CTU Journal of Innovation and Sustainable Development
ISSN 2588-1418 | e-ISSN 2815-6412

DOI:10.22144/ctujoisd.2026.065

Benchmarking encoder-decoder architectures and input resolutions for rice lodging semantic segmentation using UAV RGB imagery

Chau Thi Ngoc Loan[1,3], Nguyen Van Hoa[1,3*], and Tran Thanh Duong[2]

[1]*Faculty of Information Technology, An Giang University, Viet Nam*
[2]*College of Biotechnology and Bioengineering, Sungkyunkwan University, Republic of Korea*
[3]*Vietnam National University Ho Chi Minh City, Ho Chi Minh City, Viet Nam*
**Corresponding author (nvhoa@agu.edu.vn)*

Article info.

Received 15 Jun 2026
Revised 8 Sep 2026
Accepted 15 Oct 2026

Keywords

Deep learning, precision agriculture, rice lodging, Semantic segmentation, UAV RGB imagery

ABSTRACT

Rice lodging reduces yield and hinders mechanised harvesting; consequently, its timely delineation is important for field monitoring and damage assessment. This study benchmarks semantic segmentation models using UAV RGB imagery acquired during three field surveys in An Giang Province, Vietnam. The dataset contains 715 annotated images with four classes: Background, Lodged Rice, Ponds/Ditches, and Field Edges/Paths. Fifteen encoder–decoder configurations were first benchmarked at an input resolution of 512×512 pixels. Based on validation-set mIoU, one configuration from each decoder was then selected for further evaluation at 640 × 640 and 768 × 768 pixels: U-Net with a MiT-B1 encoder, DeepLabV3+ with an EfficientNet-B1 encoder, U-Net++ with a MobileNetV2 encoder, and SegFormer with a MiT-B1 encoder. To address class imbalance, training used a hybrid loss that combined Dice loss and weighted cross-entropy loss. At 512 × 512 pixels, SegFormer with the MiT-B1 encoder achieved the highest mIoU (66.72%) and mF1 scores (79.01%). At 640 × 640 pixels, DeepLabV3+ with the EfficientNet-B1 encoder achieved the highest mIoU (67.48%) and mF_1-score (79.60%), whereas SegFormer with the MiT-B1 encoder obtained the highest mPA (91.23%). Increasing the input resolution to 768 × 768 pixels did not yield consistent improvements. Overall, 640 × 640 pixels provided the most favourable performance in terms of mIoU and mF_1-score among the evaluated settings.

1. INTRODUCTION

Rice is one of the world's most important food crops, providing a staple food source for more than half of the global population. Maintaining stable rice production is therefore critical to food security and socioeconomic stability. However, rice crops are frequently affected by lodging throughout the growing cycle, particularly during the middle and late growing stages. Rice lodging occurs when stems become permanently displaced from their vertical orientation, commonly owing to extreme weather events such as storms, heavy rainfall, and strong winds, or to crop-management factors such as excessive planting density and nitrogen fertilisation (Zhang et al., 2024b; Guo & Tan, 2025a; Liu et al., 2026). Lodging disrupts photosynthesis, reduces grain yield and quality (Zhang et al., 2024b), and complicates mechanised harvesting because combine harvesters cannot readily access stems lying close to the ground,

thereby increasing harvest losses (Liu et al., 2025). Consequently, rapid and accurate identification of lodged rice locations and extent is essential for disaster-damage assessment, agricultural-insurance valuation, and real-time feedback to adjust cutter-head operating parameters in combine harvesters.

To support timely and spatially extensive lodging assessment, remote sensing has become a promising tool for crop-condition monitoring over large areas (Zhang et al., 2024b; Celik et al., 2026). Conventional field surveys are time-consuming, labour-intensive, subjective, and may cause secondary crop damage during direct measurement (Zhang, et al., 2024b). Although satellite remote-sensing data provide broad spatial coverage, their use for timely and fine-scale rice lodging monitoring is constrained by limited spatial resolution, revisit intervals, and cloud cover. UAVs address many of these limitations through their operational flexibility, relatively low cost, and ability to acquire high-resolution imagery at flexible spatial and temporal scales (Wilke et al., 2019; Zhao et al., 2019; Azizi et al., 2024). Early UAV-based studies commonly relied on hand-crafted spectral, textural, or structural features together with conventional machine-learning methods to identify or quantify lodged areas (Wilke et al., 2019; Wang et al., 2021; Bai et al., 2022). However, these approaches depend heavily on feature engineering and may be sensitive to variations in crop structure, illumination, soil backgrounds, and weed interference.

Rice lodging semantic segmentation has emerged as a key research direction in agricultural computer vision to address these limitations. Its primary objective is to assign a semantic label to each pixel in UAV imagery, thereby enabling detailed delineation of lodged-rice boundaries and area estimation. Compared with image-level classification or object-level detection, semantic segmentation provides pixel-level information that is particularly relevant for quantifying the spatial extent of lodging under heterogeneous field conditions. Recent studies have explored deep-learning-based segmentation approaches for rice lodging, including encoder–decoder architectures and their variants, demonstrating the feasibility of pixel-level lodging delineation from UAV imagery (Zhao et al., 2019; Zhang et al., 2024a; Guo & Tan, 2025b; Liu et al., 2026).

Despite these advances, existing studies on rice lodging segmentation have been conducted using different datasets, imaging conditions, model architectures, and evaluation settings, making it difficult to determine which encoder–decoder configurations are most suitable under a consistent experimental protocol. In particular, the comparative effects of encoder–decoder architecture and input image resolution on UAV-based rice lodging segmentation have not been systematically compared under a common experimental protocol. To address this gap, this study benchmarks U-Net, U-Net++ (Zhou et al., 2018), DeepLabV3+ (Li & Zhao, 2022), and SegFormer (Xie et al., 2021) with their available encoder combinations under a consistent experimental setting and systematically evaluates three input resolutions to examine their effects on segmentation performance under complex field conditions with ambiguous class boundaries. The objective is to identify suitable model-resolution configurations for UAV RGB-based rice lodging semantic segmentation and provide a systematic benchmark for model selection in field-scale applications.

