Journal of Agricultural Engineering

Journal of Agricultural Engineering 2026; LVII:1954
doi:10.4081/jae.2025.1954

Deep learning based agricultural remote sensing image segmentation: a review

Qinghua Ren,[1] Yanlin Wu,[1] Qingshuai Zeng,[1] Ning Yang,[2]

[1]School of Computer Science and Communication Engineering; [2]School of Electrical and Information Engineering, Jiangsu University, China

Abstract

Agricultural remote sensing image segmentation, which involves classifying each pixel of an image into a specific category, has recently been driven by deep learning methods due to their powerful feature extraction capabilities. This paper presents a comprehensive review of deep learning-based image segmentation techniques for agricultural remote sensing, along with an overview of current challenges and emerging research trends. First, it outlines the characteristics of agricultural remote sensing tasks and the requirements for remote sensing image acquisition and processing, providing an in-depth analysis of the nature of agricultural remote sensing data. Next, it systematically reviews the evolution of deep learning-based methods, with a focus on summarizing segmentation network architectures, including convolution-based models, transformer-based models, hybrid architectures, lightweight models, and vision-language models. Moreover, it discusses several deep learning paradigms designed for annotation-efficient scenarios, including semi-supervised, weakly supervised, self-supervised, and transfer learning. Then, it offers an in-depth analysis of key challenges, such as data annotation, computational cost, and model generalization. Finally, it summarizes the latest advances in deep learning for agricultural remote sensing image segmentation and outlines potential future research directions, aiming to provide technical references that promote the practical application and successful deployment of deep learning in this critical domain.

Key words: ecision agriculture; convolutional neural networks (CNNs); transformers; UAV remote sensing; satellite imagery; annotation efficiency.

Correspondence: Ning Yang, School of Electrical and Information Engineering, Jiangsu University, China. E-mail: yangn@ujs.edu.cn

Introduction

Agriculture is the foundational industry for human survival and development, playing a critical role in ensuring food security and promoting sustainable development (Luo *et al.,* 2016). The rapid development of remote sensing (RS) technology is profoundly driving the transformation of agriculture towards data-driven and intelligent models (Sishodia *et al.,* 2020). Leveraging multi-sensor systems mounted on airborne and space-based platforms-including multispectral, hyperspectral, and thermal infrared sensors-remote sensing systems can conduct comprehensive monitoring of agricultural environments across large areas, at multiple spatial resolutions, and with high temporal frequency (Tianxiang *et al.,* 2024). The high-resolution remote sensing imagery obtained contains rich information such as crop growth status, soil moisture dynamics, and pest and disease occurrence, laying a data foundation for precision agriculture that was previously unattainable (Zhang *et al.,* 2020; Chandra *et al.,* 2024). The process of data acquisition and segmentation is shown as follows (Figure 1).

Image segmentation is a core component of remote sensing information extraction, dividing images into semantically consistent regions (e.g., crop-covered areas, bare land, water bodies) to enable pixel-level classification of objects. This technology provides critical support for various agricultural applications, such as crop growth assessment model development (Dobrota *et al.,* 2021; Ntakos *et al.,* 2024; Zhao *et al.,* 2024), early pest and disease identification systems (Wan *et al.,* 2022; Zhu *et al.,* 2022; Rehman *et al.,* 2024), irrigation strategy optimisation (Zhu *et al.,* 2018; Jiang *et al.,* 2022), rational planning of farmland plots (Zhong *et al.,* 2023), precise weed segmentation (Wang *et al.,* 2021; Pei *et al.,* 2022), and efficient yield estimation (Sui *et al.,* 2018; Karlson *et al.,* 2020).

Before the rise of deep learning, the academic community had already proposed various traditional image segmentation methods, such as threshold-based segmentation (Hassanein *et al.,* 2018; Wu *et al.,* 2019; Castillo-Martínez *et al.,* 2020), clustering-based segmentation (Arango *et al.,* 2016; Di *et al.,* 2021; Khan *et al.,* 2022), wavelet transforms (Gilles, 2013; Guijarro *et al.,* 2015; Gao *et al.,* 2020; Xu *et al.,* 2021), and machine learning-based methods such as random forests (Gonzalo-Martín *et al.,* 2017; Banks *et al.,* 2019). However, these methods still exhibit significant limitations when applied to agricultural remote sensing image processing. For example, threshold-based methods primarily rely on grey-scale or colour differences for segmentation. In complex agricultural environments with uneven lighting or shadow obstruction, this often leads to blurred boundaries or classification errors. Clustering algorithms can preserve edge information to a certain extent, but their ability to distinguish between crops and soil becomes insuffi-

cient when their spectral characteristics are similar. More critically, traditional methods heavily rely on artificially designed features, making them poorly adapted to changes in field conditions such as morphological changes in crop growth stages or vegetation obstruction. Additionally, their limited generalisation capability restricts their practical application (Wang *et al.,* 2024).

In recent years, breakthroughs in deep learning (DL) technology (LeCun *et al.,* 2015) have provided new insights into agricultural remote sensing image segmentation research (Yasir *et al.,* 2023). Core architectures such as convolutional neural networks (CNNs) (O'shea *et al.,* 2015) and transformers (Vaswani *et al.,* 2017) can automatically extract deep semantic features through hierarchical non-linear transformations, significantly improving segmentation accuracy and algorithm robustness. Among these, the U-Net (Ronneberger *et al.,* 2015) effectively integrates multi-scale contextual information through its encoder-decoder architecture, enabling high-precision segmentation in various agricultural tasks such as pest and disease detection. The DeepLab series (Chen *et al.,* 2018) incorporates dilated convolutions and dilated spatial pyramid pooling (ASPP) modules to expand the receptive field and capture multi-scale semantic representations in agricultural scenes, thereby enhancing the model's generalisation capability in complex environments. Vision transformer (ViT) (Han *et al.,* 2023) uses self-attention mechanisms to model global context, addressing the shortcomings of CNNs in capturing long-range dependencies. Its variant, Swin transformer (Liu *et al.,* 2021), combines a hierarchical sliding window strategy to balance computational efficiency while preserving local details and global structure, making it particularly suitable for fine-grained segmentation tasks such as orchard canopy contour extraction. Additionally, transformer models specifically optimised for agricultural remote sensing (e.g., SegFormer; Xie *et al.,* 2021) have further improvements in adaptive multi-scale feature extraction and robustness in heterogeneous field environments. To further integrate the strengths of CNNs in local feature extraction with the global modelling capabilities of Transformers, hybrid architecture models have gradually become a research focus (Li *et al.,* 2020). Such models often use CNNs as backbone networks to extract initial features, then introduce Transformer modules to enhance global context interaction. Some architectures also achieve deep integration of the two within an encoder-decoder framework, with such joint optimisation improving segmentation performance (Padshetty *et al.,* 2024; Liu *et al.,* 2025). Driven by the demand for real-time processing from resource-constrained edge devices such as drones deployed in the field, research on lightweight models has gained widespread attention (Lei *et al.,* 2024). The core objective is to significantly reduce the number of model parameters and computational complexity while maintaining high segmentation accuracy. For example, lightweight CNNs such as MobileNetV3 (Koonce, 2021) and ShuffleNetV2 (Ma *et al.,* 2018) and other lightweight CNNs reduce computational costs through deep separable convolutions and channel shuffling mechanisms; while MobileViT (Mehta *et al.,* 2021) and other lightweight transformers embed the transformer mechanism into mobile-friendly network architectures. Additionally, advanced strategies such as model pruning, quantization, and knowledge distillation (Zheng *et al.,* 2025) are widely employed to further compress model size and improve inference efficiency. To ensure a comprehensive and relevant overview, this review was guided by a structured literature search. The search was conducted across major academic databases, including IEEE Xplore, Scopus, Web of Science, and Google Scholar. The time scale was focused on publications from 2015 to the present (August 2025), a period coinciding with the rise of deep learning in computer vision, as typified by landmark architectures such as the U-Net and the widespread adoption of Transformers. Search criteria involved a combination of keywords, including "deep

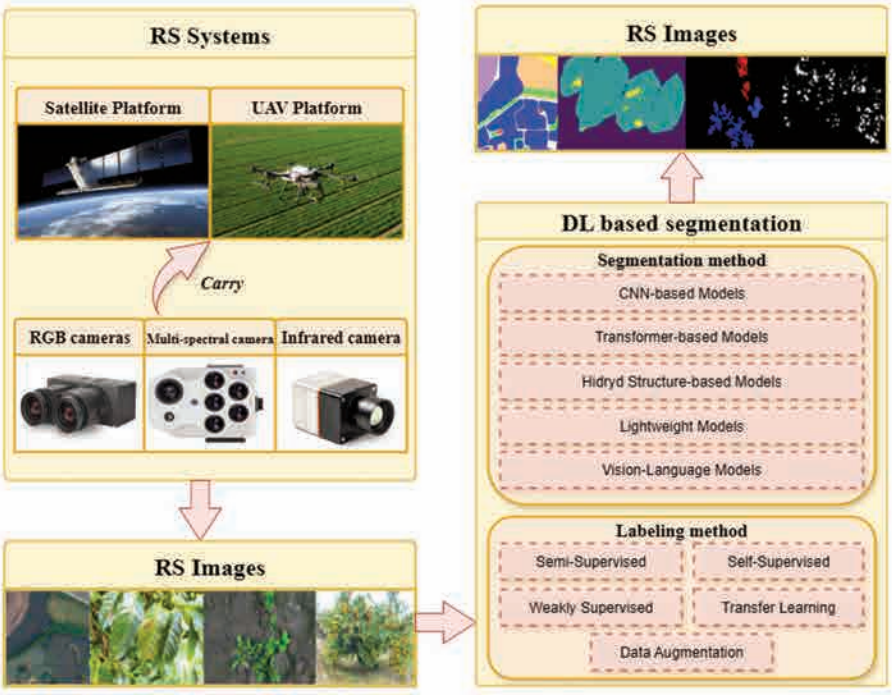

Figure 1. The process of agricultural remote sensing image segmentation.

learning," "image segmentation," "remote sensing," "precision agriculture," "UAV," "satellite imagery," and "crop monitoring." The selection of articles for inclusion was based on quality indicators such as publication in high-impact, peer-reviewed journals and conferences, direct relevance to the core topic, and the novelty of the technical contribution or application.

In summary, this paper systematically reviews the latest advances in deep learning in the field of agricultural remote sensing image segmentation. The aim is to provide comprehensive technical references for researchers in this field, promote the deep integration of deep learning and agricultural remote sensing, and accelerate the transformation of precision agriculture towards intelligentisation. Subsequent chapters will discuss key aspects such as data collection and processing, technological evolution, methodological innovation, and future research directions.

Data acquisition and processing

Remote sensing image observation platforms

The quality of remote sensing imagery is determined by three attributes: spatial resolution, spectral resolution, and temporal resolution. High-quality remote sensing imagery typically exhibits high resolution across all three dimensions. Spatial resolution refers to the actual ground area covered by a single pixel. The higher the resolution, the clearer the depiction of object details and boundaries, enabling more accurate identification of features such as shape and texture. Conversely, lower spatial resolution can result in blurred object contours and may cause adjacent targets to be confused with one another. Spectral resolution refers to a sensor's ability to distinguish between different wavelengths of electromagnetic radiation. Higher spectral resolution is more sensitive to differences in reflection and absorption within the spectral range, thereby enabling more accurate identification of target types and their physiological states. Temporal resolution is shortest interval at which a sensor can repeatedly observe the same geographical location. High temporal resolution is critical for monitoring dynamic changes on the ground and providing timely data feedback for agricultural decision-making. In the agricultural field, satellite remote sensing and unmanned aerial vehicle (UAV) remote sensing are the two main observation platforms. Satellite remote sensing is suitable for monitoring large areas, while UAV remote sensing is more suitable for high-resolution, field-level precision management (Awais *et al.,* 2022).

Satellite remote sensing platforms

Satellite remote sensing platforms offer the advantages of wide coverage and the ability to continuously and stably acquire data, making them crucial in agricultural monitoring. The currently widely used satellite systems are listed in the table, including the Landsat series, Sentinel-2, and China's Gaofen series (Table 1). Most of these systems are equipped with multispectral sensors capable of acquiring data for calculating key vegetation indices, such as the normalised difference vegetation index (NDVI). These indices play a significant role in monitoring activities such as crop health assessment (Singh *et al.,* 2020; Song *et al.,* 2021; Liepa *et al.,* 2024). Satellites operate in orbits at least 150 kilometres above the Earth's surface, capturing electromagnetic radiation reflected from Earth's surfaces to generate optical imagery and spectral datasets for subsequent analysis. Synthetic Aperture Radar (SAR) satellites, such as Sentinel-1, enable all-weather, all-time imaging, making them particularly suitable for continuous monitoring in cloudy and rainy regions (Beriaux *et al.,* 2021). The high temporal resolution of satellites is crucial for tracking crop phenological changes and monitoring soil moisture conditions. Regional-scale satellite imagery with high spatial and spectral resolution contains rich semantic information, making it suitable for applications such as land use classification, regional mapping, and agricultural disaster monitoring. These advantages make satellite imagery a reliable data source for macro-level agricultural analysis (Solangi *et al.,* 2019; Cui *et al.,* 2023; Memon *et al.,* 2023). Although satellite imagery has a wide monitoring range, it typically lacks the spatial detail provided by drone platforms.

UAV remote sensing platforms

Unmanned aerial vehicles (UAVs) offer centimetre-level spatial resolution, high operational flexibility and low cost, greatly promoting the development of precision agriculture. The table shows that UAVs can carry a variety of payloads, including RGB cameras, multispectral/hyperspectral sensors, thermal infrared sensors and lidar systems (Deng *et al.,* 2018) (Table 2). The high-resolution imagery captured by these systems can precisely identify small-scale features, making UAVs particularly suitable for detailed tasks such as weed distribution mapping, localised pest and disease detection, and field yield prediction. Compared to traditional ground-based equipment such as tripods and agricultural vehicles, low-altitude UAV platforms offer advantages in terms of monitoring coverage and data acquisition timeliness due to their compact structure, lightweight design, and ease of deployment.

Table 1. Comparison of satellite platforms for agricultural remote sensing.

Satellite platforms	TR(d)	SR(m)	Main applications
Landsat8	30	16	Crop health monitoring; land use classification
Sentinel-2	10-20	3-5	NDVI calculation; land use classification
Sentinel-1	5-20	6-12	All-weather crop monitoring
Gaofen (GF)	0.8-16	4-5	High-precision agricultural monitoring

Table 2. Comparison of UAV platforms for agricultural remote sensing.

UAV platforms	Endurance (h)	Primary sensors	Main applications
Rotary-wing UAV	<1	RGB; multispectral	Pest and disease detection; yield prediction
Fixed-wing UAV	1-2	Hyperspectral; LiDAR	Large-scale farmland monitoring
Hybrid UAV	1-3	Multispectral; thermal IR	Precision agriculture

Based on their lift generation mechanisms, UAVs are typically classified into rotorcraft platforms (capable of vertical take-off and landing), fixed-wing platforms (with longer endurance), and hybrid platforms that combine the advantages of both.

