Journal of Zhejiang University-SCIENCE B 2026 27(5):517-536
www.jzus.zju.edu.cn; www.springer.com/journal/11585
E-mail: jzus_b@zju.edu.cn

Research Article
https://doi.org/10.1631/jzus.B2500451

RCTUnet: a deep learning model for crop-residue-soil image segmentation and crop residue cover extraction

Ting LI[1], Yang LIU[1], Haikuan FENG[2,3], Meiyan SHU[1]✉, Hao YANG[2,3], Yuanyuan FU[1], Xin XU[1], Yinghao LIN[4,5]✉, Hongbo QIAO[1], Wei GUO[1], Xinming MA[1], Lei SHI[1], Jibo YUE[1,2]✉

[1]*College of Information and Management Science, Henan Agricultural University, Zhengzhou 450046, China*
[2]*Institute of Quantitative Remote Sensing and Smart Agriculture, Henan Polytechnic University, Jiaozuo 454000, China*
[3]*Key Laboratory of Quantitative Remote Sensing in Agriculture, Ministry of Agriculture and Rural Affairs, Information Technology Research Center, Beijing Academy of Agriculture and Forestry Sciences, Beijing 100097, China*
[4]*Shenzhen Research Institute of Henan University, Shenzhen 450046, China*
[5]*Henan Key Laboratory of Big Data Analysis and Processing, School of Computer and Information Engineering, Henan University, Kaifeng 475001, China*

Abstract: Accurate quantification of crop residue cover (CRC) is crucial for monitoring and evaluating conservation tillage practices, yet it poses a significant image segmentation challenge. The subtle visual distinctions between fragmented residue and soil, compounded by variable illumination and shadows in field imagery, often lead to poor segmentation performance. To overcome these limitations, we introduce RCTUnet, a novel deep learning architecture designed for robust crop-residue-soil segmentation and precise CRC estimation. RCTUnet's architecture synergistically integrates three key components: (1) a ResNet50 backbone for deep, multi-scale feature extraction; (2) a convolutional block attention module (CBAM) to adaptively focus on salient residue features across both channel and spatial dimensions; and (3) a transformer-based global context fusion module (GCFM) to model long-range spatial dependencies, which is critical for interpreting heterogeneous residue patterns. We evaluated RCTUnet on a dataset of 1220 field-acquired images spanning four typical crop rotations. Experimental results show that, compared to traditional models: (1) RCTUnet achieves significantly higher crop-residue-soil segmentation accuracy than classic models including Unet, Unet++, DeepLabV3, segmentation network (SegNet), and fully convolutional network (FCN), with improvements of 3.24%, 3.42%, 4.88%, 8.28%, and 6.05% in overall accuracy, respectively; (2) RCTUnet yields superior residue-soil segmentation performance, with increases in residue recall of 7.67%, 7.37%, 14.09%, 27.05%, and 16.91%, respectively; (3) RCTUnet shows enhanced CRC estimation accuracy, achieving a root mean square error (RMSE) of 4.875, representing a 45.5% improvement over Unet (RMSE=8.941). These results demonstrate the efficacy of our hybrid approach, which combines deep hierarchical features, dual-domain attention, and global context modeling. RCTUnet provides a robust and reliable tool for automated CRC assessment, advancing the capabilities of in-field agricultural monitoring.

Key words: Deep learning; Crop residue cover; Image segmentation; Conservation tillage

1 Introduction

Crop residues constitute a vital component of agricultural ecosystems, offering substantial benefits in reducing erosion, enhancing organic matter content, improving structure, and increasing water retention and drought resistance capacity of the soil (Liu et al., 2023; Yue et al., 2023). Crop residue cover (CRC), defined as the proportion of the soil surface area covered by crop residues relative to the total surface area, serves as a critical indicator for assessing the effectiveness of conservation tillage practices (Hively et al., 2018; Ding et al., 2020). Accurate monitoring of CRC is instrumental in understanding the spatial distribution of conservation tillage, provides a scientific basis for governmental evaluation of its adoption, and plays a crucial role in tracking the progress of sustainable

✉ Jibo YUE, yuejibo@henau.edu.cn
Meiyan SHU, smy511@henau.edu.cn
Yinghao LIN, linyh@henu.edu.cn
Jibo YUE, https://orcid.org/0000-0001-9766-5313
Meiyan SHU, https://orcid.org/0000-0002-1519-5520
Yinghao LIN, https://orcid.org/0000-0002-5048-3536
Ting LI, https://orcid.org/0009-0009-9516-1774

Received July 29, 2025; Revision accepted Nov. 25, 2025; Crosschecked May 9, 2026

agricultural practices over time (Delandmeter et al., 2024).

Traditional CRC measurement methods include a photographic method and a line-transect method (Luo et al., 2022; Yue et al., 2024). Although these methods are simple and practical, they suffer from high subjectivity and low efficiency, making it difficult to obtain large-scale CRC data within short timeframes (Wang FY et al., 2025). In recent years, techniques such as segmentation, machine learning, remote sensing, and deep learning have been increasingly applied to automate CRC estimation (Laamrani et al., 2018; Yue et al., 2019).

Threshold-based segmentation methods divide images into distinct regions based on predefined grayscale or color thresholds (Song HH et al., 2024). However, their performance is often compromised by variation in image quality and lighting conditions (Shang et al., 2021). With advances in computer vision, segmentation techniques have evolved beyond thresholding to more sophisticated edge detection approaches. Edge detection algorithms, such as the Canny detector and the Sobel operator, have gained popularity for their efficient boundary extraction capabilities (Gao PP et al., 2022; Feng et al., 2024). These methods compute image gradients to identify object boundaries, thereby facilitating downstream segmentation tasks. Nonetheless, because they rely on manually designed feature extraction rules, they struggle to fully capture the complexity of natural scenes, limiting their applicability to specific image types. Their effectiveness is particularly constrained under complex agricultural conditions. Building upon traditional methods, machine learning algorithms have introduced new paradigms for image segmentation. Support vector machine (SVM) leverages kernel functions to project features into high-dimensional spaces for constructing optimal segmentation hyperplanes (Yu et al., 2024), while random forest (RF) enhances model robustness and accuracy by integrating multiple decision trees and introducing random feature selection mechanisms (Yue and Tian, 2020; Xu et al., 2022). However, in agricultural scenarios where soil and crop residues share highly similar features, these conventional machine learning approaches exhibit limited generalization capacity. They frequently suffer from over-segmentation or under-segmentation errors due to inadequate feature representation capabilities, which hinder their ability to capture subtle yet critical class-discriminative cues (Gao et al., 2024).

In recent years, progress in remote sensing technologies has led to their widespread application in estimating CRC at regional scales (de Paul Obade and Gaya, 2020; Tao et al., 2021). Crop residues, composed mainly of lignin and cellulose, exhibit distinct absorption features at about 2100 and 2300 nm, respectively. These absorption troughs allow for effective differentiation between crop residues and background soil in hyperspectral data (Zhang et al., 2025). Serbin et al. (2009) used the SINDRI index, derived from multispectral sensor data, to estimate crop residue cover at relatively low cost. Gao LL et al. (2022) used unmanned aerial vehicle (UAV) hyperspectral imagery to acquire reflectance spectra and maize residue distribution maps, and subsequently proposed a normalized difference index adjusted for residue morphology, enabling accurate CRC estimation. Pacheco and McNairn (2010) explored a mixed-pixel spectral unmixing approach based on reflectance characteristics in agricultural remote sensing imagery to extract residue cover information. Wang et al. (2019) integrated the dead fuel index, normalized difference vegetation index, and a spectral unmixing model to estimate the proportions of vegetation, soil, and non-photosynthetic vegetation in Inner Mongolia.

In addition, deep learning-based image segmentation has emerged as a major research direction in computer vision, offering powerful alternatives for fine-grained analysis of visual data (Zhu et al., 2025). These techniques have seen widespread application in domains such as remote sensing imagery (Song WY et al., 2024; Yang et al., 2025) and medical imaging (Mahmood and Ucan, 2025; She et al., 2025). Semantic segmentation, which enables pixel-level classification, provides precise labeling for each pixel within an image. Compared to traditional segmentation methods, deep learning-based semantic segmentation algorithms leverage hierarchical feature extraction and multi-scale context fusion, thereby substantially improving segmentation performance in complex agricultural scenes. These advances provide a viable technical pathway for accurately distinguishing between crop residues and soil. Wang YY et al. (2025) proposed an improved lightweight semantic segmentation model, m-DeepLabV3+, to achieve high-precision and rapid detection of residue cover in field environments. Zhou et al. (2020) used low-altitude aerial imagery for the flexible and efficient acquisition of crop residue maps, applying

an enhanced Unet architecture and a ResNet18-UNet hybrid model to estimate CRC. Li et al. (2021) developed the Siamese domain transfer network (SDTN) architecture, based on a modified mask-region-based convolutional neural network (M-R-CNN) framework, to segment maize residue images under limited sample conditions and evaluated segmentation performance using common objects in context (COCO) metrics. Zheng et al. (2024) introduced RSU-Net, a novel deep neural network that integrates residual units with a spatial pyramid pooling mechanism to enhance the extraction of various cover types from GaoFen-1B (GF-1B) satellite imagery.

Although research has achieved high accuracy in segmenting vegetation and soil backgrounds, the segmentation of crop residue remains a persistent challenge. Fragmented crop residues often appear very similar to the surrounding soil in both color and texture, making them difficult to distinguish in the feature space (Beeche et al., 2022). Furthermore, variation in lighting, shadow interference, and uneven spatial distribution of residues significantly reduce the robustness of segmentation algorithms in complex agricultural scenes. Traditional segmentation methods struggle to effectively extract residue features due to inherent technical limitations. Threshold-based approaches rely heavily on statistical differences in pixel color or grayscale values, which are often insufficient in residue-rich environments. Traditional machine learning methods (e.g., SVM and RF), which depend on hand-crafted features (e.g., texture descriptors), are notably inadequate in capturing the subtle, complex variations between crop residues and soil, particularly when both share similar spectral and morphological characteristics. Remote sensing techniques, though valuable at broader scales, are often constrained by insufficient spatial resolution, making it difficult to accurately detect fragmented residues. Residues often appear as elongated, densely clustered structures with spectral characteristics similar to those of soil, making them prone to under-segmentation, blurred boundaries, and omission errors when processed by general-purpose models. Most deep learning segmentation models are optimized for global segmentation accuracy rather than tailored to the specific structural characteristics of crop residues. These issues critically limit the accuracy of CRC estimation, highlighting the need for more specialized segmentation models tailored to residue detection.

