IEEE Access
Multidisciplinary | Rapid Review | Open Access Journal

Date of publication xxxx 00, 0000, date of current version xxxx 00, 0000.

Digital Object Identifier 10.1109/ACCESS.2025.0429000

TerraSegNet: Bilateral Axial Attention Network for Remote Sensing Image Segmentation in Diverse Environmental Monitoring Applications

BAGUS SETYAWAN WIJAYA[1, 2], RINALDI MUNIR[1], and NUGRAHA PRIYA UTAMA[1] (Member, IEEE)

[1]School of Electrical Engineering and Informatics, Institut Teknologi Bandung, Bandung 40132, Indonesia
[2]BPS-Statistics Indonesia, Jakarta 10710, Indonesia

Corresponding author: Nugraha Priya Utama (e-mail: utama@itb.ac.id)

This work was supported by Institut Teknologi Bandung (ITB) under Grant DRI.PN-6-175-2025

ABSTRACT Semantic segmentation is crucial for environmental monitoring by enabling precise land cover mapping and change detection from satellite imagery. Although recent advances in deep learning have significantly improved segmentation performance, most existing models are designed for specific tasks. They are often trained on single-source satellite imagery, which limits their adaptability. To overcome these limitations, we propose TerraSegNet, a lightweight bilateral network that separates the extraction of spatial detail and semantic context through a dual-path architecture. The model integrates convolutional axial attention in the context path to capture long-range dependencies efficiently while maintaining compatibility with convolutional backbones. We evaluated TerraSegNet on four different remote sensing tasks using five publicly available satellite imagery datasets: cloud detection on SPARCS (Landsat 8) and WHUS2-CD+ (Sentinel-2), photovoltaic panel segmentation on PV08 (Gaofen-2), paddy rice mapping on Plot-Rice v1.0 (Sentinel-1), and ground terrain classification on AIR-PolSAR-Seg-2.0 (Gaofen-3). All training is performed independently per dataset without cross-domain transfer learning, ensuring a fair evaluation of in-domain performance. TerraSegNet achieves mean Intersection over Union (mIoU) scores of 0.8608, 0.9364, 0.9617, 0.7777, and 0.9207 on SPARCS, WHUS2-CD+, PV08, Plot-Rice v1.0, and AIR-PolSAR-Seg-2.0, respectively. These results outperform baselines and demonstrate its effectiveness for various environmental applications.

INDEX TERMS Axial attention, bilateral network, deep learning, remote sensing, satellite imagery, semantic segmentation.

I. INTRODUCTION

REMOTE sensing using satellite imagery [1]–[5] plays a critical role in a variety of environmental applications, including monitoring deforestation [6], assessing disaster impacts [7], and evaluating agricultural productivity [8]. A fundamental task in extracting meaningful geospatial information from such imagery is semantic segmentation, which involves classifying each pixel into a predefined semantic category [9], [10]. Accurate segmentation is essential for enabling reliable downstream analysis, especially in operational remote sensing workflows.

In recent years, deep learning techniques have significantly advanced computer vision by effectively capturing spatial patterns and high-level semantic features [11]–[13]. These advancements have resulted in notable improvements in semantic segmentation accuracy. Although numerous models have been developed for remote sensing applications such as land cover classification and object detection, a persistent limitation remains. However, most segmentation frameworks are tailored for task-specific scenarios and trained on data from a single satellite modality [14], [15], limiting their generalizability. Furthermore, these models often lack adaptability when applied to heterogeneous satellite datasets [16], underscoring the need for a more flexible and robust segmentation framework.

To address this limitation, we propose TerraSegNet, a lightweight semantic segmentation model designed to balance cross-sensor adaptability with computational efficiency.

TerraSegNet is inspired by bilateral network architectures, particularly BiSeNetV1 [17] and BiSeNetV2 [18], which separate spatial and contextual pathways to manage the trade-off between spatial detail preservation and semantic abstraction. Fast-SCNN [19] introduces multi-resolution branches within a streamlined encoder-decoder structure to achieve real-time performance. SeaFormer++ [20], on the other hand, employs axial attention mechanisms along the context path to efficiently capture long-range dependencies. Fig. 1 provides a comparative overview of these representative dual-path segmentation architectures and the proposed TerraSegNet.

Building upon these architectural foundations, TerraSegNet employs a dual-path design that decouples the spatial path—tasked with capturing fine-grained spectral and spatial features—from the context path, which focuses on extracting high-level semantic representations. These two branches are fused via a lightweight integration module to produce a unified and expressive feature map. To strengthen spatial representation, a spectral-spatial attention module is embedded, enabling adaptive emphasis on relevant spectral bands and spatial locations. Simultaneously, the context path incorporates a convolutional axial attention module, which applies directional attention sequentially along the height and width dimensions. This formulation efficiently captures global context while significantly reducing the computational cost compared to full self-attention. Unlike the axial attention mechanisms in models such as SeaFormer++ and Axial-DeepLab [21], which rely on resource-intensive self-attention, TerraSegNet leverages global pooling and convolutions to simulate directional attention, thereby ensuring compatibility with standard CNN backbones.

To evaluate the proposed model, we conduct extensive experiments across four remote sensing segmentation tasks: (1) cloud detection using Landsat 8 and Sentinel-2 imagery, (2) photovoltaic panel segmentation using Gaofen-2, (3) terrain classification using Gaofen-3, and (4) paddy rice mapping using Sentinel-1. Unlike cross-dataset transfer approaches, TerraSegNet is trained and evaluated independently on each dataset, providing a rigorous assessment of its adaptability across diverse imaging modalities. The model is benchmarked against several state-of-the-art segmentation architectures using standard evaluation metrics, including mean Intersection over Union (mIoU), F1-score, and pixel accuracy. The experimental results demonstrate that TerraSegNet achieves a strong balance between segmentation accuracy and computational efficiency across multiple remote sensing scenarios.

The main contributions of this study are summarized as follows:

1) We propose TerraSegNet, a lightweight bilateral segmentation framework for multi-modal remote sensing that integrates an efficient backbone with three complementary attention mechanisms, resulting in a design optimized for both accuracy and efficiency.
2) We perform comprehensive evaluations on five publicly available satellite datasets, covering four distinct remote sensing segmentation tasks and a range of imaging modalities.
3) We demonstrate that TerraSegNet provides a favorable trade-off between accuracy and computational cost, highlighting its practicality for real-world remote sensing applications.

The remainder of this paper is organized as follows. Section 2 reviews related work on semantic segmentation in remote sensing. Section 3 describes the proposed TerraSegNet architecture and details the experimental design. Section 4 presents the evaluation results. Section 5 provides a comprehensive discussion of the findings, including statistical and visual analyses. Finally, Section 6 concludes the study and outlines directions for future research.

II. RELATED WORKS

A. CONVENTIONAL SEMANTIC SEGMENTATION

A wide range of semantic segmentation models have been proposed based on distinct architectural paradigms. Early works in this area often relied on classical image processing and optimization-based methods. For instance, a simplified pulse coupled neural network (SPCNN) combined with a gbest-led gravitational search algorithm (GLGSA) was proposed in [22] to enhance image segmentation quality by optimizing cross-entropy, edge matching, and noise control measures. Although such approaches demonstrated improvements over traditional thresholding and clustering techniques, they are inherently limited in handling large-scale, multimodal satellite data due to their reliance on handcrafted objective functions and lack of feature learning.

One of the earliest and most widely adopted deep learning models is U-Net [23], originally developed for biomedical image segmentation. Its encoder–decoder structure with symmetric skip connections enables precise localization, even with limited training data. However, U-Net's reliance on skip connections for preserving spatial details may hinder its ability to capture global contextual information—an essential factor in large-scale satellite imagery analysis.

To overcome this limitation, DeepLabv3+ [24] introduces atrous spatial pyramid pooling (ASPP) for extracting multi-scale contextual features, along with a refined decoder to improve boundary delineation. Despite its effectiveness, the use of dilated convolutions may introduce grid-like artifacts in the output, potentially reducing segmentation quality. HRNet [25], in contrast, maintains high-resolution representations throughout the network by leveraging parallel convolutional streams and repeated multi-scale fusion. This architecture enhances spatial fidelity but incurs high memory and computational costs, limiting its applicability in resource-constrained settings.

More recently, SegFormer [26] has emerged as a transformer-based segmentation model that integrates a lightweight hierarchical encoder with a simple multilayer perceptron (MLP) decoder. It eliminates the need for positional encodings and employs efficient attention mechanisms, offering a balance between accuracy and efficiency.

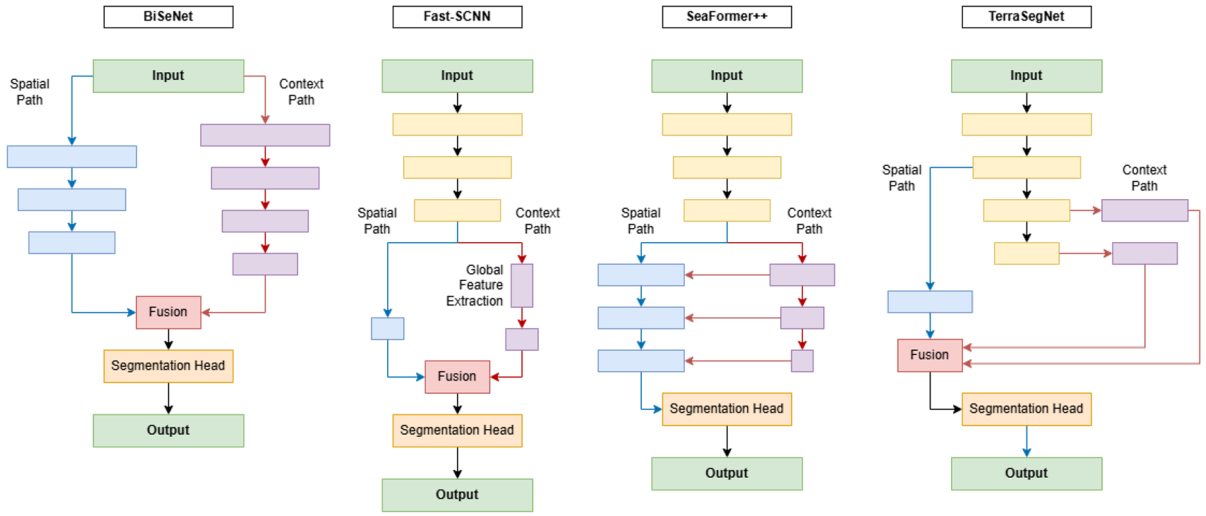


FIGURE 1. Architectural comparison of representative dual-path semantic segmentation models, including BiSeNet, Fast-SCNN, SeaFormer++, and the proposed TerraSegNet. Each model incorporates distinct design strategies to balance spatial detail preservation with semantic context extraction.

In parallel, foundation models such as the Segment Anything Model (SAM) [27] represent a paradigm shift toward general-purpose, prompt-driven segmentation frameworks. Although these models have shown remarkable performance on natural and medical image datasets, their application to remote sensing remains limited. Domain-specific fine-tuning is often necessary to address challenges such as texture variability, scale disparity, and sensor heterogeneity in satellite imagery.

In contrast to these general-purpose approaches, TerraSegNet is explicitly designed to accommodate the diverse characteristics of multi-modal satellite data. Its lightweight and adaptable architecture makes it particularly suitable for operational environmental monitoring across a wide range of remote sensing scenarios.

B. BILATERAL NETWORK ARCHITECTURE

Bilateral network architectures have emerged as a promising design paradigm for real-time semantic segmentation, effectively addressing the trade-off between preserving high-resolution spatial features and capturing high-level semantic context. A representative model in this category is BiSeNet, which introduces two separate processing branches: a spatial path that maintains fine spatial resolution, and a context path that extracts semantic features using a deep backbone with a large receptive field. This architectural decoupling enables a balance between segmentation accuracy and inference speed, making it suitable for real-time applications.

BiSeNetV2 extends this approach by independently optimizing the spatial and context paths and incorporating a lightweight feature fusion module to efficiently integrate both representations. These designs demonstrate that treating spatial and semantic features as complementary, yet decoupled, leads to improved segmentation performance with minimal computational overhead.

Several other architectures have also adopted and adapted the bilateral paradigm. For instance, Fast-SCNN introduces a learning-to-downsample module coupled with a lightweight decoder, enabling efficient segmentation of high-resolution imagery with reduced latency. SeaFormer++ further advances the context path by integrating lightweight axial attention mechanisms, enhancing global context modeling while maintaining computational efficiency.

Collectively, these works highlight the effectiveness of explicitly separating spatial detail extraction from semantic understanding—an especially beneficial strategy in remote sensing applications characterized by complex textures and large-scale spatial structures.

Building upon these architectural insights, the proposed TerraSegNet extends the bilateral framework by incorporating dedicated attention modules into both the spatial and context paths. This design enhances feature discrimination and enables efficient long-range dependency modeling, resulting in a more adaptable and computationally efficient segmentation model for heterogeneous remote sensing imagery.

C. SEMANTIC SEGMENTATION IN REMOTE SENSING

Semantic segmentation of satellite imagery plays a critical role in a wide range of remote sensing applications. Compared to natural images, satellite imagery presents unique challenges, including diverse spectral modalities, varying spatial resolutions, seasonal variations, and sensor-specific noise characteristics. These factors complicate the development of segmentation models that can maintain consistent performance across different tasks and data sources. Early approaches to segmentation in remote sensing primarily relied on handcrafted spectral indices, rule-based heuristics, or pixel-wise classifiers, which lacked scalability and struggled to capture complex scene structures [28]–[30].

The advent of deep learning has significantly advanced the state of the art in remote sensing image analysis [31]–[36].

These models have demonstrated strong performance across various tasks, including land cover mapping [37], cloud detection [38], and crop classification [39], especially when trained on task-specific datasets. Moreover, they have been effectively adapted for a variety of satellite and airborne sensors, such as MODIS [40], Landsat-8 [41], Sentinel-1 [42], Sentinel-2 [43], Gaofen-2 [44], Gaofen-3 [45], and AVIRIS [46].

Several models have been developed specifically to address segmentation challenges in remote sensing. For example, HR-Cloud-Net [47] and CDnetV2 [48] are tailored for cloud detection. HR-Cloud-Net combines pyramid pooling and cascaded feature fusion to model complex cloud textures while preserving fine details. CDnetV2 incorporates residual connections along with channel and spatial attention to enhance multi-scale feature fusion and contextual representation.

In urban scene segmentation, various hybrid CNN and Transformer models have been proposed to balance spatial detail preservation and contextual understanding. MLWNet [49] integrates a multi-scale linear self-attention mechanism to capture contextual relationships and a weighted feature fusion process to ensure spatially detailed. UNetFormer [50] combines a lightweight CNN encoder with a Transformer decoder featuring Global–Local Transformer Blocks and a Feature Refinement Head. ABCNet [51] adopts a bilateral architecture that decouples spatial detail processing from global semantic modeling. BANet [52] extends this design by incorporating a texture pathway and a Transformer-based dependency pathway, which are fused through an attentional aggregation module. CMLFormer [53] employs a CNN encoder and a multiscale local-context Transformer decoder that leverages windowed self-attention and stripe convolution for efficient global–local feature learning. Similarly, CMTFNet [54] integrates a CNN encoder with a multiscale Transformer decoder that utilizes multi-head self-attention, and a multiscale attention fusion module to capture and combine hierarchical contextual features. LAGAN [46] integrates segmentation results from two concurrent feature streams: the raw spectral features and the deep abstract spectral features unearthed by a deep autoencoder. In addition, foundation models based on SAM have also been adapted for urban scene mapping by integrating them with multiscale feature hierarchies [55].