Public datasets

Publicly available datasets play a crucial role in training and evaluating deep learning models for agricultural remote sensing image segmentation. They provide high-quality annotated images from various sensor platforms, serving as essential foundational resources for algorithm development and benchmarking. The main publicly available datasets encompass satellite and drone observation data, covering a range of tasks such as crop classification, land cover mapping, and pest and disease detection (Table 3).

Data preprocessing

Data preprocessing is a critical step in the agricultural remote sensing image segmentation process, with the core objective of optimizing the quality of raw input data to ensure that the information fed into deep learning models is accurate, consistent, and analytically valuable. Agricultural remote sensing images typically feature high spatial resolution and multispectral or hyperspectral characteristics, but they are often affected by environmental factors such as changes in lighting, cloud interference, and terrain deformation. Systematic preprocessing is crucial for noise reduction, geometric and radiometric correction, and feature enhancement. These steps significantly improve the segmentation accuracy, generalisation capability, and robustness of subsequent deep learning models. The following is an overview of the key preprocessing stages and their functions. Radiometric calibration is a fundamental step in the pre-processing of agricultural remote sensing data. This process eliminates sensor characteristic differences and system errors, converting raw digital number values (DN) into physically meaningful radiometric brightness or surface reflectance, ensuring that the spectral characteristics of image pixels accurately reflect the properties of ground objects. For multispectral and hyperspectral imagery, spectral characteristics are directly used for crop type identification (Ahmad *et al.,* 2021), plant health assessment (Zhang *et al.,* 2018; Li *et al.,* 2022), and soil property characterisation, making radiation correction particularly critical. By establishing a quantitative relationship between DN values and physical quantities, radiation correction provides a standardised foundation for multi-sensor data fusion and temporal analysis. Related studies have shown that changes in solar elevation angle and atmospheric scattering caused by cloud cover can lead to deviations in surface reflectance and vegetation indices, necessitating correction to ensure data reliability (de Souza *et al.,* 2010). Weiss *et al*. (2020) proposed that radiation correction significantly improves the accuracy of crop classification using hyperspectral data. Additionally, radiation-corrected imagery provides consistent and reliable inputs for calculating key agricultural parameters such as NDVI, thereby enhancing the precision and comparability of subsequent analysis tasks.

Geometric correction is used to eliminate spatial distortions caused by sensor imaging geometry, platform motion, and terrain undulations, enabling remote sensing images to be precise aligned with the geographic coordinate system (Bannari *et al.,* 1995). This step plays a critical role in multi-temporal image registration and spatial analysis, particularly in scenarios such as dynamic monitoring of farmland boundaries, tracking crop growth trajectories, and detecting land use changes (Ahamed *et al.,* 2012). Commonly used geometric correction methods include polynomial transformations, rational function models, and strict sensor models, which should be selected appropriately based on factors such as sensor type, the location and accuracy of ground control points (GCPs), and terrain complexity (Aguilar *et al.,* 2008).

Atmospheric correction involves modelling atmospheric scattering and absorption effects caused by aerosols, water vapour, or other atmospheric components to restore the true surface reflectance of land features. Spectral information distortion in remote sensing imagery caused by clouds, haze, or humidity may introduce systematic errors in multi-temporal analysis and multi-sensor data fusion (Gao *et al.,* 2009). Atmospheric correction can significantly mitigate these interferences, improve the accuracy of vegetation index calculations, and directly enhance the image segmentation performance between crops and background elements (Nazeer *et al.,* 2021). For example, when detecting early crop stress responses, atmospheric correction is particularly necessary because these subtle signals are highly sensitive to spectral fidelity. Hadjimitsis *et al.* (2010) used the dark object subtraction (DOS) method to demonstrate that omitting the atmospheric correction step results in significantly lower daily evapotranspiration values

Table 3. Public datasets in agricultural remote sensing segmentation tasks.

Datasets	Number of scenes	Classes	Resolution (GSD)	Devices	References
Agriculture-vision	94,986	9	RGB-NI10/15/20 cm/px	RGB-NIR	Chiu *et al*., 2020
GID-15	150	15	0.8 m/px	GF-2	Tong *et al*., 2020
PhenoBench	2179	5	1 mm/px	UAV	Weyler *et al*., 2024
EuroSAT	27,000	10	10 m/px	Sentinel-2	Helber *et al*., 2019
CropHarVest	95,186	/	10 m/px	Sentinel-2	Tseng *et al*., 2021
FLAIR	76,300	32	20 cm/px	5CIR-DSM; Sentinel-2	Garioud *et al*., 2023
PASTIS	2433	18	10 m/px	Sentinel-2	Garnot *et al*., 2021
PASTIS-HD	2433	18	1/10 m/px	SPOT 6/7; Sentinel-1; Sentinel-2	Astruc *et al*., 2025
Extended agriculture-vision	98,586	9	10 cm/px	RGB-NIR	Wu *et al*., 2023
CalCROP21	50,000	28	10 m/px	Sentinel-2	Ghosh *et al*., 2021
AgriPotential	8890	5	5 m/px	Sentinel-2	El Sakka *et al*., 2025
SICKLE	209,000	21	10/30 m	Landsat-8; Sentinel-2; Sentinel-1	Sani *et al*., 2024)
Sen4AgriNet	225,000	172	10/20/60 m	Sentinel-2	Sykas *et al*., 2022

than the actual values, emphasising the importance of atmospheric correction in the quantitative inversion of agricultural biophysical parameters.

The primary purpose of data augmentation in remote sensing image processing is to enhance the expressive power of image features and expand the training sample size. This can be achieved by calculating enhanced features such as NDVI and soil adjusted vegetation index (SAVI) to highlight the spectral response characteristics of crops. Additionally, principal component analysis (PCA) can extract principal components from high-dimensional multi/hyperspectral data, reducing redundancy, and improving feature representation efficiency (Darwin *et al.,* 2021). At the sample level, geometric and radiometric transformations such as image rotation, horizontal flipping, random cropping, and colour dithering, along with advanced sample mixing strategies like Mixup and CutMix, can significantly expand limited training datasets, thereby addressing the issue of scarce annotated data in agricultural remote sensing (Lu *et al.,* 2024). In recent years, augmentation techniques based on generative adversarial networks (GANs) have garnered increasing attention. This method generates agricultural scene images with high fidelity, demonstrating significant advantages when sample data for specific crop types or rare diseases is insufficient (Lu *et al.,* 2022).

Applications of deep learning in agricultural remote sensing image segmentation

Challenges of applying deep learning to agricultural remote sensing image segmentation

Due to the high-dimensional and heterogeneous characteristics of agricultural remote sensing data and the high complexity of agricultural application scenarios, deep learning methods face significant challenges when applied to agricultural remote sensing image segmentation (Li *et al.,* 2024). Agricultural remote sensing data typically includes multi-source, multi-modal inputs, such as multi/hyperspectral and SAR imagery (Lv *et al.,* 2023). These data not only contain detailed spectral information characterising crop physiological states, but also record the phenological dynamics of crops over time. The model must therefore possess the capability to simultaneously process high-dimensional spatial-spectral-temporal information, mitigating the dimensional disaster and information redundancy interference while extracting key features during the segmentation process (Xu *et al.,* 2023).

Agricultural remote sensing tasks heavily rely on high-quality annotated data, but such data is extremely difficult to obtain in practice. In agricultural scenarios, complex terrain structures, small differences between crop types, and agricultural activities such as irrigation and fertilisation can cause dynamic changes (Peng *et al.,* 2021). Therefore, when performing pixel-level semantic annotation, one must not only be able to interpret remote sensing imagery but also possess agricultural expertise to accurate identify crop growth stages, field facilities, or areas affected by environmental stresses such as pests and diseases (Zhu *et al.,* 2024; Raei *et al.,* 2022). This annotation process is time-consuming, requires high professional expertise, and is costly, leading to a long-standing shortage of data resources for agricultural remote sensing segmentation tasks (Luo *et al.,* 2024).

In summary, in order to promote the application of high-performance deep learning segmentation models in the agricultural field, new breakthroughs must be made in model architecture and learning paradigms. For example, new network architectures can be designed to efficiently process and integrate large-scale, high-dimensional, multi-modal temporal remote sensing data. In addition, efficient annotation learning paradigms can be explored to address the scarcity of professional annotation data.

Deep learning architecture design

CNN-based models

CNNs are the foundational architecture in the field of computer vision, and their automatic feature extraction capabilities and spatial hierarchical modelling capabilities provide robust technical support for agricultural remote sensing image segmentation (El Sakka *et al.,* 2024; Alam *et al.,* 2021). Their primary advantages stem from the synergistic interaction between convolutional layers and pooling layers: shallow layers focus on extracting local low-level features, while deep layers progressively integrate and abstract higher-level semantic information (Zhao *et al.,* 2024). This progressive transformation from local details to global

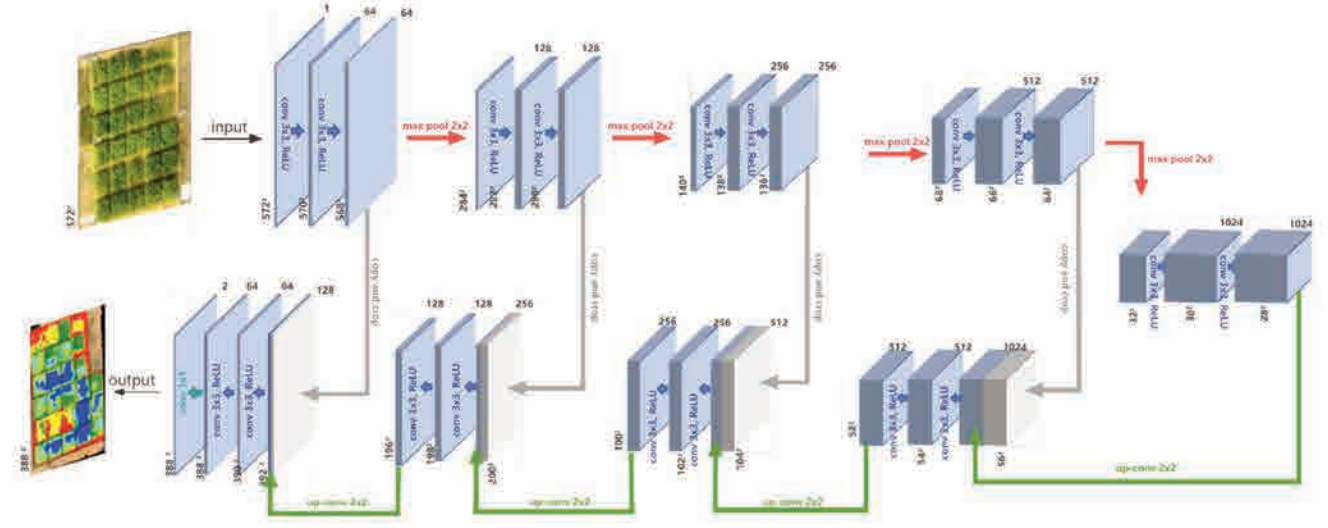

Figure 2. Schematic diagram of the U-Net network structure.

semantics ensures pixel-level precise segmentation of images.

For structured prediction tasks such as image segmentation, the CNN-based encoder-decoder architecture has been widely adopted and continuously optimised. The encoder downsamples to compress the spatial dimension and learns deep semantic representations. The decoder performs upsampling based on the latent representations to generate pixel-level outputs. The U-Net proposed by Ronneberger *et al*. (2015) features a unique symmetric design and skip-connection mechanism, enabling it to effectively capture contextual information during downsampling and fuse high-resolution spatial details from shallow encoders during upsampling (Figure 2). This mechanism mitigates the loss of spatial details during traditional CNN downsampling. To enhance model performance, researchers have proposed numerous U-Net variants. For example, SK-ResNeXt (Ramos *et al.,* 2025) introduces a cardinality and adaptive convolution kernel size design, significantly enhancing multi-scale modelling capabilities and adaptability. In the land cover classification (LCC) task, this model achieves an overall accuracy (OA) and mean intersection over union (mIoU) improvement of 5.312% and 8.906%, respectively, compared to the baseline model in the RGB configuration. AER U-Net (Jonnala *et al.,* 2025) combines multi-scale residual networks, dilated convolutions, and skip connections to significantly improve the accuracy of water body segmentation. To address the challenge of segmenting unstructured land use types in rural areas, Zhao *et al.* (2025) proposed the Land-Unet model, which incorporates a dual-branch edge-sensitive module (ESB) comprising a spatial-channel collaborative attention (SCSA) branch and dynamic upsampling (DYU) technology, effectively mitigating semantic ambiguity at boundaries. Cheng *et al.* (2024) proposed an innovative model for orchard segmentation, integrating an efficient multi-scale attention mechanism (EMA) and LayerScale adaptive scaling, combined with CycleGAN and transfer learning, to enhance generalisation across data sources (UAV imagery, Google Earth, Sentinel-2), achieving mIoU values of 97.39%, 92.08%, and 84.54%, respectively. Another classic encoder–decoder model, SegNet (Badrinarayanan *et al.,* 2017) innovates by reusing the pooling indices recorded during the encoder's max pooling operation for the decoder's upsampling operation (Figure 3). This design significantly reduces the number of parameters while preserving high-resolution details during the upsampling process. To address issues such as large parameter size, limited accuracy, and performance degradation during training, an improved SegNet variant (Weng *et al.,* 2020) introduces enhanced residual blocks in the encoder to mitigate degradation, and adopts depthwise separable convolutions and dilated convolutions to control the number of parameters and expand the receptive field, thereby improving model performance without weakening feature extraction capabilities. MASA-SegNet (Sun *et al.,* 2023), which integrates a multi-axis sequence attention mechanism, enhances feature extraction and suppresses noise through spatial-sequence propagation, significantly improving the semantic segmentation performance of PolSAR images. RWSNet (Jiang and Li, 2020) combines SegNet with optimised random walks, automatically generating seeds from network predictions and optimising weights using gradient/probability graphs, reducing computational costs while improving image segmentation accuracy.

