Remote Sensing Applications: Society and Environment 41 (2026) 101895

Contents lists available at ScienceDirect

Remote Sensing Applications: Society and Environment

journal homepage: www.elsevier.com/locate/rsase

Lightweight dual-encoder deep learning integrating Sentinel-1 and Sentinel-2 for paddy field mapping

Bagus Setyawan Wijaya [a,b,c], Rinaldi Munir [a,b], Nugraha Priya Utama [a,b,*]

[a] School of Electrical Engineering and Informatics, Bandung Institute of Technology, Bandung, 40132, West Java, Indonesia
[b] Center of AI, Bandung Institute of Technology, Bandung, 40132, West Java, Indonesia
[c] Statistics Indonesia, Central Jakarta, 10710, Jakarta, Indonesia

ARTICLE INFO

Dataset link: https://developers.google.com/earth-engine/datasets

Keywords:
Deep learning
Dual encoder
Indonesia
Lightweight model
Paddy fields mapping
Radar–optical fusion
Semantic segmentation
Sentinel-1
Sentinel-2

ABSTRACT

Timely and accurate paddy field mapping remains challenging in tropical regions due to persistent cloud cover and complex cropping patterns. We propose **DSSNet**, a lightweight dual-encoder semantic segmentation framework that fuses Sentinel-1 SAR and Sentinel-2 optical imagery. DSSNet leverages modality-specific backbones from different architectural paradigms: *EfficientNet-B0*, a convolutional, and *MaxVit-T*, a transformer-based encoder. To further enhance multimodal feature discrimination, we introduce two axial attention mechanisms — Axial Spatial Attention (ASA) and Axial Channel Attention (ACA) — to selectively emphasize directional spatial patterns and inter-channel relationships. Evaluated on imagery from Indonesia rice-growing regions during the 2019 season, DSSNet achieves an F1-score of 0.8982, pixel accuracy of 0.8998, and mIoU of 0.8156, outperforming ten benchmark models. These findings underscore the operational feasibility of lightweight dual-paradigm fusion architectures for large-scale, in-season agricultural mapping under complex environmental conditions. Our code and model will be publicly available at https://github.com/project4earth/DSSNet.

1. Introduction

Over the past two decades, global agricultural landscapes have experienced significant expansion. The harvested area of primary crops increased by 24%, reaching approximately 1.5 billion hectares by 2022 (FAO, 2024). Cereals remain the dominant crop group, accounting for more than half of the global harvested area and serving as the primary source of dietary energy across all regions. Among these, paddy rice plays a critical role, serving as a staple food for nearly half of the world's population.

Indonesia, as the fourth-largest rice producer globally, produced 53.98 million tonnes in 2023 (FAO, 2023). Within the archipelago, Indramayu Regency in West Java stands out as a key rice-producing region (Statistics Indonesia, 2025a), contributing significantly to the nation's annual supply. Accordingly, timely and spatially detailed information on paddy field extent and distribution in this area is essential for government agencies and stakeholders involved in agricultural development and food logistics. These challenges necessitate a reliable, automated, and scalable approach to paddy field mapping that accounts for local agricultural variability and diverse environmental conditions.

Recent advances in Earth observation have facilitated the use of freely available satellite data, particularly Sentinel-1 Synthetic Aperture Radar (SAR) and Sentinel-2 multispectral imagery (Gorelick et al., 2017). While Sentinel-2 provides rich spectral indicators of vegetation growth, Sentinel-1 is immune to atmospheric interference and capable of capturing surface structure and moisture dynamics—factors crucial for early-stage rice field detection. The fusion of these modalities has shown to improve classification

* Corresponding author.School of Electrical Engineering and Informatics, Bandung Institute of Technology, Bandung, 40132, West Java, Indonesia.
E-mail address: utama@itb.ac.id (N.P. Utama).

https://doi.org/10.1016/j.rsase.2026.101895
Received 15 July 2025; Received in revised form 29 November 2025; Accepted 26 January 2026
Available online 5 February 2026

performance, especially in cloud-prone agricultural regions with high spectral heterogeneity (Yang et al., 2024; Borsah et al., 2025; Fikriyah et al., 2025; Qiu et al., 2025; Therias et al., 2025). Furthermore, combining spectral and radar information has demonstrated superior performance in in-season crop classification tasks (Aashi and Vema, 2025). However, this multimodal fusion often increases data dimensionality and computational demand, highlighting the need for lightweight yet effective models that balance accuracy with efficiency.

A fundamental challenge in multimodal remote sensing is determining how best to integrate heterogeneous data sources to optimize predictive accuracy (Aashi and Vema, 2025). A common method is feature concatenation, where features extracted from both modalities are stacked and processed jointly through a unified encoder network (Altarez et al., 2023; Ghazaryan et al., 2025; Kanja et al., 2025; Therias et al., 2025). While this approach simplifies the model structure and reduces redundancy, it may lead to suboptimal learning due to the absence of modality-specific adaptation. Alternatively, dual-stream or dual-encoder architectures allocate dedicated encoders to each modality, allowing each stream to specialize in capturing unique spatial and spectral characteristics. Feature fusion then occurs in subsequent stages (Gao et al., 2021; Li et al., 2022c; Ma et al., 2022; Ren et al., 2022; Xiang et al., 2024). This paradigm is particularly effective when spatial and temporal alignment between modalities is maintained, enabling the joint exploitation of SAR's structural sensitivity and multispectral data's spectral richness to enhance segmentation performance. However, dual-encoder architectures typically introduce higher complexity and computational costs, particularly when deep backbone networks are used.

Despite recent progress, many existing models remain inefficient and insufficiently robust under persistent cloud cover, which frequently affects tropical agricultural regions. To overcome this limitation, we propose a lightweight and modality-aware segmentation framework tailored for operational-scale paddy field mapping. The model utilizes compact and complementary encoder backbones, EfficientNet-B0 for SAR data and MaxVit-T for multispectral data, offering an effective trade-off between representational capacity and computational efficiency. To compensate for the reduced capacity of lightweight encoders, we integrate a convolutional axial attention mechanism that captures long-range spatial dependencies in a parameter-efficient manner. This facilitates the aggregation of contextual information over extended receptive fields while maintaining low memory usage. As a result, the proposed model is well suited for large-scale deployment in resource-constrained and cloud-prone environments.

The remainder of this paper is organized as follows. Section 2 reviews related works in multimodal remote sensing and paddy field mapping. Section 3 describes the study area and the proposed methodology, including the dual-encoder architecture and preprocessing pipeline. Section 4 presents the experimental results and ablation studies. Section 5 discusses the key findings, implications, and limitations. Finally, Section 6 concludes the study and outlines directions for future research.

2. Related works

2.1. Remote sensing approaches for paddy field mapping

Remote sensing has been widely utilized for paddy field mapping due to its critical role in precision agriculture (Deng et al., 2025; Garcia et al., 2025; Nie et al., 2025). Optical satellite imagery is commonly used to detect rice fields based on spectral characteristics and vegetation indices (Zhao et al., 2023; Ni et al., 2021). However, such data are frequently hindered by cloud cover, particularly during key phenological stages. Synthetic Aperture Radar (SAR), especially from Sentinel-1, provides all-weather imaging capabilities and is sensitive to surface moisture and roughness—attributes essential for detecting early-stage rice transplanting (Thorp and Drajat, 2021; Jo et al., 2023; Sun et al., 2023; Hegde et al., 2025). Several studies have explored multimodal fusion of SAR and optical data to improve seasonal classification performance (Shen et al., 2023; Aashi and Vema, 2025; Kumari et al., 2025). Nonetheless, many of these approaches still rely on hand-crafted indices, rule-based thresholds, or per-pixel classifiers, which limit their generalizability across diverse agroecological zones and growing seasons.

A recent deep learning-based approach specifically designed for paddy rice mapping is FR-Net (Xia et al., 2022), which employs a multi-resolution fusion unit to integrate Landsat optical features across coarse- and fine-scale representations. While FR-Net demonstrates strong performance for mid-resolution optical imagery, its architecture assumes homogeneous single-sensor inputs and does not support cross-modal interactions between radar and optical data. In contrast, the dual-encoder fusion strategy explored in this study is explicitly developed for heterogeneous Sentinel-1/Sentinel-2 integration by separately encoding SAR backscatter and 10 m optical features prior to resolution-aligned fusion. This design leverages complementary SAR–optical information while preserving the fine spatial granularity required to delineate narrow and fragmented paddy field parcels, thereby addressing key limitations of mid-resolution optical-only fusion frameworks.

2.2. Semantic segmentation of remote sensing imagery

Deep learning has significantly advanced the performance of semantic segmentation in remote sensing, with convolutional neural networks (CNNs) and Transformer-based models being particularly influential (Boulila et al., 2024; Hussain et al., 2025; Zhang et al., 2025; Zhao et al., 2025). Classical encoder–decoder architectures such as U-Net (Ronneberger et al., 2015) remain widely used due to their simplicity and robustness in crop mapping (Wang et al., 2022c; Therias et al., 2025). Other CNN-based architectures, including TerraSegNet (Wijaya et al., 2025), extend this direction by incorporating lightweight axial or spectral–spatial attention mechanisms to enhance spatial coherence and multiscale feature representation, particularly in high-resolution optical imagery. Feature Pyramid Network variants such as A2FPN (Li et al., 2022b) have also been introduced to better capture multiscale contextual dependencies.

More recently, hybrid architectures that integrate CNNs with Transformer components have achieved promising results. For instance, ST-UNet (He et al., 2022) and DCSwin (Wang et al., 2022a) embed Swin Transformers into U-Net and FCN structures, respectively. Models such as UNetFormer (Wang et al., 2022b), CMTFNet (Wu et al., 2023), and CMLFormer (Wu et al., 2024) combine convolutional backbones (e.g., ResNet) with Transformer blocks to jointly enhance local feature representation and long-range contextual reasoning. These designs have demonstrated strong performance in high-resolution remote sensing images, where both spatial detail and global context are essential.

Despite these advancements, most existing architectures — including CNN-based, attention-augmented, and hybrid CNN–Transformer models — are primarily developed for unimodal optical data. As a result, they do not explicitly address the challenges of multimodal fusion, such as integrating heterogeneous spatial statistics and complementary backscatter–reflectance signatures across SAR and multispectral sensors. This limitation motivates the development of architectures that natively support dual-modality processing, such as the Sentinel-1 and Sentinel-2 fusion required for reliable paddy field mapping.

2.3. Dual-stream and modality-specific architectures

To better exploit the complementary properties of heterogeneous sensors, dual-stream or dual-encoder architectures have gained traction in multimodal remote sensing. Unlike early-fusion strategies that concatenate inputs and process them through a unified encoder, dual-stream models assign separate encoders for each modality, enabling modality-specific representation learning while preserving spectral and structural distinctiveness (Gao et al., 2021; Li et al., 2022c; Ren et al., 2022; Ma et al., 2022; Xiang et al., 2024). For instance, DDHRNet employs two HRNet encoders for Gaofen-2 and Gaofen-3 data, fusing their outputs via cross-modal attention mechanisms (Ren et al., 2022). Similarly, MCANet utilizes dual ResNet-101 backbones with a cross-attention module to align semantic features between modalities (Li et al., 2022c).

Transformer-based architectures have also adopted this paradigm. STransFuse integrates a ResNet-34 and Swin Transformer as dual encoders, resulting in a substantial model size of 86.99 million parameters (Gao et al., 2021). In contrast, TCNet uses a more compact configuration, combining ResNet-34 and DeiT-Small, reducing the parameter count to 34.87 million while still effectively capturing complementary features (Xiang et al., 2024). Although these architectures have demonstrated improved fusion capabilities and segmentation performance, their reliance on deep and computationally intensive backbones — such as ResNet (He et al., 2015), HRNet (Wang et al., 2021), and Swin Transformer (Liu et al., 2021) — poses limitations for real-time or large-scale applications.

To address this gap, recent efforts have explored lightweight dual-encoder architectures that prioritize computational efficiency without significantly compromising performance. In this context, our proposed DSSNet introduces a highly compact dual-stream framework comprising EfficientNet-B0 for SAR data and MaxVit-T for multispectral inputs, totaling only 2.79 million parameters. Compared to heavier models like STransFuse and TCNet, DSSNet offers a more scalable and cost-effective solution for large-area paddy field mapping under tropical cloud-prone conditions, where rapid and reliable inference is critical for operational decision-making.

3. Methodology

3.1. Study area

This study was conducted in Indramayu Regency, situated in the northern part of West Java Province, Indonesia, between 107.52°E–108.36°E longitude and 6.15°S–6.40°S latitude (Fig. 1). As one of the most productive rice-growing regions in the country, Indramayu plays a critical role in ensuring national food security. According to official statistics (Statistics Indonesia, 2025a,b), the region recorded approximately 230,457 hectares of harvested rice area in 2023, yielding around 1,424,303 tonnes of paddy. This high productivity is attributed to the region's capacity for cultivating two rice crops annually, supported by extensive irrigation systems and favorable agroclimatic conditions.

Climatically, Indramayu experiences a tropical monsoon climate characterized by distinct wet and dry seasons. In 2024, the region recorded approximately 2199 mm of annual precipitation, spread over 283 rainy days. Persistent cloud cover during the wet season poses significant challenges for optical satellite observation. These conditions make Indramayu an ideal case study for evaluating satellite-based paddy field mapping approaches. The combination of frequent cloud cover, dynamic cropping patterns, and heterogeneous field structures underscores the need for robust multi-sensor fusion techniques capable of delivering accurate and scalable mapping in tropical rice ecosystems.