However, previous approaches are typically limited to single-domain training, which reduces their robustness when applied to heterogeneous satellite imagery. Although transfer learning [16] and domain adaptation techniques [56] have been explored to address this problem, they often require additional training stages or annotated data from the target domain. Consequently, there remains a pressing need for segmentation models that are not only compact and accurate but also inherently adaptable across a broad spectrum of remote sensing scenarios. This need forms the core motivation behind the development of TerraSegNet.

III. METHODOLOGY

A. DATASETS

This study utilizes five publicly available satellite image segmentation datasets, selected to represent a broad range of satellite platforms, spatial resolutions, geographic regions, and segmentation objectives. Each dataset provides pixel-level semantic annotations suitable for benchmarking model performance under diverse remote sensing conditions.

The SPARCS dataset [57] is used for cloud detection and comprises Landsat-8 imagery with a spatial resolution of 30 meters per pixel and image dimensions of 1000×1000 pixels. For model training, each image is partitioned into overlapping 512×512 patches. To simplify class structure and balance data distribution, the original categories "Cloud Shadow" and "Cloud Shadow over Water" are merged into a single Cloud Shadow class, while "Water," "Ice/Snow," "Land," and "Flooded" are grouped into a unified Clear Sky class. SPARCS encompasses a wide range of geographic areas, supporting evaluation across varied land cover types and atmospheric conditions.

The WHUS2-CD+ dataset [15] consists of Sentinel-2 multispectral imagery with spatial resolutions ranging from 10 to 60 meters per pixel. It includes scenes from various regions in mainland China, captured between 2018 and 2020. Designed for cloud detection, the dataset defines two semantic classes: Cloud (including thin and thick clouds) and Clear Sky, enabling model assessment under challenging atmospheric conditions.

The PV08 dataset [58] is employed for photovoltaic panel segmentation and comprises high-resolution Gaofen-2 imagery at 3.24-meter spatial resolution. Each image (1024×1024 pixels) is divided into non-overlapping 512×512 patches. All samples originate from Jiangsu Province, China, offering a consistent urban–rural distribution for rooftop and ground-mounted solar panel segmentation.

The Plot-Rice v1.0 dataset [59] focuses on paddy rice mapping and uses Sentinel-1 SAR imagery at a resolution of 10 meters. It includes time-series data collected across agricultural zones in multiple countries throughout the 2023 planting season. The dataset is annotated into two semantic classes: Rice Area and Non-Rice Area, capturing seasonal and regional variability in radar backscatter.

The AIR-PolSAR-Seg-2.0 dataset [45] provides full-polarimetric SAR imagery from the Gaofen-3 satellite, acquired on April 29, 2019, with an 8-meter resolution. The scenes cover the Hangzhou region in China and are labeled for terrain classification into six categories: Water Bodies, Vegetation, Bare Land, Buildings, Roads, and Mountains.

A summary of dataset characteristics is provided in Table 1, and representative examples of input imagery and corresponding ground truth annotations are illustrated in Fig. III-A. These examples highlight the diversity in spectral modalities, spatial resolutions, and class definitions across the datasets.

To improve model robustness and mitigate overfitting, several preprocessing and data augmentation techniques were applied. Training images were normalized using min–max

TABLE 1. Summary of the benchmark datasets used for segmentation evaluation. The table reports key characteristics including geographic coverage, satellite platform, sensor modality, spatial resolution, number of semantic classes, patch configuration (size and total number of patches), and data format.

Dataset	Location	Satellite	Sensor Type	Spatial Resolution	Number of Classes	Patch Size	Number of Patches	Format
SPARCS	Global	Landsat 8	Optic	30.00 m	3	512×512	1,580	PNG
WHUS2-CD+	China	Sentinel-2	Optic	10.00 m	2	512×512	15,876	TIF
PV08	China	Gaofen-2	Optic	3.24 m	2	512×512	3,052	BMP
Plot-Rice v1.0	Global	Sentinel-1	Radar	10.00 m	2	256×256	13,552	NPZ
AIR-PolSAR-Seg-2.0	China	Gaofen-3	Radar	8.00 m	6	512×512	24,672	TIFF

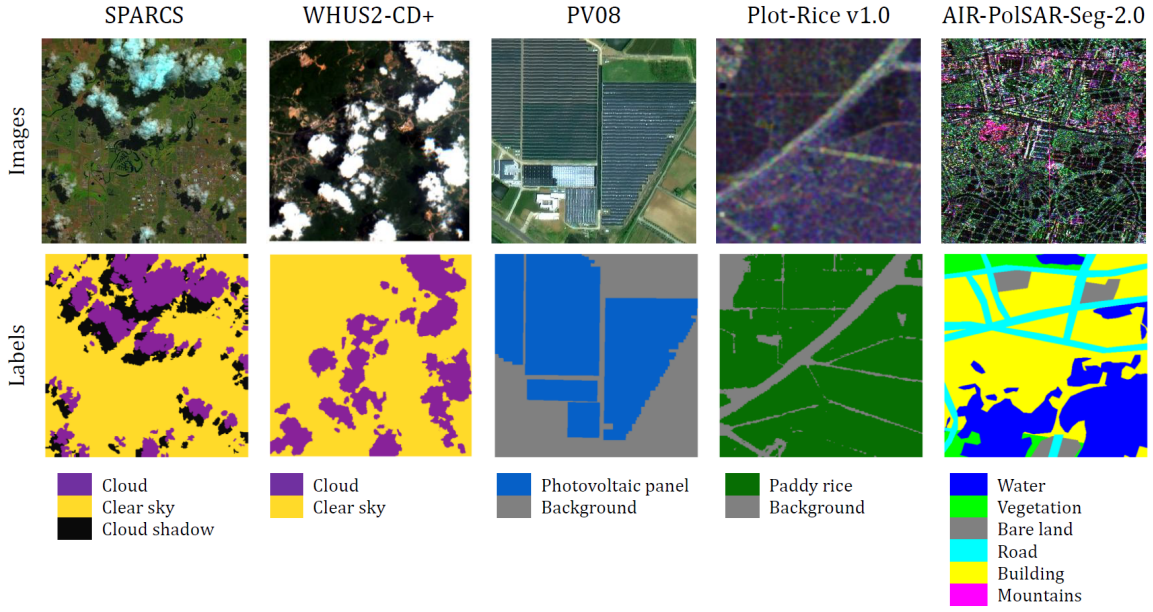


FIGURE 2. Representative samples of satellite imagery and their corresponding ground truth labels from the five benchmark datasets used in this study. The figure illustrates the diversity of imaging modalities (optical and radar), spatial resolutions, geographic contexts, and semantic categories, highlighting the heterogeneity of the evaluation scenarios.

scaling to the [0, 1] range to preserve relative spectral intensities. Augmentations included random horizontal and vertical flips (probability 0.3), 90-degree rotations (probability 0.3), and color jitter to simulate sensor-induced distortions and enhance robustness. Additionally, speckle noise was applied to simulate sensor-induced distortions, particularly for synthetic-aperture radar (SAR) datasets. All preprocessing and augmentation operations were performed with a fixed random seed of 42 to ensure reproducibility. These strategies enhance data diversity and improve model generalization across unseen satellite imagery.

B. BENCHMARK MODELS

To ensure a comprehensive and rigorous evaluation, we compare TerraSegNet against a diverse set of semantic segmentation models categorized into three benchmark groups, as summarized in Table 2:

1) **Conventional segmentation models:** This category includes widely adopted general-purpose architectures not explicitly designed for dual-path processing. Examples are *U-Net*, *DeepLabv3+*, *HRNet*, and *SegFormer*. These models serve as strong baselines in computer vision and have been extensively applied to remote sensing tasks.
2) **Bilateral segmentation models:** These models adopt a dual-path design that decouples spatial detail and semantic abstraction. Representative examples include *BiSeNetV1*, *BiSeNetV2*, *Fast-SCNN*, and *SeaFormer++*, which emphasize real-time segmentation and efficient architecture design.
3) **Remote sensing-specific models:** This group consists of architectures customized for remote sensing imagery, incorporating domain priors such as atmospheric effects, sensor characteristics, and spatial–spectral variability. Models include *HR-Cloud-Net*, *CDnetV2*, *UNetFormer*, *CMLFormer*, and *CMTFNet*.

This categorization enables a structured comparison of TerraSegNet against both general-purpose and remote sensing-optimized models with varying parameter sizes, backbone choices, and computational complexities. The diversity of models ensures robust and balanced benchmarking across multiple evaluation scenarios (see Table 2).

C. PROPOSED MODEL: TERRASEGNET

TerraSegNet is a lightweight bilateral semantic segmentation architecture explicitly designed to balance segmentation ac-

TABLE 2. Benchmark models employed for comparative evaluation, categorized into conventional architectures, bilateral designs, and remote sensing-specific networks. The table reports backbone configurations, parameter counts, and distinctive design characteristics, providing context for assessing the trade-offs between accuracy, efficiency, and domain adaptation.

Model	Backbone	Number of Parameters	Remarks
Conventional Models			
U-Net	ResNet-34	11.38 M	Encoder–decoder architecture widely used in biomedical and RS segmentation
DeepLabv3+	ResNet-34	11.14 M	Employs atrous spatial pyramid pooling for multi-scale context
HRNet	-	9.90 M	Maintains high-resolution representations throughout
SegFormer	MiT-B0	3.71 M	Transformer-based encoder with lightweight MLP decoder
Bilateral Models			
BiSeNetV1	ResNet-34	23.38 M	Dual-path design for spatial and context feature separation
BiSeNetV2	-	3.34 M	Improved lightweight version optimized for real-time inference
Fast-SCNN	-	1.20 M	Mobile-friendly model for edge devices with fast inference
SeaFormer++	-	8.90 M	Combines efficiency with attention-based modules for semantic segmentation
Remote Sensing-Specific Models			
HR-Cloud-Net	-	74.53 M	High-capacity model tailored for cloud segmentation in satellite imagery
CDnetV2	-	67.59 M	Deep model specialized for cloud detection and atmospheric correction
UNetFormer	ResNet-18	11.72 M	Combines U-Net structure with Transformer elements for scene understanding
CMLFormer	ResNet-18	12.53 M	Local-context transformer decoder with ResNet18 encoder for scene mapping
CMTFNet	ResNet-18	11.61 M	CNN + multiscale transformer with attention fusion for scene segmentation
TerraSegNet	EfficientNetV2-S	5.80 M	Proposed model with CAAM, SSAM, and EAM modules for balanced accuracy and efficiency

curacy and computational efficiency in multi-modal remote sensing tasks. Inspired by previous concepts such as efficient backbone [60], [65], [66], bilateral networks [17]–[20], axial attention [21], and spectral-spatial attention, TerraSegNet introduces a novel combination and adaptation of these ideas tailored to the unique challenges of remote sensing imagery.

First, the backbone employs EfficientNetV2-S [60], selected for its favorable accuracy–efficiency trade-off and reconfigured to operate seamlessly with variable spatial resolutions and heterogeneous spectral channels (e.g., SAR, multispectral, and fused data). This flexible input design enables TerraSegNet to generalize beyond Red-Green-Blue (RGB) inputs, a capability often missing in previous bilateral architectures.

Second, three complementary attention modules are integrated in a manner that is both synergistic and computationally lightweight:

- **Convolutional Axial Attention Module (CAAM):** Unlike standard axial attention, CAAM incorporates convolutional projections before directional attention along height and width. This reduces redundancy and computational overhead, yielding a context-aware yet efficient representation suitable for high-resolution satellite scenes.
- **Spectral–Spatial Attention Module (SSAM):** Although prior works have explored spectral–spatial attention in hyperspectral tasks, SSAM in TerraSegNet is specifically adapted to multi-modal satellite imagery, where spectral characteristics vary drastically (e.g., SAR vs. optical bands). The module jointly re-weights spectral channels and spatial locations to capture fine-grained textures and modality-specific cues.
- **Edge Attention Module (EAM):** Applied during feature fusion, EAM explicitly emphasizes object boundaries, leading to sharper delineation of narrow and irregular structures such as paddy field plots, rivers, and roads—structures that are particularly important in remote sensing applications but often blurred in conventional segmentation pipelines.

Finally, the novelty of TerraSegNet lies not only in the individual modules but also in their integration within a unified bilateral framework. By combining EfficientNetV2-S, a dual-path design, and the three attention mechanisms, TerraSegNet achieves a favorable balance between accuracy and efficiency (5.80M parameters) while delivering consistent gains across diverse remote sensing datasets. This design goes beyond incremental reuse by addressing three key challenges jointly: multi-modal input heterogeneity, global–local context integration, and edge-aware precision.

An overview of the complete architecture is illustrated in Fig. 3.

1) Convolutional Axial Attention Module (CAAM)

To enhance contextual representation while maintaining computational efficiency, we propose the *Convolutional Axial Attention Module (CAAM)*. This module independently models feature interactions along the horizontal and vertical axes, enabling the network to capture long-range dependencies with reduced complexity—particularly beneficial for dense prediction tasks in remote sensing.

Let $\mathbf{X} \in \mathbb{R}^{C\times H\times W}$ denote the input feature map. The first stage performs channel compression via a 1×1 convolution followed by batch normalization:

$$\mathbf{X}_s = \mathrm{BN}(\mathrm{Conv}_{1\times 1}(\mathbf{X})) \in \mathbb{R}^{C_s\times H\times W}, \quad C_s = \left\lfloor \frac{C}{r} \right\rfloor, \quad (1)$$

where r is a predefined channel reduction ratio.

Height-wise Attention.

To capture vertical dependencies, average and max pooling are applied along the width axis:

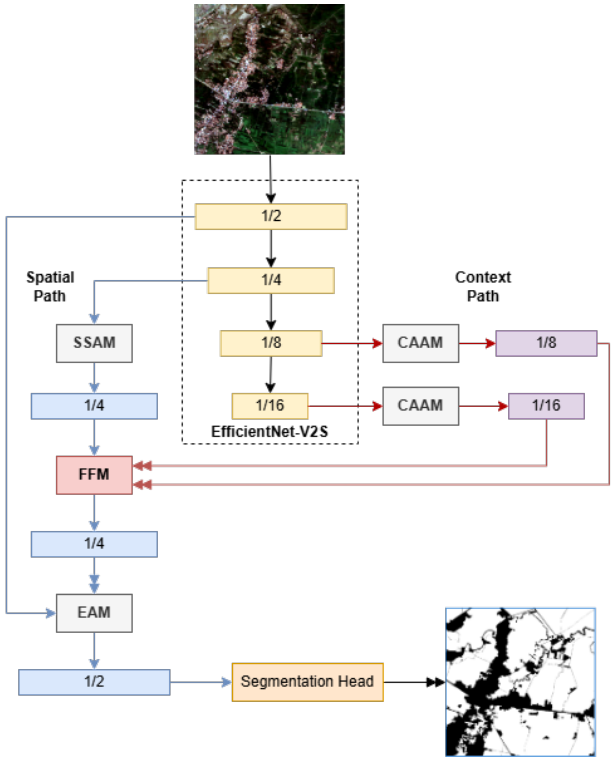


FIGURE 3. The architecture of TerraSegNet, highlighting its key components: Convolutional Axial Attention Module (CAAM), Spectral-Spatial Attention Module (SSAM), and Edge Attention Module (EAM).

$$\mathbf{F}_{\mathrm{h}} = \mathrm{Cat}(\mathrm{AvgPool}_{w}(\mathbf{X}_{s}), \mathrm{MaxPool}_{w}(\mathbf{X}_{s})) \in \mathbb{R}^{2C_{s}\times H\times 1}. \tag{2}$$

The resulting feature is convolved using a shared 1×3 convolution:

$$\mathbf{A}_{\mathrm{h}} = \mathrm{Conv}_{1\times 3}(\mathbf{F}_{\mathrm{h}}) \in \mathbb{R}^{C_{s}\times H\times 1}, \tag{3}$$

and then broadcast along the width to match the spatial dimensions of the input:

$$\hat{\mathbf{A}}_{\mathrm{h}} \in \mathbb{R}^{C_{s}\times H\times W}. \tag{4}$$

Width-wise Attention.

