Transformer-based Detection of Abnormal Rice Growth using Drone-based Multispectral Imaging

Yanfen Li[a,1], L. Minh Dang[b,c,d,1], Hanxiang Wang[a,*], Muhammad Fayaz[e], Sufyan Danish[e], Junliang Shang[a], Hyoung-Kyu Song[d], Hyeonjoon Moon[e,*]

[a] *School of Computer Science, Qufu Normal University, Rizhao, China*
[b] *Institute of Research and Development, Duy Tan University, Da Nang, 550000, Viet Nam*
[c] *Faculty of Information Technology, Duy Tan University, Da Nang, 550000, Viet Nam*
[d] *Department of Information and Communication Engineering and Convergence Engineering for Intelligent Drone, Sejong University, Seoul 05006, South Korea*
[e] *Department of Computer Science and Engineering, Sejong University, Seoul 05006, South Korea*

Abstract

Rice is a vital staple food for global food security and a primary income source for millions of farmers worldwide. However, abnormal rice growth poses a serious threat to both yield stability and grain quality, undermining agricultural productivity. Early detection of such anomalies is therefore essential to mitigate yield losses. However, existing methods either targeted only one symptom at a time, or failed to generalize under various field conditions. Moreover, lightweight real-time inference is needed for on-board UAV deployment, yet most high-accuracy models incur prohibitive computational cost. In this study, we propose ARG-TR model, a lightweight transformer-based semantic segmentation framework built on the SegFormer architecture, which utilizes long-range dependencies to identify complex growth anomalies. The model is trained and validated on a large-scale, drone-captured multi-spectral dataset. By integrating a hierarchical transformer encoder with a lightweight decoder, ARG-TR achieves rapid convergence during training and demonstrates strong generalization to unseen data. The experimental results on a challenging dataset of abnormal rice growth

*Corresponding authors
Email addresses: hanxiang@qufu.edu.cn (Hanxiang Wang), hmoon@sejong.ac.kr (Hyeonjoon Moon)

[1] These authors contributed equally to this work.

patterns show that ARG-TR achieves a robust Intersection over Union (IoU) of 64.8, which outperforms state-of-the-art baselines such as MaskFormer and KNet in both accuracy and computational efficiency.

Keywords: abnormal growth, rice, transformers, semantic segmentation, lodging

1. Introduction

The Food and Agriculture Organization (FAO) of the United Nations (Food & of the United Nations, 2025) projects that the global population will reach 9.2 billion by 2050. To meet the food demands of this growing population, global agricultural production must increase by 60–70% from current levels, as emphasized in multiple FAO reports (Samal et al., 2022; Stankus, 2021). Rice, a staple food for over half the world's population, predominantly in Asia, plays a critical role in global food security (Bin Rahman & Zhang, 2023). However, it is increasingly challenging to achieve the required production target due to abnormal growth patterns in rice, which manifest through various symptoms including stunted growth, delayed flowering, malformed grains, lodging, missing plants, and disease-specific damages, such as rice blast disease. These abnormalities stem from biotic (e.g., diseases, pests) and abiotic stressors (e.g., nutrient deficiencies, environmental pressures) (Dang et al., 2024), which disrupt normal crop development by reducing photosynthetic efficiency and compromising plant components (Rezvi et al., 2023). For example, rice blast disease (Magnaporthe oryzae) destroys photosynthetic tissues, while pest infestations weaken vital components, directly compromising both yield quantity and quality. These issues threaten to destabilize rice production if left untreated, undermining food security, farmers' livelihoods, and economic stability in rice-dependent regions. Therefore, it is urgent to address abnormal rice growth through targeted mitigation strategies to safeguard sustainable rice production and global food security.

Traditionally, abnormal rice growth detection and diagnosis relies heavily on manual field inspections by agricultural experts. However, this approach is labor-intensive, time-consuming, and impractical for large-scale monitoring due to the massive size of rice fields. Inspecting individual rice plants for subtle growth anomalies on vast areas is logistically unfeasible. To address these challenges, automated inspection systems are critical for enabling timely, field-scale assessment of crop health. Unmanned aerial vehicles (UAVs), equipped with RGB, multispectral, or thermal imaging sensors provide a viable solution for early-stage anomaly detection (Dang et al., 2020). By capturing high-resolution aerial imagery, UAVs offer a bird's-eye view that reveals subtle stress indicators, such as chlorosis, stunted growth, or canopy structural variations, often undetectable at ground level. UAVs generate large-scale datasets that motivate the development of advanced computer vision (CV) frameworks capable of automated feature extraction, anomaly classification, and quantifiable stress mapping to transform raw imagery into actionable interventions.

While conventional ML methods depend on handcrafted feature extraction, deep learning (DL) models, particularly convolutional neural networks (CNNs), demonstrate superior capacity for automated detection of subtle rice growth abnormalities from UAV or ground-level imagery (Alam et al., 2025). DL models automatically learn discriminative features, such as color, texture, structural, and spatial domains, to identify issues such as stunted growth, disease, or nutrient deficiencies with minimal human intervention. As a result, DL has achieved state-of-the-art performance in various tasks for precision agriculture, including classification (Li et al., 2020; Dang et al., 2020), detection (Dosset et al., 2025; Wang et al., 2024), and segmentation (Alam et al., 2025; Zhang et al., 2021a). For example, Tian et al. (Tian et al., 2021) employed partial least squares discrimination analysis on UAV multispectral data to detect rice lodging. By utilizing spectral, textural, and color features, the model achieved over 90% accuracy. However, its handcrafted spectral features exhibited limited generalization for various cultivars, growth stages, and regional conditions because the model was fine-tuned on Shanghai paddy characteristics. With a more advanced approach, Yang et al. (Yang et al., 2020) introduced an adaptive UAV-based scouting system that combines multi-altitude imaging and a deep segmentation model to detect rice and lodging. The model achieved 95.28% rice identification and 86.17% lodging detection. However, the results are based primarily on simulations and selected UAV energy profiles. On the other hand, Zhang et al. (Zhang et al., 2021a) developed Ir-UNet, a DL model for wheat yellow rust detection. By integrating irregular convolution and content-aware channel reweighting modules, Ir-UNet addressed challenges posed by irregularly shaped and blurred disease boundaries. The experimental results showed that the model achieved 97.13% overall accuracy on UAV multispectral data and maintained robustness with reduced input features. Recently, Wu et al. (Wu et al., 2025) proposed a YOLOv5-based pipeline for missing rice seedling detection using UAV images. UAV images were first stitched into a geo-referenced panoramic view and then cropped to a series 640×640 patches for dataset creation. The patches were used to train a YOLOv5, which achieved an 80% recall and 75% precision in identifying missing rice seedlings. However, GPS-dependent image stitching and predefined thresholds degraded performance in fields with irregular planting patterns or GPS drift, and the rectangular detection regions could miss seedlings in non-uniform layouts. In general, previous approaches suffer from three main limitations: (1) single-symptom detection, such as lodging or single disease detection, (2) poor generalizability due to limited labeled training data, (3) limited adaptability to irregular input due to grid layouts.

Originally developed for natural language processing (NLP), transformer models revolutionized CV by introducing a self-attention mechanism to model long-range spatial dependencies and global context (Lin et al., 2022). The Vision Transformer (ViT) (Dosovitskiy et al., 2020) pioneered this for CV by partitioning images into patch tokens, but its computational inefficiency limited dense prediction tasks. Swin Transformer (Liu et al., 2021) addressed this by introducing hierarchical feature extraction and shifted windowing scheme to improve efficiency and spatial reasoning for dense prediction tasks like semantic segmentation. Building on these innovations, SegFormer (Xie et al., 2021) emerged as a state-of-the-art semantic segmentation model. It combined transformer-based global context modeling with a lightweight, hierarchical architecture to simultaneously capture fine-grained details and broader contextual relationships. It achieved high accuracy of 84.0% on Cityscapes with only around 3.8 million parameters. Therefore, SegFormer was proved to be suitable for tasks requiring precise localization of subtle anomalies like abnormal rice growth identification.

Building on SegFormer's efficiency and robustness in agricultural applications (Spasev et al., 2024; Nuradili et al., 2024), this study proposes a light-weight transformer-based framework engineered to overcome the multi-symptom detection gap identified in Section 2. The model simultaneously identifies four different rice growth abnormalities using multispectral imaging. Key contributions include.

- Comprehensive data processing and effective post-processing to generate precise orthophotos for the collected multi-spectral dataset.
- A large-scale UAV-based remote sensing dataset containing over 378,000 images.
- An optimized spectral fusion of green, near-infrared, and red-edge to improve image quality for accurate abnormal rice growth recognition.
- A light-weight transformer-based system for the identification of four abnormal rice growth symptoms.

The rest of this paper is organized as follows. Section 3 describes the abnormal rice growth dataset used in this study. Section 4 presents the proposed ARG-TR framework for multi-spectral rice growth segmentation. Section 5 reports the experimental setup and results. Section 6 discusses the main findings, limitations, and practical implications. Finally, Section 7 concludes the paper and outlines directions for future research.

2. System Overview

Figure 1 depicts the main processes of AGR-TR, a multi-symptom abnormal rice growth segmentation framework. In this context, AGR indicates that the framework is applied to agriculture context, while TR refers to the transformer-based architecture.

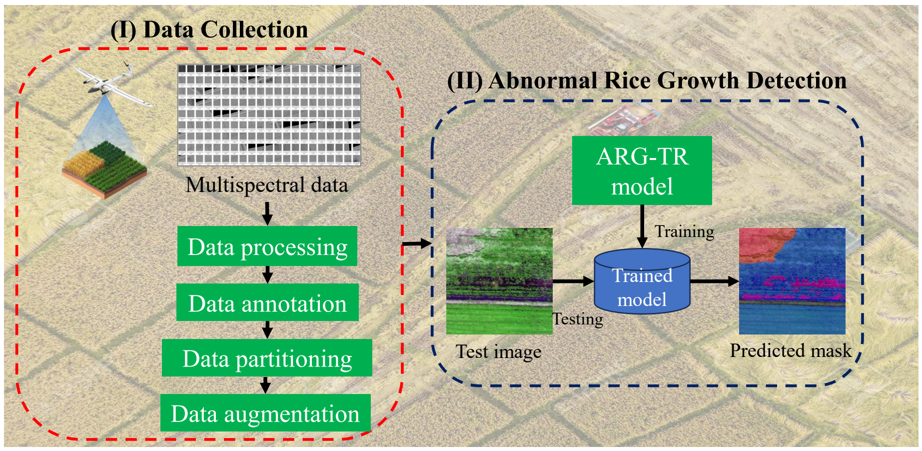

Figure 1: Depiction of the main components of the transformer-based abnormal rice growth detection framework (AGR-TR).