2. RELATED WORKS

Crop-lodging assessment has evolved from remote-sensing methods based on hand-crafted spectral, texture, and structural features toward deep-learning approaches that can directly identify lodged areas. Early studies used canopy-height models, spectral and vegetation-related features, and UAV hyperspectral imagery to characterise lodging and investigate its relationship with crop performance or yield (Wilke et al., 2019; Wang et al., 2021; Bai et al., 2022). Zhao et al. (2019) subsequently demonstrated that U-Net could effectively extract lodged-rice areas from UAV imagery. More recent studies have improved segmentation through temporal modelling and attention mechanisms (Yu et al., 2023), as well as semi-supervised learning to reduce the dependence on pixel-level annotations (Zhi et al., 2023). In parallel, lightweight architectures have been developed to improve computational efficiency and support real-time lodging detection on UAVs or edge/harvesting platforms (Feng et al., 2024; Pan et al., 2025; Liu et al., 2026).

Despite these advances, several limitations remain in existing crop-lodging studies. Existing studies have often been developed and evaluated under specific datasets, field conditions, growth stages, or sensing settings, which may limit the assessment of their generalizability across different illumination

conditions, canopy structures, and geographic locations (Yu et al., 2023; Zhang et al., 2024b; Liu et al., 2026). Moreover, semantic segmentation heavily relies on dense, pixel-level annotations, which imposes a major bottleneck due to the high annotation burden (Zhi et al., 2023). Manual delineation of lodging areas is labour-intensive and time-consuming, particularly for visually ambiguous regions and irregular transition boundaries between lodged and upright crops (Zhi et al., 2023; Zhang et al., 2025). Moreover, segmentation accuracy, computational efficiency, and inference speed are not always evaluated jointly, despite their importance for practical UAV-based monitoring and real-time deployment (Feng et al., 2024; Pan et al., 2025; Liu et al., 2026).

Despite these advances, direct comparison among existing lodging segmentation approaches remains difficult because studies differ in datasets, sensing conditions, model architecture, and experimental protocols. In particular, the relative contribution of model architecture and input image resolution to segmentation performance has not been systematically examined under a common experimental setting. Consequently, it remains unclear which encoder–decoder configurations provide the most suitable balance between segmentation accuracy and computational efficiency for UAV-based rice lodging segmentation. A controlled benchmark using a common dataset and experimental protocol can therefore provide a more reliable basis for comparing encoder–decoder architectures and identifying suitable input resolutions for practical rice lodging assessment.

3. MATERIALS AND METHOD

3.1. Overview framework

Figure 1 presents the end-to-end framework for rice lodging semantic segmentation using UAV RGB imagery. RGB images were acquired over rice fields in An Giang Province, Vietnam, during three field surveys: one in the Summer–Autumn 2024 season and two in the Summer–Autumn 2025 season. After RGB image preprocessing, including correction of optical distortion, vignetting, and exposure variations, the images were annotated to generate four-class semantic segmentation masks for Background, Lodged Rice, Ponds/Ditches, and Field Edges/Paths. The annotations were organised in COCO format, resulting in a dataset of 715 annotated images. The dataset was then partitioned into training, validation, and test subsets using a cluster-aware split to reduce data leakage among visually similar images.

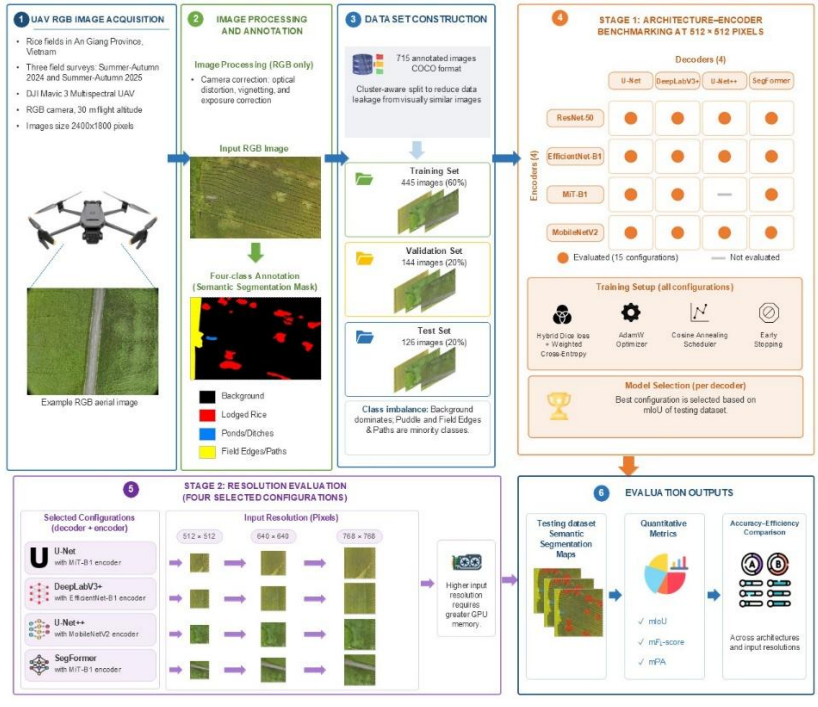


Figure 1. Overview of the proposed end-to-end framework for UAV RGB-based rice lodging semantic segmentation

In Stage 1, 15 available encoder–decoder configurations were benchmarked at an input resolution of 512×512 pixels. U-Net, DeepLabV3+, and SegFormer were evaluated with ResNet-50, EfficientNet-B1 (Tan & Le, 2019), MiT-B1 (Xie et al., 2021), and MobileNetV2 (Sandler et al., n.d.) encoders, whereas U-Net++ was evaluated with all encoders except MiT-B1. All configurations used the same training setup, including hybrid Dice loss with weighted cross-entropy, the AdamW optimiser, cosine annealing, and early stopping. Based primarily on the testing dataset's mIoU, the best configuration from each decoder was selected.