Similarly, pooling is performed along the height axis:

$$\mathbf{F}_{\mathrm{w}} = \mathrm{Cat}(\mathrm{AvgPool}_{h}(\mathbf{X}_{s}), \mathrm{MaxPool}_{h}(\mathbf{X}_{s})) \in \mathbb{R}^{2C_{s}\times 1\times W}, \tag{5}$$

followed by 3×1 convolution:

$$\mathbf{A}_{\mathrm{w}} = \mathrm{Conv}_{3\times 1}(\mathbf{F}_{\mathrm{w}}) \in \mathbb{R}^{C_{s}\times 1\times W}, \tag{6}$$

and broadcast along the height dimension:

$$\hat{\mathbf{A}}_{\mathrm{w}} \in \mathbb{R}^{C_{s}\times H\times W}. \tag{7}$$

Attention Fusion.

The final axial attention map is obtained by summing both components:

$$\mathbf{A} = \hat{\mathbf{A}}_{\mathrm{h}} + \hat{\mathbf{A}}_{\mathrm{w}}. \tag{8}$$

After applying dropout and projecting back to the original channel dimension, we obtain:

$$\mathbf{X}_{\mathrm{attn}} = \mathrm{BN}(\mathrm{Conv}_{1\times 1}(\mathrm{Dropout}(\mathbf{A}))) \in \mathbb{R}^{C\times H\times W}. \tag{9}$$

Context Gate.

To incorporate global semantic information, a context gate $\mathbf{G} \in \mathbb{R}^{C\times 1\times 1}$ is computed via:

$$\mathbf{G} = \sigma\left(W_{2} \cdot \delta(W_{1} \cdot \mathrm{GAP}(\mathbf{X}))\right), \tag{10}$$

where $\mathrm{GAP}(\cdot)$ denotes global average pooling, W_1 and W_2 are 1×1 convolution layers, $\delta(\cdot)$ is the LeakyReLU activation, and $\sigma(\cdot)$ is the sigmoid function.

The attention output is modulated by the gate and combined with the identity via residual connection:

$$\hat{\mathbf{X}} = \mathbf{X}_{\mathrm{attn}} \odot \mathbf{G}, \tag{11}$$

$$\mathbf{Y} = \hat{\mathbf{X}} + \mathbf{X}. \tag{12}$$

CAAM effectively decomposes full 2D attention into two efficient 1D attentions while preserving scene-level semantics through global modulation. This design significantly reduces memory usage compared to full attention mechanisms, making it highly suitable for remote sensing applications involving high-resolution and multi-modal satellite imagery. An architectural overview of CAAM is illustrated in Fig. 4.

2) Spectral–Spatial Attention Module (SSAM)

To enhance feature discrimination in remote sensing segmentation, we propose the *Spectral–Spatial Attention Module (SSAM)*, which integrates spectral group attention, spatial attention, and global semantic feedback in a unified and lightweight design.

Given an input feature map $\mathbf{X} \in \mathbb{R}^{C\times H\times W}$, SSAM first partitions the input into G non-overlapping channel groups $\{\mathbf{X}_1, \mathbf{X}_2, \ldots, \mathbf{X}_G\}$, where each group $\mathbf{X}_g \in \mathbb{R}^{C_g\times H\times W}$ and $C = G \cdot C_g$.

Spectral Group Attention.

For each group $\mathbf{X}_g$, spectral attention is computed via a global average pooling operation followed by a bottleneck structure consisting of two 1×1 convolutions and non-linear activation:

$$\mathbf{A}_{\mathrm{spec}}^{(g)} = \sigma\left(W_{2}^{(g)} \cdot \delta\left(W_{1}^{(g)} \cdot \mathrm{GAP}(\mathbf{X}_{g})\right)\right), \tag{13}$$

where $\mathrm{GAP}(\cdot)$ denotes global average pooling, $W_1^{(g)}$ and $W_2^{(g)}$ are learnable 1×1 convolutions, $\delta(\cdot)$ is the LeakyReLU activation function, and $\sigma(\cdot)$ is the sigmoid function.

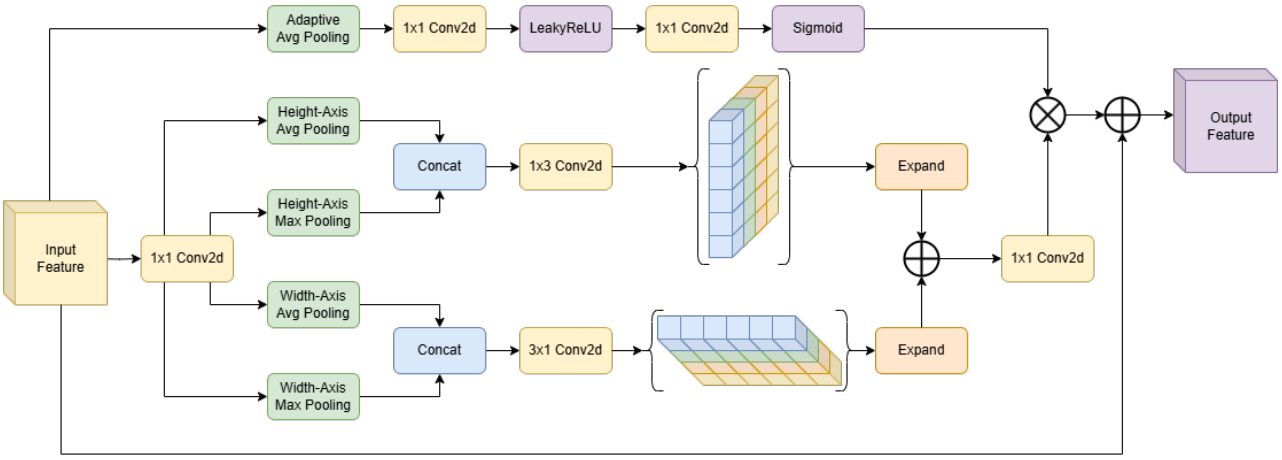


FIGURE 4. Architecture of the Convolutional Axial Attention Module (CAAM), illustrating the sequential height- and width-wise attention pathways and global context modulation for efficient long-range spatial dependency modelling.

The spectral-refined feature for group g is then:

$$\tilde{\mathbf{X}}_g = \mathbf{X}_g \odot \mathbf{A}_{\text{spec}}^{(g)}, \tag{14}$$

where $\odot$ denotes element-wise multiplication. All group-wise outputs are concatenated to reconstruct the full-channel spectral-attended feature map:

$$\tilde{\mathbf{X}} = \text{Concat}(\tilde{\mathbf{X}}_1, \ldots, \tilde{\mathbf{X}}_G). \tag{15}$$

Spatial Attention.

To model spatial dependencies, a depthwise separable convolution followed by batch normalization and a sigmoid activation is applied:

$$\mathbf{A}_{\text{spa}} = \sigma\left(\text{BN}(\text{DWConv}(\mathbf{X}))\right), \tag{16}$$

yielding a spatial attention map $\mathbf{A}_{\text{spa}} \in \mathbb{R}^{C\times H\times W}$.

Semantic Attention Gate.

To capture global semantics, a semantic gate is generated via a two-layer MLP using the original input:

$$\mathbf{A}_{\text{sem}} = \sigma\left(W_4 \cdot \delta\left(W_3 \cdot \text{GAP}(\mathbf{X})\right)\right), \tag{17}$$

where W_3 and W_4 are shared 1×1 convolution layers.

Attention Fusion and Residual Learning.

The attention-modulated feature is obtained by combining all three attention maps:

$$\mathbf{X}' = \tilde{\mathbf{X}} \odot \mathbf{A}_{\text{spa}} \odot \mathbf{A}_{\text{sem}}. \tag{18}$$

A final projection layer refines the output:

$$\mathbf{Y} = \text{Dropout}\left(\text{BN}(\text{Conv}_{1\times1}(\mathbf{X}'))\right). \tag{19}$$

Lastly, a residual connection is applied to preserve the original signal:

$$\hat{\mathbf{Y}} = \mathbf{Y} + \mathbf{X}. \tag{20}$$

This design enables the model to adaptively emphasize salient spectral and spatial patterns while maintaining computational efficiency. SSAM is agnostic to imaging modality and resolution, making it applicable to a wide range of remote sensing scenarios, including both optical and radar imagery. An overview of the SSAM architecture is shown in Fig. 5.

3) Edge Attention Module (EAM)

To refine coarse semantic predictions using fine-grained spatial cues, we propose the *Edge Attention Module (EAM)*, which emphasizes boundary-level information through spatial and edge-specific attention mechanisms. EAM is designed to enhance upsampled coarse prediction maps using guidance from intermediate high-resolution feature representations.

Let $\mathbf{F}_{\text{in}} \in \mathbb{R}^{C_{\text{in}}\times H\times W}$ denote the intermediate feature map extracted from the spatial path, and let $\mathbf{M}_{\text{coarse}} \in \mathbb{R}^{C_o\times \frac{H}{2}\times \frac{W}{2}}$ represent the coarse segmentation output from the decoder path, where C_o is the number of semantic classes.

Feature Projection and Structure Enhancement.

EAM first projects the input feature into a compact representation using 1×1 convolution, batch normalization, and LeakyReLU activation:

$$\mathbf{F}_0 = \delta(\text{BN}(\text{Conv}_{1\times1}(\mathbf{F}_{\text{in}}))) \in \mathbb{R}^{C_o\times H\times W}. \tag{21}$$

Next, a depthwise separable convolution block enriches local structural features:

$$\mathbf{F}_s = \delta\left(\text{BN}\left(\text{Conv}_{1\times1}\left(\delta\left(\text{BN}\left(\text{DWConv}_{3\times3}(\mathbf{F}_0)\right)\right)\right)\right)\right). \tag{22}$$

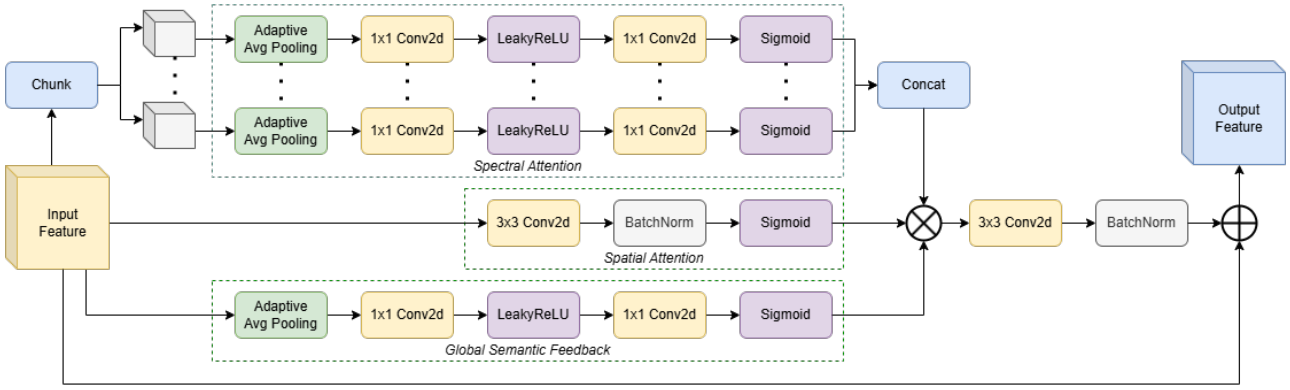


FIGURE 5. Architecture of the Spectral-Spatial Attention Module (SSAM), consisting of spectral group recalibration, spatial attention, and semantic feedback pathways. Outputs from these branches are fused to selectively enhance informative features across both spectral and spatial dimensions.

Spatial Attention.

To identify salient regions, a spatial attention map is generated by concatenating channel-wise average and max pooling, followed by a 3×3 convolution and sigmoid activation:

$$\mathbf{A}_s = \sigma\left(\mathrm{Conv}_{3\times3}\left(\mathrm{Cat}\left[\mathrm{AvgPool}_c(\mathbf{F}_s), \mathrm{MaxPool}_c(\mathbf{F}_s)\right]\right)\right), \tag{23}$$

$$\tilde{\mathbf{F}} = \mathbf{F}_s \odot \mathbf{A}_s, \tag{24}$$

where $\odot$ denotes element-wise multiplication.

Edge Attention.

To further emphasize object boundaries, an edge attention map is computed using a 1×1 convolution followed by a sigmoid function:

$$\mathbf{A}_e = \sigma(\mathrm{Conv}_{1\times1}(\tilde{\mathbf{F}})). \tag{25}$$

Edge-aware Refinement.

The coarse prediction map is first upsampled by a factor of 2:

$$\mathbf{M}_{\mathrm{up}} = \mathrm{Upsample}(\mathbf{M}_{\mathrm{coarse}}; \mathrm{scale} = 2), \tag{26}$$

and then modulated using the edge attention map to refine class boundaries:

$$\mathbf{M}_{\mathrm{refined}} = \mathbf{M}_{\mathrm{up}} \odot (1 + \mathbf{A}_e). \tag{27}$$

This formulation allows the network to selectively boost prediction confidence near edge regions without introducing significant computational overhead. EAM thus functions as an effective refinement mechanism that enhances structural fidelity—particularly valuable in dense segmentation scenarios such as rooftops, terrain classes, and agricultural field boundaries commonly observed in remote sensing imagery. The architectural overview of EAM is illustrated in Fig. 6.

D. HYPERPARAMETER OPTIMIZATION

All experiments are conducted using PyTorch 2.3.1+cu121 on a workstation equipped with an NVIDIA GeForce RTX 4060 GPU (8 GB VRAM). Careful hyperparameter optimization is essential to ensure stable training and maximize segmentation performance, particularly for challenging tasks such as cloud detection. The following strategies and settings are employed throughout the training process:

1) **Loss Function:** Boundary Loss [61] and Tversky Loss [62] are employed to enhance the delineation between foreground and background classes. Boundary Loss is particularly effective in remote sensing applications where precise boundary localization is critical, whereas Tversky Loss provides a flexible trade-off between false positives and false negatives, making it well-suited for handling class imbalance.
2) **Optimizer:** The AdamW optimizer [63] is selected with an initial learning rate of $1e^{-3}$ and a weight decay of $1e^{-4}$. This configuration helps maintain balanced weight updates, especially when dealing with class imbalance in remote sensing datasets.
3) **Learning Rate Scheduling:** Cosine annealing with warm restarts [64] is applied to periodically reset the learning rate following a cosine decay curve. This strategy encourages the model to escape local minima and explore alternative solutions during the optimization process.
4) **Batch Strategy and Memory Efficiency:** To address GPU memory limitations, gradient accumulation is employed. The model processes 8 samples at a time and combines their gradients over 4 iterations. This approach effectively simulates training with 32 samples, without requiring larger memory. In addition, Automatic Mixed Precision (AMP) is enabled. AMP reduces memory usage and speeds up training by using lower precision where possible, while keeping critical calculations in full precision to maintain accuracy.
5) **Early Stopping:** Training is terminated early if the validation mean Intersection over Union (mIoU) does

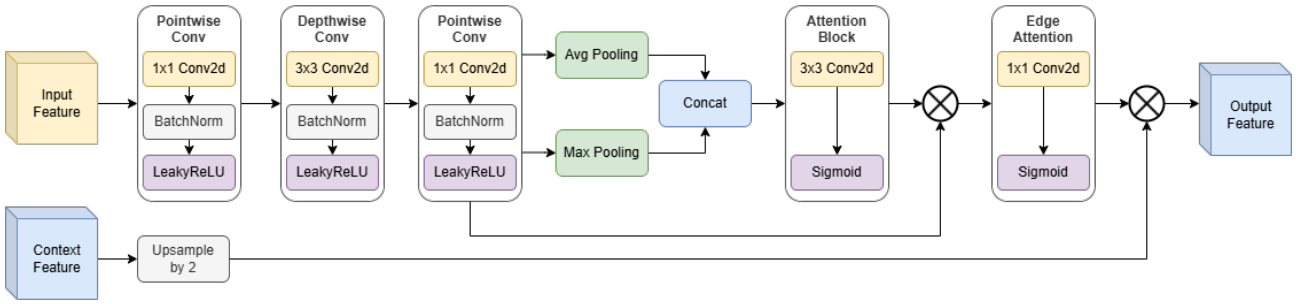


FIGURE 6. **Architecture of the Edge Attention Module (EAM), which refines the coarse segmentation prediction by combining low-level spatial features and attention-guided edge enhancement. The module integrates spatial attention and channel-wise edge maps to highlight object boundaries and improve prediction accuracy at fine scales.**

not improve over a predefined number of epochs. This prevents overfitting and avoids unnecessary computational cost.