3.2. Data acquisition

This study utilized dual-polarized Sentinel-1 Ground Range Detected (GRD) products (VV and VH) and surface reflectance imagery from Sentinel-2 Level-2A (blue, green, red, and near-infrared bands), covering the entire 2019 calendar year. All Sentinel-1 scenes were acquired in Interferometric Wide (IW) swath mode and included both ascending and descending orbits to maximize temporal coverage. The resulting backscatter images were resampled to a spatial resolution of 10 m and partitioned into non-overlapping patches of 512 × 512 pixels. Similarly, Sentinel-2 reflectance data were resampled to the same spatial resolution and co-registered with Sentinel-1 images acquired within a temporal window of ±3 days to ensure both spatial and phenological alignment across modalities. Only the 10 m Sentinel-2 bands (Blue, Green, Red, and NIR) were included in this study to maintain spatial consistency with Sentinel-1 SAR imagery, which also operates at a 10 m resolution. Short-wave infrared (SWIR) bands were

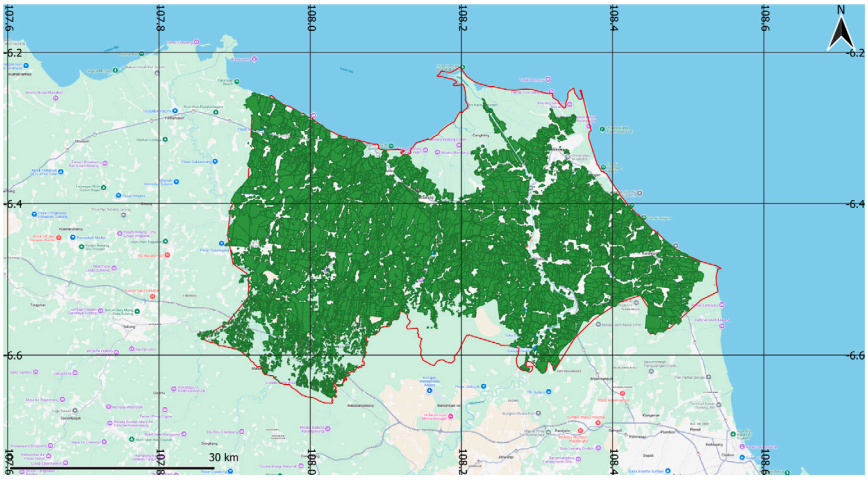

Fig. 1. Study area map of Indramayu Regency, West Java, Indonesia. The administrative boundary is shown in red, and the spatial distribution of paddy fields is overlaid in green based on the 2019 reference map from the Ministry of Agriculture. This region represents a complex agricultural landscape with intensive rice cultivation under varying irrigation and planting conditions. (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

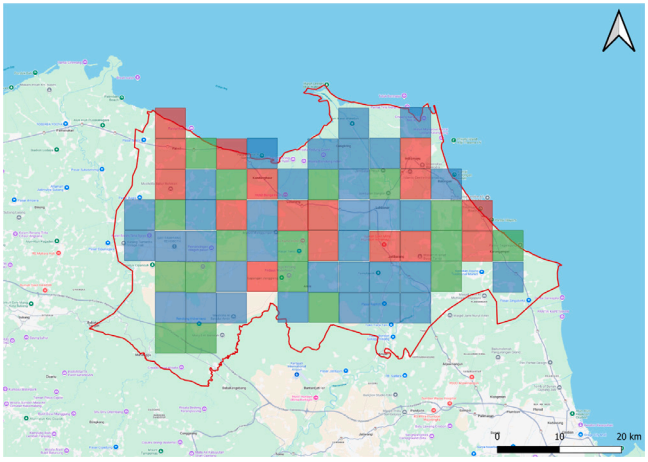

Fig. 2. Spatial partitioning of the dataset used in this study. Each color represents a different subset: training (blue), validation (green), and test (red). This spatial division mitigates data leakage and ensures reliable generalization assessment. (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

excluded because their native 20 m resolution would require upsampling, potentially introducing interpolation artefacts that can distort fine-scale boundaries of narrow and fragmented paddy fields typical in the study region. Ensuring uniform spatial resolution across modalities was particularly important for the subsequent multimodal fusion and cloud-reconstruction procedures.

A total of 73 Sentinel-1 and Sentinel-2 image pairs were collected from 58 unique acquisition dates throughout the 2019 rice-growing season, resulting in 4234 patches per modality and covering an area of approximately 1913 square kilometers. To avoid data leakage and ensure reliable generalization, the dataset was spatially divided into training (50%), validation (25%), and test (25%) subsets. An overview of the spatial partitioning is presented in Fig. 2, while the monthly distribution of scene acquisitions is summarized in Table 1. The relatively uniform temporal distribution of imagery ensures that the model is trained under diverse seasonal and phenological conditions.

Ground truth labels were derived from the official 2019 paddy field map produced by the Ministry of Agrarian Affairs and Spatial Planning/National Land Agency of Indonesia. This version was selected to ensure temporal consistency with the satellite data. A summary of the dataset composition is provided in Table 2, and representative examples of input modalities and corresponding ground truth labels are illustrated in Fig. 3.

Table 1
Monthly distribution of Sentinel-1 and Sentinel-2 paired scenes collected in 2019.

Month	Jan	Feb	Mar	Apr	May	Jun	Jul	Aug	Sep	Oct	Nov	Dec	Total
No. of images	4	4	5	5	5	5	6	5	5	5	5	4	58

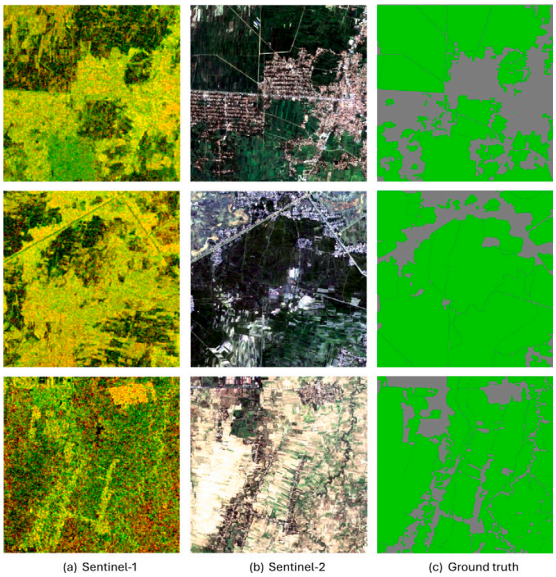

Fig. 3. Representative samples of Sentinel-1, Sentinel-2 images and corresponding ground truth labels.

Table 2
Summary of satellite and reference datasets used in this study. Sentinel-1 SAR and Sentinel-2 MSI images were acquired in 2019 to ensure consistency with the reference paddy field map.

Dataset	Sensor/Source	Type	Spatial resolution	Date range	Size/Number of patches	Remarks
Sentinel-1 GRD	ESA/Copernicus	Radar (C-band)	10 m	Jan–Dec 2019	512 × 512/4234	VV, VH
Sentinel-2 MSI Level-2A	ESA/Copernicus	Multispectral	10 m	Jan–Dec 2019	512 × 512/4234	Blue, Green, Red, NIR
Paddy field Reference map	National Land Agency (ID)	Vector Multipolygon	10 m	2019 (static)	512 × 512/73 (static)	Ground truth

3.3. Data preprocessing

A comprehensive preprocessing pipeline was implemented to enhance data quality, ensure geometric and radiometric consistency, and provide cloud-free multimodal inputs prior to model training. For Sentinel-1 SAR imagery, a Lee filter was applied to reduce speckle noise while preserving edge structures and backscatter characteristics, which are essential for paddy field delineation (Novresiandi et al., 2024).

For Sentinel-2 optical imagery, a two-stage cloud removal strategy was adopted to address persistent monsoon-season cloud contamination. In the first stage, cloud-affected pixels were identified using CMLFormer (Wu et al., 2024), a transformer-based cloud segmentation model pre-trained on the WHUS2-CD+ dataset (Li et al., 2022a). In the second stage, instead of performing inpainting strictly within the optical domain, cloud-masked regions were reconstructed using a cross-modal Pix2Pix model (Isola et al., 2017) conditioned on co-registered Sentinel-1 backscatter. This reconstruction approach exploits the complementary sensitivity of SAR to surface roughness and moisture to infer plausible optical reflectance under cloud cover.

To minimize the risk of artifacts influencing downstream learning, the reconstructed optical pixels were validated through visual inspection, local histogram consistency against nearby cloud-free patches, and checks for temporal or spatial discontinuities. Importantly, cloud removal and reconstruction were performed *before* dataset partitioning. The final cloud-free mosaics were then split into training, validation, and test subsets while preserving class proportions, ensuring that all subsets originate from a consistent data domain and preventing domain shift that would occur if synthetic pixels were present only in the training set.

An overview of the full preprocessing workflow is presented in Fig. 4, summarizing speckle filtering, cloud segmentation, cross-modal reconstruction, and subsequent dataset preparation.

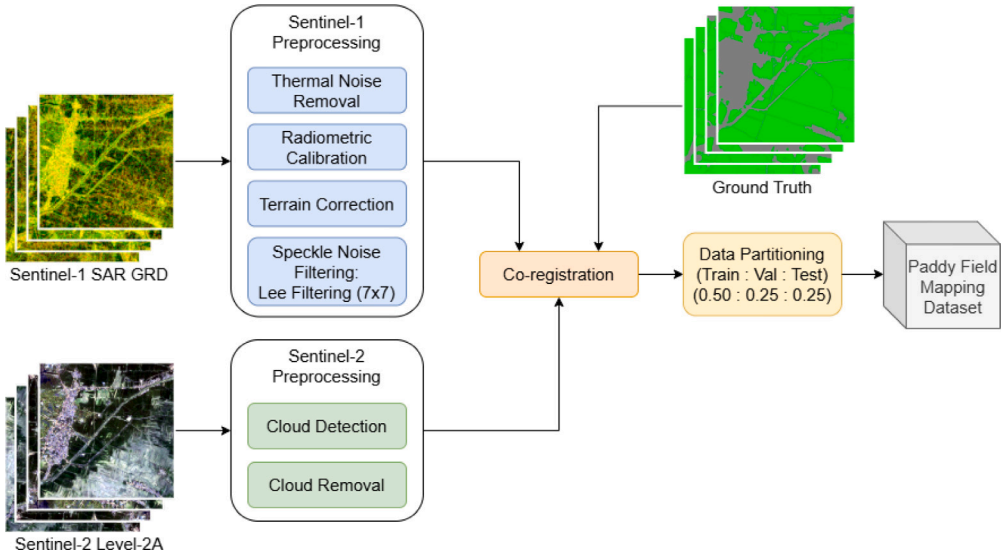

Fig. 4. Preprocessing workflow for integrating Sentinel-1 SAR GRD and Sentinel-2 Level-2A imagery. The pipeline includes speckle filtering using a Lee filter for SAR data, cloud removal in Sentinel-2, and spectral reconstruction.

3.4. Model architecture: Dual-Stream Segmentation Network (DSSNet)

We propose DSSNet, a lightweight dual-stream semantic segmentation architecture specifically designed for multimodal fusion of Sentinel-1 SAR and Sentinel-2 multispectral imagery. The network employs two modality-specific encoders: a convolutional encoder based on EfficientNet-B0 (Tan and Le, 2020) for SAR data, and a transformer-based encoder using MaxVit-T (Tu et al., 2022) for optical data. These backbones were selected due to their favorable trade-off between computational efficiency and representational capacity, which is essential for large-scale deployment in resource-constrained environments. EfficientNet-B0 enables compact feature extraction from SAR imagery, effectively capturing structural and textural cues with low parameter overhead. MaxVit-T, which integrates convolutional and self-attention operations, allows the extraction of rich contextual information from multispectral inputs by capturing both local and global dependencies. To harmonize spatial resolution and feature abstraction differences between modalities, DSSNet introduces an Adaptive Fusion Module (AFM), which aligns and fuses multi-scale features from both encoder branches. The fused feature representation is subsequently refined using Axial Spatial Attention (ASA) to enhance spatial detail, and Axial Channel Attention (ACA) to recalibrate semantic-level responses. These modules are inspired by Axial-DeepLab (Wang et al., 2020) and SeaFormer++ (Wan et al., 2025), but re-designed using convolutional axial attention rather than full self-attention to reduce computational overhead. The convolutional axial attention mechanism decomposes attention along horizontal and vertical axes, enabling the model to efficiently capture elongated spatial structures such as field boundaries. This design not only improves spatial sensitivity but also ensures scalability for high-resolution input data.

Beyond architectural efficiency, DSSNet introduces a new insight into multimodal fusion by leveraging a direction-aware axial refinement mechanism tailored for the structural geometry of paddy fields. Unlike prior hybrid fusion architectures such as MCANet (Li et al., 2022c) and DDHRNet (Ren et al., 2022), DSSNet employs convolutional axial attention to model cross-modal interactions along horizontal and vertical axes. This decomposition allows the network to capture long-range dependencies while maintaining linear memory complexity, enabling efficient processing of 10 m resolution Sentinel-1/Sentinel-2 imagery. From a theoretical perspective, axial attention restricts information flow to structured directional paths, which is advantageous for agricultural landscapes where field boundaries exhibit anisotropic spatial patterns. Empirically, this refinement mechanism enhances boundary fidelity and improves the alignment of SAR backscatter with optical reflectance gradients, providing clearer fusion dynamics than those observed in full self-attention or convolution-only settings. Thus, the novelty of DSSNet lies not merely in architectural simplification but in the introduction of a resolution-consistent, direction-aware fusion refinement strategy that strengthens multimodal feature coherence with minimal overhead. The complete DSSNet architecture is illustrated in Fig. 5.

3.4.1. Adaptive Fusion Module (AFM)

To align and integrate heterogeneous feature representations extracted from Sentinel-1 and Sentinel-2 modalities, we design an *Adaptive Fusion Module* (AFM) that simultaneously processes low-level and high-level features from each stream. This module addresses potential mismatches in spatial resolution and channel dimensions, enabling effective multimodal feature fusion.