With the deepening of research, the multi-scale target problem commonly found in remote sensing images has not yet been adequately resolved, including scattered plots, dense crop rows, and large-scale field facilities. This has driven the development of multi-scale feature fusion architectures represented by the DeepLab series. DeepLab (Chen *et al.,* 2014) introduced dilated convolutions, a technique that expands the receptive field without reducing the resolution of feature maps, effectively capturing broader spatial contextual information. DeepLabV2 (Chen *et al.,* 2018) further proposed dilated spatial pyramid pooling (ASPP), which achieves comprehensive modelling of multi-scale information by applying dilated convolutions and pooling operations with different dilation rates in parallel to feature maps. This figure illustrates the comparison between the DeepLab and DeepLabV2 architectures (Figure 4). DeepLabV3+ (Chen *et al.,* 2018) is an optimised version based on DeepLabV3 (Chen *et al.,* 2017), using the Xception network as its backbone (Figure 5). Wang *et al*. (2022) proposed CFAMNet, which improves DeepLabV3+ by integrating a category feature attention mechanism, effectively addressing the issues of inaccurate boundary segmentation and category inconsistency in high-resolution images, significantly enhancing semantic segmentation accuracy while reducing the number of model parameters. To address the challenges of blurred object boundaries and missing contextual information in ultra-high-resolution images, Du *et al.* (2021) proposed a fusion framework combining DeepLabV3+ with object-based image analysis (OBIA). By integrating deep network predictions with handcrafted features, geometric information (DSM), and object-constrained optimization using higher-order CRFs, the framework achieves highly precise segmentation. Given the importance of soil erosion monitoring in environmental assessment and management, the DD-DA model (Zhang *et al.,* 2024) was specifically optimised based on DeepLabV3+. This model incorporates a convolutional block attention module (CBAM) that combines spatial and channel attention, significantly enhancing the model's ability to extract the

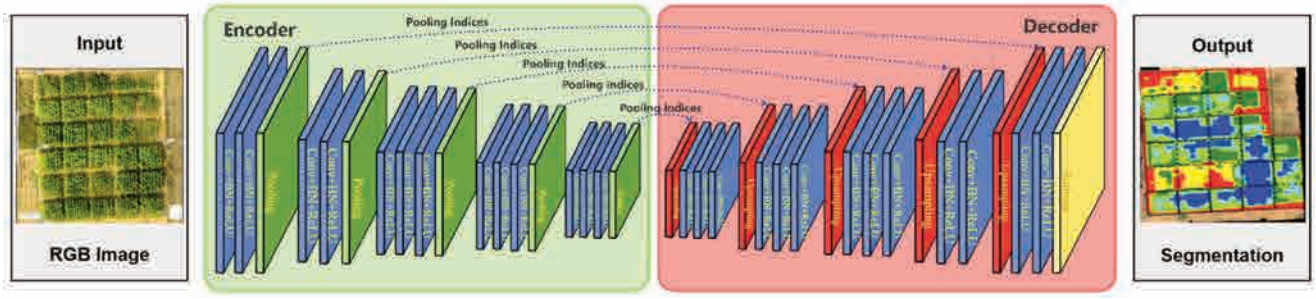

Figure 3. Schematic diagram of the SegNet Network structure.

unique morphological structures and texture features of gully erosion. CNNs are naturally adept at handling local spatial correlations in images, which is crucial for analysing complex details in satellite and drone remote sensing data (Tao *et al.,* 2022). Through continuous iteration, CNN models have evolved of powerful tools that can adaptively expand their receptive fields and effectively fuse multi-scale hierarchical information flows. This enables them to accurately express and segment complex content in agricultural remote sensing images.

Transformer-based models

The self-attention mechanism of the transformer model can efficiently model long-range dependencies and has achieved breakthrough progress in the field of natural language processing (NLP). Inspired by this, this mechanism is introduced into the agricultural remote sensing image segmentation task (Wang *et al.,* 2024), and its core structure is shown as follows (Figure 6). The self-attention mechanism can dynamically calculate the attention weights between any two spatial regions in an image, effectively addressing the inherent limitations of CNNs in modelling distant contextual information due to their restricted receptive fields (Khan *et al.,* 2022).

As a basic architecture, the Vision Transformer (ViT) (Dosovitskiy *et al.,* 2020) was originally used for image classification tasks. Its core idea is to divide the input image into patches of fixed size and treat them as a sequence, allowing the self-attention mechanism to capture the global context information of the image (Bazi *et al.,* 2021; Naseer *et al.,* 2021). Since then, researchers have continued to improve ViT. For example, the MeViT model (Panboonyuen *et al.,* 2023) combines a medium-resolution multi-

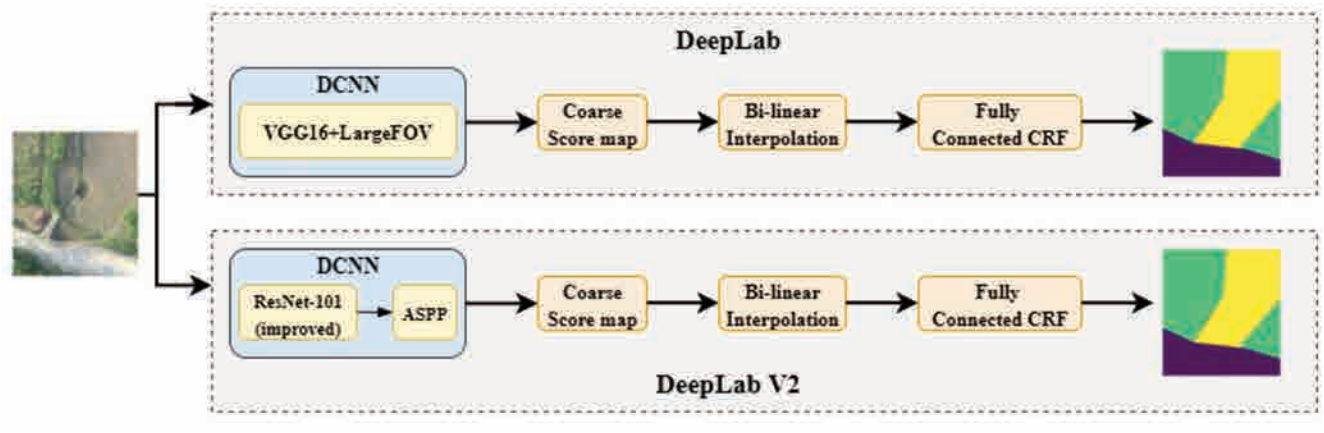

Figure 4. Comparison of the DeepLab and DeepLabV2 network structures.

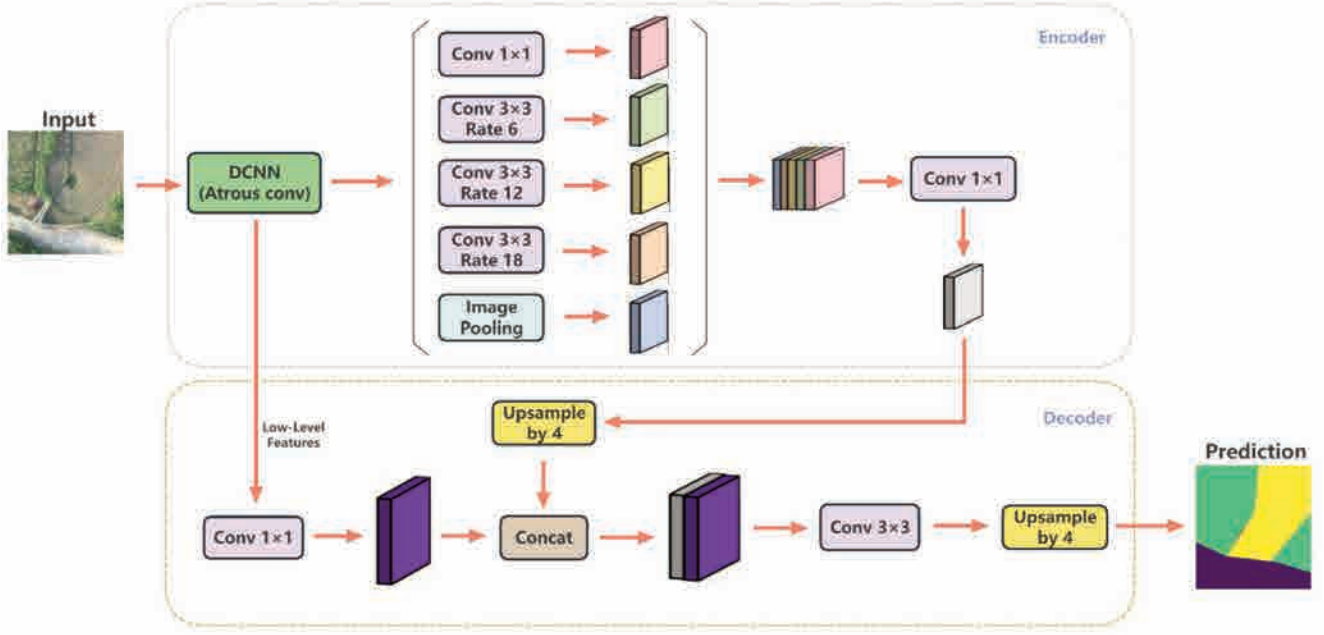

Figure 5. Schematic diagram of the DeepLabV3 network structure.

branch design with ViT, enabling it to learn semantically rich and spatially precise multi-scale representations. Additionally, the model enhances the capability of the mixed-scale convolutional feedforward network by introducing multiple deep convolutional branches, enabling it to extract multi-scale local details and effectively balance model performance and computational efficiency. However, the standard ViT architecture still has notable limitations, including insufficient ability to capture local spatial details, computational complexity that grow quadratically with input size, and strong dependence on large-scale pre-training in performance competition. To overcome the above limitations, researchers proposed the Swin transformer (Liu *et al.,* 2021), which introduces an innovative hierarchical structure and movable window mechanism. Unlike traditional transformers, the Swin transformer limits self-attention calculations to local non-overlapping windows while enabling cross-window information exchange by shifting window positions between adjacent layers. This strategy retains global modelling capabilities while significantly reducing computational complexity by approximately 40%. Its hierarchical downsampling structure generates multi-scale feature maps, making it particularly suitable for capturing phenological patterns during crop growth (Xu *et al.,* 2021). As a result, the Swin transformer was quickly applied to remote sensing image segmentation tasks. For example, ST-MDAMNet (Liu *et al.,* 2024) introduces a multi-dimensional attention mechanism and a feature enhancement module (FAM) to strengthen key feature representation. In water resource applications, the WISTE model (Ma *et al.,* 2023) utilises Swin transformer to extract water body information and designs a dual-branch encoder structure to fuse spatial details captured by a fully convolutional network (FCN) with global semantics and neighbourhood relationships obtained through multi-head self-attention. Additionally, Meng *et al.* (2022) proposed the class-guided Swin transformer (CG-Swin), which adopts a transformer-based encoder–decoder framework, using Swin as the encoder backbone and introducing a class-guided transformer module in the decoder, achieving excellent performance in land cover classification tasks.

In addition to the mainstream ViT and Swin architectures, a large number of Transformer variants specifically designed for high-resolution remote sensing image segmentation have emerged. In particular, hyperspectral imagery (HSI) can achieve fine-grained differentiation based on minute spectral differences thanks to its nearly continuous spectral information. At this point, the limitations of traditional CNN-based network architectures become increasingly evident. To address this issue, SpectralFormer (Hong *et al.,* 2022) introduces a cross-layer skip connection mechanism, which can adaptively learn 'soft' residuals within the hierarchical structure, transmitting 'memory'-like components from shallow layers to deep layers. This enables the model to learn local spectral sequences between adjacent hyperspectral bands, generating discriminative grouped spectral embeddings. Despite significant progress, deep learning models still face challenges in HSI classification, including limited receptive fields, insufficient flexibility, and poor generalisation capabilities. To address these issues, researchers introduced the BERT architecture, originally used for NLP, into HSI classification. The proposed HSI-BERT model (He *et al.,* 2020) has a global receptive field and can capture deep pixel-level dependencies that are not limited by spatial distance. It outperforms CNN-based models in terms of classification accuracy and computational efficiency and is well suited for precision agriculture monitoring tasks. To further capture deep pixel-level dependencies, the DSS-TRM model (Liu *et al.,* 2022) simultaneously introduces spectral self-attention modules (SpecSAM) and spatial self-attention modules (SpatSAM) to extract joint spatial-spectral features from HSI. These modules are fused through dot product operations to enhance the representation capability of spatial-spectral features.

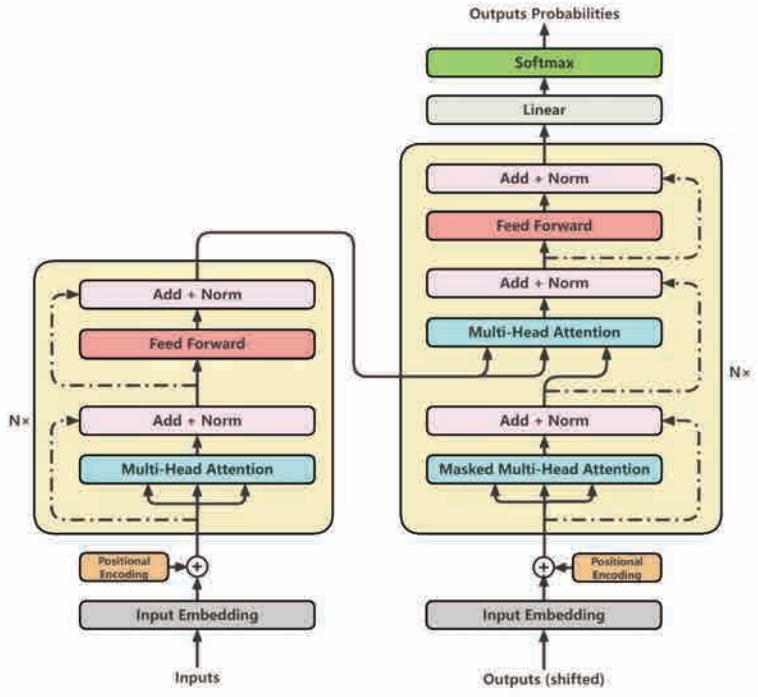

Figure 6. Schematic diagram of the transformer network structure.

Hybrid structure-based models

To overcome the inherent limitations of a single model architecture, researchers have proposed a hybrid structure that combines CNNs and transformers. These models aim to leverage the advantages of CNNs in local feature extraction and the capabilities of transformers in modelling global dependencies to improve the accuracy of image segmentation in complex agricultural scenarios (Zhu *et al.,* 2022). Based on differences in feature fusion strategies, existing hybrid frameworks can be broadly divided into two categories: single-branch serial processing architectures and dual-stream parallel processing architectures (Figures 7 and 8).