E. EVALUATION METRICS

To ensure a comprehensive performance assessment, we evaluate all segmentation models using three widely adopted metrics: F1-score, pixel accuracy, and mean Intersection over Union (mIoU). These metrics are derived from the confusion matrix, which contains the quantities of true positives (TP), false positives (FP), false negatives (FN), and true negatives (TN). Fig. 7 illustrates a sample confusion matrix, providing a breakdown of class-wise prediction outcomes. The evaluation metrics are formally defined as follows:

$$\text{Precision} = \frac{TP}{TP + FP} \tag{28}$$

$$\text{Recall} = \frac{TP}{TP + FN} \tag{29}$$

$$\text{F1-score} = \frac{2 \times \text{Precision} \times \text{Recall}}{\text{Precision} + \text{Recall}} \tag{30}$$

$$\text{Pixel Accuracy} = \frac{TP + TN}{TP + FP + FN + TN} \tag{31}$$

$$\text{IoU} = \frac{TP}{TP + FP + FN} \tag{32}$$

$$\text{mIoU} = \frac{1}{n} \sum_{i=1}^{n} \text{IoU}_i \tag{33}$$

These metrics serve complementary roles in evaluating semantic segmentation performance:

- **Precision** quantifies the proportion of correctly predicted positive samples out of all predicted positives. It is especially important in applications where false positives must be minimized.
- **Recall**, or sensitivity, measures the model's ability to correctly detect all actual positive instances. It is critical in scenarios where false negatives are costly or undesirable.

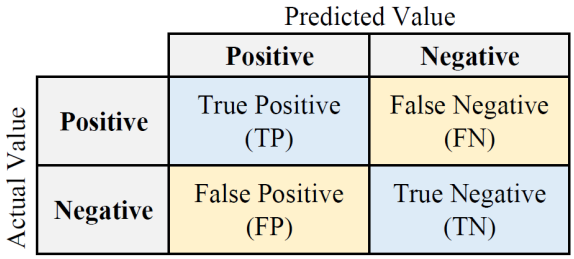


FIGURE 7. **Confusion matrix of class-wise segmentation performance, illustrating the model's effectiveness in distinguishing between different semantic categories across the evaluation datasets.**

- The **F1-score** provides a harmonic mean between precision and recall, making it well-suited for imbalanced class distributions where a trade-off must be achieved.
- **Pixel Accuracy** assesses the proportion of correctly classified pixels over the total number of pixels. However, this metric may be biased in datasets with dominant background or majority classes.
- **Intersection over Union (IoU)** evaluates the overlap between predicted and ground truth segmentation masks, offering a more discriminative metric for spatial correspondence.
- **Mean IoU (mIoU)** extends the IoU metric to the multi-class setting by averaging the IoU across all semantic categories, serving as a comprehensive performance indicator.

In this study, we adopt weighted versions of F1-score and mIoU, where each class-specific score is weighted according to the number of ground-truth pixels belonging to that class. This pixel-frequency-aware strategy is particularly important in remote sensing datasets, which are often characterized by severe class imbalance. For instance, background or non-crop regions typically occupy a large proportion of the image, whereas smaller but critical classes (e.g., water bodies, built-up areas) cover far fewer pixels. Using unweighted (macro-averaged) metrics in such cases may artificially amplify the influence of rare classes and introduce high variance in the results, especially when some categories contain only a handful of pixels in a given scene. In contrast, weighted metrics

proportionally account for the spatial extent of each class, thereby providing a more stable and realistic measure of overall segmentation quality.

Moreover, since not all semantic classes appear in every test image, averaging is performed only over the classes present in the ground truth of each image. This avoids penalizing the model for categories that are absent in a scene and ensures that the evaluation reflects the model's effectiveness on the classes that are actually observable. Consequently, the use of weighted metrics with this class-presence-aware policy yields an evaluation framework that is both reliable and representative of real-world conditions, where spatial imbalance and scene variability are inherent challenges in semantic segmentation of satellite imagery.

IV. EXPERIMENT RESULTS

A. SEGMENTATION RESULTS ON SPARCS DATASET

The SPARCS dataset is a benchmark for cloud segmentation, comprising three semantic classes: *cloud*, *clear sky*, and *cloud shadow*. All evaluated models were trained and tested using RGB input channels. Table 3 presents a detailed quantitative comparison of segmentation accuracy and computational efficiency across baseline and proposed models.

Among all the tested models, **TerraSegNet** outperforms competitors across all core metrics, achieving an **F1-score of 0.9516**, **pixel accuracy of 0.9504**, and **mIoU of 0.8608**. This superior performance underscores TerraSegNet's robustness in accurately distinguishing between cloud types and clear sky under varying atmospheric and surface conditions. Notably, TerraSegNet accomplishes this with only **5.80 million parameters**, offering a substantial reduction in model complexity compared to heavier models such as HR-cloud-Net (74.53M) and CDNetV2 (67.59M), while still surpassing them in segmentation performance.

In addition to accuracy, TerraSegNet maintains a competitive inference speed of **10.51 fps**, balancing precision with runtime efficiency. Although lightweight models such as Fast-SCNN and SeaFormer++ deliver higher throughput (>13 fps), they incur significant accuracy penalties, with mIoU values falling below 0.76. In contrast, TerraSegNet offers the most favorable trade-off between mIoU and throughput, as illustrated in Fig. 8.

Transformer-based models such as UNetFormer, CMTFNet, and CMLFormer show strong segmentation capabilities (F1-scores $\approx$ 0.94), but still fall slightly behind TerraSegNet in both mIoU and parameter efficiency. Likewise, conventional models like DeepLabv3+ and U-Net, despite their widespread adoption, exhibit limitations in scalability and generalization when benchmarked on cloud segmentation tasks with high intra-class variability.

Overall, TerraSegNet demonstrates superior performance not only in terms of segmentation quality but also in model compactness and inference efficiency. These characteristics make it a strong candidate for real-time, large-scale cloud detection systems in remote sensing pipelines, especially where computational resources and latency are operational constraints.

TABLE 3. Quantitative segmentation results on SPARCS dataset, including accuracy metrics and computational cost. TerraSegNet achieves the best trade-off between accuracy and efficiency. Best results are highlighted in bold.

Model	F1-score↑	Pixel Accuracy↑	mIoU↑	GFLOPs↓	FPS↑
Conventional Models					
U-Net	0.9344	0.9350	0.8213	32.97	13.80
DeepLabv3+	0.9413	0.9418	0.8371	30.20	13.05
HRNet	0.9170	0.9165	0.7873	17.91	11.70
SegFormer	0.9161	0.9140	0.7759	7.85	13.61
Bilateral Models					
BiSeNetV1	0.9176	0.9230	0.7924	24.47	13.05
BiSeNetV2	0.9125	0.9121	0.7703	12.26	12.14
Fast-SCNN	0.8909	0.9019	0.7587	0.73	13.56
SeaFormer++	0.8550	0.9026	0.6273	2.08	11.09
Remote Sensing-Specific Models					
HR-cloud-Net	0.9084	0.9112	0.7702	197.87	8.26
CDnetV2	0.9089	0.9151	0.7880	70.13	10.40
UNetFormer	0.9400	0.9398	0.8369	11.74	12.87
CMLFormer	0.9412	0.9409	0.8370	14.28	12.38
CMTFNet	0.9394	0.9378	0.8310	10.64	11.96
TerraSegNet	**0.9516**	**0.9504**	**0.8608**	22.30	10.51

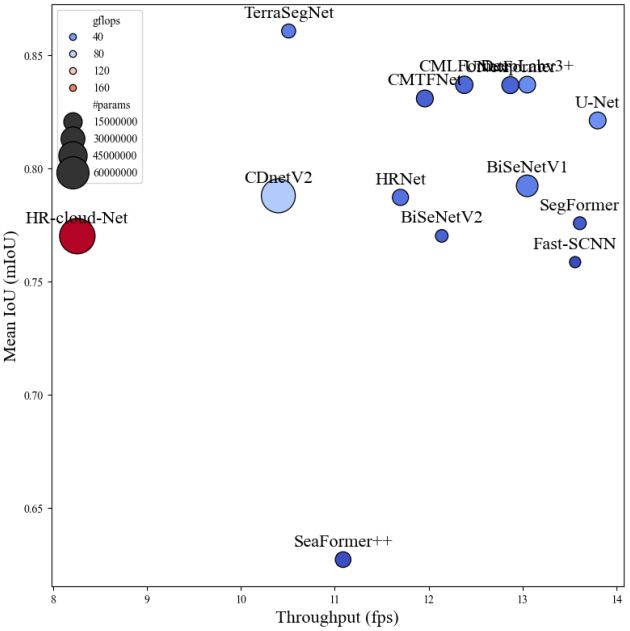


FIGURE 8. Trade-off between mIoU and throughput (fps) across segmentation models on SPARCS dataset. TerraSegNet achieves the best balance between accuracy and inference efficiency.

B. SEGMENTATION RESULTS ON WHUS2-CD+ DATASET

The WHUS2-CD+ dataset is tailored for cloud segmentation using multi-spectral Sentinel-2 imagery, which includes Red-Green-Blue (RGB), Near-Infrared (NIR), and Normalized Difference Vegetation Index (NDVI) bands. The binary segmentation task distinguishes between *cloud* and *clear sky* pixels. To ensure consistency in benchmarking, all models were trained and evaluated under the same input configura-

tion using the four input channels.

Table 4 provides a comprehensive comparison of segmentation performance and computational cost across a wide range of architectures. **TerraSegNet** achieves state-of-the-art results on this dataset, reporting an **F1-score of 0.9872**, **pixel accuracy of 0.9880**, and **mIoU of 0.9364**, thereby outperforming all other evaluated models. These results highlight the model's superior ability to generalize across different input modalities and accurately segment cloud structures under diverse atmospheric and surface conditions.

Although CMLFormer slightly surpasses TerraSegNet in F1-score (0.9873), it falls behind in mIoU and pixel accuracy, suggesting a less consistent prediction over the spatial domain. Moreover, TerraSegNet attains this performance with only **5.80 million parameters** and a moderate Giga Floating Point Operations Per Second (GFLOPs) of **22.32**, offering a more compact and efficient design than heavier alternatives like CDnetV2 (67.59M) and HR-cloud-Net (74.53M). In addition, TerraSegNet delivers a practical inference speed of **32.29 fps**, supporting near real-time deployment on edge or embedded platforms.

Lightweight models such as Fast-SCNN and SegFormer exhibit faster throughput (>37 fps), but their mIoU drops significantly (e.g., Fast-SCNN at 0.9176), underscoring the accuracy-efficiency trade-off. Models based on Transformer-CNN hybrids (UNetFormer, CMTFNet, and CMLFormer) show strong performance across all metrics but demand more computational resources than TerraSegNet for comparable results.

These findings reinforce the effectiveness of the proposed architecture in achieving high segmentation accuracy while maintaining inference efficiency and model compactness. As illustrated in Fig. 9, TerraSegNet consistently outperforms other models in the accuracy-efficiency trade-off space, confirming its suitability for real-time cloud detection applications in remote sensing pipelines.

TABLE 4. Quantitative segmentation results on WHUS2-CD+ dataset, including accuracy metrics and computational cost. TerraSegNet achieves the best trade-off between segmentation accuracy and efficiency. Best results are highlighted in bold.

Model	F1-score↑	Pixel Accuracy↑	mIoU↑	GFLOPs↓	FPS↑
Conventional Models					
U-Net	0.9864	0.9871	0.9317	33.37	48.73
DeepLabv3+	0.9865	0.9875	0.9341	30.61	52.91
HRNet	0.9859	0.9861	0.9291	17.97	30.80
SegFormer	0.9865	0.9868	0.9309	7.90	37.90
Bilateral Models					
BiSeNetV1	0.9859	0.9867	0.9308	25.29	46.01
BiSeNetV2	0.9846	0.9845	0.9214	12.35	36.88
Fast-SCNN	0.9833	0.9843	0.9176	0.76	61.87
SeaFormer++	0.9693	0.9868	0.8738	2.10	21.45
Remote Sensing-Specific Models					
HR-cloud-Net	0.9840	0.9843	0.9204	197.94	14.84
CDnetV2	0.9860	0.9863	0.9288	70.20	24.35
UNetFormer	0.9868	0.9875	0.9345	12.15	40.72
CMLFormer	**0.9873**	0.9878	0.9361	14.69	36.84
CMTFNet	0.9869	0.9873	0.9334	11.05	31.63
TerraSegNet	0.9872	**0.9880**	**0.9364**	22.32	32.29

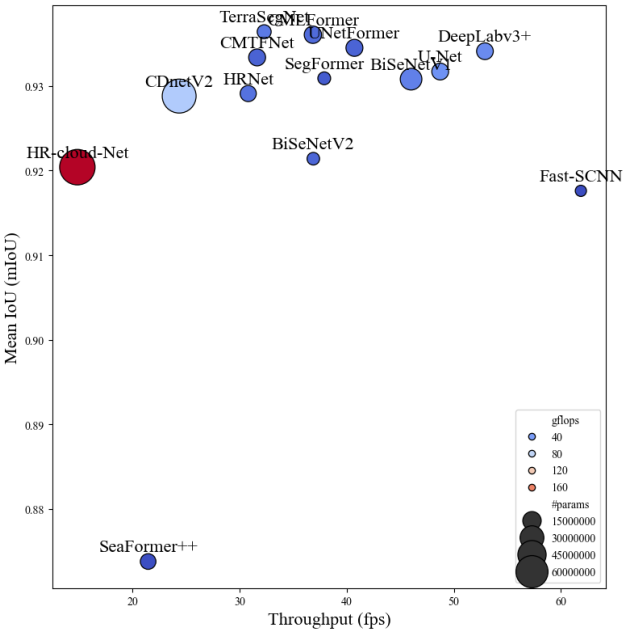


FIGURE 9. Trade-off between mIoU and throughput (fps) across segmentation models on WHUS2-CD+ dataset. TerraSegNet achieves optimal balance between accuracy and runtime performance.

C. SEGMENTATION RESULTS ON PV08 DATASET

The PV08 dataset is constructed for photovoltaic (PV) panel segmentation using high-resolution RGB imagery acquired from the Gaofen-2 satellite. The task is framed as a binary semantic segmentation problem, targeting the discrimination between *photovoltaic panel* and *background* regions, which is critical for large-scale solar energy mapping and infrastructure monitoring.

Table 5 reports the quantitative results of multiple segmentation models, highlighting both accuracy and computational efficiency. **TerraSegNet** demonstrates the highest overall performance, achieving an **F1-score of 0.9803**, **pixel accuracy of 0.9806**, and **mIoU of 0.9617**, thereby surpassing all baseline models on this dataset. These results indicate TerraSegNet's robustness in segmenting PV panels across diverse backgrounds and structural layouts, including rooftop and ground-mounted configurations.