Let $\mathbf{L}_1 \in \mathbb{R}^{C_{l1} \times H_l \times W_l}$ and $\mathbf{H}_1 \in \mathbb{R}^{C_{h1} \times H_h \times W_h}$ denote the low-level and high-level features from the Sentinel-1 encoder. Similarly, let $\mathbf{L}_2 \in \mathbb{R}^{C_{l2} \times H'_l \times W'_l}$ and $\mathbf{H}_2 \in \mathbb{R}^{C_{h2} \times H'_h \times W'_h}$ represent the corresponding features from the Sentinel-2 encoder.

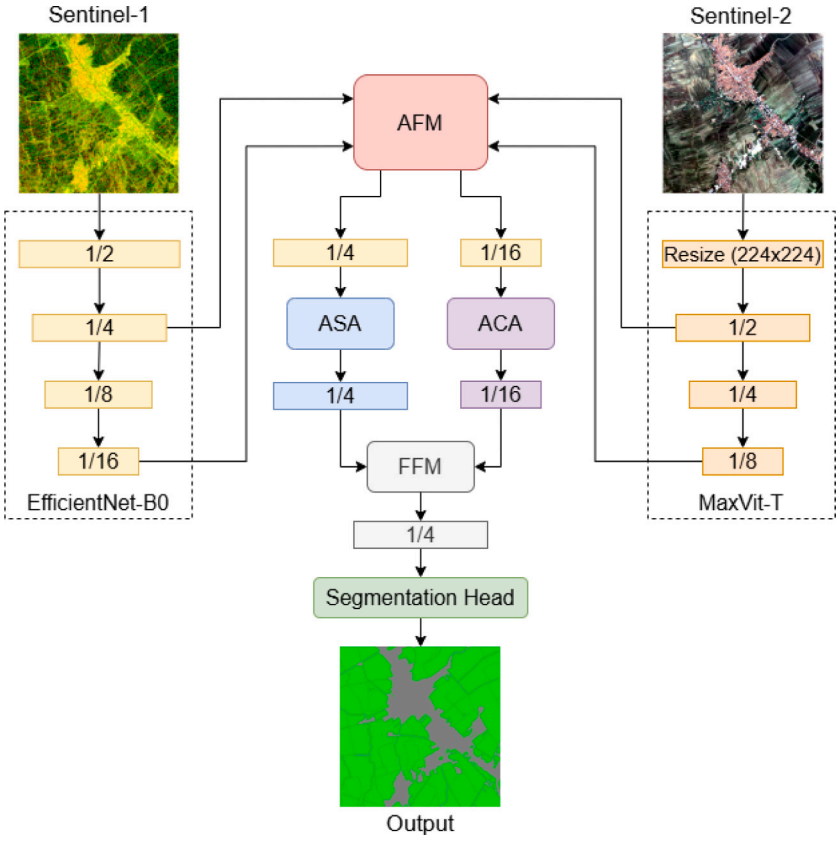

Fig. 5. The architecture of Dual-Stream Segmentation Network (DSSNet), highlighting its key components: Lighweight Dual Encoder (EfficientNet-B0 and MaxVit-T), Adaptive Fusion Module (AFM), Axial Spatial Attention (ASA), Axial Channel Attention (ACA), and Feature Fusion Module (FFM).

Each input is first projected into a common embedding space using 1×1 convolutions, batch normalization, and LeakyReLU activation:

$$\hat{\mathbf{L}}_1 = \phi_{\text{low},1}(\mathbf{L}_1), \qquad \hat{\mathbf{L}}_2 = \phi_{\text{low},2}(\text{Up}(\mathbf{L}_2)) \tag{1}$$

$$\hat{\mathbf{H}}_1 = \phi_{\text{high},1}(\mathbf{H}_1), \qquad \hat{\mathbf{H}}_2 = \phi_{\text{high},2}(\text{Up}(\mathbf{H}_2)) \tag{2}$$

where $\phi_{\cdot}$ denotes the 1×1 projection operator, and $\text{Up}(\cdot)$ represents bilinear upsampling to ensure spatial consistency if dimensions differ.

The projected features are then fused via element-wise addition:

$$\mathbf{F}_{\text{low}} = \hat{\mathbf{L}}_1 + \hat{\mathbf{L}}_2 \tag{3}$$

$$\mathbf{F}_{\text{high}} = \hat{\mathbf{H}}_1 + \hat{\mathbf{H}}_2 \tag{4}$$

These fused outputs are subsequently passed to the *Axial Spatial Attention* (ASA) and *Axial Channel Attention* (ACA) modules. Specifically, $\mathbf{F}_{\text{low}}$ is processed by ASA to capture long-range spatial dependencies in shallow features, while $\mathbf{F}_{\text{high}}$ is processed by ACA to enhance channel-level feature interactions in deeper semantic representations. The overall structure of AFM is illustrated in Fig. 6.

3.4.2. Axial Spatial Attention (ASA)

After the low-level features $\mathbf{F}_{\text{low}}$ are fused by the Adaptive Fusion Module (AFM), we refine them using the *Axial Spatial Attention* (ASA) mechanism. ASA enhances salient spatial structures by modeling long-range dependencies along the horizontal and vertical axes separately, allowing efficient context aggregation without the high computational cost of full 2D self-attention.

Given an input feature map $\mathbf{X} \in \mathbb{R}^{C \times H \times W}$, ASA computes lightweight spatial descriptors using average and max projections along each axis. The horizontal and vertical responses are then combined to form directional saliency cues, which are processed through a shallow convolutional block to generate a learned spatial attention mask $\mathbf{A}_{\text{spatial}}$. This mask highlights informative regions such as field boundaries and textured vegetation patterns.

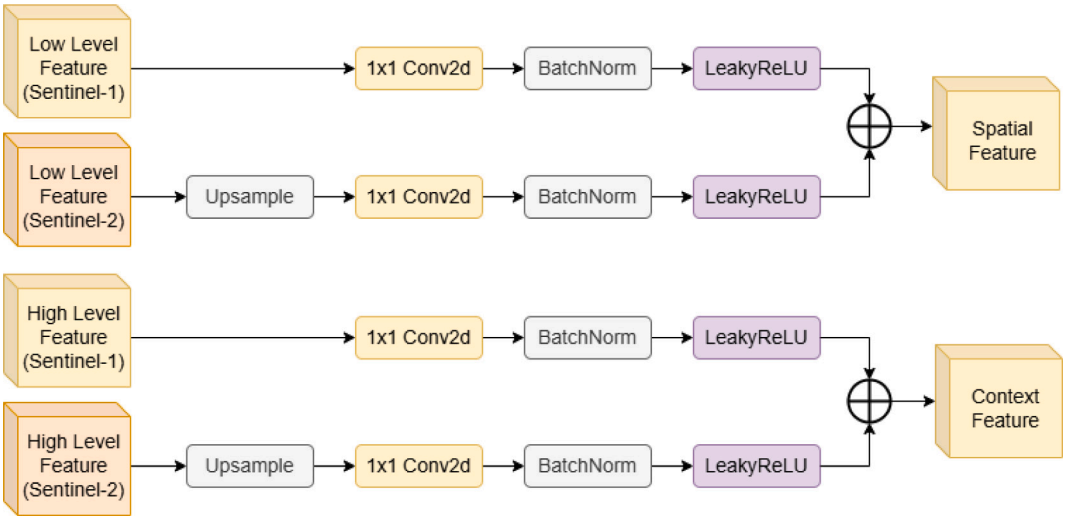

Fig. 6. Architecture of the Adaptive Fusion Module (AFM). Low-level and high-level features from both Sentinel-1 and Sentinel-2 encoders are projected into a shared space, adaptively fused.

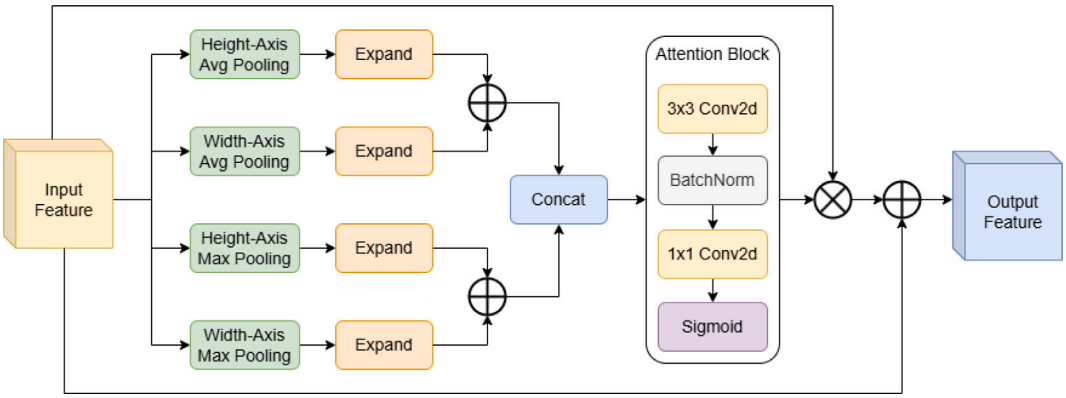

Fig. 7. Illustration of the Axial Spatial Attention (ASA) module. Spatial features are enhanced via directional mean and max pooling followed by a shared convolutional fusion to compute the spatial attention mask.

The refined output is produced in a residual manner:

$$\mathbf{X}' = \mathbf{X} \odot (1 + \mathbf{A}_{\text{spatial}}),$$

where $\odot$ denotes element-wise multiplication. The residual formulation preserves the original representation while selectively amplifying important spatial locations.

Compared to full self-attention, the axial formulation substantially reduces computational complexity while retaining strong sensitivity to elongated paddy field structures. A conceptual overview of the ASA module is illustrated in Fig. 7. Full mathematical expressions are provided in the Appendix A for reproducibility.

3.4.3. Axial Channel Attention (ACA)

To complement the spatial refinement performed by ASA, we apply the *Axial Channel Attention* (ACA) mechanism on the fused high-level feature maps $\mathbf{F}_{\text{high}}$. While ASA focuses on spatial saliency, ACA enhances discriminative channel responses by modeling long-range contextual cues along horizontal and vertical axes in a lightweight manner.

Given a feature map $\mathbf{X} \in \mathbb{R}^{C \times H \times W}$, ACA first aggregates information along each spatial axis using average and max pooling. These axial projections capture complementary channel statistics without resorting to full global pooling, thereby maintaining sensitivity to localized spectral variations commonly observed across paddy field conditions. The pooled descriptors are then processed by a shared two-layer feed-forward network to generate a compact channel representation, which is subsequently transformed into a channel attention mask $\mathbf{A}_{\text{channel}}$.

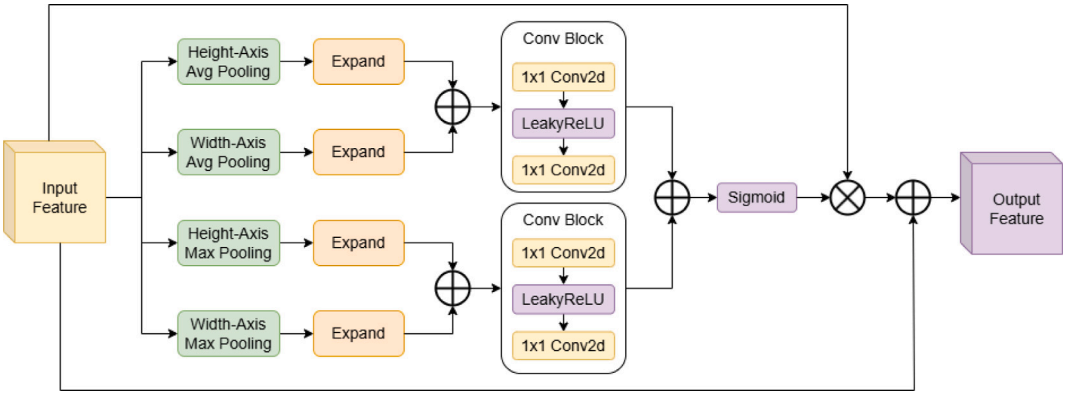

Fig. 8. Illustration of the Axial Channel Attention (ACA) module. ACA applies direction-aware pooling followed by a shared bottleneck MLP to generate channel-wise attention weights.

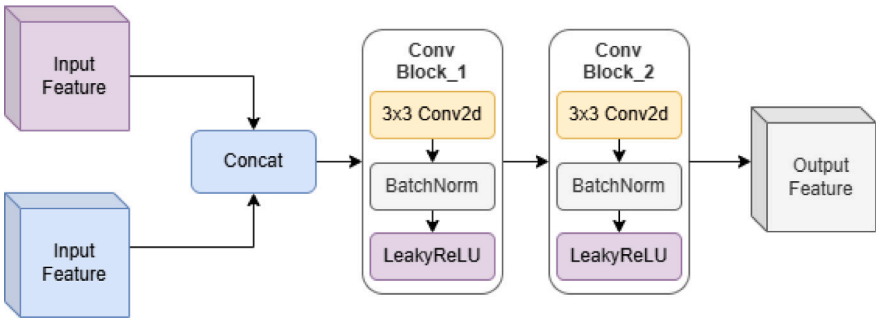

Fig. 9. Architecture of the Feature Fusion Module (FFM), which concatenates and refines spatially- and channel-attended feature maps to produce a unified semantic representation.

The refined output is computed using a residual formulation:

$$\mathbf{X}' = \mathbf{X} \odot (1 + \mathbf{A}_{\text{channel}}),$$

where $\odot$ denotes element-wise multiplication. This formulation preserves the original representation while selectively amplifying informative channels.