The framework consists of two sequential stages: data preparation and abnormal rice growth detection. In the first stage, multi-spectral UAV imagery undergoes various preprocessing steps to reduce noise and enhance quality. The preprocessed images are then annotated with pixel-level labels to distinguish abnormal growth regions. Next, the dataset is partitioned into training and validation sets. Finally, the images from the training set is augmented to improve model robustness. The second stage employs and fine-tunes a transformer-based SegFormer to segment abnormal growth areas. The model is trained on the processed dataset to automatically discriminate against spatial-spectral patterns. During inference, the trained model processes an input image to generate an output mask that highlights regions of abnormal rice growth. This end-to-end pipeline integrates advanced CV techniques with agronomic insights to support real-time rice health monitoring.

3. Abnormal Rice Growth Dataset

This study utilizes a large-scale abnormal rice growth dataset comprising approximately 378,000 multi-spectral images capturing four distinct patterns of abnormal growth. The dataset was made available for research purposes by the National Information Society Agency of Korea (NIA)[2], which ensures robustness and practical relevance for real-world agricultural applications. The dataset was developed through a collaborative initiative led by Geomatic Limited[3] in partnership with various organizations. Sunyoungeng Limited[4] manages data collection, whereas NEWLAYER Limited [5] and Muhanit Limited [6] handles data annotation and preprocessing.

3.1. Data collection

Abnormal rice growth data were collected from 2022 to 2023 in a 100-hectare experimental crop field in Jangan-ri, Jangan-myeon, Hwaseong City, Gyeonggi Province, South Korea (Figure 2). The site features a temperate monsoon climate ideal for rice cultivation, characterized by warm, humid summers (25–30 °C) and annual rainfall of 1,100–1,400 mm concentrated during the summer monsoon season. Fertile loamy to clay-loamy soils (pH 5.5–7.0) ensure strong water retention and nutrient availability (Ju et al., 2022), while the flat topography supports efficient irrigation and uniform field management.

The *Oryza sativa* 'Odae' cultivar (a widely cultivated Japonica variety) was transplanted on 26 May 2022 at a density of 30 cm × 17 cm. Fertilization followed regional standards with applications of nitrogen ($89\,\mathrm{kg/hm^2}$), phosphorus ($40\,\mathrm{kg/hm^2}$), and potash ($53\,\mathrm{kg/hm^2}$). Data collection covered five critical growth stages, including tillering, panicle initiation, booting, heading & flowering, grain filling. This study specifically targets four high-impact abnormal

[2] https://www.nia.or.kr/site/nia_kor/main.do
[3] https://www.geomatic.co.kr/
[4] http://nonghyup.ac.kr/e_main.asp
[5] http://egis.everlinks.co.kr/
[6] https://muhanit.kr/

growth groups: (1) missing plants (indicating seedling establishment failure), (2) lodging (stem collapse compromising harvest efficiency), (3) rice blast disease (Magnaporthe oryzae infection causing necrotic lesions), and (4) poor growth (exhibiting chlorosis, reduced tillering, and diminished vigor).

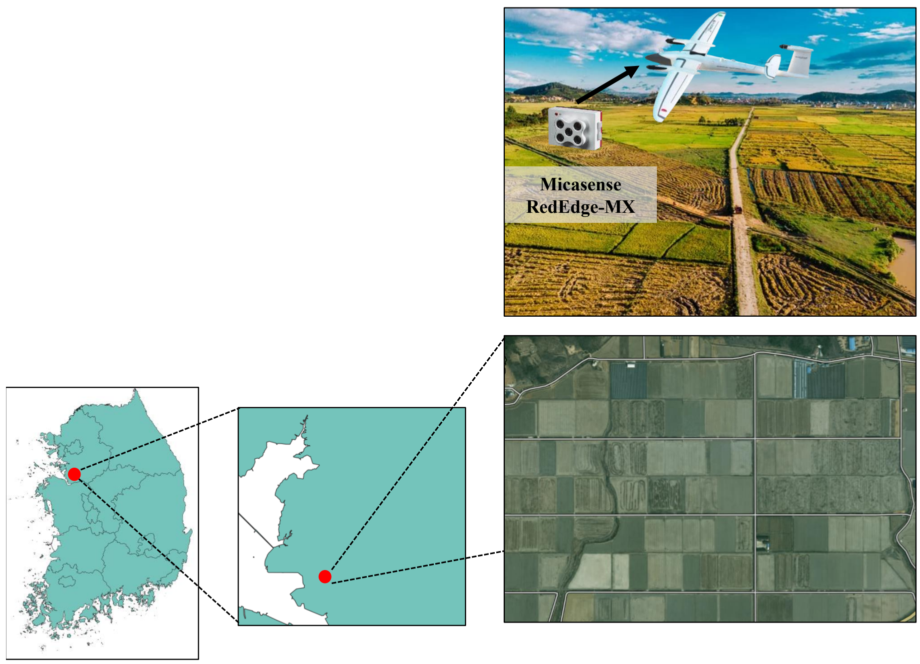

Figure 2: Abnormal rice growth test bed.

The 100 hectares experimental rice field was divided into 40 zones, with data collected from 10 representative plots per zone, leading to a total of 400 monitored plots. Data acquisition main focus was on capturing high-resolution RGB and multispectral imagery to identify rice field abnormalities using TRINITY F90+ (Measurusa, 2025). The TRINITY F90+ is a certified vertical take-off and landing mapping drone, which features a 2.394 m wingspan and a 5.0 kg maximum take-off weight. With a maximum flight time of 90 minutes and operational range of up to 100 km, it can cover approximately 700 hectares in a single flight. The drone is compatible with various payloads, including the Micasense RedEdge-MX multispectral camera (MicaSense, 2025), which captures imagery on five narrow spectral bands (blue, green, red, red-edge, and near-infrared). The integration of RedEdge-MX with the TRINITY F90+ enabled efficient, high-fidelity multispectral data acquisition and provided detailed insights into crop health, stress detection, and growth dynamics. Flights were performed at 120 m altitude and 5 m/s ground speed, which achieved a ground sample distance (GSD) of 8 cm per pixel.

Environmental variables, such as wind gusts, lighting conditions and phenological factors, can significantly influence spectral interpretation. For example, midday sun creates strong shadows that exaggerate canopy gaps, while variable solar angles alter reflectance baselines for chlorosis detection. On the other hand, late-season tillering changes the reflectance baseline against which stunting or chlorosis is detected. To address environmental variables affecting spectral interpretation, the data acquisition implemented three critical controls: (1) All flights conducted between 09:00-11:00 KST under clear-sky conditions to minimize solar angle variation, (2) Geometric correction using calibration and ground control point, (3) collection of five growth stages (tillering to grain filling) to enable robustness against canopy architecture changes.

3.2. Data processing

Figure 3 illustrates the end-to-end workflow for analyzing abnormal rice growth using drone-based multispectral imagery from the experimental field. Initially, all raw imagery undergoes internal quality assurance review by the lead data collector before transmission to the processing team. The workflow begins with raw multispectral data, which is aligned to ensure consistent spatial overlap between images. Next, spectral calibration corrects environmental variability (e.g., lighting, atmospheric conditions) and sensor inconsistencies. Ground control point (GCP) correction then enhances positional accuracy by aligning image coordinates with real-world locations (Agüera-Vega et al., 2017), followed by geometric correction to address distortions from sensor tilt or terrain variations. These preprocessing steps collectively produce precise, georeferenced orthophotos, used for subsequent analysis.

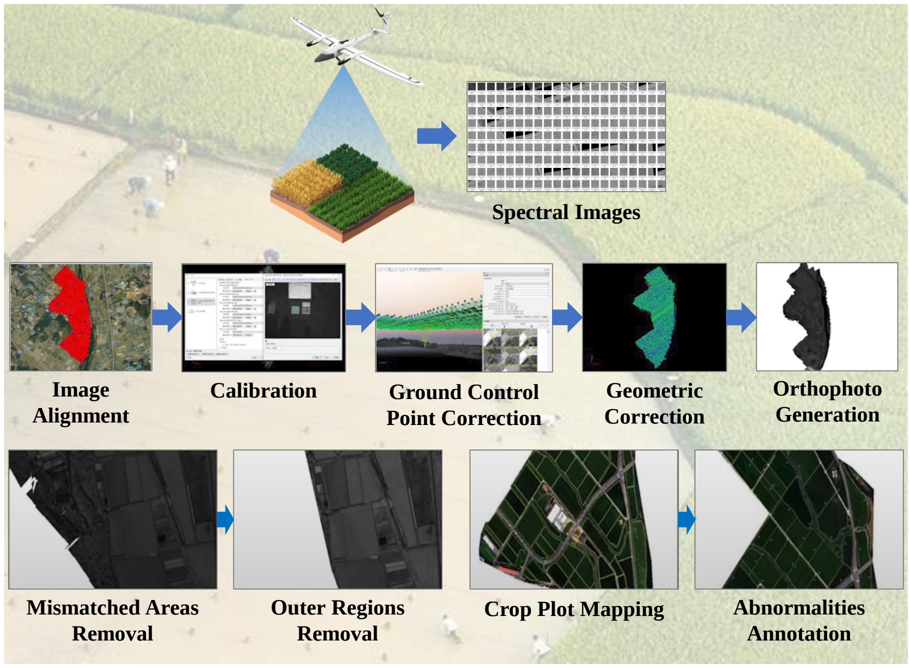

Figure 3: Depiction of the main processing steps for the collected abnormal rice growth dataset.