Although competing architectures such as DeepLabv3+, CMLFormer, and UNetFormer also exhibit strong segmentation performance with F1-scores above 0.976, they fall short in mIoU and incur either higher model complexity or lower inference efficiency. For instance, CMLFormer reaches an F1-score of 0.9784 and mIoU of 0.9581 but requires over 12 million parameters and 14.28 GFLOPs. In contrast, TerraSegNet achieves superior accuracy with a more compact design of **5.80 million parameters** and modest GFLOPs (**22.29**), reflecting an efficient architecture optimized for real-world deployment.

Lightweight models such as Fast-SCNN, BiSeNetV2, and SegFormer provide faster throughput (up to 14.93 fps), but

TABLE 5. Quantitative segmentation results on PV08 dataset including accuracy metrics and computational cost. TerraSegNet consistently outperforms other models while maintaining a compact and efficient architecture. Best results are highlighted in bold.

Model	F1-score↑	Pixel Accuracy↑	mIoU↑	GFLOPs↓	FPS↑
Conventional Models					
U-Net	0.9757	0.9760	0.9530	32.96	11.75
DeepLabv3+	0.9772	0.9775	0.9559	30.20	12.76
HRNet	0.9755	0.9758	0.9526	17.90	11.85
SegFormer	0.9720	0.9723	0.9461	7.85	13.59
Bilateral Models					
BiSeNetV1	0.9745	0.9748	0.9508	24.47	14.22
BiSeNetV2	0.9724	0.9727	0.9468	12.26	13.00
Fast-SCNN	0.9696	0.9705	0.9418	0.73	14.93
SeaFormer++	0.9690	0.9698	0.9406	2.08	11.34
Remote Sensing-Specific Models					
HR-cloud-Net	0.9643	0.9649	0.9319	197.87	8.00
CDnetV2	0.9731	0.9735	0.9481	70.13	10.58
UNetFormer	0.9768	0.9771	0.9550	11.74	12.20
CMLFormer	0.9784	0.9787	0.9581	14.28	14.12
CMTFNet	0.9772	0.9775	0.9558	10.64	14.37
TerraSegNet	**0.9803**	**0.9806**	**0.9617**	22.29	12.01

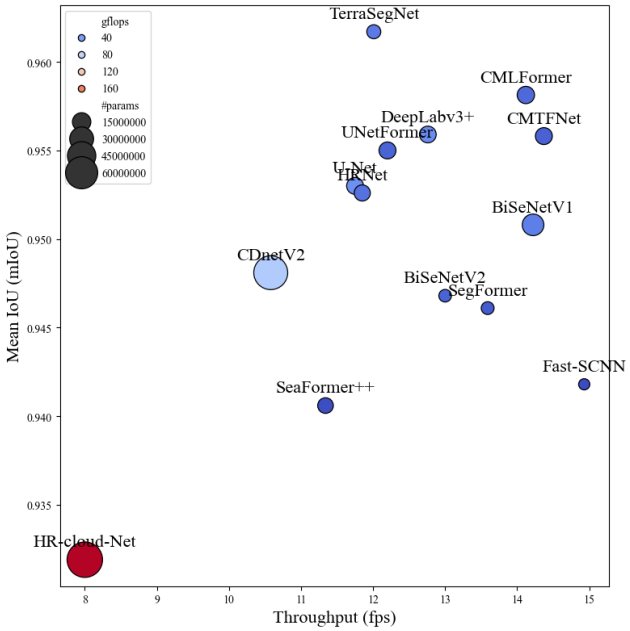


FIGURE 10. Trade-off between mIoU and throughput (fps) across segmentation models on PV08 dataset. TerraSegNet achieves the best balance of segmentation accuracy and inference efficiency.

their mIoU drops by more than 2–3%, revealing the inherent trade-off between model compactness and fine-grained segmentation accuracy. Moreover, larger models like HR-cloud-Net and CDNetV2, despite their substantial parameter budgets, exhibit underwhelming accuracy and inferior efficiency, further underscoring the advantage of task-specific architectural design.

As visualized in Fig. 10, TerraSegNet achieves the most favorable balance in the accuracy-speed trade-off, placing it firmly as a state-of-the-art solution for PV panel segmentation. Its low latency and high precision make it particularly attractive for scalable remote sensing applications that require both spatial detail and real-time inference capabilities.

D. SEGMENTATION RESULTS ON PLOT-RICE V1.0 DATASET

The Plot-Rice v1.0 dataset is developed for rice field mapping using dual-polarized SAR data, particularly DpRVIc—a vegetation index derived from Sentinel-1 backscatter. This dataset poses a binary segmentation task with two semantic classes: *rice area* and *background*. Its relevance lies in enabling accurate agricultural monitoring under all-weather conditions, especially in regions with persistent cloud cover.

As shown in Table 6, **TerraSegNet** achieves the highest performance across all evaluation metrics, attaining an **F1-score of 0.8748**, **pixel accuracy of 0.8757**, and a **mean IoU of 0.7777**. These results underscore the model's capability to effectively capture spatial and texture patterns in radar-derived vegetation structures, which are often subtle and complex in backscatter imagery.

When compared to classical architectures, U-Net and DeepLabv3+ yield competitive F1-scores (0.8689 and 0.8681, respectively) but lag behind in mIoU and computational efficiency. Although both models exceed TerraSegNet in parameter count and GFLOPs, they fail to convert this complexity into a performance advantage, suggesting inefficiencies in exploiting SAR features.

Lightweight models such as SegFormer, BiSeNetV2, and Fast-SCNN offer high throughput, with Fast-SCNN leading at 67.49 fps. However, this speed advantage comes at the expense of reduced accuracy, with mIoU values falling below 0.75. This pattern reflects the trade-off often observed in real-time segmentation tasks: compact architectures must carefully balance model expressiveness and efficiency to maintain robustness in high-frequency spectral or backscatter domains.

Meanwhile, heavier models like HR-cloud-Net and CDNetV2 exhibit neither top-tier accuracy nor acceptable runtime performance for large-scale inference. In contrast, TerraSegNet, with only **5.80 million parameters** and moderate FLOPs (5.61), offers a strong balance between accuracy and speed (**32.82 fps**). The effectiveness of TerraSegNet is further highlighted in Fig. 11, which visualizes the trade-off between throughput and mIoU.

Overall, these results validate the suitability of TerraSegNet for operational agricultural monitoring using radar remote sensing. Its ability to maintain high accuracy while remaining computationally lean makes it a promising candidate for national-scale rice mapping pipelines and adaptive crop monitoring systems under cloudy or monsoon-dominated conditions.

E. SEGMENTATION RESULTS ON AIR-POLSAR-SEG-2.0 DATASET

The AIR-PolSAR-Seg-2.0 dataset is a benchmark for high-resolution terrain segmentation using full-polarimetric SAR imagery acquired from Gaofen-3. The dataset includes six terrain categories: *water*, *vegetation*, *bare land*, *roads*, *build-*

TABLE 6. Quantitative segmentation results on Plot-Rice v1.0 dataset including accuracy metrics and computational cost. TerraSegNet outperforms all baselines with an optimal trade-off between accuracy and efficiency. Best results are highlighted in bold.

Model	F1-score↑	Pixel Accuracy↑	mIoU↑	GFLOPs↓	FPS↑
Conventional Models					
U-Net	0.8689	0.8691	0.7683	8.70	61.93
DeepLabv3+	0.8681	0.8692	0.7670	8.01	57.00
HRNet	0.8597	0.8605	0.7542	4.56	38.57
SegFormer	0.8601	0.8609	0.7547	1.82	52.40
Bilateral Models					
BiSeNetV1	0.8610	0.8621	0.7562	7.04	52.76
BiSeNetV2	0.8522	0.8530	0.7427	3.17	46.08
Fast-SCNN	0.8564	0.8581	0.7491	0.22	67.49
SeaFormer++	0.8499	0.8560	0.7392	0.55	31.55
Remote Sensing-Specific Models					
HR-cloud-Net	0.8665	0.8670	0.7647	49.56	12.73
CDnetV2	0.8682	0.8690	0.7674	18.47	25.42
UNetFormer	0.8650	0.8654	0.7622	3.40	43.13
CMLFormer	0.8645	0.8652	0.7616	4.03	33.16
CMTFNet	0.8687	0.8696	0.7680	3.14	29.01
TerraSegNet	**0.8748**	**0.8757**	**0.7777**	5.61	32.82

TABLE 7. Quantitative segmentation results on AIR-PolSAR-Seg-2.0 dataset including accuracy metrics and computational cost. TerraSegNet delivers top-tier accuracy with an efficient parameter footprint. Best results are highlighted in bold.

Model	F1-score↑	Pixel Accuracy↑	mIoU↑	GFLOPs↓	FPS↑
Conventional Models					
U-Net	0.9165	0.9389	0.7662	33.00	13.90
DeepLabv3+	0.9477	0.9629	0.8818	30.22	15.48
HRNet	0.8804	0.9125	0.6971	17.93	12.77
SegFormer	0.8263	0.8406	0.5941	7.87	10.21
Bilateral Models					
BiSeNetV1	0.9298	0.9462	0.8371	24.48	14.62
BiSeNetV2	0.8546	0.9013	0.6386	12.28	13.43
Fast-SCNN	0.8440	0.8643	0.6444	0.73	12.62
SeaFormer++	0.7859	0.8566	0.4629	2.08	8.44
Remote Sensing-Specific Models					
HR-cloud-Net	0.8549	0.8961	0.6567	197.90	7.02
CDnetV2	0.8149	0.8274	0.5966	70.15	8.64
UNetFormer	0.9355	0.9427	0.8320	11.75	9.99
CMLFormer	0.9494	0.9553	0.8705	14.29	10.36
CMTFNet	0.9404	0.9515	0.8493	10.65	8.36
TerraSegNet	**0.9663**	**0.9725**	**0.9207**	22.34	11.12

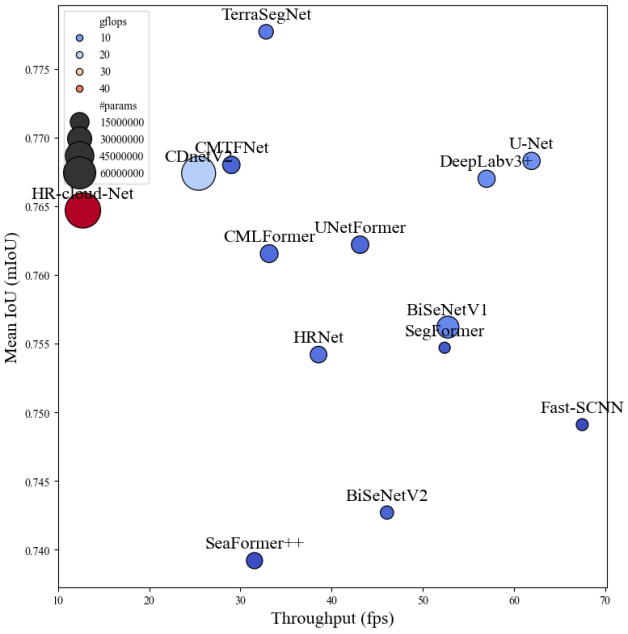


FIGURE 11. Trade-off between mIoU and throughput (fps) across segmentation models on Plot-Rice v1.0 dataset. TerraSegNet achieves both the highest accuracy and competitive inference speed.

ings, and *mountains*, posing a challenging multi-class classification task due to the complex scattering behavior and heterogeneous backscatter patterns in PolSAR data.

Table 7 reports the quantitative results of all models. **TerraSegNet** significantly outperforms all baselines, achieving an **F1-score of 0.9663**, **pixel accuracy of 0.9725**, and a **mean IoU of 0.9207**. These results confirm its strong capacity to distinguish diverse terrain types in the presence of speckle noise and backscatter ambiguity commonly encountered in PolSAR-based classification.

Although DeepLabv3+ also delivers robust performance (F1-score: 0.9477, mIoU: 0.8818), it exhibits higher computational cost. Likewise, BiSeNetV1 and CMLFormer show competitive accuracy but do not surpass TerraSegNet in either accuracy or efficiency. UNetFormer and CMTFNet provide a reasonable trade-off but are still outperformed in core segmentation metrics.

Lightweight models such as SegFormer, BiSeNetV2, and Fast-SCNN excel in inference speed, but suffer from notably lower mIoU scores (below 0.65), indicating that model simplicity compromises discriminative power on complex PolSAR textures. Moreover, heavy models such as HR-cloud-Net and CDNetV2 incur high GFLOPs yet underperform in both accuracy and throughput, highlighting poor parameter efficiency.

Despite its compact architecture with only **5.80 million parameters**, TerraSegNet achieves high segmentation fidelity and moderate computational cost (**22.34 GFLOPs, 11.12 fps**), making it well-suited for large-scale terrain mapping applications using polarimetric SAR. Fig. 12 visually confirms TerraSegNet's superior balance between accuracy and efficiency across the benchmark.

The overall results demonstrate that TerraSegNet effectively generalizes to complex remote sensing modalities like PolSAR and offers practical viability for operational deployment in ground terrain analysis and land-use monitoring.

F. ABLATION STUDY

To evaluate the contributions of individual components within **TerraSegNet**, we conducted a series of ablation experiments on the AIR-PolSAR-Seg-2.0 dataset. Table 8 summarizes the results obtained using various combinations of the proposed attention modules—Convolutional Axial Attention Module (CAAM), Spectral-Spatial Attention Module (SSAM), and Edge Attention Module (EAM)—along with different encoder backbones including both CNN-based and transformer-

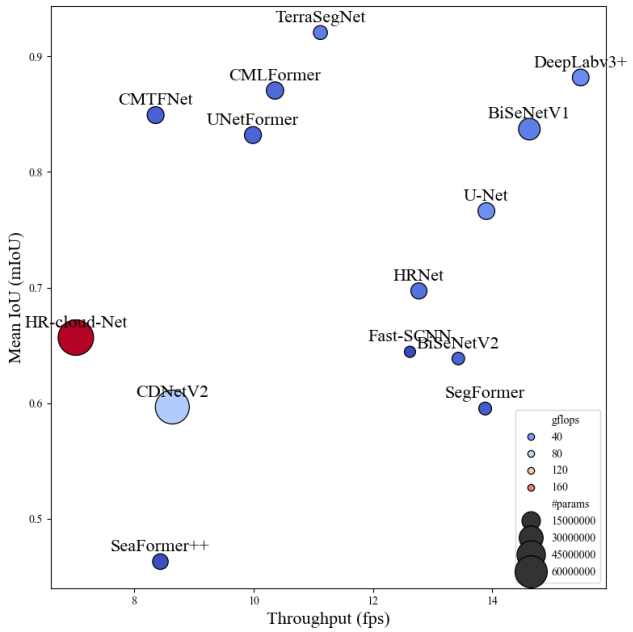


FIGURE 12. Trade-off between mIoU and throughput (fps) across segmentation models on AIR-PolSAR-Seg-2.0 dataset. TerraSegNet consistently delivers the best accuracy-efficiency trade-off.

based alternatives.

1) Contribution of Attention Modules. The baseline configuration using *EfficientNetV2-S* without any attention modules already yields strong results, achieving an F1-score of 0.9593, pixel accuracy of 0.9668, and mean IoU (mIoU) of 0.9017. This highlights the strength of EfficientNetV2-S as a compact yet expressive feature extractor. When CAAM is added, pixel accuracy improves to 0.9681 and mIoU slightly increases to 0.9022, suggesting better context aggregation. SSAM alone offers comparable pixel-level benefits, whereas EAM alone does not yield consistent improvements, indicating its effectiveness is maximized when used in combination. The full configuration integrating CAAM, SSAM, and EAM delivers the best performance, with an F1-score of **0.9663**, pixel accuracy of **0.9725**, and mIoU of **0.9207**, validating the synergistic effects of combining multi-level attention mechanisms.