By combining directional pooling with lightweight channel recalibration, ACA strengthens the semantic discrimination of multimodal features with minimal computational overhead. A conceptual overview of ACA is illustrated in Fig. 8. Full mathematical detail is provided in the Appendix B for reproducibility.

3.4.4. *Feature Fusion Module (FFM)*

After processing the modality-specific features through ASA and ACA, the next step involves combining them to form a unified semantic representation. The *Feature Fusion Module (FFM)* is designed to integrate the refined low-level and high-level features, denoted respectively by $\mathbf{F}'_{\text{low}}$ and $\mathbf{F}'_{\text{high}}$, which have been spatially and channel-wise attended. To perform the fusion, the two feature maps are concatenated along the channel dimension and passed through a series of convolutional blocks:

$$\mathbf{F}_{\text{concat}} = \text{Concat}(\mathbf{F}'_{\text{low}}, \mathbf{F}'_{\text{high}}) \tag{5}$$

The concatenated features are then refined by two stacked convolutional layers with non-linear activation and dropout:

$$\mathbf{F}_{\text{fused}} = \delta\left(\text{BN}\left(\text{Conv}_1(\mathbf{F}_{\text{concat}})\right)\right) \tag{6}$$

$$\mathbf{F}_{\text{fused}} = \delta\left(\text{BN}\left(\text{Conv}_2\left(\text{Dropout}(\mathbf{F}_{\text{fused}})\right)\right)\right) \tag{7}$$

Here, $\delta(\cdot)$ represents the LeakyReLU activation function, Conv_i denotes a 3×3 convolution, BN refers to batch normalization, and dropout is used to mitigate overfitting with a dropout rate of $p = 0.1$.

This module ensures that both low-level spatial details and high-level semantic context are effectively integrated before final segmentation. The output $\mathbf{F}_{\text{fused}}$ is subsequently passed to the segmentation head for classification at the pixel level. A structural overview of this module is illustrated in Fig. 9.

Table 3
Summary of baseline models included in the comparative evaluation. The selected architectures span convolutional, Transformer-based, and hybrid designs, and comprise both general-purpose segmentation models and methods specifically developed for remote sensing imagery.

Model	Architecture type	Remote sensing specific
UNet	CNN	No
DeepLabv3+	CNN	No
HRNet	CNN	No
SegFormer	Transformer	No
A2FPN	CNN (FPN-based)	Yes
UNetFormer	Hybrid (CNN + Transformer)	Yes
CMLFormer	Hybrid (CNN + Transformer)	Yes
CMTFNet	Hybrid (CNN + Transformer)	Yes
DCSwin	Transformer (Swin-based)	Yes
TerraSegNet	CNN (Bilateral network)	Yes

3.5. Benchmark architectures

To comprehensively assess the performance of the proposed dual-stream segmentation model, we conducted comparative evaluations using a diverse set of state-of-the-art semantic segmentation architectures. These models span various design paradigms, including convolutional neural networks (CNNs), Transformer-based approaches, and hybrid architectures. For clarity, the benchmark models are categorized into two groups: (i) general-purpose semantic segmentation models widely adopted in computer vision, and (ii) models specifically developed for remote sensing imagery (see Table 3).

The first group comprises UNet (Ronneberger et al., 2015), DeepLabv3+ (Chen et al., 2018), HRNet (Wang et al., 2021), and SegFormer (Xie et al., 2021), all of which have demonstrated strong performance across a range of semantic segmentation tasks. The second group includes A2FPN (Li et al., 2022b), UNetFormer (Wang et al., 2022b), CMLFormer (Wu et al., 2024), CMTFNet (Wu et al., 2023), DCSwin (Wang et al., 2022a), and TerraSegNet (Wijaya et al., 2025)—architectures that are explicitly designed to address the spectral heterogeneity and spatial complexity of remote sensing data.

3.6. Evaluation metrics

To quantitatively evaluate model performance in segmenting paddy fields from multimodal satellite imagery, we employed three standard metrics widely used in semantic segmentation: pixel accuracy (PA), mean intersection over union (mIoU), and F1-score. These metrics capture complementary aspects of model performance and are particularly relevant in agricultural contexts, where class imbalance and boundary delineation are critical.

Pixel Accuracy (PA) measures the proportion of correctly classified pixels relative to the total number of pixels and is defined as:

$$\mathbf{PA} = \frac{TP}{TP + TN + FP + FN} \tag{8}$$

where TP, TN, FP, and FN denote true positives, true negatives, false positives, and false negatives, respectively. While PA offers a general indication of classification correctness, it may be biased in class-imbalanced datasets, such as those dominated by non-paddy (background) pixels.

Mean Intersection over Union (mIoU) quantifies the average overlap between predicted and reference segments. For a binary classification task, IoU is calculated as:

$$\mathbf{IoU} = \frac{TP}{TP + FP + FN} \tag{9}$$

Given the binary nature of the task, mIoU simplifies to a single class-wise value. It is regarded as a robust and interpretable metric for semantic segmentation, penalizing both over- and under-segmentation errors.

F1-score, the harmonic mean of precision and recall, is especially suitable for evaluating performance in imbalanced settings. It is defined as:

$$\mathbf{F1-score} = \frac{2 \times \text{Precision} \times \text{Recall}}{\text{Precision} + \text{Recall}} = \frac{2 \times TP}{2 \times TP + FP + FN} \tag{10}$$

This metric is particularly important in paddy field mapping, where omission errors (e.g., undetected fields) and commission errors (e.g., confusion with wetlands or vegetation) can significantly impact downstream applications. In this study, we report the binary F1-score for the paddy field class as the primary performance indicator.

Table 4
Training hyperparameters used for all models in this study. These settings ensure full reproducibility across baselines and the proposed DSSNet.

Category	Hyperparameter setting
Hardware	NVIDIA RTX 4060 (8 GB VRAM)
Framework	PyTorch v2.3.1 + CUDA 12.1
Precision	Mixed precision (torch.cuda.amp)
Batch size	8
Gradient accumulation	4 steps (effective batch size = 32)
Patch size	512 × 512 pixels
Input channels	6 (VV, VH, R, G, B, NIR)
Optimizer	AdamW
Learning Rate (LR)	1×10^{-3}
Betas (β_1, β_2)	(0.9, 0.999)
Weight decay	1×10^{-4}
Epochs	20
LR scheduler	Cosine Annealing with Warm Restarts
Warm restart period T_0	10 epochs
Restart multiplier T_{mult}	1
Minimum LR	1×10^{-6}
Loss function	Tversky Loss
Tversky parameters (α, β)	(0.6, 0.4)
Data augmentation	Gaussian blur, speckle noise, rotation
Normalization	Per-band min–max scaling
Weight initialization	He normal init for modified layers
Random seeds	42

3.7. Training strategy

All training and evaluation procedures were implemented in `PyTorch` (v2.3.1 + CUDA 12.1) and executed on a single NVIDIA RTX 4060 GPU with 8 GB VRAM. Mixed-precision training (`torch.cuda.amp`) was used to reduce memory consumption and accelerate computation. Due to hardware constraints, gradient accumulation over four steps was applied, giving an effective batch size of 32 while maintaining a physical mini-batch size of 8.

All models, including the baselines and the proposed DSSNet, were trained under identical optimization settings to ensure fair comparison. We employed the AdamW optimizer (Loshchilov and Hutter, 2019) with a cosine annealing learning rate schedule with warm restarts (Loshchilov and Hutter, 2017). To address class imbalance, we adopted the Tversky Loss function (Salehi et al., 2017), which allows flexible weighting of false positives and false negatives, making it particularly effective for detecting small or fragmented paddy fields. Early stopping based on validation F1-score was used to mitigate overfitting, and the final checkpoint for each model corresponded to the highest validation F1-score.

To fully support multimodal inputs, the first convolutional layer of each baseline model was reconfigured to accept six input channels (VV, VH, R, G, B, NIR). These layers were reinitialized using He normal initialization, while all other pretrained weights were kept intact at the start of training and updated normally during optimization. A complete summary of all hyperparameter settings is provided in Table 4 to support reproducibility.

4. Experimental results

4.1. Quantitative results

The comparative results offer critical insights into the trade-offs among segmentation accuracy, model complexity, and computational efficiency (see Fig. 10). As summarized in Table 5, the proposed DSSNet outperforms all benchmark models across the three evaluation metrics. It achieved an F1-score of 0.8982, pixel accuracy of 0.8998, and mean Intersection over Union (mIoU) of 0.8156, indicating superior segmentation capability in delineating paddy field boundaries from multimodal satellite imagery.

Notably, TerraSegNet ranked second in both F1-score (0.8933) and mIoU (0.8076); however, it incurs a higher model complexity (5.80 million parameters) and computational load (22.34 GFLOPs) compared to DSSNet (2.79 million parameters, 11.18 GFLOPs). This comparison underscores the ability of DSSNet to achieve high accuracy with substantially fewer parameters and reduced computational cost.

DCSwin, despite being one of the most computationally intensive models (45.33 million parameters, 47.69 GFLOPs), exhibited the lowest performance among the evaluated models (F1-score: 0.8479, mIoU: 0.7365). This result suggests that increasing model size does not necessarily yield better performance, particularly in applications with strong spatial heterogeneity and modality-specific characteristics, such as paddy field mapping.

The overall trend highlights the importance of architectural efficiency in real-world remote sensing applications. While larger models may offer theoretical improvements in feature representation, they often suffer from diminishing returns in performance

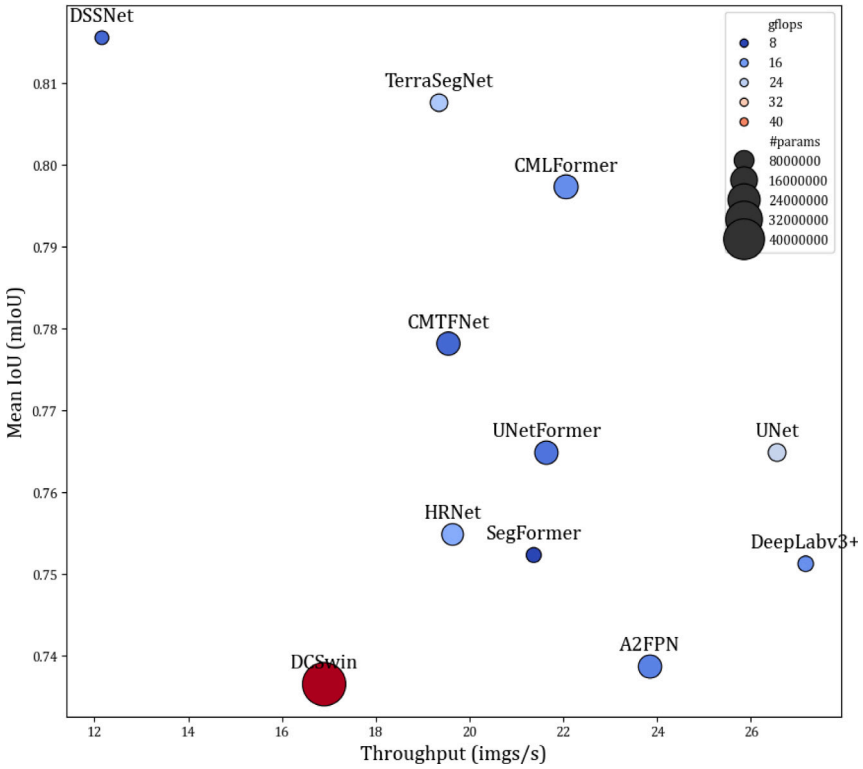

Fig. 10. Trade-off between mean Intersection over Union (mIoU) and throughput (images per second) across benchmark models.

Table 5
Quantitative comparison of DSSNet against ten state-of-the-art semantic segmentation models. The table reports accuracy metrics (F1-score, pixel accuracy, mIoU) and computational cost (number of parameters, GFLOPs, and inference throughput in images per second) under a unified multimodal training setup using paired Sentinel-1 and Sentinel-2 inputs.

Model	Backbone	Number of parameters	F1-score	Pixel accuracy	mIoU	GFLOPs	Throughput (img/s)
UNet	ResNet-18	6.00 M	0.8667	0.8674	0.7649	25.12	26.56
DeepLabv3+	ResNet-18	4.18 M	0.8576	0.8593	0.7513	14.98	27.17
HRNet	–	9.90 M	0.8600	0.8612	0.7549	18.01	19.64
SegFormer	MiT-B0	3.72 M	0.8583	0.8591	0.7523	7.92	21.37
A2FPN	ResNet-18	11.65 M	0.8492	0.8521	0.7387	13.82	23.85
UNetFormer	ResNet-18	11.73 M	0.8665	0.8679	0.7649	12.35	21.64
CMLFormer	ResNet-18	12.54 M	0.8870	0.8880	0.7974	14.90	22.06
CMTFNet	ResNet-18	11.62 M	0.8750	0.8764	0.7782	11.26	19.55
DCSwin	Swin-T	45.33 M	0.8479	0.8500	0.7365	47.69	16.90
TerraSegNet	EfficientNetV2-S	5.80 M	0.8933	0.8946	0.8076	22.34	19.35
DSSNet	EfficientNet-B0 & MaxVit-T	2.79 M	**0.8982**	**0.8998**	**0.8156**	11.18	11.89

relative to their computational cost. In contrast, DSSNet demonstrates a more favorable trade-off, achieving top-tier segmentation accuracy while maintaining a compact structure and low inference time. These characteristics position DSSNet as a practical solution for operational paddy field mapping, especially in tropical regions where rapid updates and limited computational resources are common constraints.