In Stage 2, the four selected configurations were evaluated at 512 × 512, 640 × 640, and 768 × 768 pixels to examine the effect of input resolution. We assessed performance on the test set using mean Intersection over Union (mIoU), mean F1-score (mF1-score), and mean pixel accuracy (mPA), along with inference time and GPU memory consumption, to examine the accuracy-efficiency trade-off. Overall, 640 × 640 pixels provided the most suitable input resolution within the evaluated experimental setting.

3.2. Image collection and data processing

This study was conducted in rice fields in An Giang Province, Vietnam. UAV imagery was acquired during three field surveys: one in the Summer–Autumn 2024 season in Dat Hiep and two surveys during the Summer–Autumn 2025 season in Kenh Cong Dien, on June 2 and June 10, 2025. To obtain clear canopy imagery and maintain relatively consistent illumination conditions, flights were conducted between 09:30 and 14:30 local time.

A DJI Mavic 3 Multispectral (M3M) quadcopter UAV equipped with a 20-megapixel RGB camera and a 5-megapixel multispectral sensor was used. The original RGB images had a spatial dimension of 5280 × 3956 pixels and were stored in JPG or RAW format, whereas multispectral imagery was stored as multi-band TIFF files. All flights were conducted at an altitude of 30m. Only UAV RGB imagery was used in this study.

The original RGB images were first corrected for optical distortion, vignetting, and exposure variations using camera intrinsic parameters and image metadata (Cai et al., 2024). The corrected images were subsequently resized to 2400 × 1800 pixels for manual annotation and dataset construction. Four semantic classes were defined for pixel-level annotation: Background, Lodged Rice, Ponds/Ditches, and Field Edges/Paths. We manually annotated the images to generate semantic segmentation masks and organised the annotations in COCO format. The resulting dataset comprised 715 annotated images.

3.3. Dataset splitting

Table 1 presents the class-wise pixel distributions across the three dataset splits. The data exhibit a substantial class imbalance: Background accounts for 79.7%–87.4% of pixels, whereas Lodged Rice accounts for 12.1%–18.4%. Ponds/Ditches and Field Edges/Paths each account for less than 2% of pixels in every split. This distribution motivates the use of class weighting during training.

The data were organised in COCO format. The area of each annotated class was calculated from the area field of the corresponding COCO annotations, and pixels not belonging to any annotation were assigned to Background. The 715 images were divided into training, validation, and test sets with target proportions of 60%, 20%, and 20%, respectively. To reduce data leakage, visually similar images were grouped into clusters before splitting. The clusters were allocated using a greedy procedure based on the total annotated area of each class to approximate the target class distributions. The final split comprised 445, 144, and 126 images for training, validation, and testing, respectively.

Table 1. Image counts and class-wise pixel proportions across the dataset splits. Pixel proportions are calculated within each split

Split	Images	Image proportion (%)	Background pixels (%)	Lodged Rice pixels (%)	Ponds/Ditches pixels (%)	Field Edges/Paths pixels (%)
Training	445	60	79.7	18.0	0.47	1.8
Validation	144	20	87.4	12.1	0.25	0.23
Test	126	20	79.9	18.4	0.45	1.3
Total	715	100	81.66	16.63	0.41	1.3

3.4. Experimental setup

All experiments were conducted on a workstation running Ubuntu 22.04.5 LTS, equipped with an Intel(R) Xeon(R) Gold 6152 processor clocked at 2.10 GHz, 128 GB RAM, and a 1 TB SSD. Deep-learning training was accelerated using an NVIDIA GeForce RTX 2060 GPU with 12 GB VRAM. The software environment was implemented in Python using PyTorch 2.3.1 and CUDA 12.1.

To mitigate class imbalance, in which Background pixels substantially outnumber those of the remaining classes, training used a hybrid loss that linearly combines Dice loss (Sudre et al., 2017) and weighted cross-entropy (WCE) loss (Phan & Yamamoto, 2020) at a 1:1 ratio:

$$L_{Total} = 0.5 \times L_{Dice} + 0.5 \times L_{WCE} \quad (1)$$

The WCE class weights were computed from the training-set pixel distribution using inverse frequency weighting:

$$W_C = \sum \frac{(N_C + \varepsilon)'}{(N_K + \varepsilon)^{-1'}}, \varepsilon = 10^{-6} \quad (2)$$

where N_C denotes the total pixel count of class c and ε is a small constant added for numerical stability. This formulation assigns larger weights to under-represented classes, so that the gradient contribution of minority is not dominated by the majority class.

A two-stage experimental strategy was adopted to identify suitable encoder–decoder combinations for rice lodging semantic segmentation. In the first stage, all available encoder–decoder combinations were trained and evaluated at the base input resolution of 512 × 512 pixels using the same training protocol. AdamW was used with a cosine annealing learning-rate scheduler. The learning rate was explored from 10^{-5} to 10^{-3}, while weight decay ranged from 10^{-6} to 10^{-3}. Training was limited to 50 epochs and incorporated early stopping with a patience of 10 epochs. Batch sizes of {4, 8, 12} were explored at this stage.

Based primarily on validation-set mIoU at 512 × 512 pixels, the best configuration from each decoder was selected for Stage 2. The selected configurations were U-Net with a MiT-B1 encoder, DeepLabV3+ with an EfficientNet-B1 encoder, U-Net++ with a MobileNetV2 encoder, and SegFormer with a MiT-B1 encoder. These four configurations were then trained and evaluated at 640 × 640 and 768 × 768 pixels to examine the effect of increased input resolution. Because higher resolutions require greater GPU memory consumption, batch sizes were limited to {4, 8} during the second stage. To account for stochastic variation during model training, each experimental configuration was independently trained three times with random initialisation while keeping the dataset split, training configuration, and evaluation protocol unchanged. For each metric, the reported value represents the mean ± standard deviation across the three independent runs.