Prior research has established the critical role of Green, Near-infrared (NIR), and Red-edge (RE) spectral bands for vegetation analysis (Biswal et al., 2024; Kang et al., 2021). Biswal et al. (Biswal et al., 2024) demonstrated the exclusive use of these three bands for estimating paddy crop aboveground biomass, while Kang et al. (Kang et al., 2021) highlighted the role of features derived from RE-NIR-Green band combinations in crop classification. Building on this foundation, Green, NIR, and RE bands were merged to segment abnormal rice growth, as the merged version enhanced detection of plant stress, growth anomalies, and terrain characteristics (Dang et al., 2024). Figure 4 illustrates the creation of combined RGB-like images from these bands within multispectral orthophotos. This process merges individual channel images into a 3-channel format compatible with standard CV algorithms and DL models.

Finally, a post-processing pipeline was carried out to ensure the usability and

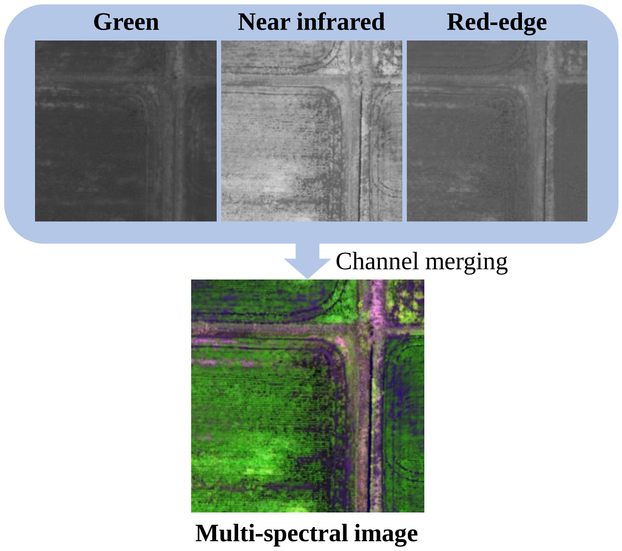

Figure 4: Example of RGB-like image generation through channel merging .

accuracy of the dataset. Irrelevant or distorted sections, such as mismatched areas and outer regions, were removed. The refined orthophoto was divided into crop plots for localized analysis. Finally, abnormalities, such as rice blast disease, lodging, poor growth, and missing plants, were labeled to support model training and validation.

The drone's high-resolution camera delivered a GSD of 8 cm/pixel, sufficient to resolve individual plants and small abnormal growth patches. For annotation process, we overlaid the farm-plot map onto the orthomosaics and annotated each plot showing abnormal growth. To ensure the accuracy of annotations in drone imagery captured at 120 meters altitude, a multi-stage verification protocol was implemented. Vegetation indices, including Normalized Difference Vegetation Index (NDVI)/Enhanced Vegetation Index (EVI), were computed to highlight potential anomalies invisible in visible spectra. For example, missing plants were identified by marking regions with NDVI values below a predefined threshold, while poor growth was annotated using EVI. To reduce ambiguity and improve annotation consistency, the missing plants class is strictly defined as continuous bare-soil regions within the planted area rather than isolated gaps. In practice, annotators marked a region as missing plants only when the bare-soil patch formed a coherent area spanning multiple adjacent planting rows or otherwise appeared as a continuous discontinuity in the crop canopy. Single-plant gaps, isolated pixels, thin shadows, wheel tracks, and other narrow non-crop features were annotated as healthy crop. Lodging was labeled by converting imagery to RGB format and marking the flattened rice locations. However, rice blast disease showed no distinctive spectral signatures that could be initially labeled. Therefore, it was labeled immediately post-flight by trained crowd workers using handheld RGB cameras and GPS markers during field surveys. All annotators completed rigorous training in multispectral image interpretation prior to labeling. Moreover, annotated images were validated through random field surveys to confirm the presence of annotated abnormalities. This integrated approach ensured annotation reliability despite the challenges of high-altitude aerial observation.

The drone-based multispectral imaging approach offers valuable insights into abnormal rice growth but faces several data acquisition limitations. (1) Operational constraints such as limited drone flight time, altitude restrictions, and narrow camera field of view complicated data collection and processing. (2) Variations in solar illumination required complicated post-collection data processing, and data were collected only at five discrete growth stages with 7-10 day intervals, potentially missing rapid rice blast disease developments. (3) The study's focus on a single geographic location with specific soil and climate conditions limits generalizability to other regions.

3.3. Dataset description

Figure 5 presents the class distribution of the abnormal rice growth dataset. The dataset contains 378,074 annotated images in five classes: normal conditions, rice blast disease, lodging, poor growth, and missing plants. For model development, 80% of the dataset (302,450 images) was used for training and validation purposes (226,853 images for training and 75,597 images for validation), 20% of the dataset was used for testing (75,624 images).

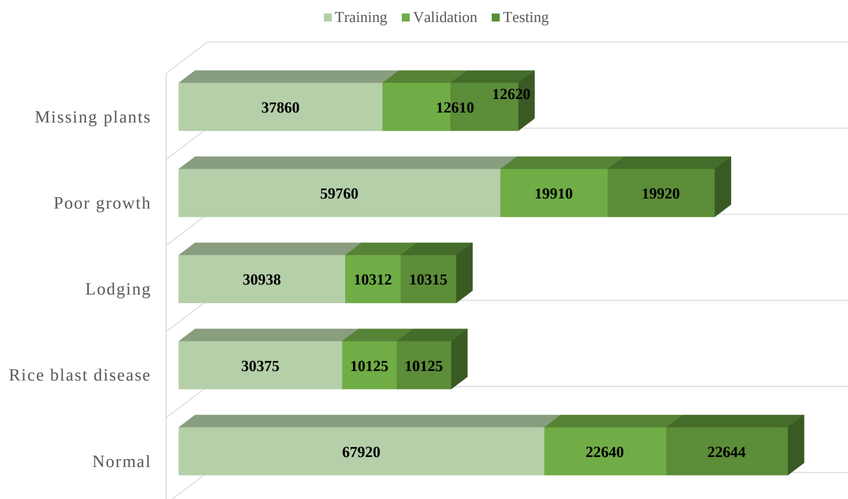

Figure 5: A bar chart showing the distribution of images across different classes in the collected dataset, including normal conditions, rice blast disease, lodging, poor growth, and missing plants.

3.4. Data augmentation

Data augmentation plays a vital role in enhancing model robustness for rare anomalies, including rice blast disease, lodging, and missing plants, by mitigating class imbalance in the training set. A multi-stage augmentation pipeline was implemented to improve generalization in spatial, spectral, and scale variation. Figure 6 provides examples of augmented images using the augmentation pipeline.

The images were first scaled by a random factor ranging from 0.5 to 2.0 followed by resizing to 512×512 pixels. This step aims to improve the model multi-scale robustness by simulating variations in object scale and distance. After that, a RandomCrop operation was implemented to sample various regions to increase diversity in spatial composition. Next, the images were flipped randomly to introduce invariance to orientation. Finally, color jittering (brightness, contrast, saturation) was applied to mimic diverse lighting conditions and sen-

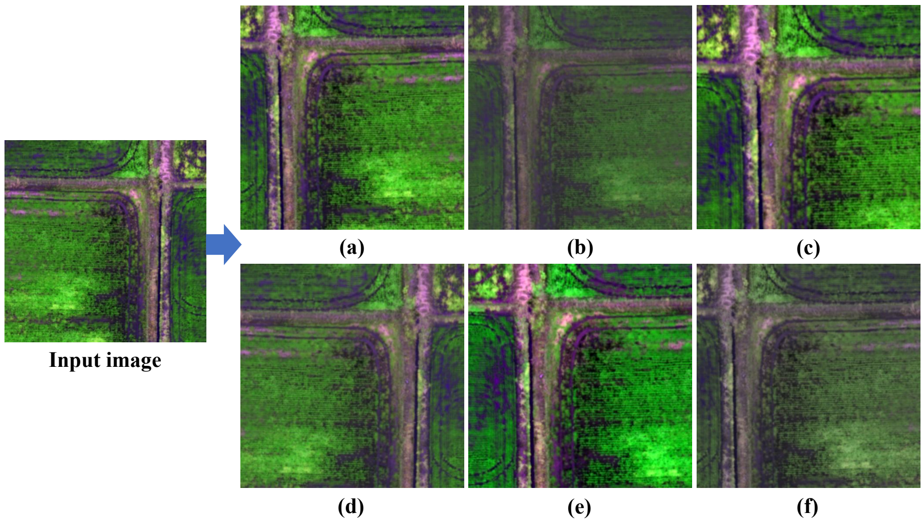

Figure 6: Examples of augmented images (a-f) used for training the ARG-TR model. Augmentations include a combination of scaling, cropping, flipping, and photometric adjustments.

sor variations. The augmentation techniques expanded the original training set of 226,853 images by three-fold to 680,550 images.

The large-scale and diverse nature of the dataset in capturing multiple types of rice growth abnormalities in different growth stages and environmental conditions presented both opportunities and challenges for analysis. The need to effectively process high-resolution multispectral imagery while accurately segmenting and classifying various abnormal growth patterns in real-time prompted the authors to choose a light-weight framework, which utilizes the rich spectral information in the dataset through self-attention mechanisms while maintaining computational efficiency without sacrificing scalability.

4. Methodology

Figure 3 illustrates the ARG-TR framework, a Transformer-based system for detecting and segmenting abnormal rice growth. In this context, ARG denote

Abnormal Rice Growth, while TR refers to a light-weight Transformer-based segmentation model.