This study aimed to address two key challenges in field-based residue segmentation: (1) the morphological complexity of fragmented crop residues, which makes them difficult to detect and segment accurately; (2) lighting variability, which increases the likelihood of misclassification between residue and soil. We developed a novel deep learning-based model named RCTUnet. Based on the Unet architecture, RCTUnet incorporates three key innovations to enhance segmentation performance:

(1) Integration of ResNet50 as the backbone. Leveraging its deep residual structure, the model enhances its multi-scale feature extraction capabilities, enabling the precise delineation of fragmented and elongated residue structures.

(2) A convolutional block attention module (CBAM). The CBAM applies both channel-wise and spatial attention, allowing the model to focus on key residue regions while suppressing interference from complex background textures.

(3) A global context fusion module (GCFM) based on transformer architecture. This module uses self-attention mechanisms to model global contextual relationships, enabling consistent recognition of fragmented residues under variable lighting conditions.

Experimental results show that the proposed RCTUnet model significantly outperforms conventional segmentation methods in accurately extracting residue features under real-world, complex field conditions.

2 Materials and methods

2.1 Materials

2.1.1 Study area

The study area (Fig. 1) was located in Yuanyang County, Xinxiang City, Henan Province, China (113°36′–114°15′E, 34°55′–35°11′N). Situated at the southern end of the North China Plain, this region features a typical warm temperate continental monsoon climate. The region predominantly follows a crop rotation system, with major grain crops such as wheat, rice, and maize, alongside economically important crops like soybean, peanut, and cotton.

2.1.2 Farmland digital image collection

We collected digital images of farmland covering various crop rotation systems. Four mainstream

Fig. 1 Study area. (a) Experimental farmland; (b) Crop-residue-soil images.

smartphones were used for image acquisition—Xiaomi 10S (4344×5792 pixels), Honor Magic5 (3024×4032 pixels), Realme Q3 (3000×3000 pixels), and Apple 14 Pro (3024×4032 pixels)—with the camera fixed at a height of 1 m above the ground to ensure standardization and repeatability in data collection. The dataset encompassed four representative crop rotation systems: (1) rice residue–wheat rotation; (2) wheat residue–soybean rotation; (3) wheat residue–maize rotation; (4) maize residue–wheat rotation.

During field digital image acquisition, two types of interference were observed due to practical field conditions: (1) anthropogenic disturbances, such as footprints of photographers and device shadows; and (2) environmental disturbances, including lens flare caused by intense sunlight. To maintain dataset quality, images containing obvious occlusions or non-agricultural objects were excluded. Each retained image was cropped into a square by taking half the length of the shortest edge, centered on the original image, to minimize irrelevant content while preserving key visual elements.

2.1.3 Digital image preprocessing

Several preprocessing steps were implemented to ensure data quality and enhance model training performance:

(1) Image normalization. To balance preservation of image detail with computational efficiency, original high-resolution images were uniformly resized to 256×256 pixels. This resolution retains residue morphological features while ensuring efficient model training and inference.

(2) Data selection. To ensure representativeness and balance, we selected field images from the four crop rotation systems, covering different plots and cultivation regimes. For each rotation type, 305 representative images were chosen, yielding a total of 1220 high-quality samples, providing a robust foundation for model training.

(3) Data annotation. Pixel-wise semantic annotation was performed using the Labelme tool, categorizing each image into three target classes—vegetation (including associated weeds), crop residue, and bare soil. Annotations were stored in JSON format and subsequently converted into single-channel PNG label maps using color encoding. The three classes were assigned fixed colors to facilitate class identification during training.

(4) Data augmentation. After completing data annotation, we applied multiple augmentation strategies to the 1220 original images to enhance diversity and complexity, thereby improving the representativeness of the dataset. Specifically, we used six augmentation techniques: random flipping, random rotation, random cropping, brightness adjustment, Gaussian noise addition, and grayscaling. After augmentation, we

obtained a total of 7320 farmland images along with their corresponding label maps.

(5) Dataset partitioning. To ensure rigorous model evaluation, we applied stratified random sampling to split the 7320 annotated images into training, validation, and test sets, in a 60% (4392 images), 20% (1464 images), and 20% (1464 images) ratio, respectively. Each crop rotation system was evenly represented across all subsets (25% per subset), thereby avoiding class imbalance that could bias model performance.

Fig. 2 illustrates representative field images and their corresponding semantic label maps, clearly showing the spatial distribution of crops, residues, and soil through color-coded annotation.

2.2 RCTUnet

2.2.1 RCTUnet architecture

RCTUnet integrates multi-scale local feature extraction, channel-spatial attention mechanisms, and global context modeling to improve segmentation performance for crop residue-related targets (Fig. 3). Specifically, RCTUnet comprises three modules:

(1) ResNet50 is introduced as the encoder to enhance feature extraction capabilities. The hierarchical structure, constructed from multi-scale residual blocks, enables effective capture of edge details and texture information associated with crop residue objects.

(2) CBAM is embedded within the decoding path. These modules learn adaptive attention weights to guide the network in focusing on crop residue-relevant features while suppressing background noise, thereby enhancing segmentation accuracy.

(3) Transformer-based GCFM is integrated into the higher semantic layers of the encoder to compensate for the convolutional networks' limitations in global semantic modeling. Leveraging self-attention mechanisms, GCFM enhances the model's capacity to handle complex scenarios, such as blurred boundaries between soil and crop residue.

RCTUnet effectively combines local detail awareness with global contextual reasoning. The model fully utilizes the five-layer features (feat1–feat5) extracted by ResNet50 to construct a feature pyramid for cross-scale information fusion. The CBAM module further enhances mid- and high-level semantic features by highlighting salient regions. Transformer is a neural network architecture introduced by Vaswani for natural language processing. The GCFM introduces a transformer-based structure into high-level semantic abstraction, thereby improving the model's robustness in challenging environments.

2.2.2 ResNet50 backbone

We adopted ResNet50 as the backbone encoder network in this study (Qiang et al., 2023; Shahi et al., 2023). ResNet50 is composed of stacked bottleneck modules, each consisting of a 1×1 convolution for dimensionality reduction, a 3×3 convolution for feature extraction, and a 1×1 convolution for dimensionality restoration (Fig. 4). The network outputs feature maps

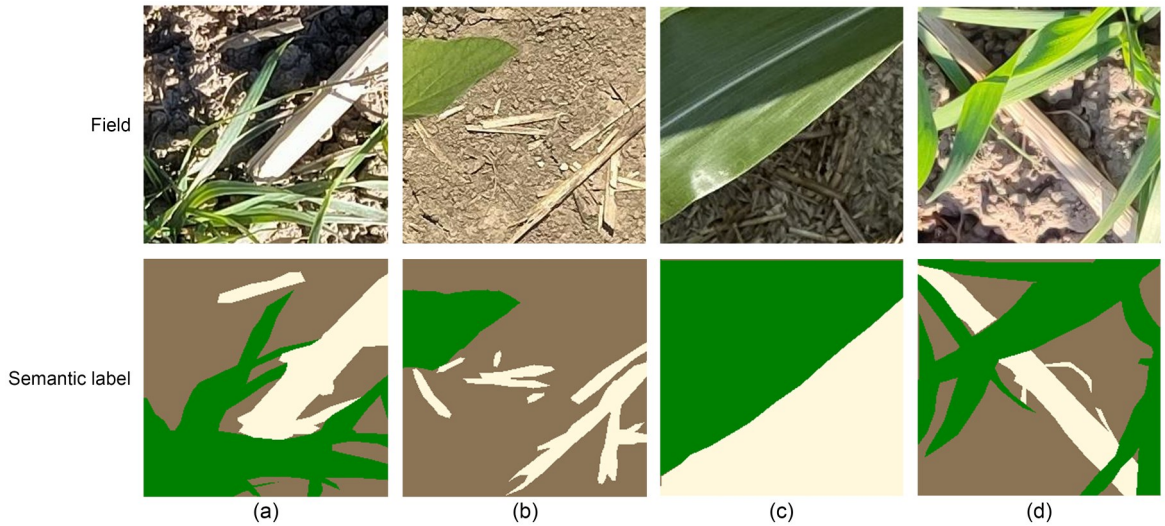

Fig. 2 Crop-residue-soil images and annotated images. (a) Rice residue–wheat; (b) Wheat residue–soybean; (c) Wheat residue–maize; (d) Maize residue–wheat.

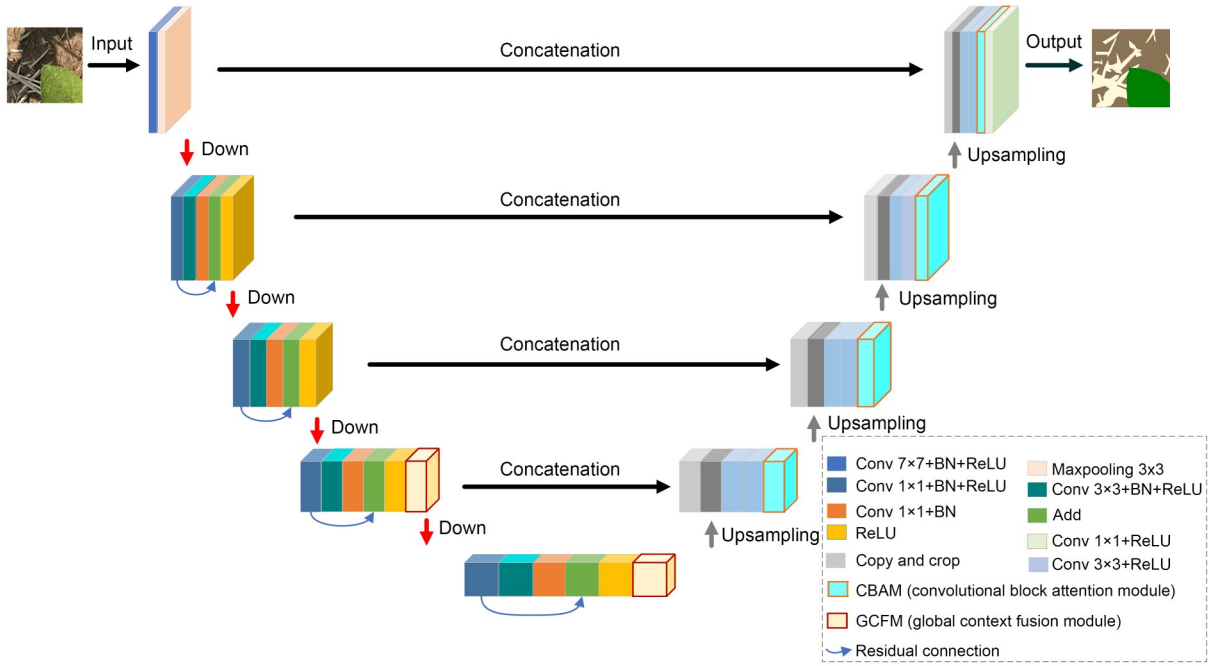

Fig. 3 RCTUnet architecture. Conv: convolution; BN: batch normalization; ReLU: rectified linear unit.

through four residual stages, progressively expanding the receptive field and increasing channel depth. This design allows the model to capture both low-level texture details (e.g., crop residue edges) and high-level semantic information (e.g., the overall morphology of crop residue). The residual connections facilitate the stable extraction of continuous features of slender crop residue structures and enhance the model's ability to distinguish between soil and crop residue at boundary regions.