4. RESULTS AND DISCUSSION

4.1. Comparison of encoder–decoder configurations at 512 × 512

Across the 15 evaluated encoder–decoder configurations on the testing dataset at a resolution of 512 × 512 pixels, overall mIoU ranged from 62.18% to 66.15% in Table 2, Table 3. A consistent class-difficulty ordering emerged across all architectures. Background (IoU: 86.73% – 88.48%) and Lodging Rice (IoU: 68.57% – 71.89%) were segmented reliably by every model. Lodged Rice, the primary class of agronomic interest, varied by less than 3.5 percentage points across all 15 configurations. In contrast, Ponds/Ditches (IoU: 50.64% – 58.57%) and Field Edges/Paths (IoU: 41.20% – 49.53%) remained substantially more challenging. Because mIoU is the unweighted mean of all four classes, performance variations in Ponds/Ditches and Field Edges/Paths contributed most to the differences in overall mIoU across configurations.

For U-Net (Table 2), the highest performance was achieved using the MiT-B1 encoder (mIoU: 65.24 ± 0.34%, mF_1: 77.95 ± 0.38%). ResNet-50 reached an mIoU of 64.65 ± 0.31% (0.59 percentage points below MiT-B1) but attained the highest mPA within U-Net (89.86 ± 0.33%). ResNet-50 performed strongly on Lodging Rice (71.56%) and Background (88.45%), whereas MiT-B1 recovered Field Edges/Paths more effectively (48.91% IoU). MobileNetV2 obtained an mIoU of 64.32 ± 0.31%. EfficientNet-B1 was a clear outlier with an mIoU of 62.18 ± 0.77%, placing it roughly 2–3 percentage points below the other three encoders. Consequently, U-Net with MiT-B1 was carried forward to Stage 2 based on overall mIoU and mF_1-score.

Table 2. Testing dataset performance (%) of U-Net with different encoders at 512 × 512 pixels

Class	Metric	ResNet-50	EfficientNet-B1	Mit-B1	MobileNetV2
Lodging Rice	IoU	**71.56 ± 0.78**	69.97 ± 0.84	70.07 ± 1.02	70.62 ± 0.95
	F_1-Score	**83.42 ± 0.53**	82.33 ± 0.58	82.40 ± 0.70	82.78 ± 0.65
	PA	91.18 ± 0.98	**92.43 ± 1.41**	91.41 ± 1.65	90.04 ± 1.52
Ponds/Ditches	IoU	50.82 ± 0.97	50.64 ± 1.30	**54.07 ± 1.79**	53.12 ± 1.37
	F_1-Score	67.39 ± 0.85	67.23 ± 1.14	**70.17 ± 1.52**	69.37 ± 1.18
	PA	**97.07 ± 0.32**	95.16 ± 1.81	94.85 ± 0.78	95.91 ± 0.80
Field Edges/Paths	IoU	47.78 ± 0.59	41.20 ± 3.42	**48.91 ± 1.05**	45.49 ± 1.32
	F_1-Score	64.66 ± 0.54	58.30 ± 3.47	**65.68 ± 0.95**	62.52 ± 1.24
	PA	80.64 ± 1.84	**80.79 ± 2.24**	80.81 ± 2.12	79.74 ± 1.29
Background	IoU	**88.45 ± 0.43**	86.93 ± 0.64	87.91 ± 0.51	88.07 ± 0.36
	F_1-Score	**93.87 ± 0.24**	93.01 ± 0.36	93.56 ± 0.29	93.65 ± 0.21
	PA	**90.56 ± 0.65**	88.73 ± 0.85	90.01 ± 0.80	90.48 ± 0.59
Overall	**mIoU**	64.65 ± 0.31	62.18 ± 0.77	**65.24 ± 0.34**	64.32 ± 0.31
	mF_1	77.34 ± 0.21	75.22 ± 0.73	**77.95 ± 0.38**	77.08 ± 0.34
	mPA	**89.86 ± 0.33**	89.28 ± 0.78	89.27 ± 0.85	89.04 ± 0.25

Encoder choice had a noticeable impact on DeepLabV3+, where mIoU spanned 2.24 percentage points across encoders (63.91% to 66.15%) in Table 3. EfficientNet-B1 achieved the best result for this decoder, leading in mIoU (66.15 ± 1.42%) and mF_1-score (78.64 ± 1.25%). This advantage was mainly driven by Field Edges/Paths, where EfficientNet-B1 achieved an IoU of 49.53 ± 3.55%, outperforming MiT-B1 (46.00%) by 3.53 percentage points. MiT-B1 ranked second with an mIoU of 65.51 ± 1.61%, performing marginally better on Ponds/Ditches (57.26%). MobileNetV2 achieved an mIoU of 64.64 ± 1.92% and the highest mPA in this group (89.49 ± 0.69%). ResNet-50 recorded the lowest mIoU at 63.91 ± 2.62%. DeepLabV3+ with EfficientNet-B1 was selected for further evaluation in Stage 2.