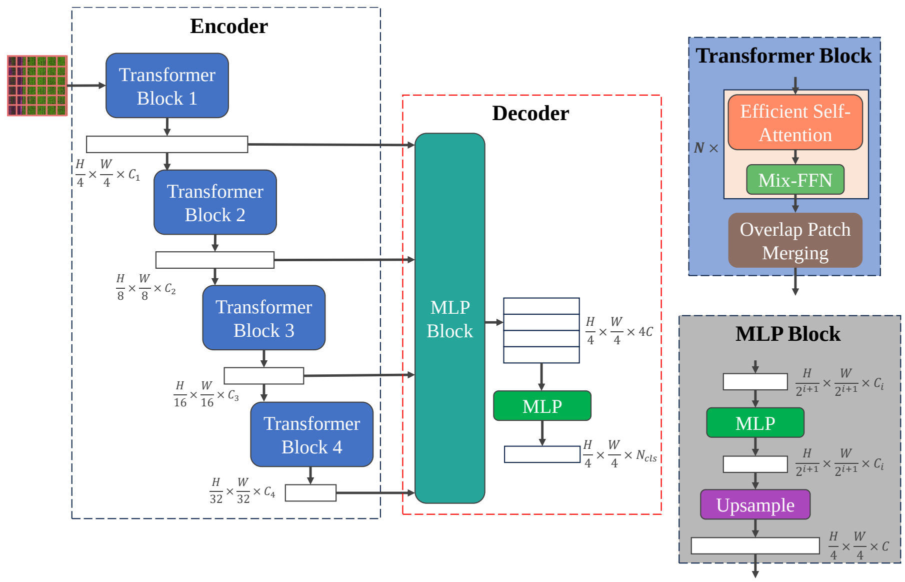

Figure 7: Schematic overview of the ARG-TR framework for abnormal rice growth segmentation. Figure adapted from (Xie et al., 2021)

SegFormer (Xie et al., 2021) is an efficient semantic segmentation architecture that combines a hierarchical transformer encoder and a lightweight multi-layer perceptron (MLP) decoder. Unlike CNN-based models, SegFormer eliminates positional embeddings through overlapped patch merging, which enables consistent performance on variable input sizes while preserving computational efficiency. This study utilizes SegFormer Mix Transformer(MiT)-b3 variant as the foundation. Its key innovations include:

- Multi-scale encoder: The encoder extracts both coarse and fine-grained features at four resolutions (1/4, 1/8, 1/16, and 1/32 input scale) via overlapping 4×4 patches. Unlike traditional approaches, it does not require positional embeddings to progressively capture fine details and contextual semantics.

- Efficient decoder: The decoder aggregates multi-scale features through MLP layers and upsamples them to produce a high-resolution segmentation map. The decoder ensures precise localization of abnormal growth patterns by fusing coarse (contextual) and fine-grained (detail-rich) features. Channel dimensionality is reduced from 1,024 to 128 via MLP blocks before generating 5-class logits (normal, blast, lodging, poor growth, missing plants.

Table 1 describe the detailed network structure of the ARG-TR model. It begins with a 7×7 convolutional patch embedding (stride=4) to downsample the input to H/4 × W/4 resolution with 64 channels, followed by LayerNorm for normalization and GELU for non-linearity. The encoder consists of four hierarchical stages of Transformer layers: Stage 1 with 3 layers (64 channels, H/4 × W/4), Stage 2 with 3 layers (128 channels, H/8 × W/8), Stage 3 with 18 layers (320 channels, H/16 × W/16), and Stage 4 with 3 layers (512 channels, H/32 × W/32). In the decoder, features from all encoder stages are upsampled to H/4 × W/4, concatenated, and processed through an MLPBlock that reduces the channel dimension from 1024 to 256 to 128, followed by a 1×1 convolution to generate 5 class logits. The head applies a softmax operation to convert logits into probabilities and resizes the output to the original image resolution. This design efficiently captures both fine and coarse details for multispectral rice growth anomaly detection.

4.1. Hierarchical Transformer Encoder

The Mix Transformer (MiT) backbone in SegFormer (Xie et al., 2021) serves as a hierarchical encoder customized for efficient semantic segmentation. It implements a four-stage pyramid structure to generate multi-scale feature maps at resolutions of 1/4, 1/8, 1/16, and 1/32 of the input image. This design enables robust segmentation of objects for varying scales, from fine-grained details to broader contextual patterns. Between transformer layers, a Mixed Feed-Forward Network (Mix-FFN) integrates depthwise 3×3 convolutions with standard MLP

Table 1: Network structure of the ARG-TR

Module	Layer / Operation	Channels	Output Size
Patch Embedding	Conv 7×7, stride 4	64	H/4×W/4
	LayerNorm + GELU	—	H/4×W/4
Encoder	3 layers (Stage 1)	64	H/4×W/4
	3 layers (Stage 2)	128	H/8×W/8
	18 layers (Stage 3)	320	H/16×W/16
	3 layers (Stage 4)	512	H/32×W/32
Decoder	Upsampling	—	H/4×W/4
	Concat → MLPBlock	1024→256→128	H/4×W/4
	1×1 Conv → 4 logits	128→4	H/4×W/4
Head & Loss	Softmax + resize	—	H/4×W/4×5 classes
	IoU	—	—

operations to enhance local spatial feature interactions. Finally, a patch merging module downsamples feature maps by concatenating neighboring patches and linearly projecting channel dimensions to establish a coarse-to-fine feature hierarchy. Moreover, overlapped patch embedding in early stages maintains local continuity without the need for positional encodings.

4.1.1. Hierarchical Feature Representation

SegFormer's encoder generates multi-scale feature maps at (1/4, 1/8, 1/16, 1/32 input spatial resolution), a crucial improvement from traditional ViTs, which produce single-scale representations. This hierarchical structure enables high-resolution feature maps to capture fine-grained details (early-stage rice blast lesions), while low-resolution feature maps encode coarse contextual information (lodging propagation). For rice growth analysis, hierarchical feature representation is essential as fine-grained features detect subtle spectral deviations in individual plants, whereas coarse features model spatial dependencies across field conditions.

4.1.2. Overlapped Patch Merging (OPM)

Overlapped Patch Merging (OPM) is an important component of SegFormer's Mix Transformer (MiT) encoder that enables hierarchical feature extraction while preserves local spatial continuity. Unlike standard ViTs with non-overlapping patches, OPM generates overlapping patches to maintain fine-grained spatial relationships essential for segmenting subtle abnormalities.

Given multi-spectral input $\mathbf{X} \in \mathbb{R}^{H \times W \times C}$, where $C = 3$ for Green/NIR/RE bands, H and W are the height and width. OPM slides a patch window across the image with a stride S smaller than the patch size K and with padding P, so adjacent patches overlap. The stride S being smaller than the patch size K is the key element that creates the overlap and shared context. Each window is flattened and linearly projected to form a token for the next hierarchical level. Repeating this merging yields hierarchical feature maps whose spatial resolution is reduced (for example from $H/4 \times W/4$ to $H/8 \times W/8$) while the channel dimension increases. The overlap size in each dimension is calculated as:

$$\text{Overlap} = K - S$$

The overlapping design reduces blocky artifacts and better preserves boundaries and fine details because pixels near patch edges contribute to multiple patch vectors

Table 2: OPM parameters for hierarchical feature maps

Stage	Patch Size (K)	Stride (S)	Padding (P)	Output Resolution
1	7	4	3	$\frac{H}{4} \times \frac{W}{4}$
2	3	2	1	$\frac{H}{8} \times \frac{W}{8}$
3	3	2	1	$\frac{H}{16} \times \frac{W}{16}$
4	3	2	1	$\frac{H}{32} \times \frac{W}{32}$

The overlapping design improves segmentation performance because it supplies the transformer with smoother, more informative multi-scale features. After patch merging, each stage's feature maps are passed through transformer blocks, which include efficient self-attention and Mix-FFN layers.

4.1.3. Efficient Self-Attention (ESA)

SegFormer employs ESA, a computationally optimized adaptation of standard self-attention used in ViTs. ESA is applied independently within each of the four stages of the MiT encoder. Given an input feature map $F_i \in \mathbb{R}^{H_i \times W_i \times C_i}$ from stage i, ESA flatten spatial locations into a token sequence $X \in \mathbb{R}^{N \times C_i}$, where $N = H_i \cdot W_i$ represents the number of spatial locations. In standard multi-head self-attention the per-head queries, keys and values are computed as

$$Q = XW_Q, \qquad K = XW_K, \qquad V = XW_V, \tag{1}$$

where $W_Q, W_K, W_V \in \mathbb{R}^{C_i \times d_{\text{head}}}$ and d_{head} denotes the per-head dimension. Standard attention computes $\text{softmax}(\frac{QK^\top}{\sqrt{d_{\text{head}}}})V$, which requires forming an $N \times N$ affinity matrix and therefore has quadratic complexity in the number of tokens.

ESA reduces this cost by shortening the key/value sequence by a reduction ratio R. A sequence-reduction operator (denoted $\text{SeqReduce}(\cdot)$) produces downsampled keys and values

$$K' = \text{SeqReduce}(K) \in \mathbb{R}^{\frac{N}{R} \times d_{\text{head}}}, \qquad V' = \text{SeqReduce}(V) \in \mathbb{R}^{\frac{N}{R} \times d_{\text{head}}}. \tag{2}$$

$\text{SeqReduce}(\cdot)$ can be implemented by reshaping and linear projection. ESA then computes attention from full-resolution queries to the reduced keys/values:

$$\text{EfficientAttention}(Q, K', V') = \text{softmax}\left(\frac{QK'^\top}{\sqrt{d_{\text{head}}}}\right)V', \tag{3}$$

where the softmax is taken along the reduced key dimension (length N/R) so that each of the N queries attends over the $\frac{N}{R}$ reduced positions. The computational cost becomes $\mathcal{O}(N \cdot \frac{N}{R} \cdot d_{\text{head}})$, which is significantly lower than the complexity of standard self-attention when $R > 1$ (Xie et al., 2021).

4.1.4. Mix Feed-Forward Network (Mix-FFN)

The Mix-FFN is a modification of the standard feed-forward network (FFN) that injects local spatial context into token-wise MLPs by inserting a 3×3

convolution between the two linear projections. This provides local positional information while preserving the global modelling capability of the FFN.