2.2.3 CBAM attention

The CBAM (Fig. 5) consists of channel attention and spatial attention (Du et al., 2022; Wang et al., 2024). CBAM was integrated into the decoder to improve the model's capability for detecting crop residue in complex farmland environments. The channel attention module (Fig. 6) captures global semantic descriptors for each channel through global max pooling and average pooling operations. These descriptors are passed through shared fully connected layers to generate channel-wise attention weights, thereby enhancing feature selection and emphasizing semantically meaningful channels. This mechanism enhances the semantic differentiation between crop residue and soil, thereby improving sensitivity to small-scale features. The spatial attention module operates on the channel-refined feature maps.

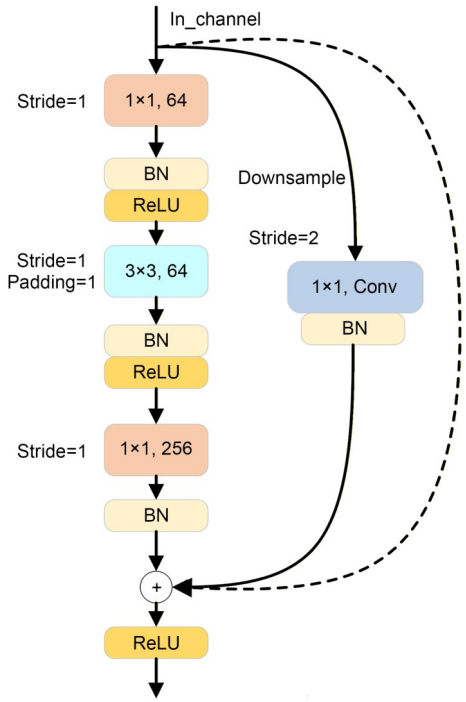

Fig. 4 Bottleneck block. Conv: convolution; BN: batch normalization; ReLU: rectified linear unit.

It performs max pooling and average pooling along the channel dimension to produce two spatial attention maps. These are fused through convolution to generate a final spatial attention map, which guides the model's

focus toward sparse regions likely to contain crop residue, thus improving spatial localization. The output of the CBAM module combines salient information across both channel and spatial dimensions, enhancing response in key target areas while effectively suppressing background interference. CBAM adaptively modulates multi-scale features passed from the encoder. In conjunction with upsampling and convolution operations, this integration provides the model with more robust shallow and deep feature representations during decoding, enabling high-precision segmentation of crop, crop residue, and soil regions in field images.

2.2.4 GCFM

Semantic segmentation of crop residue in agricultural imagery poses significant challenges, particularly under extreme lighting conditions or from occlusions that lead to blurred boundaries. Traditional convolutional neural networks, constrained by their limited local receptive fields, often fail to adequately capture global semantic context in such complex scenes. To address this limitation, we propose a GCFM—a transformer-based enhancement component—specifically introduced into the deep semantic layers of the network (feat4 and feat5) (Fig. 7). GCFM integrates global context modeling with local feature enhancement, significantly improving segmentation performance in challenging scenarios. Transformer is a self-attention-based architecture capable of modeling long-range dependencies. Leveraging this mechanism, the GCFM is designed with three critical processing stages:

(1) Feature space projection and positional encoding. The high-level feature maps (feat4 and feat5) are first compressed via a 1×1 convolution into a unified 256-dimensional embedding space, effectively reducing computational overhead. Subsequently, two-dimensional (2D) positional encoding is applied to inject spatial location information into the features, enabling transformer to perceive and preserve spatial topology within the image. The positional encoding (PE) is defined as:

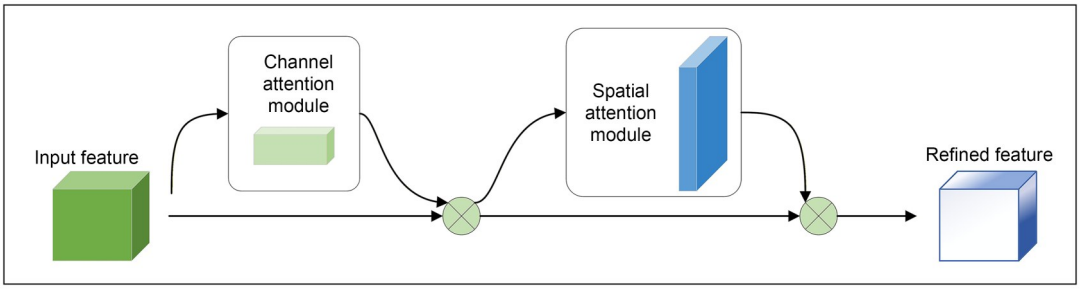

Fig. 5 CBAM architecture.

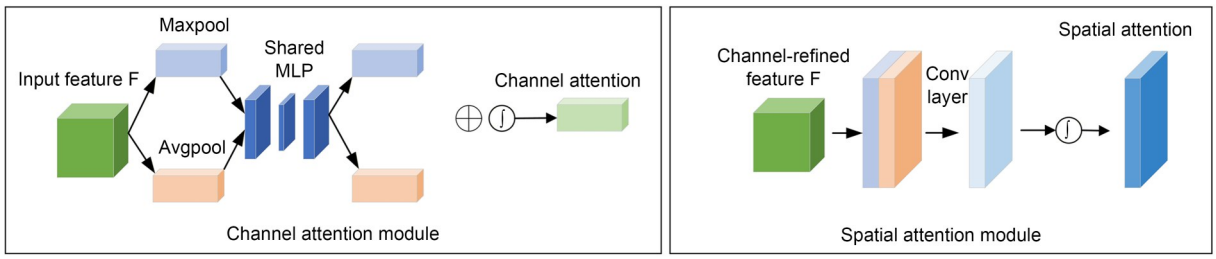

Fig. 6 Channel attention and spatial attention modules. MLP: multilayer perceptron; Conv: convolution.

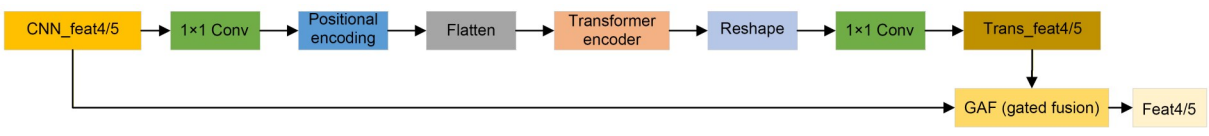

Fig. 7 Transformer-based global context fusion module (GCFM) architecture. CNN: convolutional neural network; Conv: convolution.

$$\mathrm{PE}_{(p,2i)}=\sin\left(\frac{p}{10\,000^{\frac{2i}{d_{\mathrm{model}}}}}\right), \tag{1}$$

$$\mathrm{PE}_{(p,2i+1)}=\cos\left(\frac{p}{10\,000^{\frac{2i}{d_{\mathrm{model}}}}}\right), \tag{2}$$

where p denotes the spatial index, i is the channel index, and d_{model} represents the embedding dimensionality. This encoding preserves spatial feature distribution, allowing the network to differentiate crop residue features across variable locations precisely.

(2) Global context modeling via transformer. The position-encoded features are flattened into token sequences and passed through a standard transformer encoder consisting of three layers and eight attention heads. This structure models global relationships via self-attention, computing interactions between all spatial locations:

$$\mathrm{Attention}(\boldsymbol{Q},\boldsymbol{K},\boldsymbol{V})=\mathrm{softmax}\left(\frac{\boldsymbol{Q}\boldsymbol{K}^{\mathrm{T}}}{\sqrt{d_{\mathrm{k}}}}\right)\boldsymbol{V}, \tag{3}$$

where $\boldsymbol{Q}$ and $\boldsymbol{K}$ denote the query and key matrices, respectively, representing the correlation between input features, $\boldsymbol{V}$ refers to the value matrix carrying contextual information, and d_{k} represents the dimensionality of the key vectors.

This mechanism enables the model to establish spatial dependencies across distant regions of the image. It significantly enhances global perception of crop residue structure, maintains semantic coherence even under occlusion or fragmentation, and improves discrimination in regions affected by uneven lighting. Furthermore, the global attention mitigates local noise interference and enriches the expressiveness of the learned features, leading to more accurate and stable segmentation outcomes.

(3) Feature reconstruction and gated fusion (GAF). The output of transformers (Trans_feat4 and Trans_feat5) is projected back to the original channel dimensions using a 1×1 convolution. To prevent information redundancy or conflicts during fusion with CNN features, we introduce a GAF mechanism. This mechanism adaptively balances the contribution of global context and local texture information, retaining critical edge and contour features while enhancing semantic understanding. The fusion ensures robust and context-aware segmentation of crop residue across diverse agricultural environments.

2.3 Benchmark models and accuracy evaluation

2.3.1 Benchmark models

Numerous representative deep learning architectures have been proposed for semantic segmentation tasks. The fully convolutional network (FCN) was the first to introduce an end-to-end fully convolutional architecture, enabling pixel-wise prediction through hierarchical feature fusion across variable receptive fields (Li et al., 2023). Building upon FCN, the Unet model introduced a symmetric encoder-decoder structure with skip connections that facilitate efficient multi-scale feature fusion. Unet has demonstrated exceptional performance in small-sample tasks such as medical image segmentation (Brosch et al., 2016; Li et al., 2025).

Unet architecture suffers from feature dilution, wherein high-level semantic features in deep layers may be contaminated by low-level noise during the skip connection process. To address this limitation, Unet++ introduces a nested and densely connected topology, enhancing semantic consistency across feature hierarchies through dense skip connections at multiple scales. While this design mitigates semantic gaps between shallow and deep layers, it considerably increases the number of parameters and the computational overhead.

The segmentation network (SegNet) model adopts an upsampling mechanism based on pooling indices. By recording the spatial locations of maximum activations during the pooling operation, SegNet reconstructs high-resolution feature maps during decoding without learnable parameters (Zhang et al., 2023). This non-parametric upsampling strategy significantly reduces memory consumption, making SegNet suitable for real-time segmentation tasks.