Table 3. Testing dataset performance (%) of DeepLabV3+ with different encoders at 512 × 512 pixels

Class	Metric	ResNet-50	EfficientNet-B1	Mit-B1	MobileNetV2
Lodging Rice	IoU	69.50 ± 1.90	71.08 ± 0.29	70.96 ± 0.27	68.57 ± 0.53
	F_1-Score	82.00 ± 1.33	83.10 ± 0.20	83.01 ± 0.18	81.36 ± 0.38
	PA	92.80 ± 1.63	91.55 ± 2.27	94.13 ± 0.15	92.87 ± 1.09
Ponds/Ditches	IoU	53.55 ± 2.22	55.56 ± 2.58	57.26 ± 3.90	56.67 ± 4.69
	F_1-Score	69.73 ± 1.88	71.41 ± 2.15	72.77 ± 3.16	72.26 ± 3.87
	PA	93.35 ± 2.45	91.69 ± 2.04	91.45 ± 1.86	92.50 ± 1.82
Field Edges/Paths	IoU	45.42 ± 5.72	49.53 ± 3.55	46.00 ± 2.65	46.57 ± 3.35
	F_1-Score	62.33 ± 5.38	66.20 ± 3.18	62.99 ± 2.48	63.50 ± 3.14
	PA	80.01 ± 3.38	81.32 ± 2.74	80.37 ± 3.89	84.12 ± 4.90
Background	IoU	87.19 ± 1.65	88.43 ± 0.32	87.81 ± 0.03	86.73 ± 0.24
	F_1-Score	93.15 ± 0.95	93.86 ± 0.18	93.51 ± 0.01	92.89 ± 0.14
	PA	88.96 ± 2.09	90.56 ± 0.71	89.38 ± 0.05	88.48 ± 0.51
Overall	**mIoU**	63.91 ± 2.62	66.15 ± 1.42	65.51 ± 1.61	64.64 ± 1.92
	mF_1	76.80 ± 2.18	78.64 ± 1.25	78.07 ± 1.39	77.50 ± 1.69
	mPA	88.78 ± 0.32	88.78 ± 0.64	88.83 ± 0.65	89.49 ± 0.69

U-Net++ was the least sensitive to encoder selection, with overall mIoU spanning only 0.21 percentage points across the three evaluated encoders (Table 4). ResNet-50 attained the highest mIoU (65.16 ± 0.46%) and mF_1-score (77.77 ± 0.43%), leading in Lodging Rice (71.72%) and Background (88.48%). MobileNetV2 achieved a close mIoU of 65.09 ± 1.05% and the highest mF_1-score (77.78 ± 0.87%), supported by the highest Ponds/Ditches IoU in this group (54.96 ± 1.25%, 0.50 percentage points above EfficientNet-B1). EfficientNet-B1 reached an mIoU of 64.95 ± 0.50% and the highest mPA (90.26 ± 0.52%).

Table 4. Testing dataset performance (%) of U-Net++ with different encoders at 512 × 512 pixels

Class	Metric	ResNet-50	EfficientNet-B1	MobileNetV2
Lodging Rice	IoU	71.72 ± 0.60	71.49 ± 0.84	70.57 ± 0.55
	F_1-Score	83.53 ± 0.41	83.37 ± 0.57	82.74 ± 0.38
	PA	91.73 ± 0.17	93.00 ± 0.37	91.39 ± 1.72
Ponds/Ditches	IoU	52.28 ± 0.67	54.46 ± 1.39	54.96 ± 1.25
	F_1-Score	68.66 ± 0.58	70.51 ± 1.17	70.93 ± 1.04
	PA	96.65 ± 0.52	95.11 ± 0.94	95.13 ± 0.73
Field Edges/Paths	IoU	48.15 ± 1.77	45.91 ± 2.13	46.99 ± 2.00
	F_1-Score	64.99 ± 1.62	62.91 ± 1.99	63.92 ± 1.86
	PA	79.37 ± 2.49	83.26 ± 3.29	84.08 ± 2.80
Background	IoU	88.48 ± 0.14	87.95 ± 0.62	87.84 ± 0.54
	F_1-Score	93.89 ± 0.08	93.59 ± 0.35	93.52 ± 0.31
	PA	90.55 ± 0.13	89.68 ± 0.73	89.92 ± 0.94
Overall	**mIoU**	65.16 ± 0.46	64.95 ± 0.50	65.09 ± 1.05
	mF$_1$	77.77 ± 0.43	77.60 ± 0.40	77.78 ± 0.87
	mPA	89.57 ± 0.67	90.26 ± 0.52	90.13 ± 0.40

Table 5. Testing dataset performance (%) of SegFormer with different encoders at 512 × 512 pixels

Class	Metric	ResNet-50	EfficientNet-B1	Mit-B1	MobileNetV2
Lodging Rice	IoU	71.89 ± 0.26	70.25 ± 1.27	71.05 ± 0.85	70.20 ± 0.32
	F_1-Score	83.65 ± 0.18	82.52 ± 0.88	83.07 ± 0.58	82.49 ± 0.22
	PA	91.33 ± 0.85	94.08 ± 0.83	93.08 ± 1.01	91.21 ± 0.61
Ponds/Ditches	IoU	52.50 ± 2.78	55.02 ± 1.80	58.57 ± 2.58	53.48 ± 1.65
	F_1-Score	68.82 ± 2.37	70.98 ± 1.49	73.85 ± 2.07	69.68 ± 1.41
	PA	95.64 ± 1.00	93.02 ± 1.29	91.35 ± 2.49	96.12 ± 0.46
Field Edges/Paths	IoU	43.61 ± 2.36	45.82 ± 2.37	45.76 ± 1.73	45.90 ± 1.56
	F_1-Score	60.71 ± 2.28	62.82 ± 2.25	62.77 ± 1.64	62.91 ± 1.46
	PA	79.83 ± 5.63	84.38 ± 3.72	90.88 ± 2.13	82.71 ± 1.91
Background	IoU	88.35 ± 0.26	87.23 ± 0.81	87.53 ± 0.51	87.61 ± 0.26
	F_1-Score	93.81 ± 0.15	93.18 ± 0.46	93.35 ± 0.29	93.40 ± 0.15
	PA	90.44 ± 0.47	88.71 ± 0.99	89.20 ± 0.68	89.72 ± 0.32
Overall	**mIoU**	64.09 ± 1.16	64.58 ± 1.43	65.73 ± 1.21	64.30 ± 0.18
	mF$_1$	76.75 ± 1.09	77.37 ± 1.15	78.26 ± 0.99	77.12 ± 0.15
	mPA	89.31 ± 1.57	90.05 ± 0.66	91.13 ± 0.38	89.94 ± 0.54

SegFormer with MiT-B1 (Table 5) produced an mIoU of 65.73 ± 1.21%, an mF1-score of 78.26 ± 0.99%, and the highest mPA across all 15 configurations at 512 × 512 pixels (91.13 ± 0.38%). It achieved the single highest Ponds/Ditches IoU observed in the entire benchmark (58.57 ± 2.58%). However, Field Edges/Paths remained its weakest class (45.76 ± 1.73% IoU). This demonstrates that the contextual representation of the Vision Transformer encoder is particularly beneficial for delineating small, low-contrast water regions. The other encoders combined with SegFormer achieved mIoUs of 64.58 ± 1.43% (EfficientNet-B1), 64.30 ± 0.18% (MobileNetV2), and 64.09 ± 1.16% (ResNet-50).