2) Comparison Across CNN-Based Encoders. We compare EfficientNetV2-S with alternative convolutional backbones, including EfficientNet-B2 [65], ShuffleNetV2 [66], RegNet-Y800MF [67], and ResNet-34 [68]—each integrated with all three attention modules. Among these, ResNet-34 provides reasonably high accuracy (F1: 0.9470; mIoU: 0.8674) but at the cost of 9.04M parameters and 31.12 GFLOPs. ShuffleNetV2, although lightweight, struggles to maintain segmentation quality (F1: 0.8908; mIoU: 0.7403). EfficientNetV2-S achieves the best trade-off, balancing accuracy and efficiency with only 5.80M parameters and 22.34 GFLOPs.

3) Comparison Against Transformer-Based Encoders. To further explore architectural efficiency, we evaluate two transformer-based encoders—*MaxViT-T* [69] and *Swin-V2-T* [70]—each coupled with CAAM, SSAM, and EAM. Although Swin-V2-T achieves competitive accuracy (F1: 0.9366; mIoU: 0.8486), it incurs the highest computational burden (37.18 GFLOPs). MaxViT-T offers lower complexity (7.45 GFLOPs) but still falls short of EfficientNetV2-S in segmentation performance. These findings underscore that although transformer-based models are effective, their advantages diminish when computational efficiency and real-time applicability are prioritized.

Overall, the ablation results confirm the effectiveness of each proposed module and reinforce the selection of EfficientNetV2-S as the optimal encoder. Its superior performance relative to both deeper CNNs and transformer-based backbones illustrates its robustness and suitability for operational remote sensing applications requiring a balance between accuracy, compactness, and runtime efficiency (see Table 8).

V. DISCUSSION

TerraSegNet has demonstrated robust and consistent performance across a diverse set of remote sensing segmentation tasks. These include global-scale cloud detection on SPARCS (Landsat-8, 30 m, 3 classes), regional cloud mapping on WHUS2-CD+ (Sentinel-2, 10 m, 2 classes), photovoltaic panel segmentation on PV08 (Gaofen-2, 3.24 m, 2 classes), rice field mapping on Plot-Rice v1.0 (Sentinel-1, 10 m, 2 classes), and terrain classification on AIR-PolSAR-Seg-2.0 (Gaofen-3, 8 m, 6 classes). Despite the inherent differences in spatial resolution, spectral modality, and semantic granularity across these datasets, TerraSegNet consistently achieves high performance—recording F1-scores up to 0.9872 and mIoUs up to 0.9364—while maintaining a compact model size of only 5.8 million parameters.

TerraSegNet employs a dual-path feature processing strategy. The encoder extracts shared representations, which are then bifurcated into two specialized branches: (i) a context path, leveraging CAAM to capture long-range semantic dependencies; and (ii) a spatial path, enhanced by SSAM for preserving fine-grained details. These two paths are subsequently fused and refined using an EAM to improve object boundary delineation. This architectural formulation enables TerraSegNet to effectively balance contextual abstraction and spatial fidelity.

From a computational standpoint, TerraSegNet operates efficiently across a range of hardware profiles, with a computational load of 0.7–22.3 GFLOPs and throughput ranging from 9.0 to 61.0 frames per second (fps). This positions TerraSegNet as a competitive solution for both high-resolution image analysis in cloud-based servers and real-time inference in embedded systems.

1) Statistical Analysis

To evaluate whether the observed performance differences among semantic segmentation models are statistically significant, we conducted paired t-tests using the mIoU metric

TABLE 8. Ablation study on the AIR-PolSAR-Seg-2.0 dataset using different backbone networks and combinations of attention modules. Results are reported in terms of F1-score, pixel accuracy, and mIoU. The upper block presents variations of EfficientNetV2-S with and without individual modules, while the lower blocks compare CNN-based and Transformer-based encoders integrated with all three modules. Best results are highlighted in bold.

Backbone	CAAM	SSAM	EAM	Number of Parameters	F1-score ↑	Pixel Accuracy ↑	mIoU ↑	GFLOPS ↓
EfficientNetV2-S	-	-	-	5.68 M	0.9593	0.9668	0.9017	18.26
EfficientNetV2-S	✓	-	-	5.73 M	0.9592	0.9681	0.9022	18.28
EfficientNetV2-S	-	✓	-	5.68 M	0.9587	0.9678	0.9013	18.31
EfficientNetV2-S	-	-	✓	5.73 M	0.9546	0.9652	0.8908	22.27
EfficientNetV2-S	-	✓	✓	5.74 M	0.9533	0.9669	0.8856	22.32
EfficientNetV2-S	✓	-	✓	5.80 M	0.9550	0.9669	0.8907	22.29
EfficientNetV2-S	✓	✓	-	5.74 M	0.9576	0.9669	0.9000	18.33
EfficientNetV2-S	✓	✓	✓	5.80 M	**0.9663**	**0.9725**	**0.9207**	22.34
CNN-based Encoder								
EfficientNet-B2	✓	✓	✓	1.78 M	0.9099	0.9383	0.7763	10.67
ShuffleNet V2-X1-0	✓	✓	✓	2.51 M	0.8908	0.9253	0.7403	26.19
RegNet-Y800MF	✓	✓	✓	3.68 M	0.9284	0.9452	0.8252	20.37
ResNet-34	✓	✓	✓	9.04 M	0.9470	0.9585	0.8674	31.12
Transformer-based Encoder								
MaxVit-T	✓	✓	✓	13.48 M	0.9331	0.9536	0.8429	7.45
Swin-V2-T	✓	✓	✓	13.58 M	0.9366	0.9531	0.8486	37.18

across the test sets of all five datasets: SPARCS (395 samples), WHUS2-CD+ (3514 samples), PV08 (335 samples), Plot-Rice v1.0 (2712 samples), and AIR-PolSAR-Seg-2.0 (309 samples). Specifically, the tests were performed using per-image metric results obtained from the inference outputs.

Table 9 reports the t-statistic, p-value, and mean difference in mIoU for pairwise comparisons between our proposed TerraSegNet and other baselines. Across most comparisons, the p-values were extremely small (all < 0.000001), providing strong statistical evidence that TerraSegNet significantly outperforms a wide range of baselines. These results confirm TerraSegNet's consistent superiority over conventional models (e.g., U-Net, HRNet, SegFormer), bilateral models (e.g., BiSeNetV2, Fast-SCNN, SeaFormer++), and remote sensing-specific models (e.g., HR-Cloud-Net, CD-NetV2, UNetFormer).

A notable exception was observed in the comparison with CMLFormer on the WHUS2-CD+ dataset, where the improvement was not statistically significant ($p = 0.454071$, mean difference = -0.000127). This indicates that, on this specific dataset, TerraSegNet and CMLFormer perform at a comparable level. Nevertheless, for all other datasets and baselines, including strong models such as DeepLabv3+ and CMTFNet, the improvements remain statistically significant. This comprehensive analysis highlights the robustness of TerraSegNet's performance gains while also identifying scenarios where performance is on par with the strongest baselines.

2) Stability Analysis

The quantitative results in Table 10 demonstrate the consistent superiority of the proposed TerraSegNet compared to two strong baselines, DeepLabv3+ and CMLFormer, across diverse remote sensing datasets. The evaluation was conducted under three different random seeds (42, 678, and 1406) to ensure statistical robustness, with the results reported as mean $\pm$ standard deviation. The choice of DeepLabv3+ and CMLFormer as baselines was grounded on paired t-test analysis (see Table 9), confirming their competitiveness against other mainstream architectures.

Across all datasets, TerraSegNet achieved the highest performance in terms of F1 score and mIoU while maintaining competitive pixel accuracy. Improvements were observed consistently on SPARCS (0.8592 ± 0.0021 mIoU), WHUS2-CD+ (0.9344 ± 0.0024 mIoU), PV08 (0.9607 ± 0.0010 mIoU), and Plot-Rice v1.0 (0.7787 ± 0.0010 mIoU), all surpassing the baselines despite dataset-specific challenges. On AIR-PolSAR-Seg-2.0, TerraSegNet further delivered the strongest results (0.8992 ± 0.0187 mIoU), highlighting robustness in handling polarimetric SAR imagery despite slightly higher variability. These consistent gains across optical and SAR benchmarks confirm its ability to generalize effectively across diverse modalities and scene complexities, while offering statistically significant improvements over strong baselines with relatively low variance across independent runs.

3) Visual Analysis of Segmentation Performance Across Datasets

As depicted in Fig. 13, TerraSegNet demonstrates consistently superior segmentation quality across diverse remote sensing datasets when compared to baseline models. In SPARCS and WHUS2-CD+, TerraSegNet achieves the lowest misclassification regions (in red), particularly along complex object boundaries, indicating robust edge preservation. For PV08 and Plot-Rice v1.0, the model effectively delineates fine-grained object shapes and regular plot structures, preserving both spatial consistency and class integrity. Meanwhile, in the AIR-PolSAR-Seg-2.0 dataset, which involves high-frequency noise and speckle typical of SAR imagery, TerraSegNet exhibits strong resilience, producing segmentation masks with minimal artifacts and accurate class transitions. This performance can be attributed to its dual-path

TABLE 9. Paired t-test results comparing TerraSegNet with other state-of-the-art models across five benchmark datasets (SPARCS, WHUS2-CD+, PV08, Plot-Rice v1.0, and AIR-PolSAR-Seg-2.0). Reported values include the t-statistic, p-value, and mean difference in mIoU. TerraSegNet demonstrates statistically significant improvements over most baselines, with significance assessed at $p < 0.05$.

Dataset	Comparison Model	t-statistic	p-value	Mean Difference	Interpretation
SPARCS	U-Net	15.096	0.0	0.022441	Significant
SPARCS	DeepLabv3+	12.215	0.0	0.011910	Significant
SPARCS	HRNet	15.249	0.0	0.042428	Significant
SPARCS	SegFormer	17.433	0.0	0.042756	Significant
SPARCS	BiSeNetV1	21.304	0.0	0.036948	Significant
SPARCS	BiSeNetV2	21.190	0.0	0.048249	Significant
SPARCS	Fast-SCNN	21.313	0.0	0.061344	Significant
SPARCS	SeaFormer++	21.275	0.0	0.060146	Significant
SPARCS	HR-Cloud-Net	17.033	0.0	0.050670	Significant
SPARCS	CDNetV2	20.329	0.0	0.038739	Significant
SPARCS	UNetFormer	12.988	0.0	0.014571	Significant
SPARCS	CMLFormer	13.532	0.0	0.013686	Significant
SPARCS	CMTFNet	14.207	0.0	0.017289	Significant
WHUS2-CD+	U-Net	8.668	0.0	0.001355	Significant
WHUS2-CD+	DeepLabv3+	4.851	0.000001	0.000821	Significant
WHUS2-CD+	HRNet	4.564	0.000005	0.001543	Significant
WHUS2-CD+	SegFormer	5.602	0.0	0.001558	Significant
WHUS2-CD+	BiSeNetV1	9.172	0.0	0.001585	Significant
WHUS2-CD+	BiSeNetV2	8.015	0.0	0.003034	Significant
WHUS2-CD+	Fast-SCNN	13.143	0.0	0.004243	Significant
WHUS2-CD+	SeaFormer++	9.462	0.0	0.002223	Significant
WHUS2-CD+	HR-Cloud-Net	8.084	0.0	0.003056	Significant
WHUS2-CD+	CDNetV2	5.286	0.0	0.001353	Significant
WHUS2-CD+	UNetFormer	13.846	0.0	0.073721	Significant
WHUS2-CD+	CMLFormer	-0.749	0.454071	-0.000127	Not Significant
WHUS2-CD+	CMTFNet	2.158	0.030983	0.000450	Significant
PV08	U-Net	6.269	0.0	0.007825	Significant
PV08	DeepLabv3+	3.509	0.000512	0.004751	Significant
PV08	HRNet	6.804	0.0	0.008150	Significant
PV08	SegFormer	7.447	0.0	0.013771	Significant
PV08	BiSeNetV1	7.169	0.0	0.009984	Significant
PV08	BiSeNetV2	10.282	0.0	0.013561	Significant
PV08	Fast-SCNN	9.835	0.0	0.017238	Significant
PV08	SeaFormer++	9.904	0.0	0.017638	Significant
PV08	HR-Cloud-Net	10.737	0.0	0.026113	Significant
PV08	CDNetV2	6.040	0.0	0.011828	Significant
PV08	UNetFormer	5.043	0.000001	0.005889	Significant
PV08	CMLFormer	4.725	0.000003	0.003748	Significant
PV08	CMTFNet	5.494	0.0	0.005330	Significant
Plot-Rice v1.0	U-Net	14.782	0.0	0.009640	Significant
Plot-Rice v1.0	DeepLabv3+	14.751	0.0	0.009502	Significant
Plot-Rice v1.0	HRNet	27.262	0.0	0.021082	Significant
Plot-Rice v1.0	SegFormer	23.565	0.0	0.019921	Significant
Plot-Rice v1.0	BiSeNetV1	31.392	0.0	0.019833	Significant
Plot-Rice v1.0	BiSeNetV2	35.833	0.0	0.031246	Significant
Plot-Rice v1.0	Fast-SCNN	33.253	0.0	0.025361	Significant
Plot-Rice v1.0	SeaFormer++	35.686	0.0	0.027410	Significant
Plot-Rice v1.0	HR-Cloud-Net	17.689	0.0	0.012224	Significant
Plot-Rice v1.0	CDNetV2	15.475	0.0	0.009717	Significant
Plot-Rice v1.0	UNetFormer	23.437	0.0	0.014488	Significant
Plot-Rice v1.0	CMLFormer	23.104	0.0	0.014702	Significant
Plot-Rice v1.0	CMTFNet	15.560	0.0	0.008743	Significant
AIR-PolSAR-Seg-2.0	U-Net	25.619	0.0	0.056086	Significant
AIR-PolSAR-Seg-2.0	DeepLabv3+	12.781	0.0	0.016642	Significant
AIR-PolSAR-Seg-2.0	HRNet	31.979	0.0	0.099566	Significant
AIR-PolSAR-Seg-2.0	SegFormer	30.332	0.0	0.195943	Significant
AIR-PolSAR-Seg-2.0	BiSeNetV1	24.508	0.0	0.045339	Significant
AIR-PolSAR-Seg-2.0	BiSeNetV2	31.944	0.0	0.111697	Significant
AIR-PolSAR-Seg-2.0	Fast-SCNN	33.934	0.0	0.165434	Significant
AIR-PolSAR-Seg-2.0	SeaFormer++	33.154	0.0	0.159105	Significant
AIR-PolSAR-Seg-2.0	HR-Cloud-Net	29.824	0.0	0.119841	Significant
AIR-PolSAR-Seg-2.0	CDNetV2	30.525	0.0	0.224062	Significant
AIR-PolSAR-Seg-2.0	UNetFormer	26.353	0.0	0.051091	Significant
AIR-PolSAR-Seg-2.0	CMLFormer	17.422	0.0	0.029527	Significant
AIR-PolSAR-Seg-2.0	CMTFNet	17.369	0.0	0.036527	Significant

TABLE 10. Quantitative comparison of semantic segmentation performance on five benchmark datasets using DeepLabv3+, CMLFormer, and the proposed TerraSegNet. Results are reported as mean ± standard deviation across three independent runs with random seeds 42, 678, and 1406. The best results for each dataset and metric are highlighted in bold.