4.2. Qualitative results

Fig. 11 presents a qualitative comparison of the ten semantic segmentation models across five representative paddy field scenes captured on different dates. Each scene consists of Sentinel-1 imagery (top row) and Sentinel-2 imagery (second row), followed by corresponding ground truth, and predicted segmentation masks from UNet, DeepLabv3+, HRNet, SegFormer, A2FPN, UNetFormer, CMLFormer, CMTFNet, DCSwin, TerraSegNet, and the proposed DSSNet.

Several noteworthy observations emerge from the visual comparison:

Table 6
Ablation study results on DSSNet with different lightweight backbones and modules. All configurations were trained under identical conditions on paired Sentinel-1 and Sentinel-2 imagery.

Backbone-1	Backbone-2	ASA	ACA	Number of parameters	F1-score	Pixel accuracy	mIoU	GFLOPs	Throughput (img/s)
EfficientNet-B0	MaxVit-T	–	–	2.78 M	0.8747	0.8805	0.7778	11.17	12.08
EfficientNet-B0	MaxVit-T	✓	–	2.78 M	0.8964	0.8972	0.8126	11.17	11.95
EfficientNet-B0	MaxVit-T	–	✓	2.79 M	0.8967	0.8994	0.8131	11.18	11.77
EfficientNet-B0	MaxVit-T	✓	✓	2.79 M	0.8982	0.8998	0.8156	11.18	11.89
Swin-T	ResNet-18	✓	✓	15.75 M	0.8660	0.8669	0.7640	38.73	11.44
MaxVit-T	ConvNeXt-Tiny	✓	✓	14.54 M	0.8548	0.8558	0.7469	26.01	11.26
MaxVit-T	MaxVit-T	✓	✓	3.62 M	0.8838	0.8856	0.7925	13.72	11.08
RegNet-Y400MF	RegNet-Y400MF	✓	✓	2.93 M	0.8886	0.8892	0.8001	22.74	13.01
EfficientNet-B0	EfficientNet-B0	✓	✓	2.07 M	0.8962	0.8967	0.8124	8.60	13.38
MobileNetV3-Large	MobileNetV3-Large	✓	✓	1.93 M	0.8948	0.8952	0.8100	7.37	14.48
ShuffleNetV2-x0.5	ShuffleNetV2-x0.5	✓	✓	1.37 M	0.8925	0.8933	0.8064	18.59	14.71

- **Scene consistency:** DSSNet consistently generates segmentation outputs that closely match the ground truth across diverse scenes. Its predictions exhibit reduced false positives and false negatives, particularly along paddy field boundaries, where delineation is often challenging due to fragmented plots and mixed land use.
- **Temporal generalization:** DSSNet demonstrates robust performance across different acquisition dates, maintaining high-quality segmentation throughout varying seasonal conditions. In contrast, models such as HRNet, A2FPN, and DCSwin show reduced accuracy during transitional or atypical phenological stages, often leading to misclassification of active paddy areas as non-paddy land.
- **Challenging scenarios:** In more complex scenes — such as those featuring coastal or wetland environments (e.g., Scene#22) — all models exhibit varying degrees of segmentation degradation. However, DSSNet shows superior resilience to such errors, likely due to its modality-specific encoders and fusion mechanism, which better capture subtle textural and spectral cues.

These qualitative results reinforce the findings of the quantitative evaluation and highlight the importance of tailored multimodal fusion strategies for accurate and reliable paddy field segmentation.

4.3. Ablation study

4.3.1. Impact of backbone combination and axial attention modules

To assess the contribution of architectural components to overall segmentation performance, we conducted an ablation study focusing on different combinations of encoder backbones and attention modules. The results are summarized in Table 6. We first evaluated the impact of incorporating the ASA and ACA modules into the dual-encoder design using EfficientNet-B0 and MaxVit-T. The inclusion of both ASA and ACA significantly improved performance, increasing the F1-score from 0.8747 to 0.8982 and the mIoU from 0.7778 to 0.8156, without introducing additional model complexity (remaining at 2.79 million parameters). These improvements highlight the effectiveness of axial attention in enhancing feature representation while maintaining a compact model footprint.

Next, we compared DSSNet to configurations using heavier backbone combinations, such as Swin-T + ResNet-18 (15.75M parameters) and MaxVit-T + ConvNeXt-Tiny (Liu et al., 2022) (14.54M parameters). Despite their larger sizes and higher computational demands (up to 38.73 GFLOPs), these configurations yielded lower segmentation accuracy, confirming that increased complexity does not necessarily guarantee better performance in multimodal remote sensing tasks.

Furthermore, DSSNet outperformed other lightweight alternatives, including MobileNetV3-Large, ShuffleNetV2-X0.5, and RegNet-Y400MF, across all evaluated metrics. An interesting observation from this study is that the dual EfficientNet-B0 configuration performs nearly as well as the proposed EfficientNet-B0 + MaxVit-T combination. This indicates that a purely CNN-based dual encoder is also highly effective for multimodal fusion, particularly when parameter efficiency is the primary constraint. Nonetheless, the CNN–Transformer pairing in DSSNet was deliberately chosen to leverage complementary inductive biases: EfficientNet-B0 captures fine-grained local textures and structural patterns characteristic of SAR signals, while MaxVit-T provides stronger global context modeling and long-range relational reasoning for multispectral reflectance. This diversity in architectural behavior leads to more balanced cross-modal representations and slightly improved boundary fidelity, especially in heterogeneous and fragmented paddy field areas. As further illustrated in Fig. 12, the ASA and ACA modules refine these complementary representations by enhancing spatial coherence and selectively amplifying semantically informative channels, providing qualitative support for the quantitative gains observed in Table 6.

The results in Table 7 highlight a clear advantage of the proposed late-fusion dual-encoder design over early-fusion baselines. Early-fusion models using either EfficientNet-B0 or MaxVit-T achieve competitive accuracy with low parameter counts and high throughput; however, their F1-score and mIoU remain consistently below those of the late-fusion configuration. By processing Sentinel-1 and Sentinel-2 inputs in modality-specific streams before fusion, the late-fusion architecture achieves the best overall performance, indicating more effective extraction of complementary radar–optical features. These findings confirm that modality

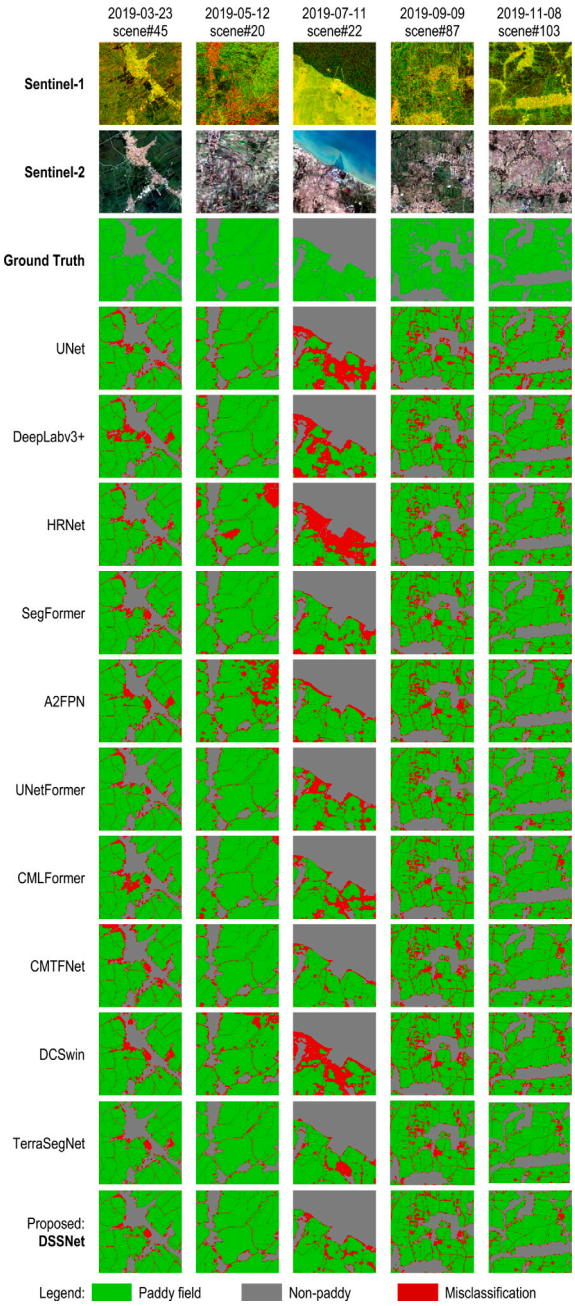

Fig. 11. Qualitative comparison of paddy field segmentation results across multiple models. From top to bottom: Sentinel-1 image, Sentinel-2 image, Ground Truth, predictions from baseline models, and the proposed DSSNet model. Red pixels in the misclassification map indicate disagreement between prediction and reference labels, including non-paddy areas incorrectly classified as paddy (false positives) and paddy areas incorrectly classified as non-paddy (false negatives). (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

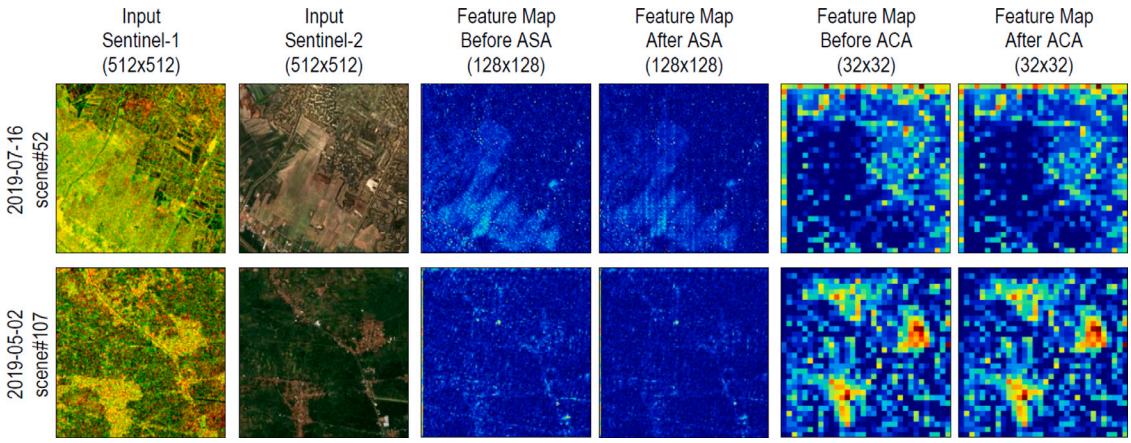

Fig. 12. Example visualization of DSSNet's multimodal feature refinement pipeline. From left to right: Sentinel-1 input patch (512 × 512), Sentinel-2 input patch (512 × 512), feature maps before and after Axial Spatial Attention (ASA) at 128 × 128 resolution, and feature maps before and after Axial Channel Attention (ACA) at 32 × 32 resolution. ASA enhances spatial coherence while ACA selectively amplifies semantically informative channels.

Table 7
Ablation study of fusion strategies in DSSNet, comparing early-fusion (channel concatenation) and late-fusion (dual-encoder) designs under identical training settings. Results highlight the quantitative benefit of modality-specific encoders, where late-fusion achieves higher segmentation accuracy (F1-score and mIoU) despite slightly higher computational cost.

Fusion method	Backbone	ASA	ACA	Number of parameters	F1-score	Pixel accuracy	mIoU	GFLOPs	Throughput (img/s)
Early-fusion	EfficientNet-B0	✓	✓	1.22 M	0.8831	0.8850	0.7911	7.33	27.62
Early-fusion	MaxVit-T	✓	✓	2.17 M	0.8755	0.8839	0.7791	11.30	24.80
Late-fusion	EfficientNet-B0 & MaxVit-T	✓	✓	2.79 M	0.8982	0.8998	0.8156	11.18	11.89

decoupling in early layers is crucial for multimodal representation learning, and that the dual-encoder fusion adopted in DSSNet provides a superior accuracy–complexity trade-off.

These findings underscore the importance of strategic backbone selection and the integration of efficient attention mechanisms to enable high-performing yet computationally tractable models suitable for operational-scale paddy field mapping in resource-limited settings.

4.3.2. Effect of multisensor input feature composition

To further evaluate the influence of input feature selection on segmentation performance, we conducted an ablation study involving five combinations of radar and optical inputs, comprising both raw bands and derived vegetation indices. The results are summarized in Table 8.

The best performance was achieved using only the core dual-polarization backscatter channels from Sentinel-1 (VV and VH) and the four primary reflectance bands from Sentinel-2 (R, G, B, and NIR), yielding an F1-score of 0.8982 and a mean IoU of 0.8156. This configuration, which leverages the full spatial and spectral richness of raw measurements, proved sufficient for accurate paddy field segmentation without relying on engineered indices.

In contrast, when using only vegetation indices—such as Radar Vegetation Index (RVI) and Radar NDVI (RNDVI) for Sentinel-1, and Normalized Difference Vegetation Index (NDVI), Green Normalized Difference Vegetation Index (GNDVI), and Soil Adjusted Vegetation Index (SAVI) for Sentinel-2, a substantial performance degradation was observed (F1-score: 0.8537, mIoU: 0.7450). These indices, while useful for indicating vegetation vigor, often abstract away critical spatial details necessary for precise field delineation, particularly during early phenological stages.

Adding vegetation indices to the base raw features offered marginal improvements in certain configurations (e.g., mIoU = 0.8017), yet none outperformed the simpler raw-only setting. Interestingly, including an excessive number of features led to a slight drop in accuracy, suggesting that feature redundancy may introduce noise and hinder discriminative learning unless properly regularized.