Based on mIoU, the configurations carried forward to Stage 2 were U-Net with MiT-B1, DeepLabV3+ with EfficientNet-B1, U-Net++ with MobileNetV2, and SegFormer with MiT-B1.

4.2. Effect of input resolution on the selected configurations

Based on validation-set mIoU at 512 × 512 pixels, the best configuration from each decoder was selected for further evaluation at 640 × 640 and 768 × 768 pixels: U-Net with a MiT-B1 encoder, DeepLabV3+ with an EfficientNet-B1 encoder, U-Net++ with a MobileNetV2 encoder, and SegFormer with a MiT-B1 encoder. Figure 2 reports the test-set results of these four configurations across the three input resolutions.

Increasing the input resolution from 512 × 512 to 640 × 640 pixels improved the mIoU for all selected configurations. At 640 × 640 pixels, DeepLabV3+ with the EfficientNet-B1 encoder achieved the highest mIoU (67.48%) and mF_1-score (79.60%), whereas SegFormer with the MiT-B1 encoder attained the highest mPA (91.23%). SegFormer–MiT-B1 also showed the smallest mIoU improvement, increasing by less than 0.1 percentage points relative to 512 × 512 pixels. The absence of further improvement at 768 × 768 pixels may have several possible explanations. At higher resolution, the same fixed-capacity encoder processes a smaller ground area per input, which may provide less contextual information relative to the spatial extent of lodging patterns. The additional fine-grained visual detail may also introduce greater within-class variability without increasing the number of annotated training images. GPU memory constraints also restricted the batch size to {4, 8} at this resolution, so the independent effect of resolution cannot be completely separated from the batch-size setting.

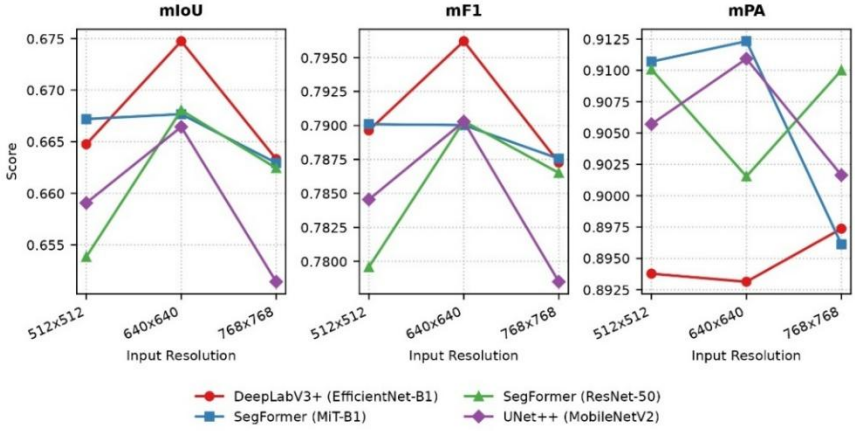


Figure 2. Testing dataset performance across input resolutions for U-Net with a MiT-B1 encoder, DeepLabV3+ with an EfficientNet-B1 encoder, U-Net++ with a MobileNetV2 encoder, and SegFormer with a MiT-B1 encoder

4.3. Visual comparison of segmentation results

At 640 × 640 pixels (Figure 3), DeepLabV3+ with the EfficientNet-B1 encoder provides the highest overall mask fidelity. Its predictions most consistently preserve the extent and irregular contours of lodged-rice regions across the representative ROIs, including large connected areas and several smaller, spatially separated patches. The model exhibits comparatively limited leakage into neighbouring standing rice and better retains the internal background gaps within complex lodging patterns. This indicates an effective balance between local boundary preservation and contextual interpretation. The result is consistent with the quantitative resolution comparison in Figure 2where this configuration achieves the highest mIoU and mF_1-Score at 640 × 640.

U-Net++ with MobileNetV2 recovers the dominant lodged-rice regions reasonably well, but it occasionally omits small isolated components and produces simplified boundaries. SegFormer with ResNet-50 shows relatively high sensitivity to lodging regions; however, it tends to enlarge predicted regions and connect nearby components that remain separated in the ground truth. This behaviour produces over-segmentation, particularly along narrow boundaries between lodged and standing rice. DeepLabV3+ with EfficientNet-B1 is more resistant to these errors, although minor boundary irregularities remain around highly fragmented regions.

SegFormer with MiT-B1 presents the most evident limitation in preserving fine spatial details. Small lodged-rice patches are frequently missed, smoothed, or merged into a single larger region, leading to under-segmentation and loss of object separation. The model also shows greater confusion near Ponds/Ditches and Field-Edge/Path structures, where visually similar textures or abrupt intensity changes may interfere with the prediction. These observations align with Figure 2: increasing the input resolution does not substantially improve the mIoU or mF_1-Score of MiT-B1, and performance decreases at 768 × 768.