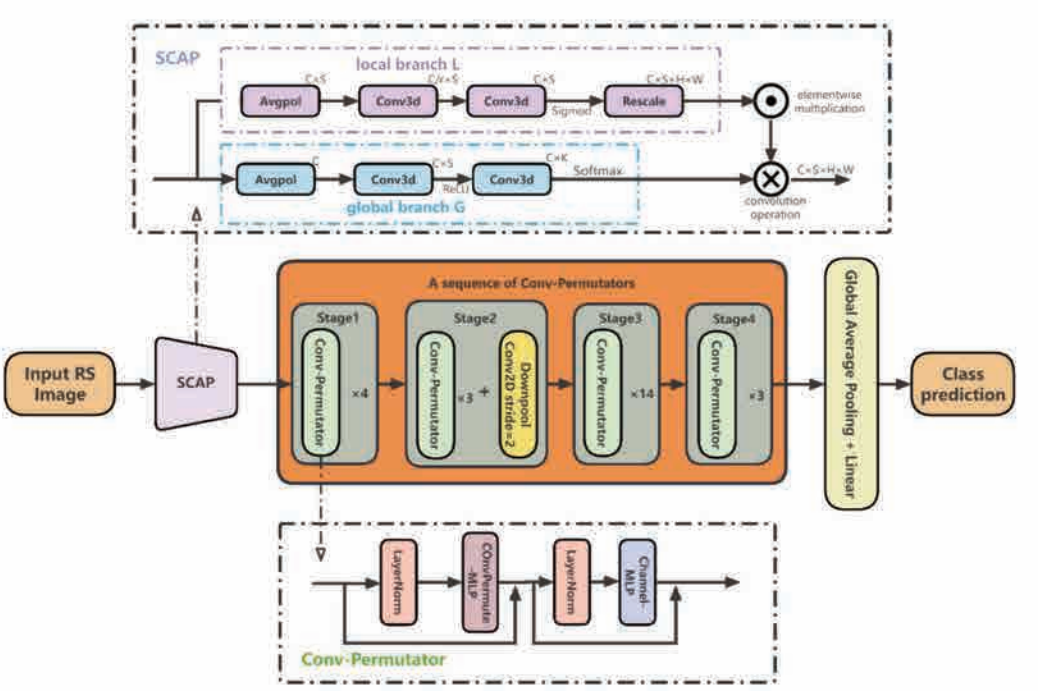

Figure 7. Schematic diagram of the serial CNN-transformer hybrid architecture.

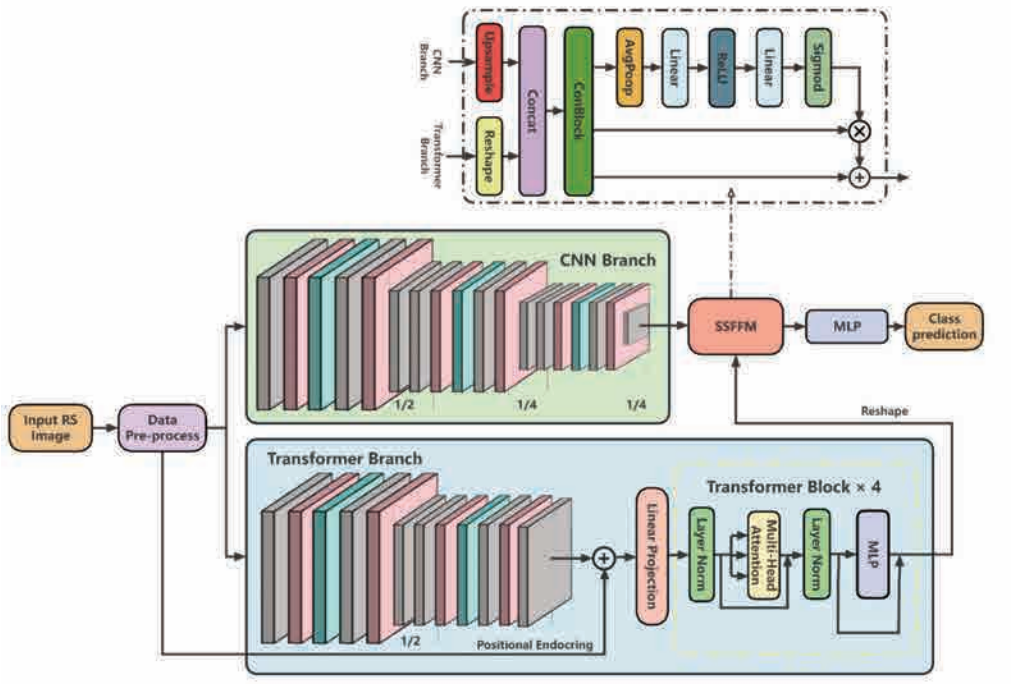

Figure 8. Schematic diagram of the parallel CNN-transformer hybrid architecture.

In the HSI classification task, Hybrid-TransCD (Ke and Zhang, 2022) adopts a single-branch hybrid architecture design, embedding convolutional layers into the Transformer encoder to achieve joint modelling of spectral features and spatial context. However, this single feature stream has limitations in constructing multi-scale long-range interactions, restricting its performance in change detection tasks that require fine-grained differentiation of complex scenes. To address this issue, STransFuse (Gao *et al.,* 2021) introduces a staged fusion paradigm: a shallow CNN is used to capture surface geometric details, while a deep Swin transformer is employed to model global contextual dependencies, leveraging an adaptive attention mechanism to fully integrate cross-scale semantic information. However, such single-branch architectures still have obvious limitations: they struggle to fully decouple the rich inter-band dependencies in HSI data and fail to fully leverage the complementarity between local features and global semantics, resulting in suboptimal segmentation accuracy for fine-grained farmland classification.

To address these shortcomings, the dual-stream parallel architecture simultaneously leverages CNNs to extract local features and transformers to model global context. For example, the ST-U network (He *et al.,* 2022) constructs a dual-encoder structure composed of parallel Swin transformer and CNN branches, using a spatial interaction module (SIM) to enhance the representation capability of occluded targets, a feature compression module (FCM) to retain small target features, and a relationship aggregation module (RAM) to hierarchically fuse global and local information. Hyper-LGNet (Zhang *et al.,* 2022) uses a CNN branch to extract local spatial features and a Transformer branch to extract global spectral dependencies, and employs a spatial–spectral feature fusion module (SSFFM) to adaptively integrate these two types of features. This is also one of the earliest classification models to simultaneously capture both spatial and spectral information from hyperspectral imagery. CTFuseNet (Xiang *et al.,* 2023) improves crop type segmentation in drone remote sensing images through a multi-scale CNN–Transformer fusion strategy, with a specially designed fusion module that aggregates information across scales. Similarly, MBT-UNet (Liu *et al.,* 2024) processes features of different scales in a multi-branch structure, effectively improving segmentation accuracy and model robustness in agricultural imagery. However, despite the promising prospects of hybrid architectures, they still face numerous challenges in practical applications, particularly in achieving efficient real-time inference and developing lightweight models suitable for edge deployment. For example, Transformer-based models typically contain a large number of parameters, and their computational complexity increases quadratically with image size, leading to significant latency. Additionally, the limited receptive field of the Swin Transformer weakens its ability to model long-range global dependencies, resulting in poor performance in complex agricultural field scenarios.

Lightweight models

Driven by the actual needs of precision agriculture, lightweight models have become a hot topic of research due to their efficiency and low resource consumption (Zhang *et al.,* 2024). Although traditional deep learning models can achieve high accuracy, their large parameter size and high computational complexity limit their real-time deployment on resource-constrained platforms such as unmanned aerial vehicle edge devices. Lightweight models that reduce parameter size and computational overhead through architectural optimisation are an ideal choice for agricultural remote sensing applications.

CNNs are highly efficient in local feature extraction and have relatively low computational requirements, so lightweight models based on CNNs are widely used in agricultural remote sensing. These models combine key technologies such as separable convolutions, parameter pruning, and attention mechanisms to significantly reduce computational costs and memory usage while maintaining high segmentation accuracy. Mainstream lightweight models based on CNNs can be broadly divided into three categories: U-Net variants, MobileNet-derived models, and YOLO-derived models. The encoder-decoder structure and skip connections of U-Net have proven highly effective in high-resolution image segmentation. However, the standard U-Net model is computationally intensive and has a large parameter size, limiting its application on edge devices. To address the challenges of annotation and real-time monitoring for large-scale segmentation of wheat lodging, Feng *et al.* (2024) proposed the ultra-lightweight L-U2NetP model, innovatively embedding a dual cross-attention (DCA) module within the U-structure unit to bridge the semantic gap and replacing complex operators with a cross-attention (CCA) module to enhance feature extraction capabilities, ultimately achieving a segmentation accuracy of 95.45% on drone aerial images. Ir-UNet (Zhang *et al.,* 2021) improves the U-Net architecture to address the issue of irregular shapes in wheat stripe rust lesions, reducing computational costs while maintaining high accuracy. Zhang *et al*. (2022) proposed DF-UNet for detecting the severity of wheat stripe rust in multispectral drone imagery, reducing computational load by over half through dual-stream feature fusion and achieving the highest overall accuracy among other state-of-the-art deep learning models. In 2025, ConvNeXt-U (Liu *et al.,* 2025) was proposed, introducing a simplified ConvNeXt backbone network and CBAM attention mechanism into the U-Net structure, effectively improving segmentation performance for complex field boundaries.

The MobileNet series, with its lightweight design and efficient inference capabilities, is highly suitable for embedded and mobile visual tasks. In agricultural remote sensing, MobileNet achieves real-time image segmentation by significantly reducing the number of parameters through separable convolutions. Lan *et al.* (2021) proposed MobileNetV2-UNet and FFB-BiSeNetV2 for real-time weed

detection in rice fields using drones. MobileNetV2-UNet combines a lightweight backbone network with a U-Net decoder, achieving nearly a 3x improvement in inference speed compared to the standard U-Net while significantly reduce the number of parameters (Figure 9). FFB-BiSeNetV2 optimises the bilateral segmentation network, outperforming traditional models in both pixel accuracy and mean intersection over union (mIoU). MST-DeepLabv3+ (Wang *et al.,* 2024) replaces the Xception backbone network in DeepLabv3+ with MobileNetV2, reducing the parameter size from 208.7MB to 22.19MB. It also introduces SENet attention modules and transfer learning to improve accuracy, offering potential applications in land cover classification under resource-constrained environments.

The high efficiency of the YOLO series models has also led to the development of variants for lightweight segmentation tasks. YOLO-Weed Nano (Wang *et al.,* 2025) is based on an optimised YOLOv8n backbone network, using depth-separable convolutions, a lightweight backbone, and an efficient detection head to reduce computational and memory requirements while maintaining high weed detection accuracy (Figure 10). Similarly, LBDC-YOLO (Zuo *et al.,* 2024) also originates from YOLOv8n, integrating Slim-neck design, triple attention mechanisms, and BiFPN architecture to achieve high-precision detection of cauliflower heads in complex field environments. Compared to the baseline YOLOv8n,

this model reduces computational load by 32.1% and model size by 44.4%. YOLOPC (Qing *et al.,* 2023) is a lightweight segmentation network that uses partial convolutions, a lightweight backbone, and CBAM attention mechanisms to enhance spatial and texture feature perception capabilities, making it suitable for use in embedded agricultural devices.

Transformers excel at global context modelling, but their high computational cost poses challenges for deployment on resource-constrained devices. To address this, researchers have developed lightweight Transformer models specifically tailored for agricultural remote sensing tasks. For example, RSegformer (Li *et al.,* 2022) is a lightweight Transformer model for plant disease segmentation, which replaces the backbone network with a compact Segformer structure, incorporates attention mechanisms, and improves the upsampling operation. This model achieved an average intersection-over-union (mIoU) of 85.38% with only 14.36 million parameters. Efficient Transformer (Xu *et al.,* 2021) is designed for remote sensing image segmentation, using an efficient backbone and MLP head to reduce computational burden, achieving an inference speed of 47.9 images *per* second, while enhancing edge segmentation accuracy through explicit and implicit edge enhancement strategies.

In addition to the aforementioned models, there are also some other lightweight models that can be used for agricultural image segmentation. Zheng *et al.* (2024) proposed the PP-LiteSeg model for crop classification using drone-based visible light imagery. This model achieved an average intersection-over-union (mIoU) of 94.79% in rice, soybean, and wheat classification tasks by introducing a pyramid pooling module (SPPF-CSPC-A) and a sparse self-attention mechanism, while maintaining a real-time inference speed of 12.3 frames per second, providing a practical solution for lightweight field image analysis. Hybrid models that combine the advantages of CNNs and Transformers, or enhance performance through architectural innovations, are emerging as an important research direction in remote sensing segmentation. For example, BEMS-UNetFormer (Zheng *et al.,* 2024) enhances the UNetFormer architecture by introducing a boundary-aware module (BAM) and a boundary-guided fusion module (BFM), achieving an mIoU of 86.12% on the Potsdam dataset. Additionally, lightweight models may lose some feature extraction capabilities during model compression, leading to a decrease in segmentation accuracy. To balance the lightweight nature of the model with segmentation accuracy, researchers are actively exploring new technologies. For example, knowledge distillation techniques can

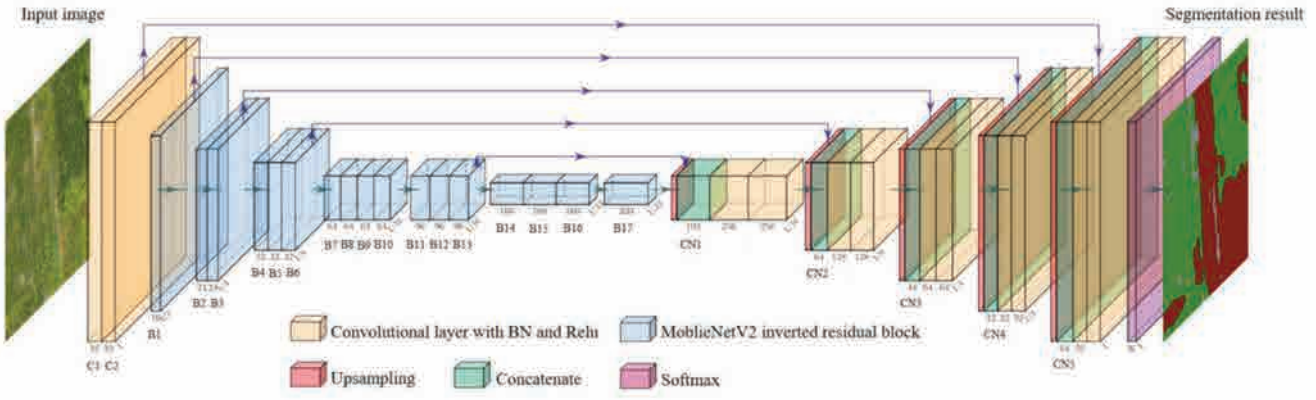

Figure 9. The structure of the MobileNetV2-UNet semantic segmentation model proposed in this study.

