics. Although this frequency domain module increases the number of parameters to 5.95 million, the overall FLOPs of the final model (12.40 G) remains lower than the original SegFormer (13.53 G) due to the efficient streamlining achieved by the front-end modules. The combination of all four improvement strategies yielded the best results. Specifically, the proposed MTFormer model achieved F1, mIoU, Accuracy, and MAP scores that were 6.17, 5.37, 1.75, and 3.68 percentage points higher than the original SegFormer, respectively. Overall, the proposed improvements not only significantly enhance potato segmentation performance in sticky soil environments but also reduce computational complexity (FLOPs) by 1.13 G while maintaining manageable parameter counts (+2.24 M). This achieves an optimal balance between accuracy and computational efficiency, demonstrating high potential for engineering deployment.

Table 2

Ablation experiment

RC_Block	ms_cam	tbfe	ffcm	F1(%)	mIoU(%)	Accuracy(%)	mAP(%)	Model Size/MB	FLOPs(G)
×	×	×	×	79.02	79.25	93.92	86.58	3.71	13.53
√	×	×	×	81.24	81.18	94.72	86.78	3.47	10.98
√	√	×	×	83.83	83.41	95.33	89.12	3.57	11.02
√	√	√	×	84.67	84.20	95.55	89.83	3.65	11.49
√	√	√	√	85.19	84.62	95.67	90.26	5.95	12.40

Comparative analysis of different algorithms

To validate the segmentation performance of this improved model, comparative experiments were conducted under identical conditions between the proposed MTFormer algorithm and SegFormer-B1, SegFormer-B2, PSPNet, DeepLabV3+, and U-Net algorithms. The segmentation results are shown in Figure 8, with the evaluation metrics of each model presented in Table 3.

The visual segmentation results in Figure 8 reveal significant performance differences among models against the sticky soil background. While the SegFormer series (B1, B2) and DeepLabV3+ successfully identified the main potato regions, they exhibited noticeable blurring and adhesion when handling edge details, struggling to effectively distinguish potato boundaries covered by soil. PSPNet generally locates the target but exhibits noticeable missed detections when encountering small objects or partially buried potatoes, resulting in incomplete segmentation. U-Net, benefiting from its skip-connection architecture, demonstrates better edge restoration but shows weaker resistance to interference in complex textured backgrounds, occasionally misclassifying soil clumps as potatoes. In contrast, the proposed MTFormer achieves the most accurate potato segmentation. It not only effectively suppresses interference from soil textures in the background but also produces smooth and clear potato edge contours, most closely matching the ground truth annotations. This demonstrates its significant advantage in complex, unstructured environments.

Further quantitative analysis in Table 3 reveals that MTFormer achieves the best overall performance among mainstream semantic segmentation models. It attains an mIoU of 84.62%, an F1 score of 85.19%, an MPA of 90.26%, and an accuracy of 95.67%. Specifically, compared to SegFormer-B1, SegFormer-B2, PSPNet, and DeepLabV3+, MTFormer achieves mIoU improvements of 3.98, 3.68, 3.71, and 5.05 percentage points, respectively, even outperforming the well-performing U-Net by 0.92 percentage points in mIoU and 0.85 percentage points in F1 score. In terms of model lightweighting and real-time inference capability, MTFormer features a compact model size of only 35 MB and achieves an inference speed (FPS) of 35 frames per second. Notably, despite using the lighter SegFormer-B0 as the base model, MTFormer achieves significant reductions in feature refinement through modules like MS-CAM and TBFE. Compared to SegFormer-B1 (51 MB, 26 FPS), SegFormer-B2 (105 MB, 24 FPS), and U-Net (95 MB, 26 FPS). Although its model size is slightly larger than lightweight models like PSPNet (9.4 MB) and DeepLabV3+ (23 MB), and its inference speed is marginally lower than PSPNet (40 FPS), MTFormer fully meets the performance requirements for agricultural machinery dynamic field operations and real-time video frame processing tasks. This is because 30 FPS typically suffices for achieving smooth operation in such scenarios. In precision farming tasks, segmentation accuracy directly determines the reliability of yield estimation and automated sorting. MTFormer achieves a significant leap in accuracy while maintaining a low parameter count and meeting real-time inference requirements, striking an optimal balance between precision, efficiency, and lightweight design.