Let the input tokens be $x_{\text{in}} \in \mathbb{R}^{N\times C}$ with $N = H_i W_i$. The Mix-FFN proceeds as follows:

$$
\begin{aligned}
z &= W_1 x_{\text{in}} + b_1 \in \mathbb{R}^{N\times d_{\text{exp}}}, && (4)\\
Z &= \text{reshape}(z) \in \mathbb{R}^{H_i\times W_i\times d_{\text{exp}}}, && (5)\\
U &= \text{Conv}_{3\times 3}(Z;\ \text{padding} = 1) \in \mathbb{R}^{H_i\times W_i\times d_{\text{exp}}}, && (6)\\
V &= \text{GELU}(U), && (7)\\
v &= \text{flatten}(V) \in \mathbb{R}^{N\times d_{\text{exp}}}, && (8)\\
x_{\text{out}} &= W_2\, v + b_2 + x_{\text{in}} \in \mathbb{R}^{N\times C}. && (9)
\end{aligned}
$$

where $W_1 : \mathbb{R}^C \to \mathbb{R}^{d_{\text{exp}}}$ and $W_2 : \mathbb{R}^{d_{\text{exp}}} \to \mathbb{R}^C$ are the two linear projections (MLPs) of the FFN and $d_{\text{exp}} = r \cdot C$ is the expansion dimension (commonly $r = 4$) (Xie et al., 2021). The 3×3 convolution uses padding 1 to preserve spatial resolution. The residual connection $+x_{\text{in}}$ is applied as in standard transformer blocks. After processing, the output is flattened back to $N \times C$ for subsequent layers.

4.2. Lightweight All-MLP Decoder

SegFormer's decoder eliminates the complexity of traditional convolutional decoders by relying entirely on MLPs for efficient feature fusion and segmentation. The decoder unifies feature channel dimensions, upsamples features to a common spatial resolution, fuses them via a pointwise linear layer, and predicts per-pixel class logits with a final linear projection. For four-level encoder feature maps F_i the decoder proceeds as follows:

1. **Feature unification.** Each encoder feature map F_i (with C_i channels) is projected to a unified channel dimension C by a pointwise linear layer:

$$\hat{F}_i \;=\; \text{Linear}(C_i, C)\big(F_i\big) \qquad \text{for all } i. \tag{10}$$

2. **Upsampling.** Each unified feature map $\hat{F}_i$ is upsampled to the common spatial resolution $\frac{H}{4} \times \frac{W}{4}$:

$$\tilde{F}_i \;=\; \text{Upsample}\big(\tfrac{H}{4}, \tfrac{W}{4}\big)\big(\hat{F}_i\big) \qquad \text{for all } i, \tag{11}$$

where $\tilde{F}_i$ denotes the upsampled version of $\hat{F}_i$.

3. **Concatenation and fusion.** The upsampled features are concatenated along the channel dimension. For four encoder levels this yields $4C$ channels, which are fused back to C channels by a pointwise linear layer:

$$F \;=\; \text{Linear}(4C, C)\Big(\text{Concat}_i\big(\tilde{F}_i\big)\Big). \tag{12}$$

4. **Segmentation prediction.** A final linear layer maps the fused feature F to per-pixel class logits for N_{cls} classes:

$$M \;=\; \text{Linear}(C, N_{\text{cls}})(F), \tag{13}$$

so that M has shape $\frac{H}{4} \times \frac{W}{4} \times N_{\text{cls}}$. M is typically upsampled (e.g., bilinear) to the original image resolution $H \times W$ for evaluation and visualization.

4.3. Implementation description

Our framework uses the MiT-B3 backbone as the foundation. It was configured with four stages containing $[3, 4, 18, 3]$ Transformer layers, respectively. The number of attention heads for the stages is $[1, 2, 5, 8]$ and the corresponding embedding dimensions are $[64, 128, 320, 512]$. The lightweight all-MLP decoder employs a hidden dimension of 768 to fuse multi-scale features and produce segmentation outputs.

The model was trained for 3,000 iterations with a batch size of 4 using the AdamW optimizer with $\beta_1 = 0.9$, $\beta_2 = 0.999$ and $\varepsilon = 10^{-8}$. The initial learning rates were set to $\eta_0^{\text{backbone}} = 6 \times 10^{-5}$ for the backbone and $\eta_0^{\text{decoder}} = 6 \times 10^{-4}$ for the decoder. A polynomial learning-rate schedule was applied:

$$\eta_t \;=\; \eta_0 \left(1 - \frac{t}{T}\right)^{0.9},$$

where η_0 is the initial learning rate, t is the current iteration, and T is the total number of iterations.

The AGR-TR framework was implemented in PyTorch (v1.7.1) and trained on a Linux workstation equipped with two NVIDIA RTX A6000 GPUs (48 GB each). Model performance was evaluated on the original validation set to assess real-world applicability. For comparison, we implemented five baseline segmentation models using MMSegmentation (Contributors, 2020): DeepLabV3 (Chen et al., 2017), Segmenter (Strudel et al., 2021), K-Net (Zhang et al., 2021b) (K-Net), MaskFormer (Cheng et al., 2021), and U-Net (Ronneberger et al., 2015). All baselines were reimplemented within the same training and evaluation pipeline to ensure a fair comparison.

4.4. Evaluation metrics

The performance of the abnormal rice growth segmentation framework is evaluated using the Intersection over Union (IoU) metric (Wang et al., 2020), a standard measure for semantic segmentation that quantifies the pixel-wise overlap between predicted and ground-truth labels. For each class c we compute the pixel-level counts: true positives (TP_c), false positives (FP_c), and false negatives (FN_c). Here, TP_c is the number of pixels correctly predicted as class c, FP_c is the number of pixels incorrectly predicted as class c, and FN_c is the number of pixels belonging to class c but predicted as another class. The IoU for class c is defined as

$$\mathrm{IoU}_c = \frac{\mathrm{TP}_c}{\mathrm{TP}_c + \mathrm{FP}_c + \mathrm{FN}_c}. \tag{14}$$

The mean IoU (mIoU) over N abnormality classes is computed as

$$\mathrm{mIoU} = \frac{1}{N} \sum_{c=1}^{N} \mathrm{IoU}_c, \tag{15}$$

where N denotes the total number of classes in the dataset. The mIoU penalizes both over- and under-segmentation and therefore provides a robust measure of segmentation accuracy.

To quantify uncertainty in the estimated performance, we report a 95% confidence interval (CI) for the mean mIoU computed across n independent experimental runs. Let x_i denote the mIoU observed in the i-th run, and define the sample mean and sample standard deviation by

$$\bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_i, \tag{16}$$

$$s = \sqrt{\frac{1}{n-1}\sum_{i=1}^{n}(x_i - \bar{x})^2}. \tag{17}$$

Under the usual assumption that the sample mean is approximately t-distributed, a two-sided 95% CI for the true mIoU is given by

$$95\%\ \mathrm{CI} \ = \ \bar{x} \pm t_{\alpha/2,\,df} \cdot \frac{s}{\sqrt{n}}, \tag{18}$$

with $\alpha = 0.05$, degrees of freedom $df = n - 1$, and $t_{\alpha/2,\,df}$ the corresponding critical value from the Student's t-distribution.

5. Experimental results

5.1. Data augmentation

To assess the effect of data augmentation on segmentation performance, each experiment (training with and without augmentation) was repeated for $n = 5$ independent runs using different fixed seeds. Table 3 summarizes ARG-TR's segmentation performance on the original and augmented datasets. In the table "±" denotes the sample standard deviation across the n runs, and the 95% confidence intervals (CIs) for the mean mIoU were computed using the Student's t-distribution with degrees of freedom $df = n - 1 = 4$.

The mean mIoU increased from 60.39% (original) to 62.88% (augmented), with corresponding 95% CIs [57.53%, 63.25%] and [60.15%, 65.61%], respectively. In addition, the consistent improvements on all evaluation metrics emphasize the critical role of data augmentation in mitigating class imbalance challenges, refining feature learning, and enhancing generalization to diverse

Table 3: ARG-TR segmentation performance on original data and augmented data. Note: ± indicates standard deviation.

	mIoU	mIoU 95% CI	Precision	Recall
Original data	60.39±2.3	[57.54, 63.24]	62.18±2.1	59.43±2.4
Augmented data	62.88±2.2	[60.15, 65.61]	65.03±2.7	63.82±2.9

field conditions. For example, the improvement in precision and recall suggests that the augmentation pipeline reduces both false positives and false negatives, where ambiguous or rare symptoms often challenge model robustness.

5.2. Spectral band contribution analysis

Figure 8 shows sample images for three different spectral-band configurations: green (G), green + near-infrared (G+NIR), and green + near-infrared + red-edge (G+NIR+RE).

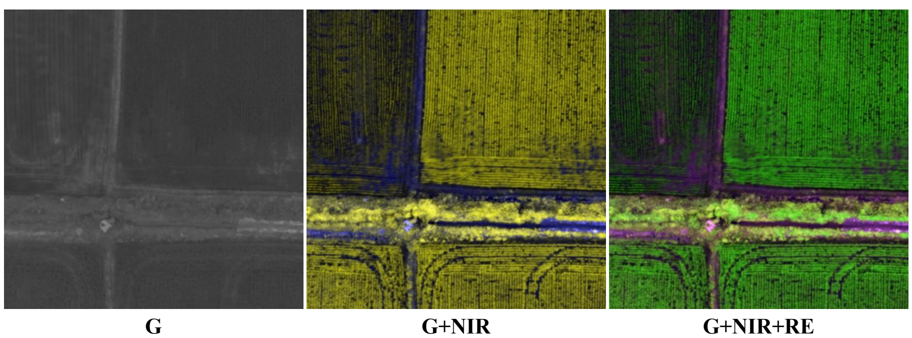

Figure 8: Illustration of the three spectral-band settings used as input to the model. "G" denotes the green band; "NIR" denotes near-infrared; "RE" represents the red-edge band.

To quantify the contribution of each spectral band to anomaly detection, we performed an ablation study using three input configurations: G only, G+NIR, and G+NIR+RE. Table 4 reports the class-wise IoU (%) for each configuration.

With only the green channel the model obtains moderate performance (IoU between 47.1% and 57.3%). The combination of G and NIR bands yield substantial gains for all anomaly types. For example, IoU for L increases from

Table 4: Ablation of spectral bands on class-wise IoU (%).