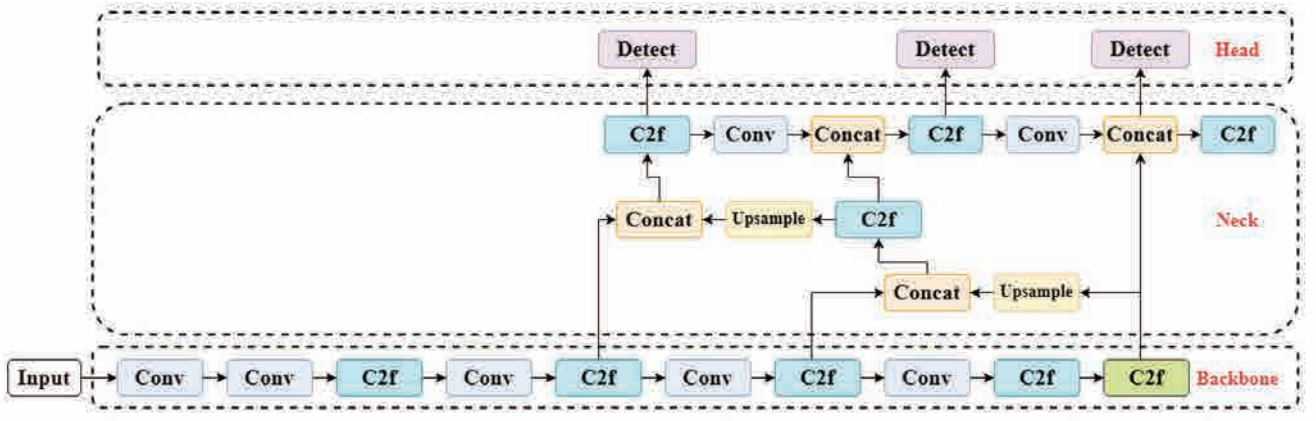

Figure 10. Schematic diagram of the YOLOv8 network structure.

transfer knowledge from large, high-performance models (teacher models) to lightweight models (student models), thereby improving the performance of the latter. Typical methods such as DP-CTNet (Zheng *et al.,* 2025) include a feature refinement module for optimising Transformer feature learning and a feature fusion module for effectively integrating CNN and Transformer features. It adopts an innovative angle-aware distillation strategy to enhance the feature transfer capability during the training process of the student model EDP-CTNet. To address the issue of poor recognition and segmentation performance caused by low pixel weights and small sizes of small objects, a study proposed a lightweight semantic segmentation network-KD-MSANet (Yang *et al.,* 2024) - based on knowledge distillation, multi-scale pyramid pooling modules, and attention mechanisms. Compared to the uncompressed version, the student model's size is reduced by 43.6%, training efficiency improves by 22.3%, and accuracy reaches 99.30% of the teacher model.

Compared to traditional CNN or transformer architectures, the core advantage of lightweight models lies in their efficiency and low resource consumption, making them ideal for running on resource-constrained devices and capable of meeting the stringent real-time requirements of practical applications. Their potential limitation is that, in order to achieve lightweight goals, their segmentation accuracy may be insufficient.

Vision-language models

Visual-language models (VLMs) have made significant breakthroughs in cross-modal learning, accelerating the development of artificial intelligence (Zhu *et al.,* 2024). These models are pre-trained using large-scale image-text paired datasets, integrating visual perception with natural language understanding to support a range of complex tasks, including image description, visual question answering, zero-shot recognition, and image segmentation (Zhu *et al.,* 2024). VLMs hold great potential in field of remote sensing, enabling precise interpretation of high-resolution satellite and aerial imagery through natural language prompts for image processing (Li *et al.,* 2024; Weng *et al.,* 2025). Their core advantage lies in achieving fine-grained alignment between visual features and textual semantics to construct a dynamic and interpretable multimodal fusion framework. This capability can enhance efficiency and accuracy of remote sensing data interpretation, breaking through the limitations of traditional single-visual methods (Tao *et al.,* 2025; Xiao *et al.,* 2025).

VLMs in agricultural image segmentation are primarily divided into two pre-training paradigms (Figure 11). Generative learning models employ mask reconstruction and text modelling methods to capture deeper semantic representations. This paradigm is particularly suitable for analysing complex crop physiological states and developmental stages. Contrastive learning-based models leverage data augmentation and cross-modal feature alignment strategies to enhance zero-shot generalisation capabilities. These models demonstrate strong adaptability when identifying new crop varieties or unfamiliar geographical regions. These paradigms each have distinct advantages in remote sensing applications, enabling fine-grained segmentation through the deep integration of multi-source heterogeneous data.

VLMs built based on the generative pre-training paradigm possess powerful instruction understanding and content generation capabilities, making them particularly suitable for addressing complex and dynamic segmentation requirements. Among these, derivative models represented by GPT stand out. For example, SpectralGPT (Hong *et al.,* 2024) focuses on in-depth analysis of hyperspectral remote sensing imagery and demonstrates outstanding performance in interpreting spectral data related to crop physiological characteristics. It can accurately delineate nitrogen-deficient areas and support high-resolution identification of soil conditions, water stress, and nutrient distribution. In contrast, SkyEyeGPT (Zhan *et al.,* 2025) employs a unified instruction parsing framework, enabling it to automatically execute segmentation tasks based on natural language queries and generate real-time text summaries. This design is particularly suited for dynamic monitoring applications and operational scenarios requiring interactive human-machine feedback.

VLMs built using contrastive learning achieve robust align-

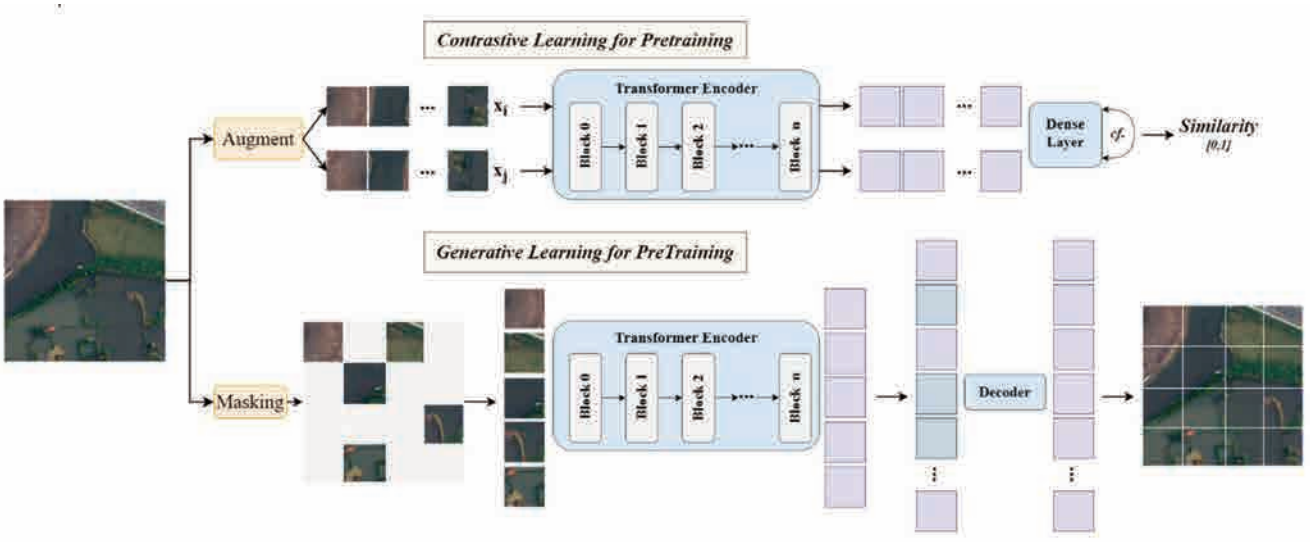

Figure 11. The difference between contrastive learning and generative learning in pre-training.

ment of image and text features through contrastive pre-training on large-scale paired image-text datasets, and perform well in few-shot and zero-shot learning tasks. A typical example is the CLIP model. This capability is particularly critical for rapid deployment in new geography areas or when dealing with unfamiliar crop types. For example, RemoteCLIP (Liu *et al.,* 2024) combines multiple remote sensing annotation strategies to enable zero-shot segmentation based on text prompts, providing an effective solution for large-scale crop mapping. GeoRSCLIP (Zhang *et al.,* 2024), specifically designed for remote sensing applications, can accurately respond to commands such as 'show rice-growing areas,' demonstrating outstanding performance in land use monitoring and farmland change detection. However, geographical biases in training data may limit its generalisation ability in diverse regions. Unlike typical CLIP-based contrast models, FSVLM (Wu *et al.,* 2025) does not rely on large-scale image-text contrast training but instead enhances semantic understanding of complex agricultural landscapes through a dedicated visual-language fusion module and few-shot learning paradigms, particularly excelling in scenarios with limited data.

VLMs are gradually driving the development of agricultural remote sensing image segmentation towards smarter and more interactive directions. GPT-based models excel at handling complex and highly customised tasks, particularly in scenarios with high semantic interpretability requirements, such as hyperspectral data processing and supporting user-interactive segmentation. In contrast, architectures like CLIP, with their efficient zero-shot transfer capabilities, excel at quickly adapting to new target categories, making them well-suited for crop identification in diverse agricultural environments. To fully leverage VLMs in driving modern agricultural transformation, optimisations and improvements are still needed in certain areas. This may involve enhancing VLMs' dense prediction capabilities and improve multi-modal data fusion strategies that integrate spectral, SAR, and time-series information, thereby strengthening the model's robustness and intelligence in dynamic monitoring throughout the entire crop growth cycle.

Efficient annotation methods

Semantic segmentation of agricultural remote sensing imagery is a key technology in precision agriculture. However, remote sensing data covers a wide area and has high resolution, making it costly and time-consuming to obtain pixel-level annotated datasets. Efficient annotation methods enhance annotation efficiency and model performance by reducing the need for manual annotation, leveraging unannotated data, and enabling cross-domain knowledge transfer. This section will provide a detailed analysis of the primary efficient annotation strategies in agricultural remote sensing segmentation, including semi-supervised learning, weakly supervised learning, self-supervised learning, transfer learning, and data augmentation (Figure 12).

Semi-supervised learning combines a small amount of labelled data with a large amount of unlabelled data to train models, making it particularly suitable for agricultural applications where labelled data is scarce. Li *et al.* (2023) proposed a self-learning-based semi-supervised method that uses UNet and DeepLabV3 to generate pseudo labels for unlabelled samples and adds these highly consistent pseudo labels to the training set. With only 20 labelled samples (4.4% of the total samples), the model achieved an F1 score of 79.45% and an average intersection-over-union (mIoU) of 68.24%, demonstrating the low dependency of semi-supervised methods in labelled data. In another study, Wang *et al.* (2020) proposed a semi-supervised method for remote sensing identification of winter wheat planting areas, achieving 82.50% mean pixel accuracy (MPA) and 76.01% mIoU with only 1/16 of the dataset labelled, outperforming fully supervised models. The advantage of semi-supervised learning lies in significantly reducing reliance on labelled data and enhancing performance through unlabelled samples. Low-quality pseudo labels may negatively impact the final results, so it is necessary to design a reliable pseudo label generation strategy.

Weakly supervised learning uses weak labels, such as single-pixel or image-level labels, instead of pixel-level annotations, thereby reducing the annotation workload. Wang *et al.* (2020) used weakly supervised learning to process 450,000 square kilometres of Landsat 8 imagery from the Midwestern United States. The Masked U-Net they proposed achieved 88% accuracy rate for farmland segmentation using only 1,000 annotated samples. The weakly supervised tree species classification method based on explainability techniques proposed by Ahlswede *et al.* (2022) demonstrates its potential in agricultural tasks. Although weakly supervised learning significantly reduces annotation costs in large-scale monitoring areas, but its segmentation accuracy in complex scenarios still needs to be optimised. Additionally, these methods

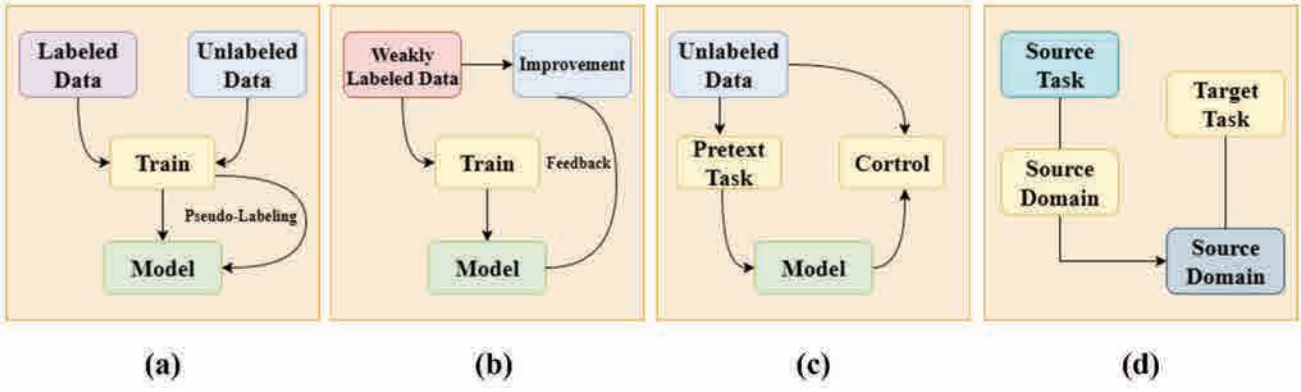

Figure 12. Schematic diagram of the efficient labelling method. **a**) Semi-supervised learning. **b**) Weakly supervised learning. **c**) Self-supervised learning. **d**) Transfer learning.

typically require specialised techniques such as category activation maps (CAM) or mask loss to properly handle weak labels, increasing the complexity of model design.

Self-supervised learning extracts features from unlabelled data by defining pre-training tasks, reducing reliance on manually labelled samples (Marsocci *et al.,* 2021). S4 (Shenoy *et al.,* 2024) is a self-supervised pre-training framework for semantic segmentation of satellite image time series (SITS). This method utilises multi-modal data (such as radar and optical images) and time alignment information to learn representations across modal and temporal dimensions. Use only 10% of labelled data in the PASTIS-R dataset, it achieved 36.5% mIoU in the SAR modality and 33.7% mIoU in the optical modality. RingMo (Sun *et al.,* 2023) is a remote sensing foundation model trained using masked image modelling, supporting downstream tasks such as segmentation through fine-tuning. The advantage of self-supervised learning lies in its ability to fully utilise unlabelled data; however, the effectiveness of this method depends on the relevence between the pre-training task and the downstream objective. If the two are mismatched, it may hinder performance improvement.

Transfer learning leverages models pre-trained on large, general-purpose datasets and fine-tunes them to adapt to specific agricultural remote sensing tasks, thereby reducing the need for task-specific annotated data. Zhang *et al.* (2020) used the EfficientNet model pre-trained on ImageNet and fine-tuned it for remote sensing segmentation tasks, achieving good results. Kerner *et al.* (2024) explored cross-region transfer learning for farmland boundary segmentation, validating its effectiveness in scenarios with limited local labelled data. The advantage of transfer learning lies in accelerating model training speed, but it may fail to fully capture domain-specific features of agricultural data, necessitating reasonable fine-tuning to adapt to practical applications (Zhu *et al.,* 2024). Additionally, Kerner's research found that model performance varies across different climatic regions, emphasising the necessity of domain-specific optimisation in transfer learning.