A closer examination of the results reveals why the inclusion of NDVI, GNDVI, and SAVI can unexpectedly reduce performance. These indices are algebraic transformations of the Red, Green, and NIR bands that were already present in the raw feature stack. When added to the multimodal input space, they introduce highly correlated and partially redundant features, which amplify greenness-related spectral characteristics while providing limited new information beyond what the raw bands already encode. In multimodal fusion settings — particularly those involving radar backscatter — the vegetation indices can also become unstable

Table 8
Ablation study on input feature combinations from Sentinel-1 and Sentinel-2 for paddy field mapping in DSSNet. The baseline configuration uses raw spectral and backscatter bands.

Sentinel-1 features	Sentinel-2 features	F1-score	Pixel accuracy	mIoU
VV, VH	R-G-B, NIR	0.8982	0.8998	0.8156
RVI, RNDVI	NDVI, GNDVI, SAVI	0.8537	0.8621	0.7450
VV, VH, RVI, RNDVI	R-G-B, NIR	0.8949	0.8958	0.8101
VV, VH	R-G-B, NIR, NDVI, GNDVI, SAVI	0.8732	0.8739	0.7755
VV, VH, RVI, RNDVI	R-G-B, NIR, NDVI, GNDVI, SAVI	0.8897	0.8947	0.8017

Table 9
Monthly inference performance of DSSNet on Indramayu paddy fields during 2019.

Month	F1-score	Pixel accuracy	mIoU
Jan	0.8861	0.8882	0.7958
Feb	0.9061	0.9075	0.8285
Mar	0.9010	0.9024	0.8202
Apr	0.8921	0.8938	0.8057
May	0.8902	0.8920	0.8025
Jun	0.8997	0.9014	0.8180
Jul	0.9136	0.9149	0.8413
Aug	0.9000	0.9013	0.8185
Sep	0.8984	0.8999	0.8160
Oct	0.8970	0.8985	0.8136
Nov	0.9000	0.9016	0.8186
Dec	0.8883	0.8900	0.7995

in cloudy or heterogeneous illumination conditions, thereby injecting additional spectral noise. The dual-encoder architecture in DSSNet is optimized to learn complementary information from Sentinel-1 and Sentinel-2; thus, redundant or noisy indices may dilute the inter-modal contrast, reduce feature distinctiveness, and ultimately hinder segmentation accuracy. This phenomenon explains why extending the feature set from VV, VH, R-G-B, NIR to a larger combination including NDVI, GNDVI, and SAVI led to a decrease in F1-score (from 0.8982 to 0.8732) despite the increased dimensionality.

Overall, these findings underscore the importance of carefully curated feature sets in multimodal segmentation tasks. Raw Sentinel-1 and Sentinel-2 bands, when temporally aligned and properly fused, provide a robust and information-rich foundation for crop mapping. While vegetation indices may offer complementary insights, their use should be judicious to avoid compromising spatial fidelity and model efficiency.

5. Discussion

Rather than explicitly modeling temporal dynamics using sequence-based architectures, DSSNet leverages spatially diverse satellite observations acquired on different dates to implicitly capture phenological variability and surface condition changes throughout the rice-growing season (Aashi and Vema, 2025). This strategy allows the model to maintain a low computational cost while achieving high segmentation accuracy, making it well suited for regional-scale and real-time agricultural monitoring applications.

Fig. 13 and Table 9 present the month-wise performance of DSSNet across the 2019 calendar year. The model demonstrates remarkable stability, with F1-scores ranging between 0.8861 (January) and 0.9136 (July), and mIoU ranging between 0.7958 and 0.8413. Performance peaks in July, which corresponds to the peak vegetative stage of the first rice season, when crop canopies are well developed and exhibit strong backscatter and spectral contrast. February also shows above-average performance, likely due to the early vegetative stages of the second crop cycle, where field conditions such as inundation and newly transplanted seedlings produce distinct radar and optical signals.

The narrow variation in monthly metrics (±0.0082 in F1-score and ±0.0136 in mIoU) highlights the model's ability to generalize across phenological stages, surface moisture levels, and atmospheric interference. These findings further validate the robustness of DSSNet and the effectiveness of its dual-stream fusion and axial attention mechanisms in dealing with the dynamic, heterogeneous nature of paddy fields in tropical regions.

5.1. Limitations and design trade-offs

- **Inference efficiency:** Despite having a relatively low parameter count (2.79M) and moderate computational complexity (11.18 GFLOPs), DSSNet exhibits the lowest inference throughput (11.89 images per second) among all compared models. This reduced throughput stems from the dual-encoder structure, where concurrent feature extraction from both radar and optical inputs introduces additional computational overhead. While suitable for offline or near-real-time batch processing, deployment in latency-sensitive or edge environments may require lightweight acceleration techniques. Potential solutions include knowledge distillation to compress DSSNet into a single-stream student model, or structured pruning to reduce redundant encoder pathways.

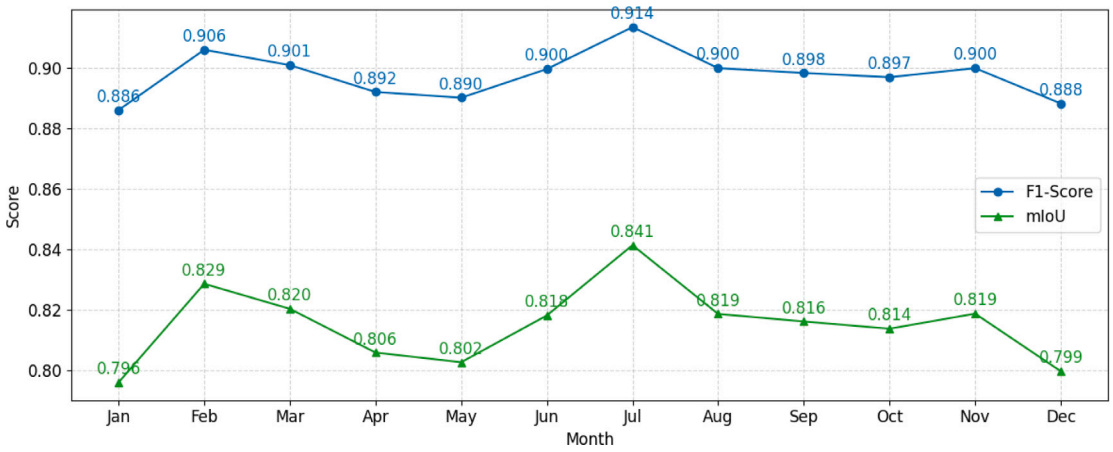

Fig. 13. Monthly variation in model performance across the 2019 rice growing season. The line chart illustrates the F1-score and mIoU achieved by DSSNet on each monthly subset, highlighting seasonal trends in segmentation accuracy and intersection-over-union. Performance generally peaks during mid-season (July), coinciding with critical phenological stages of rice crops.

- **Task generalization:** DSSNet is specialized for binary paddy field segmentation using fused radar and optical inputs, and its performance on broader remote sensing tasks — such as multiclass land-cover mapping or urban object segmentation — remains unverified. Extending the model to these domains would require careful architectural adaptation, particularly for handling complex label hierarchies and class imbalance. Strategies such as multi-task learning, or domain generalization frameworks could facilitate broader applicability.
- **Modality alignment constraints:** DSSNet relies on accurate spatial and temporal co-registration between Sentinel-1 and Sentinel-2 inputs. While this study ensured alignment using a ±3-day acquisition window and resampling to a unified 10 m grid, operational deployments may face misalignments caused by acquisition delays, geolocation inaccuracies, or heterogeneous sensor geometries. Several mitigation strategies are available. At the model level, feature-level alignment modules or domain-adaptation techniques could increase robustness to imperfect co-registration. At the data level, several advances in cross-sensor geometric correction provide practical strategies for reducing geometric bias before mapping. These include progressive spatial smoothing with graph-optimality-aware stochastic optimization for LiDAR bundle adjustment (Li et al., 2025) and coarse-to-fine non-rigid registration of laser scans with oriented imagery (Li et al., 2019). Although originally designed for LiDAR–image integration, the underlying principles — such as iterative refinement, multi-scale smoothing, and discrepancy minimization — remain broadly applicable. These techniques can be adapted to validate or improve Sentinel-1/Sentinel-2 alignment, enhance reference map quality, or support preprocessing when high-resolution auxiliary datasets (e.g., UAV LiDAR) are available.

Beyond these limitations, a key avenue for improving DSSNet lies in incorporating temporal information. Rice phenology follows a highly structured time sequence, and single-date segmentation cannot fully capture temporal dynamics such as transplanting, vegetative growth, and harvest transitions. Future extensions could integrate temporal transformers, sequence models, or multi-date fusion modules to model temporal coherence across consecutive Sentinel-1/Sentinel-2 acquisitions. Such temporal integration has the potential to further enhance segmentation stability, reduce confusion between phenological stages, and support continuous monitoring throughout the growing season.

5.2. Practical implications

The proposed DSSNet framework demonstrates strong potential for enhancing operational paddy field mapping in tropical agricultural regions through the integration of Sentinel-1 SAR and Sentinel-2 optical imagery. By leveraging the complementary properties of these sensors, DSSNet provides consistent and temporally robust segmentation performance, even under dynamic cropping conditions. This capability is particularly valuable in monsoon-affected environments, where reliance on optical data alone often results in temporal gaps and classification uncertainty.

In the context of national agricultural monitoring systems, DSSNet enables timely and accurate delineation of paddy fields at a 10-meter spatial resolution, which is sufficient for most smallholder agricultural landscapes in Southeast Asia. For government institutions such as Statistics Indonesia (BPS) that are responsible for compiling official agricultural production figures, it may reduce its dependence on labor-intensive manual surveys by incorporating satellite-derived field boundaries as input for yield estimation models.

Owing to its lightweight design and modular architecture, DSSNet is suitable for deployment in various computing environments, including cloud-based geospatial platforms and local inference engines. This flexibility facilitates integration into both centralized national-level monitoring systems and decentralized decision support tools at the provincial or regency scale.

Table 10
Stability analysis of DSSNet across four independent training runs with random seeds 42, 1465, 2544, and 3692. Results are reported as mean ± standard deviation.

Random seed	F1-score	Pixel accuracy	mIoU
42	0.8982	0.8998	0.8156
1465	0.8900	0.8932	0.8022
2544	0.8936	0.8959	0.8081
3692	0.8940	0.8954	0.8087
DSSNet	**0.8940 ± 0.0034**	**0.8961 ± 0.0027**	**0.8086 ± 0.0055**

Beyond paddy field mapping, the architectural characteristics of DSSNet also support adaptability to other crop types and agro-ecological settings. The dual-stream fusion of SAR backscatter and multispectral reflectance captures biophysical signals that are shared across crops including maize, soybean, sugarcane, and wheat. The axial refinement mechanism further enhances spatial and directional coherence, enabling the model to capture elongated planting patterns and heterogeneous canopy textures found in diverse agricultural systems. While crop-specific fine-tuning may be required to account for phenological differences, the underlying fusion and attention modules remain broadly applicable, suggesting that DSSNet can be transferred to multi-crop monitoring or extended to regions with different environmental and management conditions.

In addition to accuracy and multimodal robustness, the computational efficiency of DSSNet further supports its operational applicability. With an inference throughput of 11.89 patches per second (512 × 512 pixels), the model achieves an effective processing rate of approximately 3.12 megapixels per second. This corresponds to a processing time of about 38.7 s for a full 10-meter Sentinel-2 tile (120.56 MP) and roughly 3.2 milliseconds per km^2 of agricultural area. These tile-level and area-normalized metrics provide a more meaningful operational perspective than per-patch throughput alone, demonstrating that DSSNet is suitable for near-real-time mapping workflows in national monitoring pipelines and large-area agricultural assessments.

Although DSSNet is designed as a lightweight model with a substantially reduced parameter count (2.79M), its inference throughput is lower than that of other architectures evaluated in this study. This behavior arises from the dual-encoder design, where Sentinel-1 and Sentinel-2 inputs are processed separately before fusion. While this architecture improves multimodal feature extraction, it also introduces additional latency compared to single-stream networks. Therefore, the term lightweight in DSSNet primarily reflects its low memory footprint and parameter efficiency rather than peak inference speed. This distinction is important for deployment scenarios where GPU memory, energy budget, or model size is the dominant constraint — such as cloud-hosted batch processing or resource-limited local inference — rather than strict frame-rate requirements. In such contexts, DSSNet offers a favorable trade-off between accuracy, model compactness, and operational feasibility.

5.3. Stability analysis

Across the four independent training runs, DSSNet demonstrates strong stability in segmentation performance (see Table 10). The F1-score exhibits a mean of 0.8940 with a standard deviation of only 0.0034, indicating minimal sensitivity to random initialization. Pixel Accuracy follows a similar trend, with a mean of 0.8961 and a deviation of 0.0027, reflecting consistent global classification performance across runs. The mIoU shows slightly higher variability (0.8086 ± 0.0055), which is expected due to the metric's sensitivity to boundary alignment in narrow and fragmented paddy fields.

Overall, the low variance across all three evaluation metrics suggests that DSSNet maintains stable convergence behavior despite differences in seed initialization and stochastic training conditions. These results confirm that the dual-encoder design and axial attention refinement do not introduce instability in optimization, and that the model's performance is robust and reproducible across multiple training executions.

6. Conclusions

We proposed DSSNet, a lightweight dual-encoder network designed for multimodal fusion of Sentinel-1 SAR and Sentinel-2 multispectral imagery for paddy field segmentation. By combining EfficientNet-B0 and MaxVit-T as modality-specific encoders, and integrating Axial Spatial and Channel Attention modules, DSSNet effectively captures both spatial texture and spectral context with minimal computational cost. Comprehensive benchmarking against ten state-of-the-art segmentation models demonstrates the superior performance of DSSNet, achieving an F1-score of 0.8982, pixel accuracy of 0.8998, and mIoU of 0.8156, with only 2.79 million parameters and 11.18 GFLOPs.