Class	**G**	**G+NIR**	**G+NIR+RE**
Missing plants (MP)	49.80	60.52	65.82
Lodging (L)	57.34	66.18	68.41
Rice blast (RBD)	47.11	57.73	61.78
Poor growth (PG)	50.52	58.21	63.34

57.3% to 66.1%, and IoU for MP increases from 49.8% to 60.5%. The integration of RE band further improves performance and produces the highest IoU for every class. For example, L to 68.4% and MP to 65.8%. These results indicate that NIR provides complementary contrast useful for detecting structural and vegetation anomalies, while the RE band refines discrimination of disease- and stress-related symptoms as it is sensitive to chlorophyll content and subtle stress signals. Overall, the combination G+NIR+RE offers the most informative spectral input for abnormal rice growth segmentation in our experiments.

5.3. ARG-TR performance evaluation

We performed an ablation study to examine how encoder depth and decoder hidden size affect segmentation accuracy and model complexity. Table 5 reports three ARG-TR variants with different encoder depths, decoder hidden dimensions, total parameter counts (in millions), and the resulting mIoU.

Table 5: Ablation of encoder depths and decoder hidden size. “Hidden sizes” lists the stage-wise embedding dimensions for the encoder.

Model variant	Depths	Hidden sizes	Decoder hidden size	Params (M)	mIoU
ARG-TR (1)	[2, 2, 2, 2]	[64, 128, 320, 512]	256	14.0	61.65
ARG-TR (2)	[3, 4, 6, 3]	""	768	25.4	62.72
ARG-TR (3)	[3, 4, 18, 3]	""	768	45.2	64.86

Key observations include:

- Moving from ARG-TR (1) to ARG-TR (2) increases the parameter count by 11.4M (from 14.0M to 25.4M, about 81.4% increase) and yields a smaller mIoU gain by 1.07% (from 61.65% to 62.72%).
- Moving from ARG-TR (2) to ARG-TR (3) further increases parameters by 19.8M (from 25.4M to 45.2M, approximately 78.0% increase) and obtains a larger mIoU value of 1.96% (from 62.72% to 64.86%).

These results show that increasing model capacity consistently improves segmentation performance, and in this set of variants the largest model (ARG-TR (3)) provides the highest mIoU. Considering the balance between accuracy and computational cost, ARG-TR (3) was selected as the primary model for subsequent experiments because it achieves the highest segmentation performance. For deployment scenarios with limited memory or latency budgets, ARG-TR (1) or ARG-TR (2) are preferable due to their lower parameter counts and competitive performance.

Figure 9 presents the training progress of the ARG-TR model using two key metrics recorded over 3,000 iterations: pixel accuracy (left) and training loss (right). The validation accuracy (seg_accuracy) shows a rapid rise between approximately 600 and 1,000 iterations, reaching about 88%–90%. The loss starts near 0.9 and decreases sharply to around 0.4 by iteration 1,500, then gradually stabilizes close to 0.4 by iteration 3,000. The quick initial convergence indicates that ARG-TR efficiently learns discriminative features even with limited labeled data. After the early-stopping mark at iteration 2,000, both accuracy and loss remain stable. Therefore, 2,000 iterations are considered sufficient to achieve near-optimal generalization in our setup.

Table 6 summarizes ARG-TR's segmentation performance for four abnormal rice growth classes. Reported metrics are IoU, precision, and recall (all in percentage). The testing process was repeated for $n = 5$ independent runs for each class. Overall, ARG-TR achieves IoU over 60.8% for all classes. "Lodging" obtains the best performance (mean IoU = 67.3%, precision = 68.4%, recall = 69.13%), likely because its structural signature (bent or collapsed plants) is

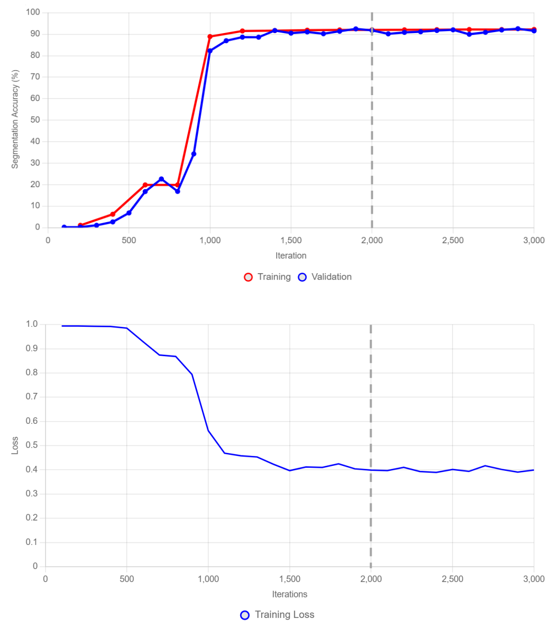

Figure 9: Training progress of the ARG-TR model on the abnormal rice growth dataset: pixel accuracy (left) and loss (right). The vertical dashed line indicates the early-stopping mark at iteration 2,000.

visually distinct. “Missing plants” follows with mean IoU = 64.1%.

“Poor growth” and “Rice blast” show lower IoU values at 61.3% and 60.8%, respectively. The relatively lower precision and recall for “Poor growth” and “Rice blast” can be attributed to the higher visual similarity of these symptoms to healthy rice plants under certain conditions, which increases the likelihood of both false positives and false negatives. For “Poor growth”, the phenotypic differences, such as slight stunting, reduced leaf area, or lighter color, can be subtle and easily confused with natural field variability or early-stage nutrient deficiencies. For “Rice blast”, the appearance of lesions may be small for early-stage disease, sparsely distributed, or partially occluded by surrounding leaves. As a consequence, it is difficult to detect them at UAV imaging resolutions.

Table 6: ARG-TR performance for each abnormal rice growth class. Note: ± indicates standard deviation.

	Missing plants	Poor growth	Lodging	Rice blast
IoU	64.11±2.2	61.29±2.5	67.37±1.8	60.87±2.8
Precision	66.80±2.5	64.27±2.1	68.42±2.5	63.47±2.3
Recall	67.34±1.8	63.98±2.4	69.13±2.4	62.18±2.0

5.4. Visualization of abnormal rice growth segmentation using the ARG-TR framework

Figure 10 and Figure 11 illustrate segmentation outputs produced by the ARG-TR framework for four abnormal rice growth classes: Missing Plants (MP), Poor Growth (PG), Rice Blast Disease (RBD), and Lodging (L). The color legend used throughout the figures is: MP (magenta), PG (red), RBD (cyan), L (orange), healthy rice (blue), and bare ground (black).

Overall, the visual alignment between model predictions and ground truth demonstrates robust segmentation performance for several categories. For MP, the model consistently detects large gaps in the field and closely matches ground truth boundaries. For L, the model successfully captures the irregular textures and patterns associated with lodged plants. For PG, the model locates small and sparse affected areas with relatively few false positives. Finally, the model correctly detects the infected RBD regions, which matches the ground truth. Figure 11 shows examples of mixed and ambiguous cases that highlight both strengths and weaknesses of the model.

In general, while the model effectively identifies large contiguous regions of L and RBD, it struggles with finer distinctions in overlapping or ambiguous cases. For example, MP areas are occasionally undersegmented or as L, particularly in regions where L regions are near the MP regions (Figure 11 c, d). Moreover, early stage RBD symptoms tend to be fragmented in predictions, which reflects the difficulty of separating infected regions from healthy areas. These errors highlight the complexity of identifying co-occurring stressors in real-world agri-

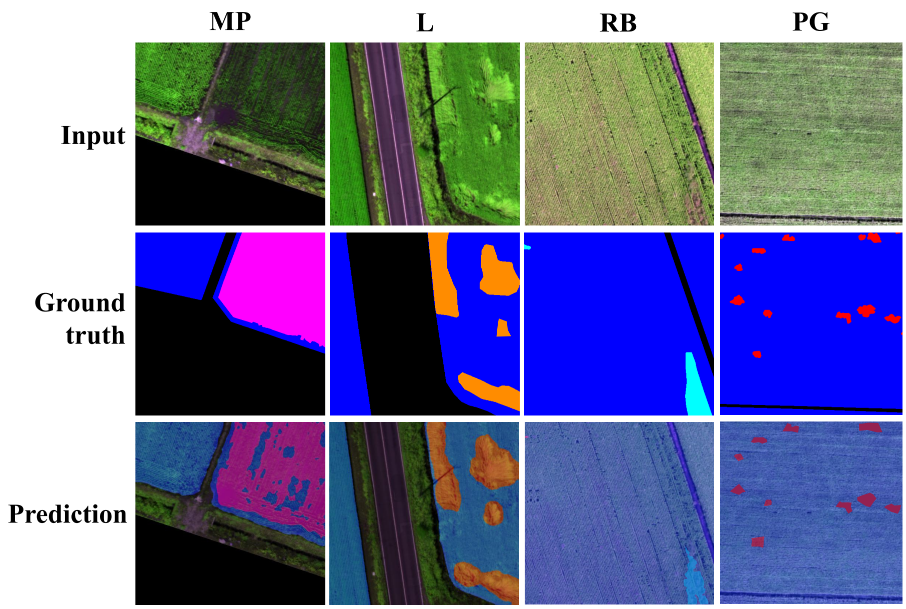

Figure 10: Predictions of the ARG-TR framework for all abnormal rice growth classes. **Note:** MP: Missing Plants (Magenta), PG: Poor Growth (Red), RBD: Rice Blast Disease (Cyan), L: Lodging (Orange), Blue indicates healthy rice, and Black represents bare ground.

cultural scenes, where symptom boundaries are often blurred by environmental variability and plant interactions.

5.5. Comparative analysis of ARG-TR and other baseline segmentation models

This section compares the ARG-TR framework with several established segmentation models: KNet (Zhang et al., 2021b), Segmenter (Strudel et al., 2021), SegFormer (Xie et al., 2021), DeepLabv3 (Chen et al., 2017), U-Net (Ronneberger et al., 2015), MaskFormer (Cheng et al., 2021), and EDANet (Yang et al., 2020). Table 7 summarizes each model's performance on the abnormal rice growth validation set using mIoU, pixel accuracy, and inference speed (frames per second (FPS).