Dataset	Model	F1 Score	Pixel Accuracy	mIoU
SPARCS	DeepLabv3+	0.9422 ± 0.0007	0.9423 ± 0.0005	0.8379 ± 0.0015
	CMLFormer	0.9418 ± 0.0007	0.9410 ± 0.0005	0.8374 ± 0.0006
	TerraSegNet	**0.9510 ± 0.0006**	**0.9496 ± 0.0007**	**0.8592 ± 0.0021**
WHUS2-CD+	DeepLabv3+	0.9861 ± 0.0004	0.9874 ± 0.0001	0.9334 ± 0.0006
	CMLFormer	0.9814 ± 0.0100	**0.9877 ± 0.0003**	0.9324 ± 0.0050
	TerraSegNet	**0.9872 ± 0.0005**	0.9876 ± 0.0005	**0.9344 ± 0.0024**
PV08	DeepLabv3+	0.9774 ± 0.0004	0.9777 ± 0.0004	0.9563 ± 0.0008
	CMLFormer	0.9785 ± 0.0001	0.9787 ± 0.0001	0.9582 ± 0.0001
	TerraSegNet	**0.9798 ± 0.0005**	**0.9800 ± 0.0005**	**0.9607 ± 0.0010**
Plot-Rice v1.0	DeepLabv3+	0.8674 ± 0.0007	0.8683 ± 0.0010	0.7660 ± 0.0011
	CMLFormer	0.8656 ± 0.0013	0.8664 ± 0.0013	0.7633 ± 0.0020
	TerraSegNet	**0.8755 ± 0.0006**	**0.8762 ± 0.0005**	**0.7787 ± 0.0010**
AIR-PolSAR-Seg-2.0	DeepLabv3+	0.9451 ± 0.0027	0.9623 ± 0.0006	0.8795 ± 0.0028
	CMLFormer	0.9484 ± 0.0031	0.9559 ± 0.0008	0.8747 ± 0.0073
	TerraSegNet	**0.9567 ± 0.0085**	**0.9674 ± 0.0044**	**0.8992 ± 0.0187**

architecture and adaptive feature fusion strategy, enabling it to capture both global context and fine spatial details.

Although the numerical improvements brought by the Edge Attention Module (EAM) appear modest in the ablation study, its qualitative benefits are evident. As illustrated in Figure V-3, the inclusion of EAM enhances the delineation of object boundaries, producing segmentation results with sharper and more coherent edges. This property is particularly important in remote sensing scenarios, where accurate boundary localization is crucial for applications such as land parcel mapping and urban structure analysis.

4) Limitations and Design Trade-off

Despite its overall effectiveness, TerraSegNet exhibits several limitations and trade-offs that merit discussion:

- **Annotation Dependency:** TerraSegNet relies on fully supervised learning, which necessitates large volumes of high-quality, pixel-level annotated data. For complex or underrepresented classes—particularly in datasets like AIR-PolSAR-Seg-2.0—manual annotation can be both time-consuming and costly.
- **Sensitivity to Class Imbalance:** Like many segmentation models, TerraSegNet is susceptible to degraded performance in the presence of class imbalance. Without loss reweighting or targeted data augmentation, minority classes may be underrepresented in the final predictions.
- **Attention Module Overhead:** While the integration of CAAM, SSAM, and EAM enhances segmentation accuracy, these modules incur additional computational cost. Compared to ultra-lightweight models such as Fast-SCNN (0.7 GFLOPs), TerraSegNet's more complex pipeline may limit its applicability on highly resource-constrained edge devices.
- **Lack of Multi-temporal Adaptability:** TerraSegNet was designed and evaluated on static image datasets. Its robustness under temporal variations (e.g., seasonal dynamics in agriculture or persistent cloud cover) remains untested, and future adaptations are needed to support spatiotemporal learning and domain generalization.

These limitations suggest that while TerraSegNet is well-suited for many operational settings, it may require further customization or optimization in scenarios with severe constraints or rapidly changing environments.

5) Practical Implications

The empirical results and architectural design of TerraSegNet underscore its strong practical applicability in remote sensing scenarios:

- **Operational Readiness:** The model's high throughput and strong accuracy enable its deployment in both offline and near real-time applications, including agricultural monitoring, infrastructure surveillance, and environmental mapping.
- **Platform Flexibility:** Its compact size and moderate GFLOPs make TerraSegNet suitable for deployment on various platforms—ranging from ground-based data centers to low-power edge devices such as UAVs, nanosatellites, and embedded boards.
- **Sensor Agnosticism:** TerraSegNet has been validated across diverse sensor modalities, including optical (RGB, NIR), radar (DpRVIc), and full-polarimetric SAR, without architecture modifications. This adaptability enhances its value for heterogeneous satellite missions.
- **Scalability for Large-scale Monitoring:** Due to its balance between accuracy and efficiency, TerraSegNet is well-positioned to support regional and national-scale Earth observation programs where consistent, automated segmentation across large datasets is essential.

In summary, TerraSegNet represents a scalable and practical semantic segmentation framework that bridges the gap between performance-driven research models and field-deployable solutions in real-world remote sensing applications.

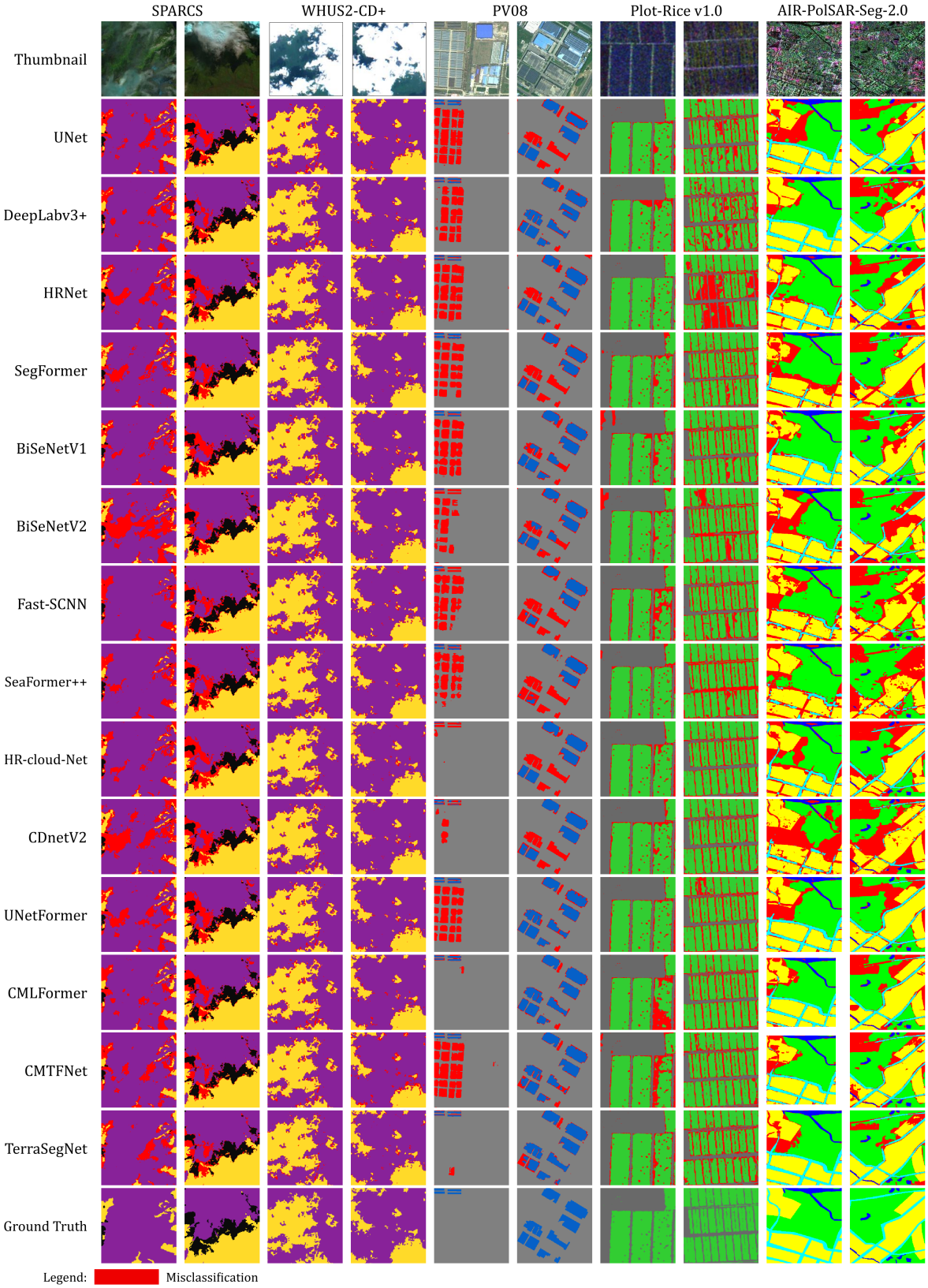


FIGURE 13. **Qualitative comparison of segmentation results on five benchmark datasets using 14 models. From top to bottom: Satellite image, predictions from UNet, DeepLabv3+, HRNet, SegFormer, BiSeNetV1, BiSeNetV2, Fast-SCNN, SeaFormer++, HR-Cloud-Net, CDnetV2, UNetFormer, CMLFormer, CMTFNet, the proposed TerraSegNet model, and ground truth. Red color indicates misclassification.**

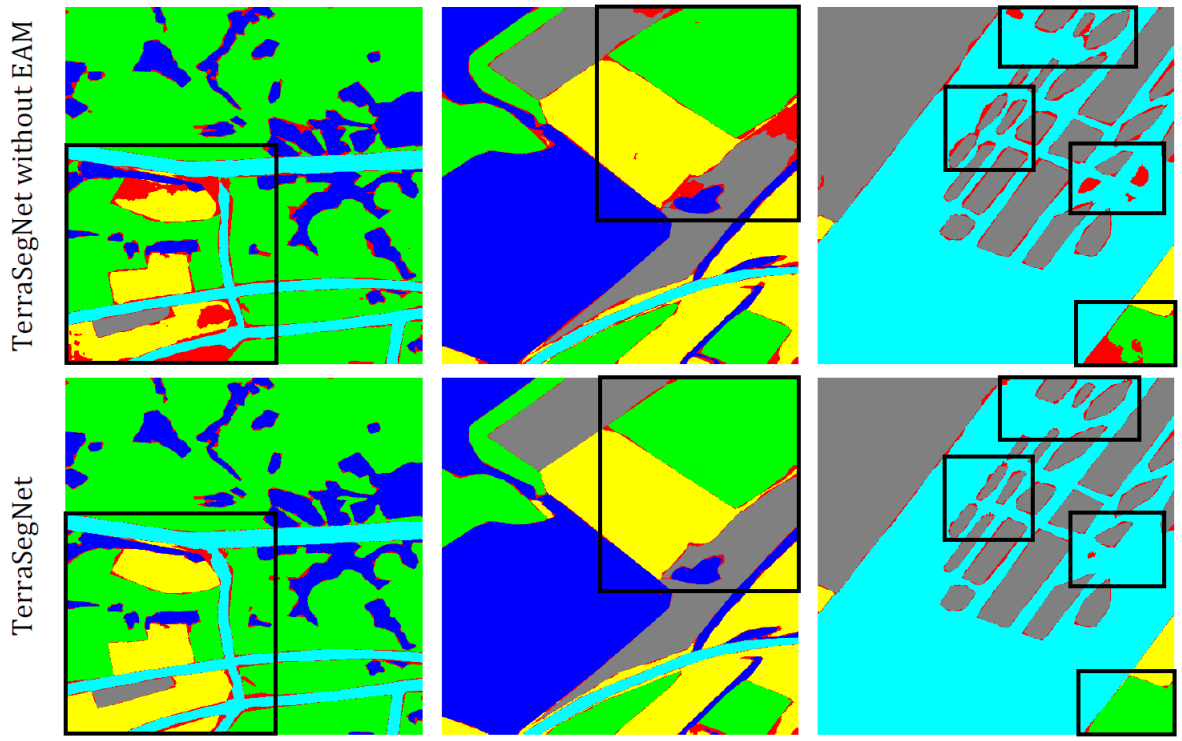


FIGURE 14. Qualitative comparison of segmentation results with and without the Edge Attention Module (EAM) on representative samples. The inclusion of EAM yields clearer boundary preservation and improved edge consistency, even when overall numerical gains are modest.

VI. CONCLUSION

In this work, we presented **TerraSegNet**, a lightweight semantic segmentation network that introduces several key innovations. First, its bilateral feature processing design explicitly separates spatial detail preservation from semantic context modeling, enabling the network to maintain fine boundaries while capturing global information. Second, the integration of CAAM, SSAM, and EAM provides complementary attention mechanisms that strengthen cross-scale interaction and feature fusion without significantly increasing complexity. Third, TerraSegNet achieves a favorable trade-off between accuracy and efficiency, requiring only **5.8M parameters** while delivering competitive performance. Finally, extensive experiments on five diverse benchmark datasets demonstrate the robustness and generalizability of the model across different modalities and spatial resolutions. The results confirm that the model consistently performs well under varying conditions. However, attention modules inevitably add some computational overhead, and the current approach still relies on pixel-level annotations. To address these challenges, future work will explore multi-temporal modeling and semi-supervised strategies to further improve scalability. Overall, TerraSegNet offers a practical and scalable solution, making clear contributions to both methodological innovation and real-world applicability in remote sensing.

DATA AVAILABILITY STATEMENT

The datasets used in this study are publicly available from the following sources: SPARCS dataset [57] at https://emapr.ceoas.oregonstate.edu/sparcs/, WHUS2-CD+ dataset [15] at https://doi.org/10.5281/zenodo.5511792, PV08 dataset [58] at https://doi.org/10.5281/zenodo.5171711, Plot-Rice v1.0 dataset [59] at https://doi.org/10.5281/zenodo.13897215, and AIR-PolSAR-Seg-2.0 dataset [45] at https://doi.org/10.57760/sciencedb.radars.00041.

DECLARATION OF COMPETING INTEREST

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

REFERENCES

[1] A. E. Swiggs, I. R. Lawrence, A. Ridout and A. Shepherd, "Detecting Sea Ice Leads and Floes in the Northwest Passage Using CryoSat-2," in *IEEE J. of Sel. Top. in Appl. Earth Obs and Remote Sens.*, vol. 18, pp. 1226-1236, 2025.

[2] J. Liu, L. Chen, D. Tang, H. Xie, X. Cui and P. Li, "Using Multimission Satellite Altimetry to Monitor Subglacial Hydrological Activities in the Totten Basin, East Antarctica," in *IEEE J. of Sel. Top. in Appl. Earth Obs and Remote Sens.*, vol. 18, pp. 795-812, 2025.

[3] D. Cerra, S. Auer, A. Baissero and F. Bachofer, "Detection and Monitoring of Floating Plastic Debris on Inland Waters From Sentinel-2 Time Series," in *IEEE J. of Sel. Top. in Appl. Earth Obs and Remote Sens.*, vol. 18, pp. 1122-1138, 2025.

[4] Y. Wang, H. Huang and B. Wu, "Evaluating the Potential of SDGSAT-1 Glimmer Imagery for Urban Road Detection," in *IEEE J. of Sel. Top. in Appl. Earth Obs and Remote Sens.*, vol. 18, pp. 785-794, 2025.

[5] H. Li et al., "In-Season Mapping of Sugarcane Planting Based on Sentinel-2 Imagery," in *IEEE J. of Sel. Top. in Appl. Earth Obs and Remote Sens.*, vol. 18, pp. 1410-1421, 2025.

[6] N. Ramachandran, J. Irvin, H. Sheng, S. Johnson-Yu, K. Story, R. Rustowicz, A. Y. Ng, and K. Austin, "Automatic deforestation driver attribution using deep learning on satellite imagery," in *Global Env.al Change*, vol. 86, pp. 102843, May 2024.