Data augmentation addresses the scarcity of labelled data by expanding datasets with synthetic data. Su *et al.* (2021) proposed a data augmentation framework based on random image cropping and patching (RICAP), which relaxes the constraints on image patches and introduces a boundary alignment cost to improve the quality of generated images. Compared with traditional augmentation techniques such as flipping, rotation, and colour jittering, RICAP demonstrates superior performance. In addition to traditional methods, generative adversarial networks (GANs) have emerged as a powerful tool for learning data distributions and generating realistic samples since their first proposal in 2014 (Lu *et al.,* 2022). Abbas and colleagues (2021) applied conditional generative adversarial networks (CGANs) to generate shape-adaptive objects for training segmentation models in crop/weed detection. However, data augmentation-generated samples may deviate from the true distribution, so validation with real data is still required.

Challenges and limitations

Deep learning-based agricultural remote sensing image segmentation technology provides critical support for agricultural automation, crop condition monitoring, and precision agriculture, and has enormous application potential. However, this technology still faces many challenges and limitations in its actual deployment and promotion, mainly including data scarcity, highly heterogeneous growth environments, high computing resource consumption, and limited model transferability (Xing *et al.,* 2014).

Scarcity of labelled data and high annotation cost

One of the most critical development bottlenecks in practical applications lies in the severe shortage of high-quality annotated data (Hua *et al.,* 2022; Rasmussen *et al.,* 2022). Existing public datasets typically focus on only a few mainstream or single crop varieties, resulting in limited coverage. This limitation leads to a decrease in segmentation accuracy and generalisation ability when models are transferred to unseen crop types or different growth environments, exposing insufficient environmental adaptability (Mamat *et al.,* 2022). A deeper issue lies in the complexity of agricultural scenarios, which significantly increases the technical difficulty of annotation. Achieving pixel-level precise annotation is inherently a labour-intensive and highly demanding task, requiring manual pixel-level identification and contour tracing, which is extremely costly (Alzubaidi *et al.,* 2024).

Environmental complexity and interference factors

The highly dynamic nature of the agricultural environment is also a major challenge affecting the accuracy of agricultural remote sensing image segmentation (Charisis and Argyropoulos, 2024). Crop phenotypic characteristics such as leaf morphology, root structure, and canopy structure are easily affected by the changes in growth conditions (Jiang *et al.,* 2020; Ou *et al.,* 2024). Image acquisition based on unmanned aerial vehicle (UAV) platforms is often affected by colour differences caused by fluctuations in light intensity and temperature, especially in scenarios such as low light conditions and extreme weather, where RGB images struggle to capture key details, severely limiting segmentation accuracy (Zhang *et al.,* 2023).

Computational demands and real-time processing bottlenecks

High-resolution remote sensing images contain huge amounts of data, requiring large amounts of memory and computing resources for processing. The demand for real-time agricultural monitoring further exacerbates this challenge (Holder and Shafique, 2022). Complex model inference has long latency, making it difficult to meet the needs of real-time decision-making and response in the field (Xie *et al.,* 2023). In addition, multi-temporal data analysis required for monitoring crop phenological changes requires continuous processing of time-series images, further increasing storage and computing burdens.

Limited model generalization

Current segmentation models often exhibit unstable performance across different regions and crop types. For example, a model trained in a specific geographical region may experience a significant drop in accuracy when transferred to another region due to differences in soil composition, farming practices, or climatic conditions (Ullah and Bais, 2022). Crop diversity further exacerbates this issue, as different species exhibit significant variations in morphological and spectral characteristics, making it is challenging for a single model to achieve consistent performance. Additionally, the scarcity of extreme or rare samples, such as pest outbreaks or drought conditions, weakens the model's generalisation ability in abnormal scenarios.

Future directions

Explainable artificial intelligence (XAI): improving the explainability of models may have certain economic and environmental impacts on agriculture-related decision-making (Ryo,

2022; Cartolano *et al.,* 2024). The 'black box' nature of deep learning models often limits user trust. Therefore, future research should prioritise the development of transparent and explainable AI technologies. Tools such as feature visualisation, saliency maps, and model distillation can reveal the prediction process, enhancing user understanding and confidence. Research indicates that explainability not only fosters trust but also aids in model optimisation and reliability enhancement (Shams *et al.,* 2024). For example, attention map visualisation techniques are particularly effective in highlighting crop areas affected by pests, providing valuable reference information for agronomists and farmers.

Integration with the Internet of Things (IoT): with the widespread deployment of IoT devices such as drones, sensors, and robots in agriculture, the agricultural sector has gained access to rich and continuous data streams (Abouzahir *et al.,* 2017; Jia *et al.,* 2018). Future research should focus on seamlessly integrating these real-time data stream with deep learning models. Deploying lightweight models on resource-constrained edge devices remains key to achieving real-time field analysis (Tsakiridis *et al.,* 2020).

Efficient annotation and training strategies: breaking the inefficiency of manual pixel-level annotation remains one of the major bottlenecks in the field of agricultural remote sensing (Hong *et al.,* 2024). Future research should focus on developing semi-supervised and self-supervised learning methods to reduce reliance on annotated data. A promising example is SpectralGPT (Hong *et al.,* 2024), a self-supervised model that successfully captures complex agricultural patterns by training on large amounts of unlabelled data without requiring extensive manual annotation. Another effective strategy is active learning, which focuses on annotating the most informative samples to improve annotation efficiency and reduce the overall data annotation burden (Flores *et al.,* 2024).

Conclusions

This review has provided a comprehensive analysis of deep learning's role in agricultural remote sensing image segmentation, a critical technology for advancing modern precision agriculture. We have systematically charted the landscape of data acquisition, from satellite platforms (e.g., Sentinel, Landsat) for large-scale monitoring to the high-resolution, field-level data provided by UAVs. A core component of this review was the detailed synthesis of deep learning architectures, tracing their evolution from foundational CNN-based models like U-Net and DeepLab, which excel at local feature extraction, to the powerful, context-aware Transformer and hybrid models that capture global dependencies. We also identified significant and persistent challenges that temper this progress, most notably the scarcity of high-quality annotated data, high computational demands, and persistent issues with model generalizability across diverse agricultural environments. The future directions identified, such as the integration of Explainable AI, fusion with IoT data streams, and the development of efficient annotation strategies, are not merely speculative but are direct, necessary responses to these identified bottlenecks. Ultimately, overcoming these challenges will be key to unlocking the full potential of this technology, driving the next wave of intelligent, data-driven agriculture and contributing to sustainable global food security.

References

Abbas A, Jain S, Gour M, Vankudothu S, 2021. Tomato plant disease detection using transfer learning with C-GAN synthetic images. Comput Electron Agric 187:106279.

Abouzahir S, Sadik M, Sabir E, 2017. IoT-empowered smart agriculture: A real-time light-weight embedded segmentation system. In: E. Sabir, A. García Armada, M. Ghogho, M. Debbah (eds.), Ubiquitous networking. UNet 2017. Cham, Springer; pp. 319-332.

Aguilar M, Agüera F, Aguilar F, Carvajal F, 2008. Geometric accuracy assessment of the orthorectification process from very high resolution satellite imagery for Common Agricultural Policy purposes. Int J Remote Sens 29:7181-7197.

Ahlswede S, Madam NT, Schulz C, Kleinschmit B, Demir B, 2022. Weakly supervised semantic segmentation of remote sensing images for tree species classification based on explanation methods. Proc. IEEE Int. Geoscience and Remote Sensing Symp, Kuala Lumpur; pp. 4846–4849.

Ahamed T, Tian L, Jiang Y, Zhao B, Liu H, Ting K, 2012. Tower remote-sensing system for monitoring energy crops; Image acquisition and geometric corrections. Biosyst Eng 112:93-107.

Ahmad H, Sun J, Nirere A, Shaheen N, Zhou X, Yao K, 2021. Classification of tea varieties based on fluorescence hyperspectral image technology and ABC-SVM algorithm. J Food Process Preserv 45:e15241.

Alam M, Wang J, Guangpei C, Yunrong L, Chen Y, 2021. Convolutional neural network for the semantic segmentation of remote sensing images. Mob Netw Appl 26:200-215.

Alzubaidi L, Chlaib H, Fadhel M, Chen Y, Bai J, Albahri A, et al., 2024. Reliable deep learning framework for the ground penetrating radar data to locate the horizontal variation in levee soil compaction. Eng Appl Artif Intell 129:107627.

Arango R, Campos A, Combarro E, Canas E, Díaz I, 2016. Mapping cultivable land from satellite imagery with clustering algorithms. Int J Appl Earth Obs Geoinf 49:99-106.

Astruc G, Gonthier N, Mallet C, Landrieu L, 2025. OmniSat: Self-supervised modality fusion for Earth observation. In A. Leonardis, E. Ricci, S. Roth, O. Russakovsky, T. Sattler, G. Varol (eds.), Computer Vision – ECCV 2024. Cham, Springer; pp. 409-427.

Awais M, Li W, Hussain S, Cheema M, Li W, Song R, et al., 2022. Comparative evaluation of land surface temperature images from unmanned aerial vehicle and satellite observation for agricultural areas using in situ data. Agriculture 12:184.

Badrinarayanan V, Kendall A, Cipolla R, 2017. SegNet: A deep convolutional encoder-decoder architecture for image segmentation. IEEE T Pattern Anal Mach Intell 39:2481-2495.

Banks S, White L, Behnamian A, Chen Z, Montpetit B, Brisco B, et al., 2019. Wetland classification with multi-angle/temporal SAR using random forests. Remote Sens 11:670.

Bannari A, Morin D, Bénié GB, Bonn FJ, 1995. A theoretical review of different mathematical models of geometric corrections applied to remote sensing images. Remote Sens Rev 13:27-47.

Bazi Y, Bashmal L, Al Rahhal M, Al Dayil R, Al Ajlan N, 2021. Vision transformers for remote sensing image classification. Remote Sens 13:516.

Beriaux E, Jago A, Lucau-Danila C, Planchon V, Defourny P. 2021. Sentinel-1 time series for crop identification in the framework of the future CAP monitoring. Remote Sens 13:2785.

Cartolano A, Cuzzocrea A, Pilato G, 2024. Analyzing and assessing explainable AI models for smart agriculture environments.

Multim Tools Appl 83:1–22.
Castillo-Martínez M, Gallegos-Funes F, Carvajal-Gámez B, Urriolagoitia-Sosa G, Rosales-Silva A, 2020. Color index based thresholding method for background and foreground segmentation of plant images. Comput Electron Agric 178:105783.
Chandra S, Hareendran S, Albaaji G, 2024. Precision farming for sustainability: An agricultural intelligence model. Comput Electron Agric 226:109386.
Charisis C, Argyropoulos D, 2024. Deep learning-based instance segmentation architectures in agriculture: A review of the scopes and challenges. Smart Agric Technol 8:100448.
Chen LC, Papandreou G, Kokkinos I, Murphy K, Yuille AL, 2014. Semantic image segmentation with deep convolutional nets and fully connected CRFs. arXiv 1412.7062v4.
Chen LC, Papandreou G, Kokkinos I, Murphy K, Yuille AL, 2018. DeepLab: Semantic image segmentation with deep convolutional nets, atrous convolution, and fully connected CRFs. IEEE T Pattern Anal Mach Intell 40:834-848.
Chen LC, Papandreou G, Schroff F, Adam H, 2017. Rethinking atrous convolution for semantic image segmentation. arXiv 1706.05587.
Chen LC, Zhu Y, Papandreou G, Schroff F, Adam H, 2018. Encoder-decoder with atrous separable convolution for semantic image segmentation. In: V. Ferrari, M. Hebert, C. Sminchisescu, Y. Weiss (eds.), Computer Vision – ECCV 2018. Cham, Springer; pp. 801–818.
Cheng J Zhu Y, Zhao Y, Li T, Chen M, Sun Q, et al., 2024. Application of an improved U-Net with image-to-image translation and transfer learning in peach orchard segmentation. Int J Appl Earth Obs Geoinf 130:103871.
Chiu M, Xu X, Wei Y, Huang Z, Schwing A, Brunner R, et al., 2020. Agriculture-Vision: A large aerial image database for agricultural pattern analysis. Proc. IEEE/CVF Conf. on Computer Vision and Pattern Recognition (CVPR), Seattle; pp. 2825–2835.
Cui X, Han W, Zhang H, Dong Y, Ma W, Zhai X, et al., 2023. Estimating and mapping the dynamics of soil salinity under different crop types using Sentinel-2 satellite imagery. Geoderma 440:116738.
Darwin B, Dharmaraj P, Prince S, Popescu D, Hemanth D, 2021. Recognition of bloom/yield in crop images using deep learning models for smart agriculture: A review. Agronomy 11:646.
de Souza E, Scharf P, Sudduth K, 2010. Sun position and cloud effects on reflectance and vegetation indices of corn. Agron J 102:734-744.
Deng L, Mao Z, Li X, Hu Z, Duan F, Yan Y, 2018. UAV-based multispectral remote sensing for precision agriculture: A comparison between different cameras. ISPRS J Photogramm Remote Sens 146:124-136.
Di S, Liao M, Zhao Y, Li Y, Zeng Y, 2021. Image superpixel segmentation based on hierarchical multi-level LI-SLIC. Opt Laser Technol 135:106703.
Dobrota C, Carpa R, Butiuc-Keul A, 2021. Analysis of designs used in monitoring crop growth based on remote sensing methods. Turk J Agric For 45:730-742.
Dosovitskiy A, Beyer L, Kolesnikov A, Weissenborn D, Zhai X, Unterthiner T, et al., 2020. An image is worth 16x16 words: Transformers for image recognition at scale. arXiv 2010.11929v2.
Du S, Du S, Liu B, Zhang X, 2021. Incorporating DeepLabV3+ and object-based image analysis for semantic segmentation of very high resolution remote sensing images. Int J Digit Earth 14:357-378.
El Sakka M, Mothe J, Ivanovici M, 2024. Images and CNN applications in smart agriculture. Eur J Remote Sens 57:2352386.
Feng G, Wang C, Wang A, Gao Y, Zhou Y, Huang S, et al., 2024. Segmentation of wheat lodging areas from UAV imagery using an ultra-lightweight network. Agriculture 14:244.
Flores C, Valenzuela A, Verschae R, 2024. Active learning for image classification: A comprehensive analysis in agriculture. In: X.S. Yang, R.S. Sherratt, N. Dey, A. Joshi. (eds), Proc. 9th Int. Cong. on Information and Communication Technology. ICICT 2024. Singapore, Springer; pp. 607-616.
Gao B, Montes M, Davis C, Goetz A, 2009. Atmospheric correction algorithms for hyperspectral remote sensing data of land and ocean. Remote Sens Environ 113: S17-S24.
Gao J, Wang B, Wang Z, Wang Y, Kong F, 2020. A wavelet transform-based image segmentation method. Optik 208:164123.
Gao L, Liu H, Yang M, Chen L, Wan Y, Xiao Z, et al., 2021. STransFuse: fusing Swin transformer and convolutional neural network for remote sensing image semantic segmentation. IEEE J Sel Top Appl Earth Obs Remote Sens 14:10990-11003.
Garioud A, Gonthier N, Landrieu L, De Wit A, Valette M, Poupée M, et al., 2023. FLAIR: A country-scale land cover semantic segmentation dataset from multi-source optical imagery. arXiv 2310.13336v1.
Garnot V, Landrieu L, 2021. Panoptic segmentation of satellite image time series with convolutional temporal attention networks. Proc. 18th IEEE/CVF Int. Conf. on Computer Vision (ICCV), Montreal; pp. 4852-4861.
Ghosh R, Ravirathinam P, Jia X, Khandelwal A, Mulla D, Kumar V, 2021. CalCROP21: A georeferenced multi-spectral dataset of satellite imagery and crop labels. Proc. 9th IEEE Int. Conf. on Big Data, Orlando; pp. 1625-1632.
Gilles J, 2013. Empirical wavelet transform. IEEE T Signal Process 61:3999-4010.
Gonzalo-Martín C, Lillo-Saavedra M, García-Pedrero A, Lagos O, Menasalvas E, 2017. Daily evapotranspiration mapping using regression random forest models. IEEE J Sel Top Appl Earth Obs Remote Sens 10:5359–5368.
Guijarro M, Riomoros I, Pajares G, Zitinski P, 2015. Discrete wavelets transform for improving greenness image segmentation in agricultural images. Comput Electron Agric 118:396-407.
Hadjimitsis D, Papadavid G, Agapiou A, Themistocleous K, Hadjimitsis M, Retalis A, et al., 2010. Atmospheric correction for satellite remotely sensed data intended for agricultural applications: Impact on vegetation indices. Nat Hazards Earth Syst Sci 10:89-95.
Han K, Wang Y, Chen H, Chen X, Guo J, Liu Z, et al., 2023. A survey on vision transformer. IEEE T Pattern Anal Mach Intell 45:87-110.
Hassanein M, Lari Z, El-Sheimy N, 2018. A new vegetation segmentation approach for cropped fields based on threshold detection from hue histograms. Sensors (Basel) 18:1253.
He J, Zhao L, Yang H, Zhang M, Li W, 2020. HSI-BERT: Hyperspectral image classification using the bidirectional encoder representation from transformers. IEEE T Geosci Remote Sens 58:165-178.
He X, Zhou Y, Zhao J, Zhang D, Yao R, Xue Y, 2022. Swin transformer embedding UNet for remote sensing image semantic segmentation. IEEE T Geosci Remote Sens 60:4408715.
Helber P, Bischke B, Dengel A, Borth D, 2019. EuroSAT: A novel