The DeepLabV3+ model systematically integrates atrous spatial pyramid pooling with multi-scale feature fusion strategies. By using atrous convolutions with variable dilation rates, it captures multi-scale contextual information while preserving spatial resolution (Shi et al., 2025). DeepLabV3+ achieves state-of-the-art performance in complex segmentation tasks, albeit at the cost of reduced computational efficiency.

The SegFormer model represents a new generation of transformer-based segmentation architectures that discard traditional decoders and instead use a

lightweight multilayer perceptron (MLP) head for feature aggregation (Xie et al., 2021). It uses a hierarchical transformer encoder to capture both global semantic dependencies and fine-grained spatial details through efficient multi-scale feature extraction.

The cross-channel residual and spatial fusion network (CCRSNet) explicitly fuses shallow and deep feature representations extracted from multiple stages of the backbone network (Gao et al., 2024). This hierarchical integration enables the model to capture both fine-grained textural cues and high-level contextual semantics, thereby improving the discrimination of visually similar regions such as crop residue and soil. The incorporation of multi-scale contextual fusion and attention refinement allows CCRSNet to more effectively extract and use rich spatial and semantic information from complex farmland scenes, leading to improved segmentation accuracy and enhanced generalization in the proportional estimation of crops, crop residue, and soil.

2.3.2 Evaluation metrics

To objectively assess the performance of the proposed model in the crop residue segmentation task, we used several commonly used metrics: mean intersection over union (MIoU), accuracy (A), precision (P), and recall (R) for each class. As a central metric in semantic segmentation, MIoU quantifies the overlap between predicted segmentation and ground truth annotations. Accuracy provides an overall measure of prediction correctness across all classes. Precision reflects the proportion of correctly identified positive samples among all samples predicted as positive. Recall evaluates the model's ability to detect actual positive instances (Zhao et al., 2023; Jin et al., 2025).

$$\mathrm{MIoU}=\frac{N_{\mathrm{TP}}}{N_{\mathrm{TP}}+N_{\mathrm{FN}}+N_{\mathrm{FP}}}, \tag{4}$$

$$A=\frac{N_{\mathrm{TP}}+N_{\mathrm{TN}}}{N_{\mathrm{TP}}+N_{\mathrm{TN}}+N_{\mathrm{FP}}+N_{\mathrm{FN}}}, \tag{5}$$

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

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

where N_{TP} denotes the number of true positives (TP), N_{TN} denotes the number of true negatives (TN), N_{FP} denotes the number of false positives (FP), and N_{FN} denotes the number of false negatives (FN).

To further assess the accuracy of class proportion estimation among the three categories—soil, crop, and crop residue—we adopted root mean square error (RMSE) and coefficient of determination (R^2) as additional evaluation indicators.

$$\mathrm{RMSE}=\sqrt{\frac{1}{m}\sum_{i=1}^{m}(y_i-\hat{y}_i)^2}, \tag{8}$$

$$R^2=1-\frac{\sum_{i=1}^{m}\left(\hat{y}_i-y_i\right)^2}{\sum_{i=1}^{m}\left(\bar{y}_i-y_i\right)^2}, \tag{9}$$

where y_i denotes the ground truth, $\bar{y}_i$ denotes the sample mean, $\hat{y}_i$ denotes the predicted value, and m denotes the sample size.

All experiments were conducted on a 64-bit Windows 11 operating system with CUDA 11.8. The models were implemented in Python 3.8 using the PyTorch 2.4.1 deep learning framework. The models were trained on an NVIDIA GeForce RTX 3080 GPU using the Adam optimizer with an initial learning rate of 1×10^{-4}, the first moment decay coefficient β_1=0.9, the second moment decay coefficient β_2=0.999, and the numerical stability term $\varepsilon=1\times10^{-8}$. The batch size was 8 and models were trained for 100 epochs. The loss function was a combination of cross-entropy (CE) loss. All segmentation models were independently trained and tested five times, and the results reported in the final table correspond to the best-performing run.

To evaluate the computational efficiency of the proposed model, we measured the average inference time per image on an NVIDIA RTX 3080 GPU. The proposed RCTUnet achieved an average inference time of 0.07 s per 256×256 images, which is comparable to mainstream segmentation architectures such as SegFormer, indicating that it is suitable for near real-time field deployment scenarios.

3 Results

3.1 Comparison of segmentation performance across different models

Table 1 presents the performance metrics of various segmentation models. The Unet model achieved a MIoU of 82.52%, outperforming DeepLabV3, SegNet, FCN, and Unet++ by 3.82%, 9.43%, 6.85%, and 0.31%,

respectively. Unet showed the highest accuracy at 90.75%, which exceeded those of DeepLabV3, SegNet, FCN, and Unet++ by 1.64%, 5.04%, 2.81%, and 0.18%, respectively. In terms of precision, Unet achieved 89.75%, surpassing DeepLabV3, SegNet, FCN, and Unet++ by 1.18%, 4.31%, 3.02%, and 0.19%, respectively (Table 1).

The results in Table 1 further indicate that recall for the crop class across all models exceeded 95%, suggesting strong recognition capabilities for crop regions. In contrast, recall values for the crop residue class were generally lower and exhibited greater variability, highlighting a significant bottleneck in current segmentation performance. This performance gap stems mainly from the visual similarity between crop residue and soil, particularly under natural conditions where crop residue and soil share similar color and texture characteristics. Such resemblance often leads to the misclassification of crop residue as background, resulting in elevated false negatives. Unet achieved 82.65% recall for the crop residue, which is significantly lower than its recall for crops (95.97%) and soil (93.59%) (Table 1). This indicates that Unet, while effective overall, still has limitations in modeling crop residue features in complex backgrounds.

To further validate the performance advantages of the proposed model, additional experiments were conducted comparing baseline models and SegFormer and CCRSNet with the proposed RCTUnet. The experimental results show that RCTUnet consistently achieved the best performance across all evaluation metrics, with an MIoU of 88.11%, accuracy of 93.99%, and precision of 93.66%. Notably, the recall for the crop residue category increased to 90.32%, representing an improvement of 7.67% compared to the baseline Unet model. In contrast, SegFormer and CCRSNet achieved MIoU values of 83.11% and 85.96%, respectively—both outperforming conventional CNN-based architectures but still falling short of RCTUnet.

These results indicate that RCTUnet exhibits superior robustness and generalization capability in complex field environments, particularly in recognizing and delineating crop residue boundaries. The model effectively mitigates segmentation challenges arising from the elongated and fragmented morphology of crop residue targets. Therefore, in this study we adopted RCTUnet, which achieved the optimal performance, as the final model for subsequent in-depth analysis and discussion.

Fig. 8 shows the results of different segmentation models. All models performed well in extracting crop regions. However, their ability to distinguish between crop residue and soil varied significantly. All models exhibited notable declines in segmentation accuracy for crop residue regions under complex lighting conditions (Figs. 8a and 8b). DeepLabV3, FCN, and SegNet showed severe over-segmentation, frequently misclassifying large portions of soil as crop residue (Figs. 8c and 8d). In contrast, Unet and Unet++ delivered more accurate discrimination between crop residue and soil, particularly under complex lighting scenarios, where they better preserved the edge details of crop residue. Beyond the conventional CNN-based architectures, SegFormer and CCRSNet showed improved global feature representation and achieved more consistent segmentation of crop residue regions, particularly in reducing the misclassification of small-scale targets. Nevertheless, both models still had difficulties in

Table 1 Performance metrics of different segmentation models

Model	MIoU (%)	Accuracy (%)	Precision (%)	Recall (%)		
				Crop	Residue	Soil
Unet	82.52	90.75	89.75	95.97	82.65	**93.59**
Unet++	82.21	90.57	89.56	96.01	82.35	93.26
DeepLabV3	78.70	89.11	88.57	96.34	76.23	90.05
SegNet	73.09	85.71	85.44	95.91	63.27	89.42
FCN	75.67	87.94	86.73	95.53	73.41	83.91
SegFormer	83.11	91.27	90.94	95.91	83.72	91.54
CCRSNet	85.96	92.92	92.29	96.42	87.19	92.90
RCTUnet	**88.11**	**93.99**	**93.66**	**97.19**	**90.32**	93.29

The highest metric values for each group are marked in bold. SegNet: segmentation network; FCN: fully convolutional network; CCRSNet: cross-channel residual and spatial fusion network. MIoU: mean intersection over union.

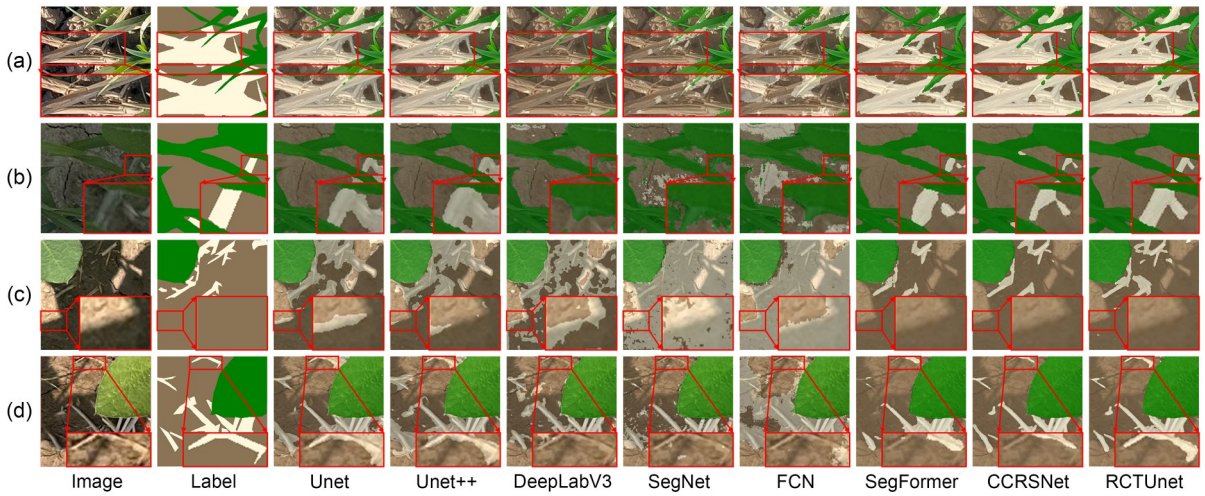

Fig. 8 Segmentation results of Unet, Unet++, DeepLabV3, segmentation network (SegNet), fully convolutional network (FCN), SegFormer, cross-channel residual and spatial fusion network (CCRSNet), and RCTUnet under complex lighting scenarios. (a) High crop residue coverage; (b) Low lighting scene; (c) Low crop residue coverage; (d) Fragmented crop residues.

capturing the fine-grained texture and boundary information of fragmented crop residues under strong illumination or occlusion.