ARG-TR achieves the highest mIoU (64.86%) and pixel accuracy (93.42%)

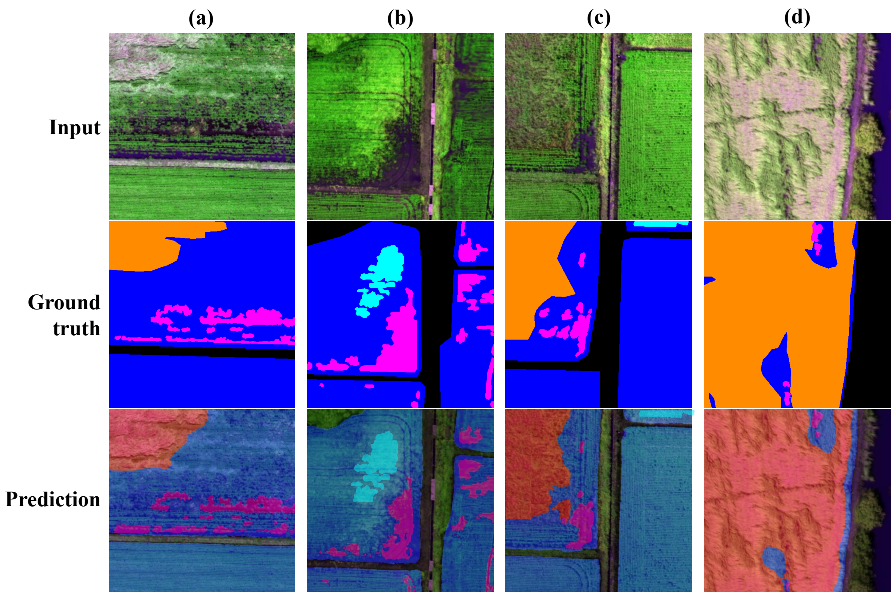

Figure 11: ARG-TR predictions for challenging mixed-condition cases. **Note:** MP: Missing Plants (Magenta), RBD: Rice Blast Disease (Cyan), L: Lodging (Orange), Blue indicates healthy rice, and Black represents bare ground.

on the dataset, outperforming strong baselines such as KNet (mIoU: 57.34%) and MaskFormer (mIoU: 60.13%). This improvement suggests that ARG-TR offers superior contextual understanding and finer feature discrimination, likely due to its transformer-based global-context modeling and the integration of targeted anomaly-aware modules.

Regarding efficiency, ARG-TR reaches a better trade-off between accuracy and speed. While it is slower than EDANet (61 FPS) and DeepLabv3 (29 FPS), its accuracy gains make it more suitable for precision agricultural monitoring where segmentation quality is prioritized. Lighter models, like UNet and some transformer variants, such as Segmenter and MaskFormer show lower segmentation performance on this task, which highlights limitations in capturing complex spatial hierarchies.

Figure 12 presents qualitative comparisons between ARG-TR, MaskFormer, and KNet on three representative UAV samples. ARG-TR consistently produces

Table 7: Performance comparison of the ARG-TR framework and baseline models on the abnormal rice growth dataset. Note: Inference speed measured on the same evaluation environment (batch size = 1).

Model	mIoU	Pixel accuracy	Inference speed (FPS)
KNet (Zhang et al., 2021b)	57.34	89.12	12
Segmenter (Strudel et al., 2021)	56.47	88.75	10
DeepLabv3 (Chen et al., 2017)	54.89	86.43	29
UNet (Ronneberger et al., 2015)	49.21	83.56	15
MaskFormer (Cheng et al., 2021)	60.13	90.58	9
EDANet (Yang et al., 2020)	56.52	89.23	61
ARG-TR (Segformer) (Xie et al., 2021)	64.86	93.42	25

masks with sharp boundaries and reduced noise. In the first two samples (distinct L and MP regions), ARG-TR's predictions align closely with ground truth annotations. In the third sample, ARG-TR shows minor over-segmentation but remains more similar to the ground truth than MaskFormer and KNet. MaskFormer tends to produce more fragmented MP regions, while KNet produces noisier and more scattered masks, especially in samples with mixed abnormalities. These qualitative differences emphasize ARG-TR's strengths in fine-grained anomaly localization and boundary adherence, both important for real-world agricultural monitoring where small or ambiguous symptoms must be detected reliably.

Through previous experiments, ARG-TR consistently outperformed state-of-the-art baselines in both mean IoU and pixel accuracy. Using G+NIR+RE input bands produced a 13.2% increase in IoU compared with only using the green channel. In addition, ARG-TR achieved real-time inference and produced more precise segmentation boundaries, particularly in mixed or ambiguous field conditions.

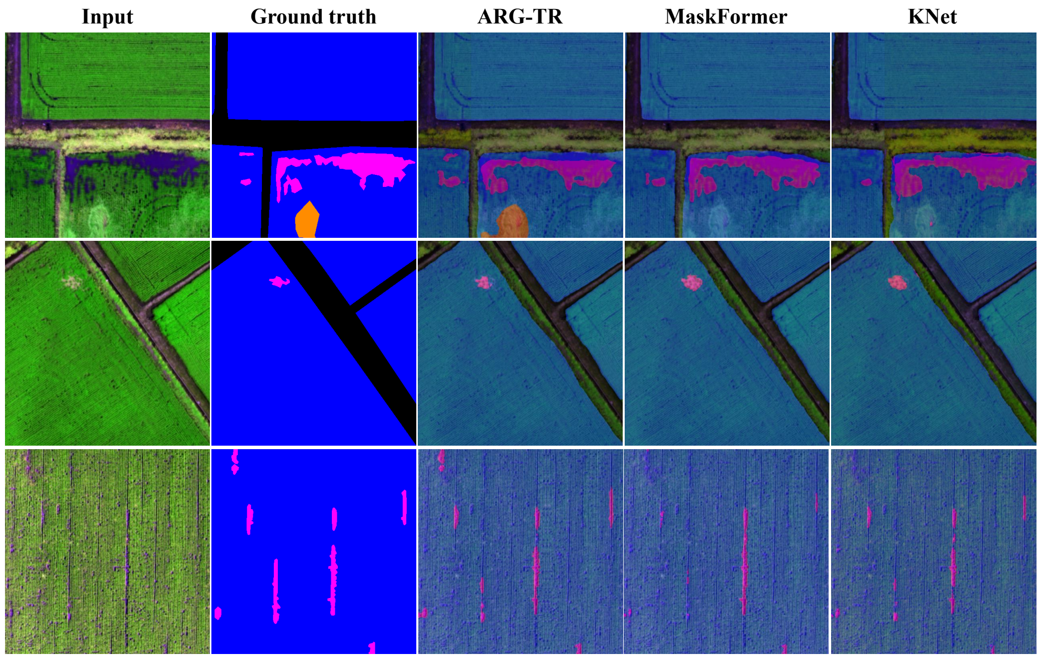

Figure 12: Comparison of the output of ARG-TR framework and two other state-of-the-art segmentation models, MaskFormer and KNet on three different input samples. **Note:** MP: Missing Plants (Magenta), L: Lodging (Orange), Blue indicates healthy rice, and Black represents bare ground.

6. Discussion

The primary goal of this study was to identify an efficient and robust DL framework for abnormal rice growth detection. We evaluated multiple segmentation architectures on a large, manually annotated UAV dataset. Through a series of experiments, transformer-based architectures, such as SegFormer and MaskFormer, achieved higher segmentation performance than other CNN-based alternatives (e.g., DeepLabv3, U-Net). These results are consistent with recent work that highlights transformers' ability to model global context and long-range dependencies for agricultural disease and stress detection (Wang et al., 2024; Kapetas et al., 2024). According the results reported in Table 7, the ARG-TR framework showed the best overall performance (mIoU = 64.86%, pixel accuracy = 93.42%). The hierarchical feature fusion and transformer-based global-context modeling improve discrimination of subtle anomalies such as lodging and rice blast.

Several aspects of UAV data acquisition greatly affect the performance of the framework. First, variation in solar illumination and viewing geometry introduces spectral shifts that reduce class separability. We addressed this by applying geometric correction, spectral band combination, and augmentations during training, but residual effects can still increase false positives/negatives in borderline cases. Second, weather constraints and the limited number of imaging dates (five discrete growth stages) create temporal gaps that can miss rapid symptom progression. For example, mIoU scores for Rice Blast Disease (60.8%) and Poor Growth (61.3%) indicate lower per-class performance compared with large, contiguous anomalies such as missing plants. This is expected because stunting and early infections produce weak, spatially dispersed spectral signatures that are difficult to distinguish from normal variability. Third, flight altitude and ground-sampling distance limit the detectability of very small or early-stage lesions; multispectral indices (e.g., NDVI, red-edge) partly compensate by highlighting physiological stress that is not obvious in RGB, but small-scale symptoms remain challenging. Finally, ARG-TR's inference speed (25 FPS measured in our evaluation setting) is suitable for many UAV-based monitoring workflows where segmentation quality is prioritized. However, in applications that require very high throughput, such as continuous video streams or large-area rapid surveys, lighter-weight models or optimized inference engines (EDANet, DeepLabv3) are preferable.

Although this study focused on G, NIR, and RE bands, the network architecture can readily accommodate different spectral combinations or higher resolution sensors with minimal modification. While we demonstrated the model mainly on rice, the framework can be retrained for other species, such as wheat or maize, because the model learns spatial spectral representations directly from data, it can adapt to diverse geographic regions and environmental conditions given representative training samples. The model is suitable for real-time UAV deployments or integration into monitoring platforms for various agricultural settings.

7. Conclusions and future works

In this work, we introduced ARG-TR, a transformer-based segmentation framework specifically, configured for identifying abnormal rice growth patterns using drone-captured imagery. The model was trained on a large-scale drone-based dataset containing 378,074 high-resolution images covering four common abnormal rice growth anomalies (lodging, rice blast disease, poor growth, and missing plants). By integrating hierarchical transformer architecture with a strategic augmentation pipeline, ARG-TR achieves rapid convergence during training and robust generalization to diverse field conditions. With a mIoU of 64.86 and 93.42% pixel accuracy, ARG-TR excels in identifying distinct anomalies like lodging and rice blast disease, while maintaining efficient inference speed (25 FPS).