dataset and deep learning benchmark for land use and land cover classification. IEEE J Sel Top Appl Earth Obs Remote Sens 12:2217-2226.

Holder CJ, Shafique M, 2022. On efficient real-time semantic segmentation: A survey. arXiv 2206.08605.

Hong D, Han Z, Yao J, Gao L, Zhang B, Plaza A, et al., 2022. SpectralFormer: Rethinking hyperspectral image classification with transformers. IEEE T Geosci Remote Sens 60:5518615.

Hong D, Zhang B, Li X, Li Y, Li C, Yao J, et al., 2024. SpectralGPT: Spectral remote sensing foundation model. IEEE T Pattern Anal Mach Intell 46:5227–5244.

Hua Y, Marcos D, Mou L, Zhu X, Tuia D, 2022. Semantic segmentation of remote sensing images with sparse annotations. IEEE Geosci Remote Sens Lett 19:3051053.

Jia W, Zheng Y, Zhao D, Yin X, Liu X, Du R, 2018. Preprocessing method of night vision image application in apple harvesting robot. Int J Agric Biol Eng 11:158-163.

Jiang J, Lyu C, Liu S, He Y, Hao X, 2020. RWSNet: A semantic segmentation network based on SegNet combined with random walk for remote sensing. Int J Remote Sens 41:487-505.

Jiang Y, Li C, 2020. Convolutional neural networks for image-based high-throughput plant phenotyping: A review. Plant Phenomics 2020:4152816.

Jiang Y, Tang Y, Li H, 2022. A review of trends in the use of sewage irrigation technology from the livestock and poultry breeding industries for farmlands. Irrig Sci 40:297-308.

Jonnala N, Siraaj S, Prastuti Y, Chinnababu P, Babu B, Bansal S, et al., 2025. AER U-Net: Attention-enhanced multi-scale residual U-Net structure for water body segmentation using Sentinel-2 satellite images. Sci Rep 15:16099.

Karlson M, Ostwald M, Bayala J, Bazié HR, Ouedraogo AS, Soro B, et al., 2020. The potential of Sentinel-2 for crop production estimation in a smallholder agroforestry landscape, Burkina Faso. Front Environ Sci 8:85.

Ke Q, Zhang P, 2022. Hybrid-TransCD: A hybrid transformer remote sensing image change detection network via token aggregation. ISPRS Int J Geo-Inf 11:263.

Kerner H, Sundar S, Satish M, 2024. Multi-region transfer learning for segmentation of crop field boundaries in satellite images with limited labels. arXiv 2404.00179.

Khan S, Narvekar M, 2022. Novel fusion of color balancing and superpixel based approach for detection of tomato plant diseases in natural complex environment. J King Saud Univ-Comput Inf Sci 34:3506-3516.

Khan S, Naseer M, Hayat M, Zamir S, Khan F, Shah M. 2022. Transformers in vision: A survey. ACM Comput Surv 54:200.

Koonce B, 2021. MobileNetV3. In: B. Koonce (ed.), Convolutional neural networks with Swift for Tensorflow: Image recognition and dataset categorization. Berkeley, Apress; pp. 125-144.

Lan Y, Huang K, Yang C, Lei L, Ye J, Zhang J, et al., 2021. Real-time identification of rice weeds by UAV low-altitude remote sensing based on improved semantic segmentation model. Remote Sens 13:4370.

LeCun Y, Bengio Y, Hinton G, 2015. Deep learning. Nature 521:36–444.

Lei L, Yang Q, Yang L, Shen T, Wang R, Fu C, 2024. Deep learning implementation of image segmentation in agricultural applications: A comprehensive review. Artif Intell Rev 57:149.

Li J, Cai Y, Li Q, Kou M, Zhang T, 2024. A review of remote sensing image segmentation by deep learning methods. Int J Digit Earth 17:2328827.

Li J, Luo W, Han L, Cai Z, Guo Z, 2022. Two-wavelength image detection of early decayed oranges by coupling spectral classification with image processing. J Food Compos Anal 111:104642.

Li Z, Chen P, Shuai L, Wang M, Zhang L, Wang Y, et al., 2022. A copy paste and semantic segmentation-based approach for the classification and assessment of significant rice diseases. Plants 11:3174.

Li Z, Chen G, Zhang T, 2020. A CNN-Transformer hybrid approach for crop classification using multitemporal multisensor images. IEEE J Sel Top Appl Earth Obs Remote Sens 13: 847-858.

Li L, Zhang W, Zhang X, Emam M, Jing W, 2023. Semi-supervised remote sensing image semantic segmentation method based on deep learning. Electronics 12:348.

Li X, Wen C, Hu Y, Yuan Z, Zhu XX, 2024. Vision-language models in remote sensing: Current progress and future trends. IEEE Geosci Remote Sens Mag 12:32–66.

Liepa A, Thiel M, Taubenböck H, Steffan-Dewenter I, Abu IO, Singh Dhillon M, et al., 2024. Harmonized NDVI time-series from Landsat and Sentinel-2 reveal phenological patterns of diverse, small-scale cropping systems in East Africa. Remote Sens Appl Soc Environ 35:101230.

Liu B, Li B, Liu H, Li S, 2024. ST-MDAMNet: Swin transformer combines multi-dimensional attention mechanism for semantic segmentation of high-resolution earth surface images. Adv Space Res 74:3691–3705.

Liu B, Li B, Sreeram V, Li S, 2024. MBT-UNet: Multi-branch transform combined with UNet for semantic segmentation of remote sensing images. Remote Sens 16:2776.

Liu B, Yu A, Gao K, Tan X, Sun Y, Yu X, 2022. DSS-TRM: Deep spatial-spectral transformer for hyperspectral image classification. Eur J Remote Sens 55:103-114.

Liu F, Chen D, Guan Z, Zhou X, Zhu J, Ye Q, et al., 2024. RemoteCLIP: A vision language foundation model for remote sensing. IEEE T Geosci Remote Sens 62:5622216.

Liu S, Cao S, Lu X, Peng J, Ping L, Fan X, et al., 2025. Lightweight deep learning model, ConvNeXt-U: An improved U-Net network for extracting cropland in complex landscapes from Gaofen-2 images. Sensors (Basel) 25:261.

Liu Y, Lan Y, Chen X, 2025. Partial convolutional biformer: A transformer architecture for diagnosing crop diseases under complex backgrounds. Crop Prot 193:107007.

Liu Z, Lin Y, Cao Y, Hu H, Wei Y, Zhang Z, et al., 2021. Swin transformer: Hierarchical vision transformer using shifted windows. Proc. IEEE/CVF Int. Conf. on Computer Vision, Montreal; pp. 10012-10022.

Lu P, Zheng W, Lv X, Xu J, Zhang S, Li Y, et al., 2024. An extended method based on the geometric position of salient image features: Solving the dataset imbalance problem in greenhouse tomato growing scenarios. Agriculture 14:1893.

Lu Y, Chen D, Olaniyi E, Huang Y, 2022. Generative adversarial networks (GANs) for image augmentation in agriculture: A systematic review. Comput Electron Agric 200:107208.

Luo Z, Yang W, Yuan Y, Gou R, Li X, 2024. Semantic segmentation of agricultural images: A survey. Inf Process Agric 11:172-186.

Luo X, Liao J, Zang Y, Zhou Z, 2016. Improving agricultural mechanization level to promote agricultural sustainable development. T CSAE 32:1-11.

Lv J, Shen Q, Lv M, Li Y, Shi L, Zhang P, 2023. Deep learning-based semantic segmentation of remote sensing images: A

review. Front Ecol Evol 11:1201125.
Ma D, Jiang L, Li J, Shi Y, 2023. Water index and Swin Transformer Ensemble (WISTE) for water body extraction from multispectral remote sensing images. GIScience Remote Sens 60:2251704.
Ma N, Zhang X, Zheng HT, Sun J, 2018. ShuffleNet V2: Practical guidelines for efficient CNN architecture design. Proc. European Conf. on Computer Vision (ECCV). Cham, Springer; pp. 122-138.
Mamat N, Othman M, Abdoulghafor R, Belhaouari S, Mamat N, Hussein S, 2022. Advanced technology in agriculture industry by implementing image annotation technique and deep learning approach: A review. Agriculture 12:1033.
Marsocci V, Scardapane S, Komodakis N, 2021. MARE: Self-supervised multi-attention REsu-Net for semantic segmentation in remote sensing. Remote Sens 13:3275.
Mehta S, Rastegari M, 2021. MobileViT: Light-weight, general-purpose, and mobile-friendly vision transformer. arXiv 2110.02178v2
Memon M, Chen S, Niu Y, Zhou W, Elsherbiny O, Liang R, et al., 2023. Evaluating the efficacy of Sentinel-2B and Landsat-8 for estimating and mapping wheat straw cover in rice-wheat fields. Agronomy 13:2691.
Meng X, Yang Y, Wang L, Wang T, Li R, Zhang C, 2022. Class-guided swin transformer for semantic segmentation of remote sensing imagery. IEEE Geosci Remote Sens Lett 19:6517505.
Naseer M, Ranasinghe, K, Khan, S, Hayat, M, Khan, F, Yang, M. 2021. Intriguing properties of vision transformers. arXiv 2105.10497v3.
Nazeer M, Ilori C, Bilal M, Nichol J, Wu W, Qiu Z, et al., 2021. Evaluation of atmospheric correction methods for low to high resolutions satellite remote sensing data. Atmos Res 249:105308.
Ntakos G, Prikaziuk E, ten Den T, Reidsma P, Vilfan N, van der Wal T, et al. 2024. Coupled WOFOST and SCOPE model for remote sensing-based crop growth simulations. Comput Electron Agric 225:109238.
O'Shea K, Nash R, 2015. An introduction to convolutional neural networks. arXiv 1511.08458v2.
Ou Y, Yan J, Liang Z, Zhang B, 2024. Hyperspectral imaging combined with deep learning for the early detection of strawberry leaf gray mold disease. Agronomy 14:2694.
Padshetty S, Umashetty A, 2024. Agricultural innovation through deep learning: A hybrid CNN-Transformer architecture for crop disease classification. J Spatial Sci 1-32.
Panboonyuen T, Charoenphon C, Satirapod C, 2023. MeViT: A medium-resolution vision transformer for semantic segmentation on Landsat satellite imagery for agriculture in Thailand. Remote Sens 15:5124.
Pei H, Sun Y, Huang H, Zhang W, Sheng J, Zhang Z, 2022. Weed detection in maize fields by UAV images based on crop row preprocessing and improved YOLOv4. Agriculture 12:975.
Peng Y, Wang A, Liu J, Faheem M, 2021. A comparative study of semantic segmentation models for identification of grape with different varieties. Agriculture 11:997.
Qing J, Deng X, Lan Y, Li Z, 2023. GPT-aided diagnosis on agricultural image based on a new light YOLOPC. Comput Electron Agric 213:108168.
Raei E, Asanjan A, Nikoo M, Sadegh M, Pourshahabi S, Adamowski J, 2022. A deep learning image segmentation model for agricultural irrigation system classification. Comput Electron Agric 198:106977.
Ramos L, Sappa A, 2025. Leveraging U-Net and selective feature extraction for land cover classification using remote sensing imagery. Sci Rep 15:784.
Rasmussen C, Kirk K, Moeslund T, 2022. The challenge of data annotation in deep learning - A case study on whole plant corn silage. Sensors (Basel) 22:1596.
Rehman M, Liu J, Nijabat A, Faheem M, Wang W, Zhao S, 2024. Leveraging convolutional neural networks for disease detection in vegetables: A comprehensive review. Agronomy 14:2231.
Ronneberger O, Fischer P, Brox T, 2015. U-Net: Convolutional networks for biomedical image segmentation. In: N. Navab, J. Hornegger, W.M. Wells, A.F. Frangi (eds.), Medical image computing and computer-assisted intervention – MICCAI 2015. Cham, Springer; pp. 234-241.
Ryo M, 2022. Explainable artificial intelligence and interpretable machine learning for agricultural data analysis. Artif Intell Agric 6:257–265.
Sakka ME, De Pourtales C, Chaari L, Mothe J, 2025. AgriPotential: A novel multi-spectral and multi-temporal remote sensing dataset for agricultural potentials. arXiv 2506.11740.
Sani D, Mahato S, Saini S, Agarwal HK, Devshali CC, Anand S, et al., 2024. SICKLE: A multi-sensor satellite imagery dataset annotated with multiple key cropping parameters. arXiv 2312.00069v1.
Shams M, Gamel S, Talaat F, 2024. Enhancing crop recommendation systems with explainable artificial intelligence: A study on agricultural decision-making. Neural Comput Appl 36:5695–5714.
Shenoy J, Zhang X, Tao B, Mehrotra S, Yang R, Zhao H, et al., 2024. Self-supervised learning across the spectrum. Remote Sens 16:3470.
Singh BM, Komal C, Victorovich KA, 2020. Crop growth monitoring through Sentinel and Landsat data based NDVI time-series. Comput Opt 44:409-419.
Sishodia R, Ray R, Singh S, 2020. Applications of remote sensing in precision agriculture: A review. Remote Sens 12:3136.
Solangi K, Siyal A, Wu Y, Abbasi B, Solangi F, Lakhiar I, et al., 2019. An assessment of the spatial and temporal distribution of soil salinity in combination with field and satellite data: A case study in Sujawal District. Agronomy 9:869.
Song XP, Huang W, Hansen MC, Potapov P, 2021. An evaluation of Landsat, Sentinel-2, Sentinel-1 and MODIS data for crop type mapping. Sci Remote Sens 3:100018.
Su D, Kong H, Qiao Y, Sukkarieh S, 2021. Data augmentation for deep learning based semantic segmentation and crop-weed classification in agricultural robotics. Comput Electron Agric 190:106418.
Sui J, Qin Q, Ren H, Sun Y, Zhang T, Wang J, et al., 2018. Winter wheat production estimation based on environmental stress factors from satellite observations. Remote Sens 10:962.
Sun J, Yang S, Gao X, Ou D, Tian Z, Wu J, et al., 2023. MASA-SegNet: A semantic segmentation network for PolSAR images. Remote Sens 15:3662.
Sun X, Wang P, Lu W, Zhu Z, Lu X, He Q, et al., 2023. RingMo: A remote sensing foundation model with masked image modeling. IEEE T Geosci Remote Sens 61:5612822.
Sykas D, Sdraka M, Zografakis D, Papoutsis I, 2022. A Sentinel-2 multiyear, multicountry benchmark dataset for crop classification and segmentation with deep learning. IEEE J Sel Top Appl Earth Obs Remote Sens 15:3323-3339.