To further assess the computational efficiency and deployment feasibility of the proposed network, a comparative analysis was conducted using three representative indicators: the number of parameters (Params), floating-point operations (FLOPs), and inference speed (frames per second (FPS)). These metrics collectively reflect the trade-off between model accuracy and computational efficiency, which is crucial for real-time applications in field environments. The results are summarized in Table 2.

Table 2 Computational complexity and inference efficiency of different segmentation models

Model	Params (M)	FLOPs (G)	Inference speed (FPS)
Unet	31.04	54.75	169.17
Unet++	32.16	34.91	151.23
DeepLabV3	30.67	30.76	284.52
SegNet	29.45	40.21	183.69
FCN	18.64	25.50	320.23
SegFormer	37.15	16.93	128.83
CCRSNet	44.02	23.09	58.91
RCTUnet	65.74	20.62	82.16

FLOPs: floating-point operations; FPS: frames per second; M: million; G: giga; SegNet: segmentation network; FCN: fully convolutional network; CCRSNet: cross-channel Residual and spatial fusion network.

The proposed RCTUnet achieved a favorable balance between representational capacity and computational efficiency. Although the parameter count (65.74 million (M)) is higher than those of conventional convolutional architectures such as Unet (31.04 M) and DeepLabV3 (30.67 M), the overall computational cost remains moderate, with only 20.62 giga (G) FLOPs, which is comparable to SegFormer (16.93 G) and substantially lower than Unet (54.75 G). This result indicates that the proposed hybrid architecture effectively enhances feature expressiveness without introducing an excessive computational burden.

In terms of inference speed, RCTUnet achieves 82.16 FPS on an NVIDIA RTX 3080 GPU, which is about 40% faster than CCRSNet (58.91 FPS) and significantly faster than most attention-based segmentation networks. This improvement stems mainly from two architectural optimizations: (1) GCFM limits transformer-based self-attention to low-resolution feature maps, thereby reducing the quadratic growth in computation; and (2) the channel-adaptive fusion mechanism efficiently integrates multi-scale contextual information while suppressing redundant convolutional operations. Overall, the results show that RCTUnet attains an optimal balance among segmentation accuracy, model complexity, and inference efficiency, achieving real-time processing capability while maintaining high segmentation precision. These characteristics confirm the practical feasibility of deploying RCTUnet in

real-world agricultural monitoring scenarios, where both computational efficiency and boundary-level accuracy are essential.

3.2 Comparative experiments

3.2.1 Comparison of different backbone networks

We integrated VGG16 and ResNet50 backbone architectures into the encoder of the Unet framework. Table 3 summarizes the segmentation results for each backbone. The baseline Unet model achieved an MIoU of 82.52% and an accuracy of 90.75%. Its symmetric encoder-decoder structure showed strong performance in soil-background identification (recall=93.59%), yet had clear limitations in crop residue detection (recall=82.65%). Replacing the encoder with VGG16 led to improvements in overall segmentation performance, with the MIoU increasing to 83.06% and accuracy to 91.27%. Notably, this configuration achieved the highest recall for the crop class (recall=96.81%). However, its contribution to crop residue recognition remained limited, yielding only a marginal improvement (recall=83.85%). When ResNet50 was adopted as the backbone network, the model achieved an MIoU of 84.62% and an overall accuracy of 92.18%. The residual connection mechanism significantly enhanced the model's ability to recognize straw residues (recall=85.05%, representing a 2.4% improvement over the baseline), while maintaining stable crop recognition performance (recall>95%). The strengthened deep feature extraction capability notably improved segmentation accuracy in the straw-soil boundary regions. Although a slight decrease was observed in soil background recall (92.18%), the concurrent improvement in precision (91.79%) indicates that the model effectively reduced false-positive predictions.

3.2.2 Comparison of attention modules

The model using ResNet50 as the backbone, hereafter referred to as RUnet, achieved the best overall performance (Table 3). We used RUnet as the baseline and incorporated different attention mechanisms into the decoder. The attention modules evaluated included CBAM, global attention mechanism (GAM), and simple attention module (SimAM). Three attention-enhanced variants outperformed the baseline model in terms of MIoU, with CBAM, GAM, and SimAM yielding improvements of 0.95%, 0.39%, and 0.24%, respectively (Table 4). The CBAM-enhanced model (RUnet) exhibited the best performance across all metrics, achieving a crop residue recall of 87.37%.

3.2.3 Ablation study

We conducted a series of ablation experiments to verify the effectiveness of each proposed module. When ResNet50 was used as the encoder backbone, the MIoU increased from 82.52% to 84.62%, indicating that its deep residual structure enhanced the representation of slender and low-contrast straw textures through multi-level feature fusion (Table 5). The introduction of the CBAM attention mechanism further improved the MIoU to 85.57%. Although the overall gain was moderate, CBAM adaptively strengthened

Table 3 Performance metrics for different backbone networks

Model	MIoU (%)	Accuracy (%)	Precision (%)	Recall (%)		
				Crop	Residue	Soil
Unet	82.52	90.75	89.75	95.97	82.65	**93.59**
Unet+VGG16	83.06	91.27	91.02	**96.81**	83.85	92.24
Unet+ResNet50	**84.62**	**92.18**	**91.79**	96.79	**85.05**	92.18

The highest metric values for each group are marked in bold. MIoU: mean intersection over union.

Table 4 Performance metrics for different attention modules

Model	MIoU (%)	Accuracy (%)	Precision (%)	Recall (%)		
				Crop	Residue	Soil
RUnet	84.62	92.18	91.79	96.79	85.05	92.18
RUnet+CBAM	**85.57**	**92.66**	**92.14**	96.64	**87.37**	92.33
RUnet+GAM	85.01	92.43	91.93	96.52	85.95	92.34
RUnet+SimAM	84.86	92.33	91.86	**96.84**	85.53	**92.60**

The highest metric values for each group are marked in bold. MIoU: mean intersection over union; CBAM: convolutional block attention module; GAM: global attention mechanism; SimAM: simple attention module.

the response to straw edges and fine structural details through joint channel-spatial attention, effectively reducing missed detections of fragmented targets.

After integrating the GCFM module, the MIoU rose significantly to 88.11%. The transformer-based self-attention mechanism within GCFM compensates for the limited global context modeling capability of CNNs, enabling the model to better capture the spatial continuity and holistic morphology of slender straw residues.

Overall, the synergy among the three modules substantially enhanced the robustness and boundary preservation of the proposed RCTUnet when handling fragmented and morphologically complex straw targets. The final RCTUnet model achieved a 7.67% improvement in recall for the straw class compared with the baseline Unet (Table 5), confirming the superior effectiveness of the proposed architecture for straw residue extraction.

We conducted visual analyses of the ablation study predictions to evaluate the individual contributions of each module (Fig. 9). The Unet model performed poorly under complex lighting interference (Figs. 9a–9c), failing to identify crop residue features robustly.

Table 5 Ablation experiments results

Unet	ResNet50	CBAM	GCFM	MIoU (%)	Accuracy (%)	Precision (%)	Recall (%)		
							Crop	Residue	Soil
√				82.52	90.75	89.75	95.97	82.65	**93.59**
√	√			84.62	92.18	91.79	96.79	85.05	92.18
√	√	√		85.57	92.66	92.14	96.64	87.37	92.33
√	√	√	√	**88.11**	**93.99**	**93.66**	**97.17**	**90.32**	93.29

The highest metric values for each group are marked in bold. CBAM: convolutional block attention module; GCFM: global context fusion module; MIoU: mean intersection over union.

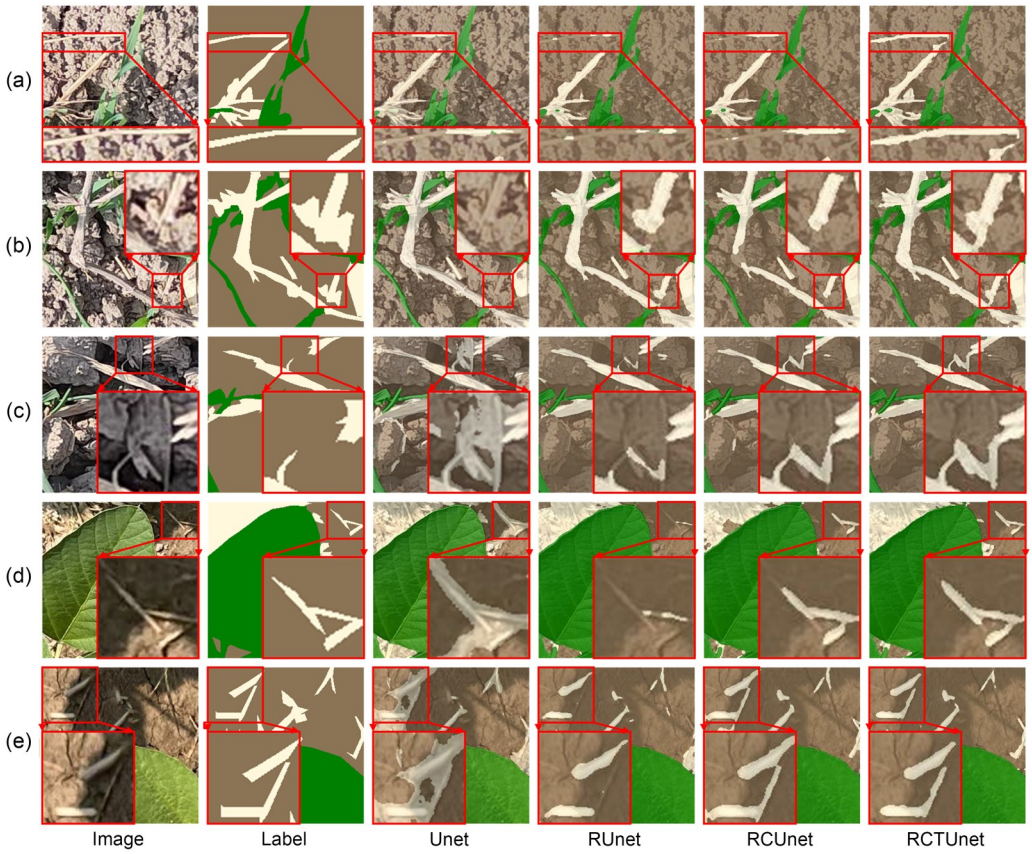

Fig. 9 Segmentation results of the ablation study. (a, b) High lighting scene; (c) Medium lighting scene; (d, e) Low lighting scene.