In summary, DSSNet provides an efficient and scalable solution for paddy field mapping in tropical environments, where persistent cloud cover and crop heterogeneity challenge traditional approaches. While not intended for strict on-board real-time execution, its lightweight design and high throughput make it well-suited for near-real-time or large-area operational processing, including cloud-based geospatial platforms and resource-limited deployment settings. Future work will explore extending DSSNet to other crop types, integrating continuous temporal sequences, and scaling the framework for national-level agricultural monitoring.

CRediT authorship contribution statement

Bagus Setyawan Wijaya: Writing – original draft, Visualization, Software, Resources, Methodology, Data curation, Conceptualization. **Rinaldi Munir:** Writing – review & editing, Validation, Supervision, Methodology. **Nugraha Priya Utama:** Writing – review & editing, Validation, Supervision, Methodology.

Funding

This work was supported by Bandung Institute of Technology under Grant DRI.PN-6-175-2025.

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.

Appendix A. Mathematical formulation of Axial Spatial Attention (ASA)

This appendix details the mathematical operations underlying the proposed Axial Spatial Attention (ASA) module, as implemented in our framework.

Let the fused low-level feature map be denoted as

$$\mathbf{X} \in \mathbb{R}^{B\times C\times H\times W},$$

where B, C, H, and W represent the batch size, number of channels, height, and width, respectively.

A.1. Axial mean projections

ASA first computes lightweight axial mean descriptors by aggregating features along channel and one spatial dimension. Specifically, the horizontal and vertical mean responses are defined as

$$\mathbf{M}_{1\times W}(b,h,w) = \frac{1}{CH}\sum_{c=1}^{C}\sum_{h'=1}^{H}\mathbf{X}(b,c,h',w),$$

$$\mathbf{M}_{H\times 1}(b,h,w) = \frac{1}{CW}\sum_{c=1}^{C}\sum_{w'=1}^{W}\mathbf{X}(b,c,h,w').$$

Both projections are broadcast to $\mathbb{R}^{B\times 1\times H\times W}$ and combined to form the axial mean response:

$$\mathbf{M}_{\text{avg}} = \mathbf{M}_{1\times W} + \mathbf{M}_{H\times 1}.$$

A.2. Axial max projections

In parallel, ASA extracts complementary high-activation cues using axial max pooling. The horizontal and vertical max responses are computed as

$$\mathbf{P}_{1\times W}(b,h,w) = \max_{c,h'}\mathbf{X}(b,c,h',w),$$

$$\mathbf{P}_{H\times 1}(b,h,w) = \max_{c,w'}\mathbf{X}(b,c,h,w').$$

After broadcasting, these responses are summed to obtain the axial max descriptor:

$$\mathbf{M}_{\max} = \mathbf{P}_{1\times W} + \mathbf{P}_{H\times 1}.$$

A.3. Spatial attention mask generation

The average and max axial descriptors are concatenated along the channel dimension:

$$\mathbf{F}_{\text{axial}} = \left[\mathbf{M}_{\text{avg}}\,;\,\mathbf{M}_{\max}\right] \in \mathbb{R}^{B\times 2\times H\times W}.$$

This fused representation is passed through a shallow convolutional fusion block consisting of a 3×3 convolution, batch normalization, a 1×1 pointwise convolution, and a sigmoid activation:

$$\mathbf{A}_{\text{spatial}} = \sigma\left(\text{Conv}_{1\times 1}\left(\text{BN}\left(\text{Conv}_{3\times 3}(\mathbf{F}_{\text{axial}})\right)\right)\right),$$

where $\mathbf{A}_{\text{spatial}} \in \mathbb{R}^{B\times 1\times H\times W}$ denotes the learned spatial attention mask.

A.4. Residual spatial refinement

Finally, the attention mask is applied to the input feature map in a residual scaling manner:

$$\mathbf{X}' = \mathbf{X} \odot \left(1 + \mathbf{A}_{\text{spatial}}\right),$$

where $\odot$ denotes element-wise multiplication. This formulation preserves the original feature representation while selectively amplifying spatially informative regions.

Overall, ASA efficiently captures long-range directional dependencies along horizontal and vertical axes with linear complexity in spatial dimensions, making it particularly well suited for elongated and structured patterns commonly observed in agricultural landscapes.

Appendix B. Mathematical formulation of Axial Channel Attention (ACA)

This appendix provides a detailed mathematical description of the proposed Axial Channel Attention (ACA) module.

Let the input high-level fused feature map be denoted as

$$\mathbf{X} \in \mathbb{R}^{B\times C\times H\times W},$$

where B, C, H, and W represent the batch size, number of channels, height, and width, respectively.

B.1. Axial average pooling

To capture direction-aware channel statistics, ACA first applies axial average pooling along the horizontal and vertical spatial dimensions. Specifically, average pooling is performed independently along height and width:

$$\mathbf{A}^{h}_{\text{avg}}(b,c,w) = \frac{1}{H}\sum_{h=1}^{H}\mathbf{X}(b,c,h,w),$$

$$\mathbf{A}^{w}_{\text{avg}}(b,c,h) = \frac{1}{W}\sum_{w=1}^{W}\mathbf{X}(b,c,h,w).$$

After pooling, both responses are broadcast back to $\mathbb{R}^{B\times C\times H\times W}$ and summed to obtain the axial average descriptor:

$$\mathbf{A}_{\text{avg}} = \text{Expand}(\mathbf{A}^{h}_{\text{avg}}) + \text{Expand}(\mathbf{A}^{w}_{\text{avg}}).$$

B.2. Axial max pooling

In parallel, ACA extracts complementary discriminative cues via axial max pooling:

$$\mathbf{A}^{h}_{\text{max}}(b,c,w) = \max_{h}\mathbf{X}(b,c,h,w),$$

$$\mathbf{A}^{w}_{\text{max}}(b,c,h) = \max_{w}\mathbf{X}(b,c,h,w).$$

These responses are similarly broadcast and combined:

$$\mathbf{A}_{\text{max}} = \text{Expand}(\mathbf{A}^{h}_{\text{max}}) + \text{Expand}(\mathbf{A}^{w}_{\text{max}}).$$

B.3. Channel attention mask generation

The axial average and max descriptors are independently passed through a shared bottleneck transformation implemented using pointwise convolutions. First, channel dimensionality is reduced by a factor r:

$$\mathbf{Z}_{\text{avg}} = \mathbf{W}_2\,\delta\left(\mathbf{W}_1\mathbf{A}_{\text{avg}}\right), \qquad \mathbf{Z}_{\text{max}} = \mathbf{W}_2\,\delta\left(\mathbf{W}_1\mathbf{A}_{\text{max}}\right),$$

where $\mathbf{W}_1 \in \mathbb{R}^{\frac{C}{r}\times C}$ and $\mathbf{W}_2 \in \mathbb{R}^{C\times\frac{C}{r}}$ denote 1×1 convolutional kernels, and $\delta(\cdot)$ represents the LeakyReLU activation.

The two transformed responses are then aggregated and passed through a sigmoid function to generate the channel attention mask:

$$\mathbf{A}_{\text{channel}} = \sigma\left(\mathbf{Z}_{\text{avg}} + \mathbf{Z}_{\text{max}}\right),$$

where $\mathbf{A}_{\text{channel}} \in \mathbb{R}^{B\times C\times H\times W}$.

B.4. Residual channel refinement

Finally, ACA applies the learned channel attention in a residual scaling manner:

$$\mathbf{X}' = \mathbf{X} \odot \left(1 + \mathbf{A}_{\text{channel}}\right),$$

where $\odot$ denotes element-wise multiplication. This residual formulation preserves the original feature representation while selectively amplifying informative channels based on direction-aware contextual cues.

By replacing global pooling with axial aggregation, ACA maintains sensitivity to localized spectral and structural variations while incurring minimal additional computational cost.

Data availability

The data that support the findings of this study are openly available from the Google Earth Engine platform (https://developers.google.com/earth-engine/datasets). The data can be accessed from the image collections `COPERNICUS/S1_GRD` for Sentinel-1 and `COPERNICUS/S2_SR_HARMONIZED` for Sentinel-2.

References

Aashi, A., Vema, V.K., 2025. The role of image aggregation techniques in enhancing in-season crop classification accuracy: A multi-sensor approach. Remote. Sens. Appl.: Soc. Environ. 39, 101629. http://dx.doi.org/10.1016/j.rsase.2025.101629.

Altarez, R.D.D., Apan, A., Maraseni, T., 2023. Deep learning U-Net classification of sentinel-1 and 2 fusions effectively demarcates tropical montane forest's deforestation. Remote. Sens. Appl.: Soc. Environ. 29, 100887. http://dx.doi.org/10.1016/j.rsase.2022.100887.

Borsah, A.A., Wong, M.S., Nazeer, M., Shi, G., 2025. Machine learning models for subtropical forest aboveground biomass mapping using combined SAR and optical satellite imagery. Remote. Sens. Appl.: Soc. Environ. 39, 101640. http://dx.doi.org/10.1016/j.rsase.2025.101640.

Boulila, W., Ghandorh, H., Masood, S., Alzahem, A., Koubaa, A., Ahmed, F., Khan, Z., Ahmad, J., 2024. A transformer-based approach empowered by a self-attention technique for semantic segmentation in remote sensing. Heliyon 10 (8), e29396. http://dx.doi.org/10.1016/j.heliyon.2024.e29396.

Chen, L.-C., Zhu, Y., Papandreou, G., Schroff, F., Adam, H., 2018. Encoder-decoder with atrous separable convolution for semantic image segmentation. In: Computer Vision – ECCV 2018. pp. 833–851. http://dx.doi.org/10.1007/978-3-030-01234-2_49.

Deng, Y., Peng, B., Guan, K., Runkle, B.R., Moreno-García, B., Wu, X., Wang, S., Zhou, Q., Reba, M.L., 2025. Detecting the onset of rice field inundation in the lower mississippi river basin via harmonized landsat sentinel-2 (HLS) satellite time series. ISPRS J. Photogramm. Remote Sens. 228, 28–43. http://dx.doi.org/10.1016/j.isprsjprs.2025.07.003.

FAO, 2023. FAOSTAT: Countries by commodity. https://www.fao.org/faostat/en/#rankings/countries_by_commodity. (Accessed 02 July 2025).

FAO, 2024. World Food and Agriculture – Statistical Yearbook 2024. Food and Agriculture Organization of the United Nations, Rome, http://dx.doi.org/10.4060/cd2971en.

Fikriyah, V.N., Darvishzadeh, R., Laborte, A., Nelson, A., 2025. Ratoon rice mapping based on sentinel-1 and sentinel-2 imagery. Remote. Sens. Appl.: Soc. Environ. 38, 101592. http://dx.doi.org/10.1016/j.rsase.2025.101592.

Gao, L., Liu, H., Yang, M., Chen, L., Wan, Y., Xiao, Z., Qian, Y., 2021. STransFuse: Fusing swin transformer and convolutional neural network for remote sensing image semantic segmentation. IEEE J. Sel. Top. Appl. Earth Obs. Remote. Sens. 14, 10990–11003. http://dx.doi.org/10.1109/JSTARS.2021.3119654.

Garcia, A.D.B., Islam, M.S., Prudente, V.H.R., Sanches, I.D., Cheng, I., 2025. Irrigated rice-field mapping in Brazil using phenological stage information and optical and microwave remote sensing. Appl. Comput. Geosci. 25, 100223. http://dx.doi.org/10.1016/j.acags.2025.100223.

Ghazaryan, G., Ernst, S., Sempel, F., Nendel, C., 2025. Large-scale irrigation mapping at field level in Northern Germany with integrated use of sentinel-2, landsat 8 and sentinel-1 time series. Remote. Sens. Appl.: Soc. Environ. 38, 101593. http://dx.doi.org/10.1016/j.rsase.2025.101593.

Gorelick, N., Hancher, M., Dixon, M., Ilyushchenko, S., Thau, D., Moore, R., 2017. Google earth engine: Planetary-scale geospatial analysis for everyone. Remote Sens. Environ. 202, 18–27. http://dx.doi.org/10.1016/j.rse.2017.06.031.

He, K., Zhang, X., Ren, S., Sun, J., 2015. Deep residual learning for image recognition. http://dx.doi.org/10.48550/arXiv.1512.03385, CoRR abs/1512.03385.

He, X., Zhou, Y., Zhao, J., Zhang, D., Yao, R., Xue, Y., 2022. Swin transformer embedding unet for remote sensing image semantic segmentation. IEEE Trans. Geosci. Remote Sens. 60, 1–15, URL: https://doi.org/10.1109/TGRS.2022.3144165.

Hegde, A., Umesh, P., Tahiliani, M.P., 2025. Automated rice mapping using multitemporal sentinel-1 SAR imagery using dynamic threshold and slope-based index methods. Remote. Sens. Appl.: Soc. Environ. 37, 101410. http://dx.doi.org/10.1016/j.rsase.2024.101410.

Hussain, Z., Congshi, J., Adrees, M., Chaudhary, H., Shafqat, R., 2025. A novel architecture for building rooftop extraction using remote sensing and deep learning. Remote. Sens. Appl.: Soc. Environ. 38, 101551. http://dx.doi.org/10.1016/j.rsase.2025.101551.

Isola, P., Zhu, J., Zhou, T., Efros, A.A., 2017. Image-to-image translation with conditional adversarial networks. CVPR. http://dx.doi.org/10.1109/CVPR.2017.632.

Jo, H.W., Park, E., Sitokonstantinou, V., Kim, J., Lee, S., Koukos, A., Lee, W.K., 2023. Recurrent U-Net based dynamic paddy rice mapping in South Korea with enhanced data compatibility to support agricultural decision making. GIScience Remote. Sens. 60, http://dx.doi.org/10.1080/15481603.2023.2206539.