Tao K, Wang A, Shen Y, Lu Z, Peng F, Wei X, 2022. Peach flower density detection based on an improved CNN incorporating attention mechanism and multi-scale feature fusion. Horticulturae 8:904.

Tao L, Zhang H, Jing H, Liu Y, Yan D, Wei G, et al., 2025. Advancements in vision-language models for remote sensing: Datasets, capabilities, and enhancement techniques. Remote Sens 17:162.

Tianxiang Z, Yuanxiu C, Peixian Z, Jiangyun L, 2024. Remotely sensed crop disease monitoring by machine learning algorithms: A review. Unmanned Syst12:161-171.

Tong X, Xia G, Lu Q, Shen H, Li S, You S, et al., 2020. Land-cover classification with high-resolution remote sensing images using transferable deep models. Remote Sens Environ 237:111322.

Tsakiridis NL, Diamantopoulos T, Symeonidis AL, Theocharis JB, Iossifides A, Chatzimisios P, et al., 2020. Versatile Internet of Things for agriculture: An eXplainable AI approach. In: I. Maglogiannis, L. Iliadis, E. Pimenidis (eds.), Artificial intelligence applications and innovations. AIAI 2020. Cham, Springer; pp. 180-191.

Tseng G, Zvonkov I, Nakalembe CL, Kerner H, 2021. CropHarvest: A global dataset for crop-type classification. Proc. 35th Conf. on Neural Information Processing Systems Datasets and Benchmarks Track.

Ullah H, Bais A, 2022. Evaluation of model generalization for growing plants using conditional learning. Artif Intell Agric 6:189–198.

Vaswani A, Shazeer N, Parmar N, Uszkoreit J, Jones L, Gomez AN, et al., 2017. Attention is all you need. arXiv 1706.03762v7.

Wan L, Li H, Li C, Wang A, Yang Y, Wang P, 2022. Hyperspectral sensing of plant diseases: principle and methods. Agronomy 12:1451.

Wang J, Ding C, Chen S, He C, Luo B, 2020. Semi-supervised remote sensing image semantic segmentation via consistency regularization and average update of pseudo-label. Remote Sens 12:3603.

Wang J, Qi Z, Wang Y, Liu Y, 2025. A lightweight weed detection model for cotton fields based on an improved YOLOv8n. Sci Rep 15:457.

Wang Q, Qin W, Liu M, Zhao J, Zhu Q, Yin Y, 2024. Semantic segmentation model-based boundary line recognition method for wheat harvesting. Agriculture 14:1846.

Wang R, Ma L, He G, Johnson B, Yan Z, Chang M, et al., 2024. Transformers for remote sensing: A systematic review and analysis. Sensors (Basel) 24:3495.

Wang S, Chen W, Xie S, Azzari G, Lobell D, 2020. Weakly supervised deep learning for segmentation of remote sensing imagery. Remote Sens 12:207.

Wang Y, Yang L, Liu X, Yan P, 2024. An improved semantic segmentation algorithm for high-resolution remote sensing images based on DeepLabV3+. Sci Rep 14:9716.

Wang Y, Zhang X, Ma G, Du X, Shaheen N, Mao H, 2021. Recognition of weeds at asparagus fields using multi-feature fusion and backpropagation neural network. Int J Agric Biol Eng 14:190-198.

Wang Z, Wang J, Yang K, Wang L, Su F, Chen X, 2022. Semantic segmentation of high-resolution remote sensing images based on a class feature attention mechanism fused with DeeplabV3+. Comput Geosci 158:104969.

Weiss M, Jacob F, Duveiller G, 2020. Remote sensing for agricultural applications: A meta-review. Remote Sens Environ 236:111402.

Weng L, Xu Y, Xia M, Zhang Y, Liu J, Xu Y, 2020. Water areas segmentation from remote sensing images using a separable residual SegNet network. ISPRS Int J Geo-Inf 9:256.

Weng X, Pang C, Xia G, 2025. Vision-language modeling meets remote sensing: Models, datasets, and perspectives. IEEE Geosci Remote Sens Mag 13:3572702.

Weyler J, Magistri F, Marks E, Chong Y, Sodano M, Roggiolani G, et al., 2024. PhenoBench: A large dataset and benchmarks for semantic image interpretation in the agricultural domain. IEEE T Pattern Anal Mach Intell 46:9583-9594.

Wu H, Du Z, Zhong D, Wang Y, Tao C, 2025. FSVLM: A vision-language model for remote sensing farmland segmentation. IEEE T Geosci Remote Sens 63:4402813.

Wu J, Pichler D, Marley D, Wilson D, Hovakimyan N, Hobbs J, 2023. Extended agriculture-vision: An extension of a large aerial image dataset for agricultural pattern analysis. arXiv 2303.02460v1.

Wu Z, Gao Y, Li L, Xue J, Li Y, 2019. Semantic segmentation of high-resolution remote sensing images using fully convolutional network with adaptive threshold. Connect Sci 31:169-184.

Xiang J, Liu J, Chen D, Xiong Q, Deng C, 2023. CTFuseNet: A multi-scale CNN-Transformer feature fused network for crop type segmentation on UAV remote sensing imagery. Remote Sens 15:1151.

Xiao A, Xuan W, Wang J, Huang J, Tao D, Lu S, et al., 2025. Foundation models for remote sensing and Earth observation: A survey. IEEE Geosci Remote Sens Mag 13:2–29.

Xie E, Wang W, Yu Z, Anandkumar A, Alvarez J, Luo P, 2021. SegFormer: Simple and efficient design for semantic segmentation with transformers. Proc. 35th Annual Conf. on Neural Information Processing Systems; pp. 12077-12090.

Xie Y, Guo Y, Mi Z, Yang Y, Obaidat M, 2023. Edge-assisted real-time instance segmentation for resource-limited IoT devices. IEEE Internet Things J 10:473-485.

Xing J, Sieber R, Kalacska M, 2014. The challenges of image segmentation in big remotely sensed imagery data. Ann GIS 20:233–244.

Xu M, Sun J, Zhou X, Tang N, Shen J, Wu X, 2021. Research on nondestructive identification of grape varieties based on EEMD-DWT and hyperspectral image. J Food Sci 86:2011-2023.

Xu S, Xu X, Zhu Q, Meng Y, Yang G, Feng H, et al., 2023. Monitoring leaf nitrogen content in rice based on information fusion of multi-sensor imagery from UAV. Precis Agric 24:2327-2349.

Xu Z, Zhang W, Zhang T, Yang Z, Li J, 2021. Efficient Transformer for remote sensing image segmentation. Remote Sens 13:3585.

Yang Y, Wang Y, Dong J, Yu B, 2024. A knowledge distillation-based ground feature classification network with multiscale feature fusion in remote-sensing images. IEEE J Sel Top Appl Earth Obs Remote Sens 17:2347-2359.

Yasir M, Wan J, Liu S, Sheng H, Xu M, Hossain M, 2023. Coupling of deep learning and remote sensing: A comprehensive systematic literature review. Int J Remote Sens 44:157-193.

Zhan Y, Xiong Z, Yuan Y, 2025. SkyEyeGPT: Unifying remote sensing vision-language tasks via instruction tuning with large language model. ISPRS J Photogramm Remote Sens 221:64-

77.
Zhang D, Liu Z, Shi X, 2020. Transfer learning on EfficientNet for remote sensing image classification. Proc. 5th Int. Conf. on Mechanical, Control and Computer Engineering (ICMCCE), Harbin; pp. 2255-2258.
Zhang N, Yang G, Pan Y, Yang X, Chen L, Zhao C, 2020. A review of advanced technologies and development for hyperspectral-based plant disease detection in the past three decades. Remote Sens 12:3188.
Zhang T, Wang W, Wang J, Cai Y, Yang Z, Li J, 2022. Hyper-LGNet: Coupling local and global features for hyperspectral image classification. Remote Sens 14:5251.
Zhang T, Xu Z, Su J, Yang Z, Liu C, Chen W, et al., 2021. Ir-UNet: Irregular segmentation U-shape network for wheat yellow rust detection by UAV multispectral imagery. Remote Sens 13:3892
Zhang T, Yang Z, Xu Z, Li J, 2022. Wheat yellow rust severity detection by efficient DF-UNet and UAV multispectral imagery. IEEE Sensors J 22:9057-9068.
Zhang X, Zhang S, Meng X, Zhang G, Zang D, Han Y, et al., 2024. Remote sensing image segmentation of gully erosion in a typical black soil area in Northeast China based on improved DeepLabV3+ model. Ecol Inform 84:102929.
Zhang Y, Lv C, 2024. TinySegformer: A lightweight visual segmentation model for real-time agricultural pest detection. Comput Electron Agric 218:108740.
Zhang Y, Sun J, Li J, Wu X, Dai C, 2018. Quantitative analysis of cadmium content in tomato leaves based on hyperspectral image and feature selection. Appl Eng Agric 34:789-798.
Zhang Z, Lu Y, Zhao Y, Pan Q, Jin K, Xu G, et al., 2023. TS-YOLO: An all-day and lightweight tea canopy shoots detection model. Agronomy 13:1411.
Zhang Z, Zhao T, Guo Y, Yin J, 2024. RS5M and GeoRSCLIP: A large-scale vision-language dataset and a large vision-language model for remote sensing. IEEE T Geosci Remote Sens 62:5642123.
Zhao Y, Xie J, Zhu H, Luo T, Xiong Y, Fan C, et al., 2025. Land-Unet: A deep learning network for precise segmentation and identification of non-structured land use types in rural areas for green urban space analysis. Ecol Inform 87:103078.
Zhao Y, Zhang X, Sun J, Yu T, Cai Z, Zhang Z, et al., 2024. Low-cost lettuce height measurement based on depth vision and lightweight instance segmentation model. Agriculture 14:1596.
Zhao X, Wang L, Zhang Y, Han X, Deveci M, Parmar M, 2024. A review of convolutional neural networks in computer vision. Artif Intell Rev 57:99.
Zheng K, Chen Y, Wang J, Liu Z, Bao S, Zhan J, et al., 2025. Enhancing remote sensing semantic segmentation accuracy and efficiency through transformer and knowledge distillation. IEEE J Sel Top Appl Earth Obs Remote Sens 18:4074–4092.
Zheng Z, Yuan J, Yao W, Yao H, Liu Q, Guo L, 2024. Crop classification from drone imagery based on lightweight semantic segmentation methods. Remote Sens 16:4099.
Zhong B, Wei T, Luo X, Du B, Hu L, Ao K, et al. 2023. Multi-Swin mask transformer for instance segmentation of agricultural field extraction. Remote Sens 15:549.
Zhu H, Lin C, Liu G, Wang D, Qin S, Li A, et al. 2024. Intelligent agriculture: Deep learning in UAV-based remote sensing imagery for crop diseases and pests detection. Front Plant Sci 15:1435016.
Zhu H, Qin S, Su M, Lin C, Li A, Gao J, 2024. Harnessing large vision and language models in agriculture: A review. arXiv:2407.19679.
Zhu H, Wang D, Wei Y, Zhang X, Li L, 2024. Combining transfer learning and ensemble algorithms for improved citrus leaf disease classification. Agriculture 14:1549.
Zhu W, Feng Z, Dai S, Zhang P, Wei X, 2022. Using UAV multispectral remote sensing with appropriate spatial resolution and machine learning to monitor wheat scab. Agriculture 12:1785.
Zhu W, Sun J, Wang S, Shen J, Yang K, Zhou X, 2022. Identifying field crop diseases using Transformer-embedded convolutional neural network. Agriculture 12:1083.
Zhu X, Chikangaise P, Shi W, Chen W, Yuan S, 2018. Review of intelligent sprinkler irrigation technologies for remote autonomous system. Int J Agric Biol Eng 11:23-30.
Zuo Z, Gao S, Peng H, Xue Y, Han L, Ma G, et al., 2024. Lightweight detection of broccoli heads in complex field environments based on LBDC-YOLO. Agronomy 14:2359.

Received: 26 August 2025; Accepted: 19 November 2025.

Contributions: all authors made a substantive intellectual contribution, read and approved the final version of the manuscript and agreed to be accountable for all aspects of the work.

Conflict of interest: the authors declare no competing interests, and all authors confirm accuracy.

Acknowledgments: this research was funded in part by the National Key Research and Development Program for Young Scientists under Grant 2022YFD2000200; in part by the Natural Science Foundation of Jiangsu Higher Education Institutions under Grant 22KJB520015; and in part by the Scientific Research Foundation of Jiangsu University under Grant 21JDG051.