It misclassified adjacent soil regions as crop residue (Figs. 9d and 9e). The use of ResNet50 significantly improved the model's ability to extract fine-scale crop residue structures, aided by residual connections and hierarchical feature fusion (Fig. 9e). Adding the CBAM module further enhanced the model's sensitivity to crop residue features by combining channel and spatial attention, resulting in more precise crop residue-soil boundaries (Fig. 9d). Integrating the GCFM module effectively mitigated misclassification under extreme lighting conditions (Figs. 9a and 9b) and improved the recognition of ambiguous crop residue regions in complex scenes (Fig. 9c).

3.2.4 Model interpretability and class-wise performance analysis

To further enhance model transparency and provide a fine-grained understanding of its performance across categories and rotation types, both visual interpretability analysis and class-wise quantitative evaluation were conducted.

To further validate the segmentation effectiveness of RCTUnet, a class-wise comparison with the baseline Unet was conducted in terms of IoU and balanced F score (F1-score) (Table 6). The proposed RCTUnet achieved consistent improvements across all categories, with particularly notable gains in the residue class. Specifically, the IoU of the residue category increased from 74.05% to 83.04%, and the F1-score from 85.09% to 88.08%, indicating that the proposed modules significantly enhanced the model's ability to identify thin and fragmented straw regions.

In the crop and soil categories, the IoU improved by 1.10 and 6.65 percentage points, respectively, accompanied by moderate gains in F1-score. These results suggest that while both Unet and RCTUnet effectively capture the large, homogeneous crop regions, RCTUnet shows superior robustness in boundary delineation and small-object segmentation. Overall, the improvements across all categories confirm that the proposed enhancements, particularly the GCFM module with transformer-based global context modeling, substantially strengthen the model's feature representation and contextual awareness, leading to more precise and reliable residue segmentation under complex field conditions.

Fig. 10 compares the percentage-based confusion matrices of Unet and RCTUnet for the three surface classes (crop, residue, and soil). Overall, both models performed similarly on crop and soil, while more notable differences appeared for the residue class.

Crop: both models achieved similar true positive ratios of about 34% and true negative ratios of around 62.9%. The false negative rate of RCTUnet was slightly lower (1.0% compared to 1.4% for Unet), indicating a modest reduction in missed crop pixels and a marginal improvement in crop detection accuracy.

Residue: The Unet model attained a true positive of 21.6% and a true negative of 70.9%, whereas RCTUnet increased the true negative to 74.9% while slightly reducing the true positive to 20.8%. Simultaneously, the false positive rate decreased significantly from 6.1% to 2.0%, while the false negative rate rose modestly from 1.5% to 2.2%. These results suggest that RCTUnet substantially reduces false detections—misclassifying fewer non-residue regions as residue—while exhibiting slightly more conservative predictions that marginally increase missed detections. Overall, the model shows stronger discriminative capability and improved precision in identifying residue regions.

Soil: the true positive ratio of RCTUnet increased from 34.9% (Unet) to 38.5%, while the false negative ratio decreased from 6.3% to 2.8%, reflecting a notable improvement in soil classification accuracy.

In summary, RCTUnet markedly reduced false positives in the residue category and enhanced true positive detection in the soil category. These findings indicate that the improved model exhibits greater robustness and precision in distinguishing crop residue from the surrounding soil background.

Table 6 Category promotion in RCTUnet versus Unet

Category	IoU (%)		ΔIoU (%)	F1 (%)		ΔF1 (%)
	Unet	RCTUnet		Unet	RCTUnet	
Crop	92.35	93.45	+1.10	95.99	96.02	+0.03
Residue	74.05	83.04	+8.99	85.09	88.08	+2.99
Soil	81.17	87.82	+6.65	89.61	92.31	+2.70

IoU: intersection over union; F1: balanced F score.

Fig. 10 Confusion matrix. FP: false positive; TP: true positive; TN: true negative; FN: false negative.

To validate the model's capability in handling straw texture and boundary constraints, we used gradient-weighted class activation mapping (Grad-CAM) contrast visualization analysis (Fig. 11). The heatmap results reveal that the improved model (RCTUnet) exhibited a more concentrated distribution of high responses within the straw regions, with particularly pronounced activation at the points where straws intertwined and along their boundaries. In contrast, the responses from the original Unet were relatively dispersed, with some concentrated in non-target areas (such as plant leaves or soil). This outcome shows that RCTUnet significantly enhances the model's ability to focus on fine-grained textures and structural boundaries, thereby achieving stronger semantic focus during the feature extraction stage.

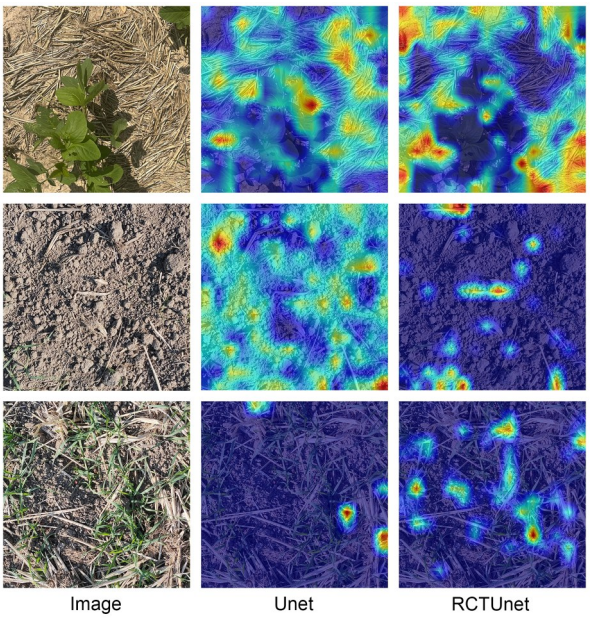

Fig. 11 Gradient-weighted class activation mapping (Grad-CAM) visualization analysis.

3.3 Crop residue segmentation and coverage information extraction based on RCTUnet

3.3.1 Application of crop residue segmentation

We evaluated the segmentation performance of the proposed RCTUnet model based on five farmland scenarios (Fig. 12). Unet misidentified adjacent soil regions as crop residues in the fragmented crop residue scenario (Figs. 12a and 12b) and small portions of soil connected to crop residues as crop residues in the high crop residue coverage scenario (Fig. 12c). Additionally, our results suggest that the Unet model was affected by complex lighting conditions (Figs. 12d and 12e). RCTUnet showed superior segmentation performance across various complex farmland scenarios (Fig. 12).

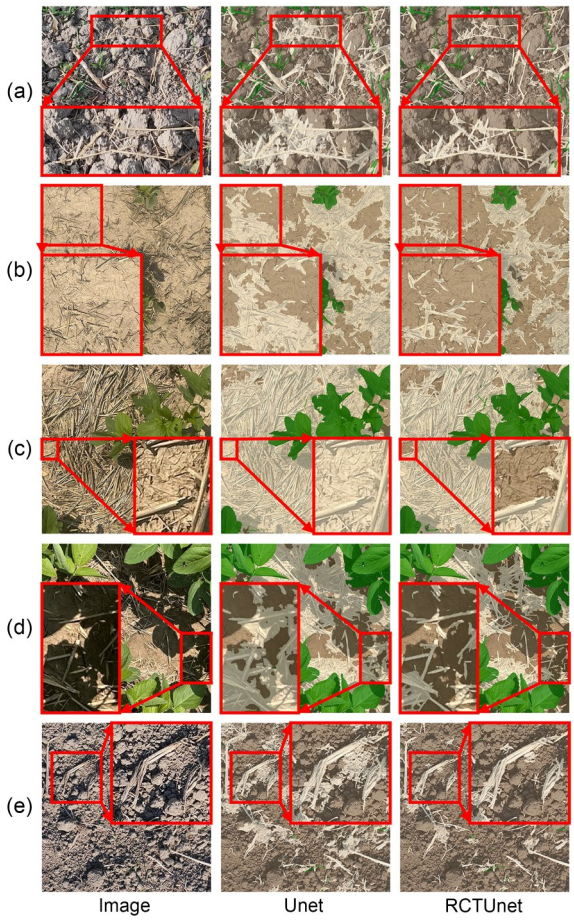

Fig. 12 Segmentation prediction results of Unet and RCTUnet. (a) Low crop residue coverage scenario; (b) Fragmented crop residues scenario; (c) High crop residue coverage scenario; (d) Complex lighting scenario; (e) Complex lighting+ fragmented crop residues scenario.

3.3.2 Crop residue coverage information extraction

We calculated the fractional vegetation coverage (FVC), CRC, and bare soil (BS) of each image. The results indicate that the RCTUnet model demonstrated superior fitting ability across all three coverage estimation tasks (Fig. 13). For CRC estimation, R^2=0.943, and RMSE=4.875, indicating strong crop residue recognition capability. The Unet model also performed well in FVC estimation (R^2=0.990, RMSE=2.181, Fig. 14). In contrast, the prediction accuracy for both CRC (R^2=0.871, RMSE=8.941) and BS (R^2=0.918, RMSE=9.266) declined significantly. Overall, the prediction accuracy of Unet was lower than that of RCTUnet, with a reduction in CRC prediction error of about 45.5%.

4 Discussion

4.1 Advantages of the RCTUnet model

RCTUnet demonstrated exceptionally superior performance in extracting fragmented crop residues and performing segmentation under complex lighting conditions. The model achieved an MIoU of 88.11% and an accuracy of 93.99%, representing improvements of 5.59 and 3.24 percentage points (Table 5), respectively, over the traditional Unet. In the crop residue category, RCTUnet achieved a recall of 90.32%, representing a 7.67% improvement over Unet. RCTUnet incorporates ResNet50 as the encoder backbone network, combined with a multi-level feature fusion strategy, which significantly enhances the model's ability to extract edge and texture features of fragmented crop residues. This approach aligns with the method used by Han et al. (2025) in crack detection tasks, where ResNet50 was used to extract microstructural features, further validating the general applicability of residual connections in complex object recognition.