The experimental results above indicate that traditional CNN models (such as DeepLabV3+ and PSPNet) tend to lose detailed information of small objects in sticky soil during downsampling, leading to limited segmentation accuracy. While the original SegFormer possesses global modeling capabilities, it struggles to suppress local high-frequency noise in isochromatic scenes where background and object textures are extremely similar. The superior overall performance of MTFormer in complex environments mainly arises from its purpose-built enhancement modules. In particular, the Fusion Fourier Convolution Mixer (FFCM) utilizes frequency-domain filtering to decouple high-frequency soil texture noise from low-frequency potato features, effectively mitigating interference caused by background homogeneity. The Twin-Branch Feature Extraction (TBFE) module enhances the model's ability to distinguish potato geometric edges and deep semantic features by combining 3D channel modeling with 2D spatial filtering. The residual CNN extractor introduced at the front end compensates for the Transformer's lack of inductive bias in shallow feature extraction, thereby achieving more accurate and faster segmentation results without significantly increasing computational overhead.

Fig. 8 –Model Segmentation Effect Comparison

Table 3

Different model experiment results

Models	F1%	mIoU%	Accuracy%	mAP%	Model Size/MB	FPS
SegFormer-B1	80.76	80.64	94.34	87.92	51	26
SegFormer-B2	80.98	80.94	94.45	87.99	105	24
PSPNet	80.92	80.91	94.59	86.94	9.4	40
DeepLabV3+	79.46	79.57	94.00	86.95	23	29
U-Net	84.34	83.70	95.32	90.20	95	26
MTFormer	85.19	84.62	95.67	90.26	35	35

Counting Test Analysis

To validate the engineering application value of the proposed MTFormer model in practical agricultural workflows, this study conducted preliminary potato yield estimation (counting) experiments based on image segmentation. The experiment randomly selected 50 representative images of potatoes in clay soil from the test dataset. These samples encompassed varying lighting conditions, weed coverage, and soil coverage levels, with each image containing one or multiple naturally distributed potato targets. Manual counting results served as ground truth, which were compared against the automatic counting results from other models and the proposed MTFormer model. To quantitatively assess counting accuracy, the Mean Absolute Error (MAE) and Root Mean Square Error (RMSE) were introduced as evaluation metrics:

$$MAE = \frac{1}{N}\sum_{i=1}^{N} | C_{m,i} - C_{gt,i}| \tag{12}$$

$$RMSE = \sqrt{\frac{1}{N}\sum_{i=1}^{N}(C_{m,i} - C_{gt,i})^2} \tag{13}$$

where: N is the total number of test images; $C_{m,i}$ is the potato count predicted by the model for the *i*^(th) image; $C_{gt,i}$ denotes the actual manual count for the i-th image.

Table 4

Comparison of Potato Counting Errors Based on Different Segmentation Models

Models	Total Manual Count (units)	Total Number of Models (units)	MAE	RMSE	Counting Accuracy Rate (%)
SegFormer-B1	349	332	0.068	0.261	95.13
SegFormer-B2	349	334	0.060	0.245	95.70
PSPNet	349	336	0.052	0.228	96.28
DeepLabV3+	349	331	0.072	0.269	94.84
U-Net	349	337	0.048	0.220	96.56
MTFormer	349	340	0.036	0.190	97.42

As shown in the counting validation results of Table 4, traditional segmentation models tend to undercount in actual yield estimation due to boundary blurring and soil clod interference caused by sticky soil. Among them, DeepLabV3+ exhibits the most severe undercounting, with an overall counting accuracy of only 94.84%; U-Net performs relatively better, achieving an accuracy of 96.56%. In contrast, the proposed MTFormer model demonstrates exceptional engineering reliability, with its automatic count of 340 being the closest to the manual ground truth count of 349. In quantitative evaluation on the test set, MTFormer achieved an average absolute error (MAE) of just 0.036 and a root mean square error (RMSE) as low as 0.190, with an overall counting accuracy of 97.42%. This not only demonstrates the model's robust noise resistance and segmentation capabilities when handling same-object-different-spectrum features but also vividly highlights its immense application potential in automated yield estimation for agricultural machinery.