Challenges exists in detecting subtle or overlapping stressors like early-stage stunting and ambiguous symptom boundaries. Future work will explore hybrid architectures that combine local texture encoders with global transformers, as well as domain-specific synthetic augmentations to enrich rare-class representations. Moreover, the integration of additional modalities, such as spectral or temporal data, may further sharpen boundary delineation and symptom discrimination. Finally, with continued enhancements in model design and training strategies, ARG-TR has the potential to power real-time and scalable agriculture systems capable of delivering timely and actionable insights for crop health management.

Acknowledgement

This work was supported by the Young Scientists Fund of the National Natural Science Foundation of China (No. 62502271), the Young Scientists Fund of the Natural Science Foundation of Shandong Province (No. ZR2025QC630), the

Natural Science Foundation of Rizhao City (Nos. RZ2024ZR33 and RZ2024ZR34) and by Institute of Information & communications Technology Planning & Evaluation (IITP) under the metaverse support program to nurture the best talents (IITP-2024-RS-2023-00254529) grant funded by the Korea government(MSIT).

Author contributions

Yanfen Li: Writing – original draft. **L. Minh Dang**: Methodology, Writing – original draft, Writing – review & editing, Data curation. **Hanxiang Wang**: Conceptualization, Writing – review & editing. **Muhammad Fayaz**: Data curation. **Sufyan Danish**: Visualization.**Junliang Shang**: Formal analysis, Supervision. **Hyoung-Kyu Song**: Funding acquisition. **Hyeonjoon Moon**: Supervision.

Declaration of Competing Interest

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

References

Agüera-Vega, F., Carvajal-Ramírez, F., & Martínez-Carricondo, P. (2017). Assessment of photogrammetric mapping accuracy based on variation ground control points number using unmanned aerial vehicle. *Measurement*, *98*, 221–227.

Alam, N., Sagar, A. S., Dang, L. M., Zhang, W., Park, H. Y., & Hyeonjoon, M. (2025). Deep learning based radish and leaf segmentation for phenotype trait measurement. *Signal, Image and Video Processing*, *19*, 178.

Bin Rahman, A. R., & Zhang, J. (2023). Trends in rice research: 2030 and beyond. *Food and Energy Security*, *12*, e390.

Biswal, S., Pathak, N., Chatterjee, C., & Mailapalli, D. R. (2024). Estimation of aboveground biomass from spectral and textural characteristics of paddy crop using uav-multispectral images and machine learning techniques. *Geocarto International*, *39*, 2364725.

Chen, L.-C., Papandreou, G., Kokkinos, I., Murphy, K., & Yuille, A. L. (2017). Deeplab: Semantic image segmentation with deep convolutional nets, atrous convolution, and fully connected crfs. *IEEE transactions on pattern analysis and machine intelligence*, *40*, 834–848.

Cheng, B., Schwing, A., & Kirillov, A. (2021). Per-pixel classification is not all you need for semantic segmentation. *Advances in neural information processing systems*, *34*, 17864–17875.

Contributors, M. (2020). Mmsegmentation: Openmmlab semantic segmentation toolbox and benchmark.

Dang, L. M., Hassan, S. I., Suhyeon, I., kumar Sangaiah, A., Mehmood, I., Rho, S., Seo, S., & Moon, H. (2020). Uav based wilt detection system via convolutional neural networks. *Sustainable Computing: Informatics and Systems*, *28*, 100250.

Dang, M., Wang, H., Li, Y., Nguyen, T.-H., Tightiz, L., Xuan-Mung, N., & Nguyen, T. N. (2024). Computer vision for plant disease recognition: a comprehensive review. *The Botanical Review*, *90*, 251–311.

Dosovitskiy, A., Beyer, L., Kolesnikov, A., Weissenborn, D., Zhai, X., Unterthiner, T., Dehghani, M., Minderer, M., Heigold, G., Gelly, S. et al. (2020). An image is worth 16x16 words: Transformers for image recognition at scale. *arXiv preprint arXiv:2010.11929*, .

Dosset, A., Dang, L. M., Alharbi, F., Habib, S., Alam, N., Park, H. Y., & Moon, H. (2025). Cassava disease detection using a lightweight modified soft attention network. *Pest Management Science*, *81*, 607–617.

Food, & of the United Nations, A. O. (2025). Food and agriculture projections to 2050. https://www.fao.org/global-perspectives-studies/food-agriculture-projections-to-2050/en/, accessed 2025-04-03.

Ju, O.-J., Choi, B.-R., Jang, E. K., Soh, H., Lee, S.-W., & Lee, Y.-S. (2022). Climate change and rice yield in hwaseong-si gyeonggi-do over the past 20 years (2001˜ 2020). *Korean Journal of Environmental Agriculture*, *41*, 16–23.

Kang, Y., Meng, Q., Liu, M., Zou, Y., & Wang, X. (2021). Crop classification based on red edge features analysis of gf-6 wfv data. *Sensors*, *21*, 4328.

Kapetas, D., Kalogeropoulou, E., Christakakis, P., Klaridopoulos, C., & Pechlivani, E. M. (2024). Multi-spectral image transformer descriptor classification combined with molecular tools for early detection of tomato grey mould. *Smart Agricultural Technology*, *9*, 100580.

Li, Y., Wang, H., Dang, L. M., Sadeghi-Niaraki, A., & Moon, H. (2020). Crop pest recognition in natural scenes using convolutional neural networks. *Computers and Electronics in Agriculture*, *169*, 105174.

Lin, T., Wang, Y., Liu, X., & Qiu, X. (2022). A survey of transformers. *AI open*, *3*, 111–132.

Liu, Z., Lin, Y., Cao, Y., Hu, H., Wei, Y., Zhang, Z., Lin, S., & Guo, B. (2021). Swin transformer: Hierarchical vision transformer using shifted windows. In *Proceedings of the IEEE/CVF international conference on computer vision* (pp. 10012–10022).

Measurusa (2025). TRINITY F90+. https://measurusa.com/pages/trinity-f90, accessed 2025-04-05.

MicaSense (2025). MicaSense RedEdge-MX. https://support.micasense.com/hc/en-us/articles/360011389334-RedEdge-MX-Integration-Guide, accessed 2025-04-05.

Nuradili, P., Zhou, J., & Melgani, F. (2024). Wetland segmentation method for uav multispectral remote sensing images based on segformer. In *IGARSS 2024-2024 IEEE International Geoscience and Remote Sensing Symposium* (pp. 6576–6579). IEEE.

Rezvi, H. U. A., Tahjib-Ul-Arif, M., Azim, M. A., Tumpa, T. A., Tipu, M. M. H., Najnine, F., Dawood, M. F., Skalicky, M., & Brestič, M. (2023). Rice and food security: Climate change implications and the future prospects for nutritional security. *Food and Energy Security*, *12*, e430.

Ronneberger, O., Fischer, P., & Brox, T. (2015). U-net: Convolutional networks for biomedical image segmentation. In *International Conference on Medical image computing and computer-assisted intervention* (pp. 234–241). Springer.

Samal, P., Babu, S. C., Mondal, B., & Mishra, S. N. (2022). The global rice agriculture towards 2050: An inter-continental perspective. *Outlook on Agriculture*, *51*, 164–172.

Spasev, V., Dimitrovski, I., Chorbev, I., & Kitanovski, I. (2024). Semantic segmentation of unmanned aerial vehicle remote sensing images using segformer. In *International Conference on Intelligent Systems and Pattern Recognition* (pp. 108–122). Springer.

Stankus, A. (2021). State of world aquaculture 2020 and regional reviews: Fao webinar series. *FAO aquaculture newsletter*, (pp. 17–18).

Strudel, R., Garcia, R., Laptev, I., & Schmid, C. (2021). Segmenter: Transformer for semantic segmentation. In *Proceedings of the IEEE/CVF international conference on computer vision* (pp. 7262–7272).

Tian, M., Ban, S., Yuan, T., Ji, Y., Ma, C., & Li, L. (2021). Assessing rice lodging using uav visible and multispectral image. *International journal of remote sensing*, *42*, 8840–8857.

Wang, H., Nguyen, T.-H., Nguyen, T. N., & Dang, M. (2024). Pd-tr: End-to-end plant diseases detection using a transformer. *Computers and Electronics in Agriculture*, *224*, 109123.

Wang, Z., Wang, E., & Zhu, Y. (2020). Image segmentation evaluation: a survey of methods. *Artificial Intelligence Review*, *53*, 5637–5674.

Wu, S., Ma, X., Jin, Y., Yang, J., Zhang, W., Zhang, H., Wang, H., Chen, Y., Lin, C., & Qi, L. (2025). A novel method for detecting missing seedlings based on uav images and rice transplanter operation information. *Computers and Electronics in Agriculture*, *229*, 109789.

Xie, E., Wang, W., Yu, Z., Anandkumar, A., Alvarez, J. M., & Luo, P. (2021). Segformer: Simple and efficient design for semantic segmentation with transformers. *Advances in neural information processing systems*, *34*, 12077–12090.

Yang, M.-D., Boubin, J. G., Tsai, H. P., Tseng, H.-H., Hsu, Y.-C., & Stewart, C. C. (2020). Adaptive autonomous uav scouting for rice lodging assessment using edge computing with deep learning edanet. *Computers and Electronics in Agriculture*, *179*, 105817.

Zhang, T., Xu, Z., Su, J., Yang, Z., Liu, C., Chen, W.-H., & Li, J. (2021a). Ir-unet: Irregular segmentation u-shape network for wheat yellow rust detection by uav multispectral imagery. *Remote Sensing*, *13*, 3892.

Zhang, W., Pang, J., Chen, K., & Loy, C. C. (2021b). K-net: Towards unified image segmentation. *Advances in Neural Information Processing Systems*, *34*, 10326–10338.