Kanja, K., Zhang, C., Lippe, M., Atkinson, P.M., 2025. Mapping tropical dry miombo woodlands into functional forest classes using sentinel-1 and sentinel-2 imagery and machine learning. Remote. Sens. Appl.: Soc. Environ. 38, 101615. http://dx.doi.org/10.1016/j.rsase.2025.101615.

Kumari, M., Chakraborty, A., Adhikari, R., Panda, A., Raju, P.S., Chowdary, V., Sreenivas, K., 2025. Potential of sentinel-1 C-band SAR data in tandem with sentinel-2 optical data towards in-season mapping of paddy residue burnt area. Remote. Sens. Appl.: Soc. Environ. 37, 101514. http://dx.doi.org/10.1016/j.rsase.2025.101514.

Li, J., Nguyen, T.-M., Cao, M., Yuan, S., Hung, T.-Y., Xie, L., 2025. Graph optimality-aware stochastic LiDAR bundle adjustment with progressive spatial smoothing. IEEE Trans. Intell. Transp. Syst. 26 (11), 19076–19091. http://dx.doi.org/10.1109/TITS.2025.3597242.

Li, R., Wang, L., Zhang, C., Duan, C., Zheng, S., 2022b. A2-FPN for semantic segmentation of fine-resolution remotely sensed images. Int. J. Remote Sens. 43, 1131–1155. http://dx.doi.org/10.1080/01431161.2022.2030071.

Li, J., Wu, Z., Hu, Z., Jian, C., Luo, S., Mou, L., Zhu, X.X., Molinier, M., 2022a. A lightweight deep learning-based cloud detection method for sentinel-2A imagery fusing multiscale spectral and spatial features. IEEE Trans. Geosci. Remote Sens. 60, 1–19. http://dx.doi.org/10.1109/TGRS.2021.3069641.

Li, J., Yang, B., Chen, C., Habib, A., 2019. NRLI-UAV: Non-rigid registration of sequential raw laser scans and images for low-cost UAV LiDAR point cloud quality improvement. ISPRS J. Photogramm. Remote Sens. 158, 123–145. http://dx.doi.org/10.1016/j.isprsjprs.2019.10.009.

Li, X., Zhang, G., Cui, H., Hou, S., Wang, S., Li, X., Chen, Y., Li, Z., Zhang, L., 2022c. MCANet: A joint semantic segmentation framework of optical and SAR images for land use classification. Int. J. Appl. Earth Obs. Geoinf. 106, 102638. http://dx.doi.org/10.1016/j.jag.2021.102638.

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: 2021 IEEE/CVF International Conference on Computer Vision. ICCV, IEEE Computer Society, http://dx.doi.org/10.1109/ICCV48922.2021.00986.

Liu, Z., Mao, H., Wu, C.-Y., Feichtenhofer, C., Darrell, T., Xie, S., 2022. A ConvNet for the 2020s. http://dx.doi.org/10.48550/arXiv.2201.03545.

Loshchilov, I., Hutter, F., 2017. SGDR: Stochastic gradient descent with warm restarts. In: ICLR 2017. http://dx.doi.org/10.48550/arXiv.1608.03983.

Loshchilov, I., Hutter, F., 2019. Decoupled weight decay regularization. In: ICLR 2019. http://dx.doi.org/10.48550/arXiv.1711.05101.

Ma, X., Zhang, X., Pun, M.-O., 2022. A crossmodal multiscale fusion network for semantic segmentation of remote sensing data. IEEE J. Sel. Top. Appl. Earth Obs. Remote. Sens. 15, 3463–3474. http://dx.doi.org/10.1109/JSTARS.2022.3165005.

Ni, R., Tian, J., Li, X., Yin, D., Li, J., Gong, H., Zhang, J., Zhu, L., Wu, D., 2021. An enhanced pixel-based phenological feature for accurate paddy rice mapping with sentinel-2 imagery in Google earth engine. ISPRS J. Photogramm. Remote Sens. 178, 282–296. http://dx.doi.org/10.1016/j.isprsjprs.2021.06.018.

Nie, H., Lin, Y., Luo, W., Liu, G., 2025. Rice cropping sequence mapping in the tropical monsoon zone via agronomic knowledge graphs integrating phenology and remote sensing. Ecol. Inform. 87, 103075. http://dx.doi.org/10.1016/j.ecoinf.2025.103075.

Novresiandi, D.A., Setiyoko, A., Indriasari, N., Veronica, K.W., Budiono, M.E., Dianovita, Amriyah, Q., Subehi, M., 2024. Evaluation of speckle filtering configurations on sentinel-1 SAR backscatter analysis ready data (S1ARD) preparation framework on the google earth engine platform for supporting rice monitoring activities. Remote. Sens. Appl.: Soc. Environ. 36, 101337. http://dx.doi.org/10.1016/j.rsase.2024.101337.

Qiu, B., Wu, F., Hu, X., Yang, P., Wu, W., Chen, J., Chen, X., He, L., Joe, B., Tubiello, F.N., Qian, J., Wang, L., 2025. A robust framework for mapping complex cropping patterns: The first national-scale 10 m map with 10 crops in China using sentinel 1/2 images. ISPRS J. Photogramm. Remote Sens. 224, 361–381. http://dx.doi.org/10.1016/j.isprsjprs.2025.04.012.

Ren, B., Ma, S., Hou, B., Hong, D., Chanussot, J., Wang, J., Jiao, L., 2022. A dual-stream high resolution network: Deep fusion of GF-2 and GF-3 data for land cover classification. Int. J. Appl. Earth Obs. Geoinf. 112, 102896. http://dx.doi.org/10.1016/j.jag.2022.102896.

Ronneberger, O., Fischer, P., Brox, T., 2015. U-Net: Convolutional networks for biomedical image segmentation. In: Medical Image Computing and Computer-Assisted Intervention – MICCAI 2015. Springer International Publishing, Cham, pp. 234–241.

Salehi, S.S.M., Erdogmus, D., Gholipour, A., 2017. Tversky loss function for image segmentation using 3D fully convolutional deep networks. http://dx.doi.org/10.48550/arXiv.1706.05721, CoRR abs/1706.05721.

Shen, R., Pan, B., Peng, Q., Dong, J., Chen, X., Zhang, X., Ye, T., Huang, J., Yuan, W., 2023. High-resolution distribution maps of single-season rice in China from 2017 to 2022. Earth Syst. Sci. Data 15, 3203–3222. http://dx.doi.org/10.5194/essd-15-3203-2023.

Statistics Indonesia, 2025a. Executive Summary of Paddy Harvested Area and Production in Jawa Barat 2024 (Final Figures). Statistics Indonesia, Bandung, URL: https://jabar.bps.go.id/id/publication/2025/06/05/51bb1554fa1ec0f7883e9c4f/ringkasan-eksekutif-luas-panen-dan-produksi-padi-di-provinsi-jawa-barat-2024.html.

Statistics Indonesia, 2025b. Indramayu Regency in Figures 2025. Statistics Indonesia, Indramayu, URL: https://indramayukab.bps.go.id/id/publication/2025/02/28/3d52dbf0a5eeef01db0edf8d/kabupaten-indramayu-dalam-angka-2025.html.

Sun, C., Zhang, H., Xu, L., Ge, J., Jiang, J., Zuo, L., Wang, C., 2023. Twenty-meter annual paddy rice area map for mainland Southeast Asia using sentinel-1 synthetic-aperture-radar data. Earth Syst. Sci. Data 15, 1501–1520. http://dx.doi.org/10.5194/essd-15-1501-2023.

Tan, M., Le, Q.V., 2020. EfficientNet: Rethinking model scaling for convolutional neural networks. In: ICML 2019. http://dx.doi.org/10.48550/arXiv.1905.11946.

Therias, A., Rafiee, A., Lhermitte, S., van der Lugt, P., Lindenbergh, R., 2025. Integrating radar and multi-spectral data to detect cocoa crops: a deep learning approach. Remote. Sens. Appl.: Soc. Environ. 39, 101652. http://dx.doi.org/10.1016/j.rsase.2025.101652.

Thorp, K., Drajat, D., 2021. Deep machine learning with sentinel satellite data to map paddy rice production stages across West Java, Indonesia. Remote Sens. Environ. 265, 112679. http://dx.doi.org/10.1016/j.rse.2021.112679.

Tu, Z., Talebi, H., Zhang, H., Yang, F., Milanfar, P., Bovik, A., Li, Y., 2022. MaxViT: Multi-axis vision transformer. URL: https://doi.org/10.48550/arXiv.2204.c.

Wan, Q., Huang, Z., Lu, J., Yu, G., Zhang, L., 2025. SeaFormer++: Squeeze-enhanced axial transformer for mobile visual recognition. Int. J. Comput. Vis. 133, http://dx.doi.org/10.1007/s11263-025-02345-2.

Wang, L., Li, R., Duan, C., Zhang, C., Meng, X., Fang, S., 2022a. A novel transformer based semantic segmentation scheme for fine-resolution remote sensing images. IEEE Geosci. Remote. Sens. Lett. 19, 1–5. http://dx.doi.org/10.1109/LGRS.2022.3143368.

Wang, L., Li, R., Zhang, C., Fang, S., Duan, C., Meng, X., Atkinson, P.M., 2022b. UNetFormer: A unet-like transformer for efficient semantic segmentation of remote sensing urban scene imagery. ISPRS J. Photogramm. Remote Sens. 190, 196–214. http://dx.doi.org/10.1016/j.isprsjprs.2022.06.008.

Wang, J., Sun, K., Cheng, T., Jiang, B., Deng, C., Zhao, Y., Liu, D., Mu, Y., Tan, M., Wang, X., Liu, W., Xiao, B., 2021. Deep high-resolution representation learning for visual recognition. IEEE Trans. Pattern Anal. Mach. Intell. 43, 3349–3364. http://dx.doi.org/10.1109/TPAMI.2020.2983686.

Wang, M., Wang, J., Cui, Y., Liu, J., Chen, L., 2022c. Agricultural field boundary delineation with satellite image segmentation for high-resolution crop mapping: A case study of rice paddy. Agronomy 12, http://dx.doi.org/10.3390/agronomy12102342.

Wang, H., Zhu, Y., Green, B., Adam, H., Yuille, A., Chen, L.-C., 2020. Axial-DeepLab: Stand-alone axial-attention for panoptic segmentation. In: European Conference on Computer Vision. ECCV, http://dx.doi.org/10.34123/icdsos.v2023i1.399.

Wijaya, B.S., Munir, R., Utama, N.P., 2025. TerraSegNet: Bilateral axial attention network for remote sensing image segmentation in diverse environmental monitoring applications. IEEE Access 13, http://dx.doi.org/10.1109/ACCESS.2025.3633712.

Wu, H., Huang, P., Zhang, M., Tang, W., Yu, X., 2023. CMTFNet: CNN and multiscale transformer fusion network for remote-sensing image semantic segmentation. IEEE Trans. Geosci. Remote Sens. 61, 1–12. http://dx.doi.org/10.1109/TGRS.2023.3314641.

Wu, H., Zhang, M., Huang, P., Tang, W., 2024. CMLFormer: CNN and multiscale local-context transformer network for remote sensing images semantic segmentation. IEEE J. Sel. Top. Appl. Earth Obs. Remote. Sens. 17, 7233–7241. http://dx.doi.org/10.1109/JSTARS.2024.3375313.

Xia, L., Zhao, F., Chen, J., Yu, L., Lu, M., Yu, Q., Liang, S., Fan, L., Sun, X., Wu, S., Wu, W., Yang, P., 2022. A full resolution deep learning network for paddy rice mapping using landsat data. ISPRS J. Photogramm. Remote Sens. 194, 91–107. http://dx.doi.org/10.1016/j.isprsjprs.2022.10.005.

Xiang, X., Gong, W., Li, S., Chen, J., Ren, T., 2024. TCNet: Multiscale fusion of transformer and CNN for semantic segmentation of remote sensing images. IEEE J. Sel. Top. Appl. Earth Obs. Remote. Sens. 17, 3123–3136. http://dx.doi.org/10.1109/JSTARS.2024.3349625.

Xie, E., Wang, W., Yu, Z., Anandkumar, A., Alvarez, J.M., Luo, P., 2021. SegFormer: Simple and efficient design for semantic segmentation with transformers. In: NeurIPS 2021. http://dx.doi.org/10.48550/arXiv.2105.15203.

Yang, J., Hu, Q., Li, W., Song, Q., Cai, Z., Zhang, X., Wei, H., Wu, W., 2024. An automated sample generation method by integrating phenology domain optical-SAR features in rice cropping pattern mapping. Remote Sens. Environ. 314, 114387. http://dx.doi.org/10.1016/j.rse.2024.114387.

Zhang, Z., Shu, D., Liao, C., Liu, C., Zhao, Y., Wang, R., Huang, X., Zhang, M., Gong, J., 2025. FlexiSAM: A flexible SAM-based semantic segmentation model for land cover classification using high-resolution multimodal remote sensing imagery. ISPRS J. Photogramm. Remote Sens. 227, 594–612. http://dx.doi.org/10.1016/j.isprsjprs.2025.05.028.

Zhao, X., Nishina, K., Akitsu, T.K., Jiang, L., Masutomi, Y., Nasahara, K.N., 2023. Feature-based algorithm for large-scale rice phenology detection based on satellite images. Agricult. Forest. Meterol. 329, 109283. http://dx.doi.org/10.1016/j.agrformet.2022.109283.

Zhao, Y., Xie, J., Zhu, H., Luo, T., Xiong, Y., Fan, C., Xia, H., Chen, Y., Zhang, F., 2025. Land-unet: A deep learning network for precise segmentation and identification of non-structured land use types in rural areas for green urban space analysis. Ecol. Inform. 87, 103078. http://dx.doi.org/10.1016/j.ecoinf.2025.103078.