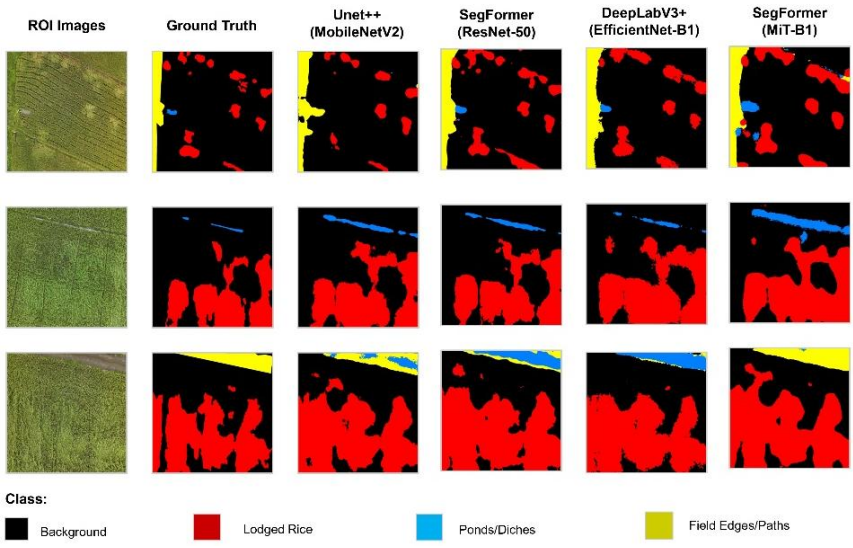


Figure 3. Qualitative comparison of segmentation predictions at 640 × 640 pixels

Within the evaluated configurations, CNN-based encoders preserve local texture and boundary cues more effectively than the MiT-B1 Transformer encoder. EfficientNet-B1, combined with the atrous spatial pyramid pooling and decoder of DeepLabV3+, benefits from multi-scale feature extraction while retaining discriminative local information. This combination explains its superior ability to distinguish irregular lodging boundaries from visually similar background structures in UAV imagery.

5. CONCLUSION

This study benchmarked 15 encoder–decoder configurations for rice lodging semantic segmentation using UAV RGB imagery under a common experimental protocol. Based on validation-set mIoU at 512×512 pixels, four configurations were selected for evaluation at higher resolutions, with DeepLabV3+ using the EfficientNet-B1 encoder achieving the best overall performance at 640×640 pixels (mIoU: 67.48%; mF_1-Score: 79.60%). Increasing the resolution to 768 × 768 pixels did not provide consistent improvements, indicating that model architecture and input resolution should be jointly considered for UAV-based rice-lodging segmentation. Performance remained limited for minority classes, emphasizing the necessity of additional annotated samples and validation across locations, seasons, rice varieties, and illumination conditions. Future work should also investigate partial and complete lodging to enable severity-aware mapping for potential yield-loss assessment and harvest management.

CONFLICT OF INTEREST

The authors declare no conflict of interest.

REFERENCES

Azizi, A., Zhang, Z., Rui, Z., Li, Y., Igathinathane, C., Flores, P., Mathew, J., Pourreza, A., Han, X., & Zhang, M. (2024). Comprehensive wheat lodging detection after initial lodging using UAV RGB images. *Expert Systems with Applications*, *238*, 121788. https://doi.org/10.1016/j.eswa.2023.121788

Bai, D., Li, D., Zhao, C., Wang, Z., Shao, M., Guo, B., Liu, Y., Wang, Q., Li, J., Guo, S., Wang, R., Li, Y., Qiu, L., & Jin, X. (2022). Estimation of soybean yield parameters under lodging conditions using RGB information from unmanned aerial vehicles. *Frontiers in Plant Science*, *13*, 1012293. https://doi.org/10.3389/fpls.2022.1012293

Cai, W., Lu, K., Fan, M., Liu, C., Huang, W., Chen, J., Wu, Z., Xu, C., Ma, X., & Tan, S. (2024). Rice Growth-Stage Recognition Based on Improved YOLOv8 with UAV Imagery. *Agronomy*, *14*(12), 2751. https://doi.org/10.3390/agronomy14122751

Celik, M. F., Nagarajan, P., Nelson, A., Unnisa, Z., Ogutu, B., Dash, J., Boschetti, M., Dobrowolska, E., Volden, E., & Darvishzadeh, R. (2026). Hyperspectral assessment of wheat lodging: From field to EnMAP satellite observations. *International*

Journal of Applied Earth Observation and Geoinformation, *149*, 105289. https://doi.org/10.1016/j.jag.2026.105289

Feng, G., Wang, C., Wang, A., Gao, Y., Zhou, Y., Huang, S., & Luo, B. (2024). Segmentation of Wheat Lodging Areas from UAV Imagery Using an Ultra-Lightweight Network. *Agriculture*, *14*(2), 244. https://doi.org/10.3390/agriculture14020244

Guo, C., & Tan, F. (2025a). SWRD–YOLO: A Lightweight Instance Segmentation Model for Estimating Rice Lodging Degree in UAV Remote Sensing Images with Real-Time Edge Deployment. *Agriculture*, *15*(15), 1570. https://doi.org/10.3390/agriculture15151570

Guo, C., & Tan, F. (2025b). SWRD–YOLO: A Lightweight Instance Segmentation Model for Estimating Rice Lodging Degree in UAV Remote Sensing Images with Real-Time Edge Deployment. *Agriculture*, *15*(15), 1570. https://doi.org/10.3390/agriculture15151570

Li, W., & Zhao, S. (2022). Semantic segmentation of buildings in high-resolution remote sensing images based on DeepLabV3+ algorithm. *Journal of Physics: Conference Series*, *2400*, 012037. https://doi.org/10.1088/1742-6596/2400/1/012037

Liu, P., Chai, X., Cui, Z., Zhu, Z., Hu, J., Yang, W., & Xu, L. (2026). StarNet-RiceSeg: An Efficient High-Dimensional Feature Mapping Network with Spatial Attention for Real-Time Rice Lodging Detection. *Agriculture*, *16*(7), 775. https://doi.org/10.3390/agriculture16070775