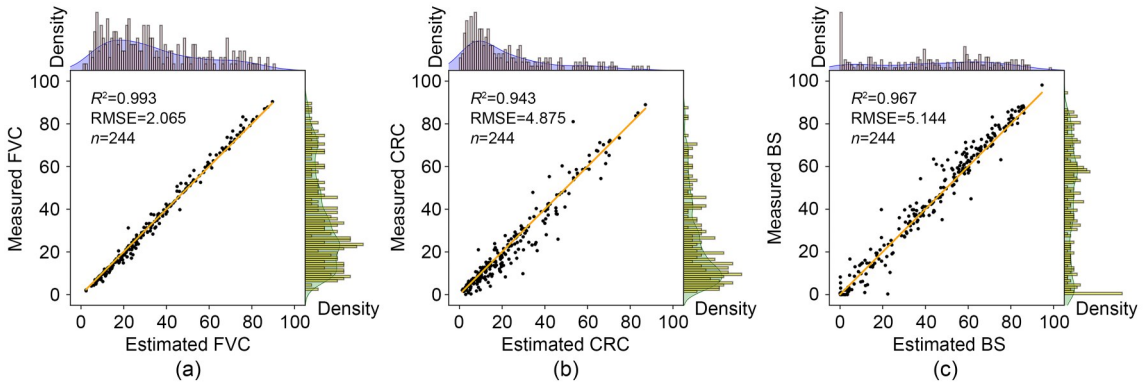

Fig. 13 Fractional vegetation coverage (FVC), crop residue cover (CRC), and bare soil (BS) estimation results from RCTUnet. (a) FVC; (b) CRC; (c) BS.

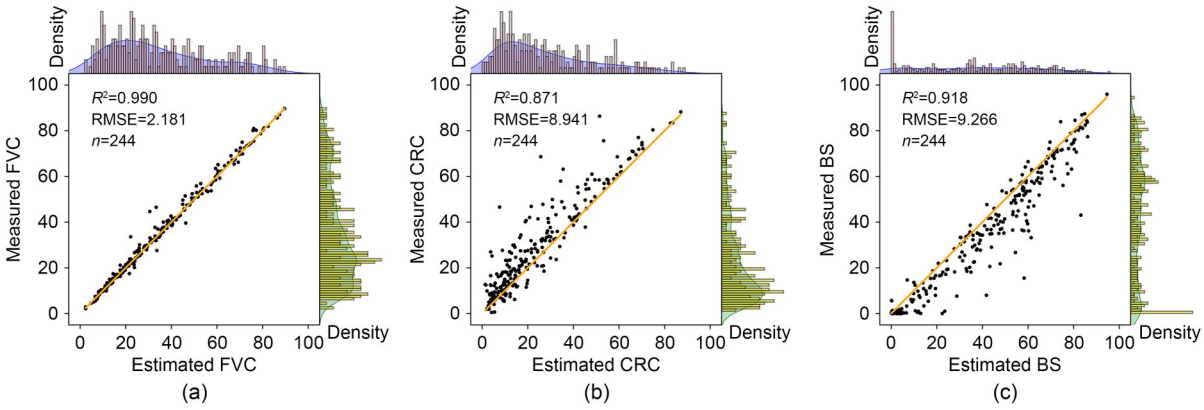

Fig. 14 Fractional vegetation coverage (FVC), crop residue cover (CRC), and bare soil (BS) estimation results from Unet. (a) FVC; (b) CRC; (c) BS.

The CBAM attention module, through its dual weighting mechanism in both the channel and spatial dimensions, effectively strengthens the expression of salient features in crop residue regions while suppressing the interference from background soil and vegetation. The GCFM module, by introducing a self-attention mechanism to model global context information, compensates for the traditional convolutional structure's lack of capability in modeling non-local semantic relationships. This improvement is particularly evident in regions with significant lighting variation or discontinuous crop residue distributions, where the model shows enhanced stability.

Compared to traditional remote sensing methods, the RCTUnet model also shows significant advantages in CRC estimation. Traditional optical remote sensing-based methods for estimating CRC rely mainly on spectral indices and linear spectral mixture analysis. Previous studies based on Sentinel-2 Multi-Spectral Instrument (MSI) images have reported a model with an RMSE of 8.3 for CRC estimation (Yue et al., 2023), which is higher than the RMSE of RCTUnet. However, due to their limited spatial resolution, such methods struggle to accurately capture the fragmented morphology and boundary features of crop residues in farmland. In contrast, RCTUnet-based crop residue coverage estimation in this study achieved an RMSE of 4.875 (Fig. 13), significantly outperforming the traditional remote sensing methods. This result underscores the potential of deep learning methods for recognizing micro-level features in farmland. Particularly in small target extraction, RCTUnet effectively identifies crop residue texture features that are difficult for traditional remote sensing methods to capture by integrating local edge information with global semantic modeling, thereby demonstrating a stronger segmentation capability.

4.2 Limitations of the RCTUnet model

Although RCTUnet performs well in farmland crop residue segmentation tasks, several limitations remain, which warrant further improvement and optimization in future research. The multi-module integrated hybrid structure used by RCTUnet enhances feature representation capabilities but significantly increases the overall complexity of the model. In particular, the self-attention mechanism introduced in the GCFM module results in a computational complexity that grows quadratically with the size of the input feature map, thus imposing higher memory and computational resource demands during both the training and inference phases. In agricultural applications with limited computational resources or when deploying the model on edge devices, this issue may become a critical bottleneck limiting the model's real-time applicability.

Despite the high performance of RCTUnet in crop residue recognition (with a recall of 90.32%) (Table 5), there remains a need for improvement when considering the practical needs of agricultural monitoring applications. The model may still fail to detect crop residues when dealing with blurry edges or regions with soil textures that resemble those of crop residues, which could impact the accuracy of subsequent crop residue coverage estimation. This limitation is especially relevant in scenarios where precise quantification of crop residues is necessary to guide agricultural management practices, as it could restrict the model's practical effectiveness. Furthermore, the model's performance

is somewhat constrained by the coverage and representativeness of the training data. Although the dataset used in this study encompassed four typical crop rotation systems, the model may still exhibit a decline in generalization performance when applied to different geographical regions, crop varieties, or soil types.

Future research should focus on model lightweighting, improving recognition accuracy, and enhancing cross-region generalization ability. To reduce the computational overhead of RCTUnet, lightweight techniques such as model pruning and knowledge distillation could be explored to improve deployment efficiency on edge devices. Additionally, constructing a collaborative monitoring framework that integrates deep learning with optical remote sensing is expected to further enhance crop residue recognition capabilities. To enhance the model's adaptability across regions, expanding the training sample pool with multi-source data and incorporating transfer learning strategies could help improve its generalization performance.

5 Conclusions

In this study, we designed the RCTUnet model to address issues such as small target detection failures and sensitivity to lighting variation in crop residue image segmentation. Building upon the Unet architecture, the model integrates the ResNet50 backbone network to enhance multi-scale feature extraction, incorporates the CBAM attention mechanism to strengthen responses to key regions, and introduces the GCFM module to improve the model's global context modeling capabilities. The RCTUnet enables accurate recognition and segmentation of crop residue targets in complex farmland backgrounds. The main conclusions of this study are as follows:

(1) Compared to classical segmentation models such as Unet, Unet++, DeepLabV3, SegNet, and FCN, RCTUnet achieved more accurate "crop-residue-soil" segmentation results. The accuracy improved by 3.24, 3.42, 4.88, 8.28, and 6.05 percentage points, respectively (Table 1).

(2) By leveraging multi-scale feature extraction, a joint spatial-channel attention mechanism, and global context modeling, RCTUnet significantly alleviates the challenges faced by traditional CNNs in extracting fragmented crop residue targets and dealing with lighting instability. The proposed RCTUnet model demonstrated superior "crop-residue-soil" segmentation performance, with residue recall improving by 7.67, 7.97, 14.09, 27.05, and 16.91 percentage points compared to Unet, Unet++, DeepLabV3, SegNet, and FCN, respectively (Table 1).

(3) The RCTUnet model achieved superior CRC extraction accuracy (RMSE=4.875; Fig. 13), with CRC extraction accuracy improving by 45.5% over Unet (RMSE=8.941; Fig. 14).

Data availability statement

The datasets generated and analyzed during this study, including the field-collected data, are available from the first author upon reasonable request.

Acknowledgments

This study was supported by the National Natural Science Foundation of China (No. 42101362), the Natural Science Foundation of Henan Province (No. 252300421158), the Shenzhen Science and Technology Program (No. JCYJ20220530162001003), and the Science and Technology Development Program of Henan Province (No. 242300421639), China.

Author contributions

Ting LI performed the experiments, analyzed the data, and wrote and edited the manuscript. Jibo YUE and Yang LIU performed the establishment of models. Haikuan FENG, Meiyan SHU, and Hao YANG contributed to the study design. Yuanyuan FU, Xin XU, and Yinghao LIN contributed to the data analysis. Hongbo QIAO, Wei GUO, Xinming MA, and Lei SHI contributed to the writing and editing of the manuscript. All authors have read and approved the final manuscript, and therefore, have full access to all the data in the study and take responsibility for the integrity and security of the data.

Compliance with ethics guidelines

Ting LI, Yang LIU, Haikuan FENG, Meiyan SHU, Hao YANG, Yuanyuan FU, Xin XU, Yinghao LIN, Hongbo QIAO, Wei GUO, Xinming MA, Lei SHI, and Jibo YUE declare that they have no conflicts of interest.

This article does not contain any studies with human participants or animals performed by any of the authors.

Declaration on the use of generative AI tools

During the preparation of this work, the authors used ChatGPT to improve language and readability, and to check for grammatical errors. After using this tool, the authors reviewed and edited the content as needed and take full responsibility for the content of the publication.