CONCLUSIONS

This study addresses the issues of edge blurring and background texture interference in potato segmentation within viscous soil environments by proposing an improved MTFormer model. Through theoretical analysis and comparative experiments, the following key conclusions were drawn.

(1) To address the limitation of the original SegFormer model in inductively missing biases during shallow feature extraction, a residual-based CNN feature extractor was introduced at the encoder front end, significantly enhancing the model's perception of potato surface texture and geometric contours. To address the misclassification problem caused by the heterogeneous spectrum (i.e., highly similar textures) between background soil clumps and potatoes, the Fourier Convolution Mixer (FFCM) is integrated with a dual-branch feature extraction module (TBFE). By leveraging frequency-domain denoising and a joint spatial–temporal modeling strategy, the model separates high-frequency soil texture noise from low-frequency potato features. This mechanism significantly enhances feature representation and improves robustness against background interference in complex field environments.

(1) MTFormer achieves an optimal balance between accuracy and computational efficiency. Experimental results demonstrate that on the self-built dataset, MTFormer achieves mean intersection over union (mIoU), F1 score, mean per-area (MPA), and accuracy of 84.62%, 85.19%, 90.26%, and 95.67%, respectively. All segmentation evaluation metrics significantly outperform the baseline comparison models. While maintaining high accuracy, the model's parameter size is only 35 MB, with total floating-point operations (FLOPs) controlled at 12.40 G and inference speed reaching 35 FPS. This demonstrates that the model significantly reduces storage requirements and computational burden while fully meeting the real-time and lightweight constraints of intelligent agricultural machinery during dynamic field operations. The automatic potato tuber counting results generated by the segmentation mask based on MTFormer showed high consistency with manual actual counts. The Mean Absolute Error (MAE) was only 0.036, the Root Mean Square Error (RMSE) was as low as 0.190, and the overall counting accuracy reached 97.42%. This further confirms that it can provide technical support for subsequent automated yield measurement.

(2) Currently, the proposed MTFormer model has demonstrated promising application potential in specific cohesive soil scenarios, fully validating the effectiveness of each improvement strategy. However, given that natural field environments are constrained by multidimensional complex physical factors, future research must further expand field datasets to encompass multiple regions, crop varieties, growth stages, and extreme light-climate conditions. This will enhance the model's robustness across vast unstructured agricultural landscapes. Concurrently, the research will incorporate deep model compression techniques such as network pruning, knowledge distillation, and tensorization to overcome frame rate limitations on low-power, computationally constrained agricultural edge computing devices. Ultimately, by advancing equipment-level closed-loop testing, this visual perception and counting module will be scaled for deployment in real potato harvesters' yield measurement systems, thereby providing comprehensive technical support for the industrialization of smart agricultural visual yield measurement equipment.

ACKNOWLEDGEMENT

This research was funded by the National Key R&D Program of China (Project no.2023YFD2000904)

REFERENCES

[1] Chen L C, Papandreou G, Kokkinos I. (2017). Deeplab: Semantic image segmentation with deep convolutional nets, atrous convolution, and fully connected CRFs [J]. *IEEE transactions on pattern analysis and machine intelligence*, 40(4): 834-848.

[2] Chen P, He W, Qian F. (2025). A synergistic CNN-transformer network with pooling attention fusion for hyperspectral image classification [J]. *Digital Signal Processing*, 160: 105070.