Liu, P., Cui, Z., Hu, J., Zhang, Q., Sun, J., Chai, X., & Xu, L. (2025). Lodge-Unet: A dual-frequency feature fusion network with boundary-aware optimisation for wheat lodging detection via autonomous harvesters. *Computers and Electronics in Agriculture*, *238*, 110769. https://doi.org/10.1016/j.compag.2025.110769

Pan, C., Xie, J., Zhang, G., Cheng, T., Han, D., Fang, Q., Ju, S., & Zhang, D. (2025). Real-time UAV-based wheat lodging detection via edge-accelerated improved Mask-RT-DETR. *Smart Agricultural Technology*, *12*, 101509. https://doi.org/10.1016/j.atech.2025.101509

Phan, T. H., & Yamamoto, K. (2020). *Resolving Class Imbalance in Object Detection with Weighted Cross Entropy Losses* (arXiv:2006.01413). arXiv. https://doi.org/10.48550/arXiv.2006.01413

Sandler, M., Howard, A., Zhu, M., Zhmoginov, A., & Chen, L.-C. (n.d.). *MobileNetV2: Inverted Residuals and Linear Bottlenecks*.

Sudre, C. H., Li, W., Vercauteren, T., Ourselin, S., & Jorge Cardoso, M. (2017). Generalised Dice Overlap as a Deep Learning Loss Function for Highly Unbalanced Segmentation. In M. J. Cardoso, T. Arbel, G. Carneiro, T. Syeda-Mahmood, J. M. R. S. Tavares, M. Moradi, A. Bradley, H. Greenspan, J. P. Papa, A. Madabhushi, J. C. Nascimento, J. S. Cardoso, V. Belagiannis, & Z. Lu (Eds.), *Deep Learning in Medical Image Analysis and Multimodal Learning for Clinical Decision Support* (pp. 240–248). Springer International Publishing. https://doi.org/10.1007/978-3-319-67558-9_28

Tan, M., & Le, Q. (2019). EfficientNet: Rethinking Model Scaling for Convolutional Neural Networks. *Proceedings of the 36th International Conference on Machine Learning*, 6105–6114. https://proceedings.mlr.press/v97/tan19a.html

Wang, J., Wu, B., Kohnen, M. V., Lin, D., Yang, C., Wang, X., Qiang, A., Liu, W., Kang, J., Li, H., Shen, J., Yao, T., Su, J., Li, B., & Gu, L. (2021). Classification of Rice Yield Using UAV-Based Hyperspectral Imagery and Lodging Feature. *Plant Phenomics*, *2021*, 9765952. https://doi.org/10.34133/2021/9765952

Wilke, N., Siegmann, B., Klingbeil, L., Burkart, A., Kraska, T., Muller, O., Van Doorn, A., Heinemann, S., & Rascher, U. (2019). Quantifying Lodging Percentage and Lodging Severity Using a UAV-Based Canopy Height Model Combined with an Objective Threshold Approach. *Remote Sensing*, *11*(5), 515. https://doi.org/10.3390/rs11050515

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. https://proceedings.neurips.cc/paper_files/paper/2021/hash/64f1f27bf1b4ec22924fd0acb550c235-Abstract.html

Yu, J., Cheng, T., Cai, N., Zhou, X.-G., Diao, Z., Wang, T., Du, S., Liang, D., & Zhang, D. (2023). Wheat Lodging Segmentation Based on Lstm_PSPNet Deep Learning Network. *Drones*, *7*(2), 143. https://doi.org/10.3390/drones7020143

Zhang, P., Niu, L., Cai, M., Chen, H., & Sun, X. (2024a). AAUConvNeXt: Enhancing Crop Lodging Segmentation with Optimised Deep Learning Architectures. *Plant Phenomics*, *6*, 0182. https://doi.org/10.34133/plantphenomics.0182

Zhang, P., Zhang, S., Wang, J., & Sun, X. (2024b). Identifying rice lodging based on semantic segmentation architecture optimisation with UAV remote sensing imaging. *Computers and Electronics in Agriculture*, *227*, 109570. https://doi.org/10.1016/j.compag.2024.109570

Zhang, X., Han, X., Qin, C., An, Z., Qi, B., Liu, J., Du, B., Liu, H., Wang, Y., Meng, L., & Wang, C. (2025). A Multi-Scale Rice Lodging Monitoring Method Based on MSR-Lodfnet. *Agriculture*, *15*(23), 2487. https://doi.org/10.3390/agriculture15232487

Zhao, X., Yuan, Y., Song, M., Ding, Y., Lin, F., Liang, D., & Zhang, D. (2019). Use of Unmanned Aerial Vehicle Imagery and Deep Learning UNet to Extract

Rice Lodging. *Sensors*, *19*(18), 3859. https://doi.org/10.3390/s19183859

Zhi, H., Yang, B., & Zhu, Y. (2023). Collaborative Wheat Lodging Segmentation Semi-Supervised Learning Model Based on RSE-BiSeNet Using UAV Imagery. *Agronomy*, *13*(11), 2772. https://doi.org/10.3390/agronomy13112772

Zhou, Z., Rahman Siddiquee, M. M., Tajbakhsh, N., & Liang, J. (2018). UNet++: A Nested U-Net Architecture for Medical Image Segmentation. In D. Stoyanov, Z. Taylor, G. Carneiro, T. Syeda-Mahmood, A. Martel, L. Maier-Hein, J. M. R. S. Tavares, A. Bradley, J. P. Papa, V. Belagiannis, J. C. Nascimento, Z. Lu, S. Conjeti, M. Moradi, H. Greenspan, & A. Madabhushi (Eds.), *Deep Learning in Medical Image Analysis and Multimodal Learning for Clinical Decision Support* (pp. 3–11). Springer International Publishing. https://doi.org/10.1007/978-3-030-00889-5_1