References

Beeche C, Singh JP, Leader JK, et al., 2022. Super U-Net: a modularized generalizable architecture. *Pattern Recognit*, 128:108669. https://doi.org/10.1016/j.patcog.2022.108669

Brosch T, Tang LYW, Yoo Y, et al., 2016. Deep 3D convolutional encoder networks with shortcuts for multiscale feature integration applied to multiple sclerosis lesion segmentation. *IEEE Trans Med Imaging* 35(5):1229-1239. https://doi.org/10.1109/TMI.2016.2528821

Delandmeter M, Colinet G, Pierreux J, et al., 2024. Combining field measurements and process-based modelling to analyse soil tillage and crop residues management impacts on crop production and carbon balance in temperate areas. *Soil Use Manage*, 40(3):e13098. https://doi.org/10.1111/sum.13098

de Obade Paul Obade V, Gaya C, 2020. Mapping tillage practices using spatial information techniques. *Environ Manage*, 66(4):722-731. https://doi.org/10.1007/s00267-020-01335-z

Ding YL, Zhang HY, Wang ZQ, et al, 2020. A comparison of estimating crop residue cover from Sentinel-2 data using empirical regressions and machine learning methods. *Remote Sens*, 12(9):1470. https://doi.org/10.3390/rs12091470

Du L, Lu ZC, Li DL, 2022. Broodstock breeding behaviour recognition based on Resnet50-LSTM with CBAM attention mechanism. *Comput Electron Agric*, 202:107404. https://doi.org/10.1016/j.compag.2022.107404

Feng YG, Han B, Wang XC, et al., 2024. Self-supervised transformers for unsupervised SAR complex interference detection using canny edge detector. *Remote Sens*, 16(2):306. https://doi.org/10.3390/rs16020306

Gao GF, Zhang SX, Shen JN, et al., 2024. Segmentation and proportion extraction of crop, crop residues, and soil using digital images and deep learning. *Agriculture*, 14(12): 2240. https://doi.org/10.3390/agriculture14122240

Gao LL, Zhang C, Yun WJ, et al., 2022. Mapping crop residue cover using Adjust Normalized Difference Residue Index based on Sentinel-2 MSI data. *Soil Tillage Res*, 220: 105374. https://doi.org/10.1016/j.still.2022.105374

Gao PP, Song Y, Song MH, et al., 2022. Extract nanoporous gold ligaments from SEM images by combining fully convolutional network and Sobel operator edge detection algorithm. *Scr Mater*, 213:114627. https://doi.org/10.1016/j.scriptamat.2022.114627

Han XJ, Cheng QB, Chen QZ, et al., 2025. Deep learning-based multi-category disease semantic image segmentation detection for concrete structures using the Res-Unet model. *J Civil Struct Health Monit*, 15(5):1369-1380. https://doi.org/10.1007/s13349-024-00893-8

Hively WD, Lamb BT, Daughtry CST, et al., 2018. Mapping crop residue and tillage intensity using WorldView-3 satellite shortwave infrared residue indices. *Remote Sens*, 10(10):1657. https://doi.org/10.3390/rs10101657

Jin ZY, Hong WJ, Wang YR, et al., 2025. A transformer-based symmetric diffusion segmentation network for wheat growth monitoring and yield counting. *Agriculture*, 15(7): 670. https://doi.org/10.3390/agriculture15070670

Laamrani A, Pardo Lara R, Berg AA, et al., 2018. Using a mobile device "app" and proximal remote sensing technologies to assess soil cover fractions on agricultural fields. *Sensors*, 18(3):708. https://doi.org/10.3390/s18030708

Li K, Zhang YJ, Wang TF, et al., 2025. FreqUNet: a lightweight dual-branch network with frequency-aware decomposition for retinal vessel segmentation. *Expert Syst Appl*, 287:128124. https://doi.org/10.1016/j.eswa.2025.128124

Li L, Li J, Lv CX, et al., 2021. Maize residue segmentation using Siamese domain transfer network. *Comput Electron Agric*, 187:106261. https://doi.org/10.1016/j.compag.2021.106261

Li YW, Zhao HS, Qi XJ, et al., 2023. Fully convolutional networks for panoptic segmentation with point-based supervision. *IEEE Trans Pattern Anal Mach Intell*, 45(4):4552-4568. https://doi.org/10.1109/tpami.2022.3200416

Liu J, Qiu TY, Peñuelas J, et al., 2023. Crop residue return sustains global soil ecological stoichiometry balance. *Global Change Biol*, 29(8):2203-2226. https://doi.org/10.1111/gcb.16584

Luo CH, Chen JP, Guo SX, et al., 2022. Development and application of a remote monitoring system for agricultural machinery operation in conservation tillage. *Agriculture*, 12(9):1460. https://doi.org/10.3390/agriculture12091460

Mahmood MT, Ucan ON, 2025. Data and image processing for intelligent glaucoma detection and optic disc segmentation using deep convolutional neural network architecture. *Discov Comput*, 28:73. https://doi.org/10.1007/s10791-025-09587-1

Pacheco A, McNairn H, 2010. Evaluating multispectral remote sensing and spectral unmixing analysis for crop residue mapping. *Remote Sens Environ*, 114(10):2219-2228. https://doi.org/10.1016/j.rse.2010.04.024

Qiang J, Liu WJ, Li XX, et al., 2023. Detection of citrus pests in double backbone network based on single shot multibox detector. *Comput Electron Agric*, 212:108158. https://doi.org/10.1016/j.compag.2023.108158

Serbin G, Hunt Jr ER, Daughtry CST, et al., 2009. An improved ASTER index for remote sensing of crop residue. *Remote Sens*, 1(4):971-991. https://doi.org/10.3390/rs1040971

Shahi TB, Dahal S, Sitaula C, et al., 2023. Deep learning-based weed detection using UAV images: a comparative study. *Drones*, 7(10):624. https://doi.org/10.3390/drones7100624

Shang CJ, Zhang D, Yang Y, 2021. A gradient-based method for multilevel thresholding. *Expert Syst Appl*, 175:114845. https://doi.org/10.1016/j.eswa.2021.114845

She QS, Sun SK, Ma YL, et al., 2025. LUCF-Net: lightweight

U-shaped cascade fusion network for medical image segmentation. *IEEE J Biomed Health Inform*, 29(3):2088-2099.
https://doi.org/10.1109/jbhi.2024.3506829

Shi JF, Ji SS, Jin HY, et al., 2025. Multi-feature lightweight DeeplabV3+ network for polarimetric SAR image classification with attention mechanism. *Remote Sens*, 17(8):1422.
https://doi.org/10.3390/rs17081422

Song HH, Wang JQ, Bei JL, et al., 2024. Modified snake optimizer based multi-level thresholding for color image segmentation of agricultural diseases. *Expert Syst Appl*, 255:124624.
https://doi.org/10.1016/j.eswa.2024.124624

Song WY, Nie FX, Wang C, et al., 2024. Unsupervised multiscale hybrid feature extraction network for semantic segmentation of high-resolution remote sensing images. *Remote Sens*, 16(20):3774.
https://doi.org/10.3390/rs16203774

Tao WC, Xie ZX, Zhang Y, et al., 2021. Corn residue covered area mapping with a deep learning method using Chinese GF-1 B/D high resolution remote sensing images. *Remote Sens*, 13(15):2903.
https://doi.org/10.3390/rs13152903

Wang FY, Lv CX, Jiang HL, et al., 2025. Efficient detection of corn straw coverage in complex agricultural scenarios. *Comput Electron Agric*, 235:110338.
https://doi.org/10.1016/j.compag.2025.110338

Wang GD, Bai D, Lin HF, et al., 2024. FireViTNet: a hybrid model integrating ViT and CNNs for forest fire segmentation. *Comput Electron Agric*, 218:108722.
https://doi.org/10.1016/j.compag.2024.108722

Wang GZ, Wang JP, Zou XY, et al., 2019. Estimating the fractional cover of photosynthetic vegetation, non-photosynthetic vegetation and bare soil from MODIS data: assessing the applicability of the NDVI-DFI model in the typical Xilingol grasslands. *Int J Appl Earth Obs Geoinf*, 76:154-166.
https://doi.org/10.1016/j.jag.2018.11.006

Wang YY, Gao XB, Sun Y, et al., 2025. Semantic segmentation-based conservation tillage corn straw return cover type recognition. *Comput Electron Agric*, 229:109792.
https://doi.org/10.1016/j.compag.2024.109792

Xie EZ, Wang WH, Yu ZD, et al., 2021. SegFormer: simple and efficient design for semantic segmentation with Transformers. Advances in Neural Information Processing Systems, 34:12077-12090.
https://doi.org/10.48550/arXiv.2105.15203

Xu JY, Zhou SY, Xu AJ, et al., 2022. Automatic scoring of postures in grouped pigs using depth image and CNN-SVM. *Comput Electron Agric*, 194:106746.
https://doi.org/10.1016/j.compag.2022.106746

Yang H, Sun H, Wang K, et al., 2025. Enhanced farmland extraction from Gaofen-2: multi-scale segmentation, SVM integration, and multi-temporal analysis. *Agriculture*, 15(10):1073.
https://doi.org/10.3390/agriculture15101073

Yu FH, Bai JC, Fang JY, et al, 2024. Integration of a parameter combination discriminator improves the accuracy of chlorophyll inversion from spectral imaging of rice. *Agric Commun*, 2(3):100055.
https://doi.org/10.1016/j.agrcom.2024.100055

Yue JB, Tian QJ, 2020. Estimating fractional cover of crop, crop residue, and soil in cropland using broadband remote sensing data and machine learning. *Int J Appl Earth Obs Geoinf*, 89:102089.
https://doi.org/10.1016/j.jag.2020.102089

Yue JB, Tian QJ, Tang SF, et al., 2019. A dynamic soil endmember spectrum selection approach for soil and crop residue linear spectral unmixing analysis. *Int J Appl Earth Obs Geoinf*, 78:306-317.
https://doi.org/10.1016/j.jag.2019.02.001

Yue JB, Tian QJ, Liu Y, et al., 2023. Mapping cropland rice residue cover using a radiative transfer model and deep learning. *Comput Electron Agric*, 215:108421.
https://doi.org/10.1016/j.compag.2023.108421

Yue JB, Li T, Feng HK, et al., 2024. Enhancing field soil moisture content monitoring using laboratory-based soil spectral measurements and radiative transfer models. *Agric Commun*, 2(4):100060.
https://doi.org/10.1016/j.agrcom.2024.100060

Zhang TT, Hu DN, Wu CX, et al., 2023. Large-scale apple orchard mapping from multi-source data using the semantic segmentation model with image-to-image translation and transfer learning. *Comput Electron Agric*, 213:108204.
https://doi.org/10.1016/j.compag.2023.108204

Zhang WQ, Li WJ, Wang C, et al., 2025. A novel index for mapping crop residue covered cropland using remote sensing data. *Comput Electron Agric*, 231:109995.
https://doi.org/10.1016/j.compag.2025.109995

Zhao JL, Li Z, Lei Y, et al., 2023. Application of UAV RGB images and improved PSPNet network to the identification of wheat lodging areas. *Agronomy*, 13(5):1309.
https://doi.org/10.3390/agronomy13051309

Zheng XX, Cao F, Ou JY, et al., 2024. RSU-Net: a new method for fine classification of corn residue coverage in black soil area using Chinese GF-1B PMS image. *Ecol Front*, 44(6):1259-1268.
https://doi.org/10.1016/j.ecofro.2024.08.001

Zhou DY, Li M, Li Y, et al., 2020. Detection of ground straw coverage under conservation tillage based on deep learning. *Comput Electron Agric*, 172:105369.
https://doi.org/10.1016/j.compag.2020.105369

Zhu QL, Wang K, Liang D, et al., 2025. WLUSNet: a lightweight wheat lodging segmentation network based on UAV image. *Comput Electron Agric*, 237:110587.
https://doi.org/10.1016/j.compag.2025.110587