[3] Dai Y, Gieseke F, Oehmcke S. (2021). Attentional feature fusion [C] // *Proceedings of the IEEE/CVF winter conference on applications of computer vision*. 3560-3569.

[4] Devaux A, Goffart J P, Petsakos A. (2020). Global food security, contributions from sustainable potato agri-food systems[M] // The potato crop: Its agricultural, nutritional and social contribution to humankind. *Cham*: *Springer International Publishing*, 3-35.

[5] Dosovitskiy A. (2020). An image is worth 16x16 words: Transformers for image recognition at scale [J]. *arXiv preprint arXiv:2010*.11929.

[6] ElMasry G, Cubero S, Moltó E. (2012). In-line sorting of irregular potatoes by using automated computer-based machine vision system [J]. *Journal of Food Engineering*, 112(1-2): 60-68.

[7] Gao N, Jiang X, Zhang X. (2024). Efficient frequency-domain image deraining with contrastive regularization[C]//European Conference on Computer Vision. *Cham: Springer Nature Switzerland*. 240-257.

[8] Geng B, Dai G, Zhang H. (2024). Accurate non-destructive testing method for potato sprouts focusing on deformable attention [J]. *INMATEH-Agricultural Engineering*, Vol.72(1): pp. 402-413. DOI: https://doi.org/10.35633/inmateh-72-36

[9] He K, Gkioxari G, Dollár P. (2017). Mask R-CNN [C] // *Proceedings of the IEEE international conference on computer vision*. pp. 2961-2969.

[10] Liao H, Wang G, Jin S. (2025). HCRP-YOLO: A lightweight algorithm for potato tuber segmentation [J]. *Smart Agricultural Technology*,10: 100849.

[11] Liu W, Anguelov D, Erhan D. (2016). SSD: Single shot MultiBox detector[C]//European conference on computer vision. *Cham: Springer International Publishing*: 21-37.

[12] Lyu J., Tian Z., Yang Y. (2015). Development status, existing problems and development trend of potato machinery (马铃薯机械发展现状、存在问题及发展趋势). *Research on Agricultural Mechanization*, 2015, 37(12): 258-263.

[13] Ma Y, Zhang G, Liu J. (2016). Research on Leaf Image Disease Spot Segmentation Algorithm Based on SSD Network [J].2019.

[14] Oppenheim D, Shani G, Erlich O. (2019). Using deep learning for image-based potato tuber disease segmentation[J]. *Phytopathology*, 109(6): 1083-1087.

[15] Razmjooy N, Mousavi B S, Soleymani F. (2012). A real-time mathematical computer method for potato inspection using machine vision[J]. *Computers & Mathematics with Applications*, 63(1): 268-279.

[16] Redmon J, Divvala S, Girshick R. (2016). You only look once: Unified, real-time object segmentation [C] // *Proceedings of the IEEE conference on computer vision and pattern recognition*. 779-788.

[17] Ren S, He K, Girshick R. (2016). Faster R-CNN: Towards real-time object segmentation with region proposal networks[J]. *IEEE transactions on pattern analysis and machine intelligence*, 39(6): 1137-1149.

[18] Ronneberger O, Fischer P, Brox T. (2015). U-net: Convolutional networks for biomedical image segmentation[C]//International Conference on Medical image computing and computer-assisted intervention. Cham: Springer international publishing, 234-241.

[19] Xi R, Hou J, Lou W. (2020). Potato bud segmentation with improved faster R-CNN [J]. *Transactions of the ASABE*, 63(3): 557-569.

[20] Xie E, Wang W, Yu Z. (2020). SegFormer: Simple and efficient design for semantic segmentation with transformers[J]. Advances in neural information processing systems, 34: 12077-12090.

[21] Zhang W., Han Y., Huang C., Chen Z. (2022). Recognition method for seed potato buds based on improved YOLOv3-tiny [J]. *INMATEH-Agricultural Engineering*, 67(2), pp. 364-373 DOI: https://doi.org/10.35633/inmateh-67-37