[7] Y. Vetrita, et al., "Monthly Mapping of Indonesia's Burned Areas: Implementation, History, Techniques, and Future Directions," in *Int. J. of Remote Sens.*, vol. 46, pp. 636–660, Nov. 2024.

[8] S. Khaki, H. Pham, and L. Wang, "Simultaneous corn and soybean yield prediction from remote sensing data using deep transfer learning." in *Sci. Rep*. 11, 11132, 2021.

[9] Q. Wang, W. Chen, Z. Huang, H. Tang and L. Yang, "MultiSenseSeg: A Cost-Effective Unified Multimodal Semantic Segmentation Model for Remote Sensing," in *IEEE Trans. Geosci. Remote Sens.*, vol. 62, pp. 1-24, 2024.

[10] F. Zhang and X. Xia, "Efficient Semantic Segmentation of Remote Sensing Images Through Global-Local Feature Integration," in *IEEE Access*, vol. 13, pp. 115653-115668, 2025.

[11] A. Dosovitskiy, L. Beyer, A. Kolesnikov, D. Weissenborn, X. Zhai, T. Unterthiner, M. Dehghani, M. Minderer, G. Heigold, S. Gelly, J. Uszkoreit, and N. Houlsby, "An Image is Worth 16x16 Words: Transformers for Image Recognition at Scale," in *Int. Conf. on Learn. Represent. ICLR 2021*, Vienna, Austria, Oct. 2020.

[12] Liu, Ze, et al. "Swin transformer v2: Scaling up capacity and resolution." in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*. 2022.

[13] A. Kirillov, E. Mintun, N. Ravi, H. Mao, C. Rolland, L. Gustafson, T. Xiao, S. Whitehead, A. C. Berg, W. Lo, P. Dollár, and R. Girshick, "Segment Anything," in *arXiv*, Apr. 2023.

[14] K. R. Thorp, D. Drajat, "Deep machine learning with Sentinel satellite data to map paddy rice production stages across West Java, Indonesia," in *Remote Sens. of Env.*, vol. 265, pp. 112679, Nov. 2021.

[15] J. Li, Z. Wu, Z. Hu, C. Jian, S. Luo, and L. Mou, "A Lightweight Deep Learning-Based Cloud Detection Method for Sentinel-2A Imagery Fusing Multiscale Spectral, Spatial Features," in *IEEE Trans. on Geosci. and Remote Sens.*, vol. 60, pp. 1–19, Apr. 2021.

[16] Y. Ma, S. Chen, S. Ermon, and D. B. Lobell, "Transfer learning in Environmental remote sensing," in *Remote Sens. of Env.*, vol. 301, pp. 113924, Feb. 2024.

[17] C. Yu, J. Wang, C. Peng, C. Gao, G. Yu, and N. Sang, "BiSeNet: Bilateral Segmentation Network for Real-Time Semantic Segmentation," in *Comp. Vis. – ECCV 2018*, Munich, Germany, pp. 334–349, Oct. 2018.

[18] C. Yu, C. Gao, J. Wang, G. Yu, C. Shen, and N. Sang, "BiSeNet V2: Bilateral Network with Guided Aggregation for Real-Time Semantic Segmentation," in *Int. J. Comput. Vis. 129*, vol. 129, pp. 3051—3068, Sep. 2021.

[19] R. P. K. Poudel, S. Liwicki, and R. Cipolla, "Fast-SCNN: Fast Semantic Segmentation Network," in *arXiv*, Feb. 2019.

[20] Q. Wan, Z. Huang, J. Lu, G. Yu, and L. Zhang, "SeaFormer++: Squeeze-Enhanced Axial Transformer for Mobile Visual Recognition," in *Int. J. of Comp. Vis.*, vol. 133, Jan. 2025.

[21] H. Wang, Y. Zhu, B. Green, H. Adam, A. Yuille, and L. C. Chen, "Axial-DeepLab: Stand-Alone Axial-Attention for Panoptic Segmentation," in *European Conf. on Comp. Vis. (ECCV)*, Glasgow, UK, Oct. 2020.

[22] K. Jiao and Z. Pan, "A Novel Method for Image Segmentation Based on Simplified Pulse Coupled Neural Network and Gbest Led Gravitational Search Algorithm," in *IEEE Access*, vol. 7, pp. 21310-21330, Jan. 2019.

[23] O. Ronneberger, P. Fischer, and T. Brox, "U-Net: Convolutional Networks for Biomedical Image Segmentation," in *Medical Image Comp., Comp.-Assisted Intervention – MICCAI 2015*, Munich, Germany, pp. 234–241, Nov. 2015.

[24] L. Chen, Y. Zhu, G. Papandreou, F. Schroff, and H. Adam, "Encoder-Decoder with Atrous Separable Convolution for Semantic Image Segmentation," in *Comp. Vis. – ECCV 2018*, Munich, Germany, pp. 833–851, Oct. 2018.

[25] J. Wang, K. Sun, T. Cheng, B. Jiang, C. Deng, Y. Zhao, D. Liu, Y. Mu, M. Tan, X. Wang, W. Liu, and B. Xiao, "Deep High-Resolution Representation Learning for Visual Recognition," in *IEEE Trans. on Patt. Anal., Machine Intell.*, vol. 43, pp. 3349–3364, Apr. 2020.

[26] E. Xie, W. Wang, Z. Yu, A. Anandkumar, J. M. Alvarez, and P. Luo, "SegFormer: Simple, Efficient Design for Semantic Segmentation with Transformers," in *NeurIPS 2021*, Virtual, May 2021.

[27] A. Kirillov, E. Mintun, N. Ravi, H. Mao, C. Rolland, L. Gustafson, T. Xiao, S. Whitehead, A.C. Berg, W.Y. Lo, and P. Dollár, "Segment Anything," in *CVF Int. Conf. on Comp. Vis.*, pp. 4015-4026, Apr. 2023.

[28] L. Sun, X. Mi, J. Wei, J. Wang, X. Tian, H. Yu, and P. Gan, "A cloud detection algorithm-generating method for remote Sensing data at visible to short-wave infrared wavelengths," in *ISPRS J. of Photogramm. and Remote Sens.*, vol. 124, pp. 70–88, Feb. 2017.

[29] D. Frantz, E. Haß, A. Uhl, J. Stoffels, and J. Hill, "Improvement of the Fmask algorithm for Sentinel-2 images: Separating clouds from bright surfaces based on parallax effects," in *Remote Sens. of Env.*, vol. 215, pp. 471–481, Sep. 2018.

[30] S. Qiu, Z. Zhu, and B. He, "Fmask 4.0: Improved cloud, cloud shadow detection in Landsats 4–8, Sentinel-2 imagery," in *Remote Sens. of Env.*, vol. 231, Sep. 2019.

[31] Y. Chen et al., "SparseFormer: A Credible Dual-CNN Expert-Guided Transformer for Remote Sensing Image Segmentation With Sparse Point Annotation," in *IEEE Trans. Geosci. Remote Sens.*, vol. 63, pp. 1-16, 2025.

[32] H. Feng et al., "FTransDeepLab: Multimodal Fusion Transformer-Based DeepLabv3+ for Remote Sensing Semantic Segmentation," in *IEEE Trans. Geosci. Remote Sens.*, vol. 63, pp. 1-18, 2025.

[33] Z. Meng, Q. Yan, F. Zhao, G. Chen, W. Hua and M. Liang,"Global–Local Multigranularity Transformer for Hyperspectral Image Classification" in *IEEE J. of Sel. Top. in Appl. Earth Obs and Remote Sens.*, vol. 18, pp. 112-131, 2025.

[34] Y. Huang, D. Jiao, X. Huang, T. Tang and G. Gui, "A Hybrid CNN-Transformer Network for Object Detection in Optical Remote Sensing Images: Integrating Local and Global Feature Fusion," in *IEEE J. of Sel. Top. in Appl. Earth Obs and Remote Sens.*, vol. 18, pp. 241-254, 2025.

[35] S. Zhang, M. Li, W. Zhao, X. Wang and Q. Wu, "Building Type Classification Using CNN-Transformer Cross-Encoder Adaptive Learning From Very High Resolution Satellite Images," in *IEEE J. of Sel. Top. in Appl. Earth Obs and Remote Sens.*, vol. 18, pp. 976-994, 2025.

[36] J. Qiu, W. Liu, X. Zhang, E. Li, L. Zhang and X. Li, "DED-SAM:Adapting Segment Anything Model 2 for Dual Encoder–Decoder Change Detection," in *IEEE J. of Sel. Top. in Appl. Earth Obs and Remote Sens.*, vol. 18, pp. 995-1006, 2025.

[37] H. Yang, Z. Jiang, Y. Zhang, Y. Wu, H. Luo, P. Zhang, and B. Wang, "A high-resolution remote sensing land use/land cover classification method based on multi-level features adaptation of segment anything model," in *Int. J. of Appl. Earth Obs. GeoInf.*, vol. 141, pp. 104659, Jul. 2025.

[38] N. Wright, J. M. A. Duncan, J. N. Callow, S. E. Thompson, and R. J. George, "Training sensor-agnostic deep learning models for remote sensing: Achieving state-of-the-art cloud, cloud shadow identification with OmniCloudMask," in *Remote Sens. of Env.*, vol. 322, pp. 114694, May 2025.

[39] X. Li, J. Li, J. Jiang, X. Pan, and X. Huang, "Spatio-temporal-text fusion for hierarchical multi-label crop classification based on time-series remote sensing imagery," in *Int. J. of Appl. Earth Obs. GeoInf.*, vol. 139, pp. 104471, May 2025.

[40] S. Jeong, J. Ko, J. Ban, T. Shin, and J. Yeom, "Deep learning-enhanced remote sensing-integrated crop modeling for rice yield prediction," in *Acologi. Inf.*, vol. 84, pp. 102886, Dec. 2024.

[41] S. Mohajerani, P. Saeedi, "Cloud, Cloud Shadow Segmentation for Remote Sensing Imagery Via Filtered Jaccard Loss Function, Parametric Augmentation," in *IEEE J. of Sel. Top. in Appl. Earth Obs. Remote Sens.*, vol. 14, pp. 4254-4266, Apr. 2021.

[42] S. Papini, S. X. Rao, S. Charyyev, M. Jiang, and P. H. Egger, "Urban growth unveiled: Deep learning with satellite imagery for measuring 3D building-stock evolution in Urban China," in *Remote Sens. Appl.: Soc. and Env.*, vol. 38, pp. 101523, Apr. 2025.

[43] I. Colkesen, M. Saygi, M. Y. Ozturk, and O. Y. Altuntas, "U-shaped deep learning networks for algal bloom detection using Sentinel-2 imagery: Exploring model performance, transferability," in *J. of Env. Management*, vol. 381, pp. 125152, May 2025.

[44] H. Liu, B. Sun, Z. Gao, Z. Chen, and Z. Zhu, "High resolution remote sensing recognition of elm sparse forest via deep-learning-based semantic segmentation," in *Acologi. Indicators*, vol. 166, pp. 1112428, Aug. 2024.

[45] W. Zhirui, Z. Liangjin, W. Yuelei, Z. Xuan, K. Jian, Y. Jian, and S. Xian, "AIR-PolSAR-Seg-2.0: Polarimetric SAR ground terrain classification dataset for large-scale complex scenes," in *J. of Radars*, vol. 14, pp. 353-–365, Apr. 2025.

[46] Z. Chen, G. Wu, H. Gao, Y. Ding, Y., D. Hong, and B. Zhang, "Local Aggregation and Global Attention Network for Hyperspectral Image

Classification with Spectral-Induced Aligned Superpixel Segmentation," in *Expert Syst. Appl.*, vol. 232, p.120828, Dec. 2023.

[47] J. Li, T. Xue, J. Zhao, J. Ge, Y. Min, W. Su, and K. Zhan, "High-resolution cloud detection network," in *J. of Electro. Imaging*, vol. 33, no. 4, Jul. 2024.

[48] J. Guo, J. Yang, H. Yue, H. Tan, C. Hou, and K. Li, "CDnetV2: CNN-Based Cloud Detection for Remote Sensing Imagery With Cloud-Snow Coexistence," in *IEEE Trans. on Geosci. and Remote Sens.*, vol. 59, pp. 700–713, May 2020.

[49] Z. Wen, H. Huang, and S. Liu, "Multi-scale Attention Fusion Network for Semantic Segmentation of Remote Sensing Images," in *Int. J. of Remote Sens.*, vol. 44, pp. 7909–7926, Dec. 2023.

[50] L. Wang, R. Li, C. Zhang, S. Fang, C. Duan, X. Meng, and P. M. Atkinson, "UNetFormer: A UNet-like transformer for efficient semantic segmentation of remote sensing urban scene imagery," in *ISPRS J. of Photogramm. and Remote Sens.*, vol. 190, pp. 196-214, Aug. 2022.

[51] R. Li, S. Zheng, C. Zhang, C. Duan, L. Wang, and P. M. Atkinson, "ABCNet: Attentive bilateral contextual network for efficient semantic segmentation of Fine-Resolution remotely sensed imagery," in *ISPRS J. of Photogramm. and Remote Sens.*, vol. 181, pp. 84-98, Nov. 2021.

[52] L. Wang, R. Li, D. Wang, C. Duan, T. Wang, and X. Meng, "Transformer Meets Convolution: A Bilateral Awareness Network for Semantic Segmentation of Very Fine Resolution Urban Scene Images," in *Remote Sens.*, vol. 13, Aug. 2021.

[53] H. Wu, M. Zhang, P. Huang and W. Tang, "CMLFormer: CNN and Multiscale Local-Context Transformer Network for Remote Sensing Images Semantic Segmentation," in *IEEE J. of Sel. Top. in Appl. Earth Obs. Remote Sens.*, vol. 17, pp. 7233-7241, 2024.

[54] H. Wu, P. Huang, M. Zhang, W. Tang and X. Yu, "CMTFNet: CNN and Multiscale Transformer Fusion Network for Remote-Sensing Image Semantic Segmentation," in *IEEE Trans. on Geosci. and Remote Sens.*, vol. 61, pp. 1-12, 2023.

[55] S. Chen, Y. Yu, Y. Li, Z. Wang, X. Li and J. Han, "Multiscale Adapter Based on SAM for Remote Sensing Semantic Segmentation," in *IEEE J. Sel. Top. Appl. Earth Obs. Remote Sens.*, vol. 18, pp. 6806-6819, 2025.

[56] M. Luo, S. Ji, "Cross-spatiotemporal land-cover classification from VHR remote sensing images with deep learning based domain adaptation," in *ISPRS J. of Photogramm. and Remote Sens.*, vol. 191, pp. 105–128, Sep. 2022.

[57] M. J. Hughes, R. Kennedy, "High-Quality Cloud Masking of Landsat 8 Imagery Using Convolutional Neural Networks," in *Remote Sens.*, vol. 11, Nov. 2019.

[58] H. Jiang, L. Yao, N. Lu, J. Qin, T. Liu, Y. Liu, and C. Zhou, "Multi-resolution dataset for photovoltaic panel segmentation from satellite and aerial imagery," in *Earth Syst. Sci. Data*, vol. 13, pp. 5389–5401, Nov. 2021.

[59] J. Ge, H. Zhang, W. Huang, Z. Guo, L. Xu, Y. Xie, M. Song, Y. Ding, and C. Wang, "Plot-Rice v1.0: A global plot-based rice benchmark dataset with spatiotemporal heterogeneity for scientific deep learning," in *Int. J. of Appl. Earth Obs. GeoInf.*, vol. 140, pp. 104569, Jun. 2025.

[60] M. Tan, Q. V. Le, "EfficientNetV2: Smaller Models, Faster Training," in *ICML 2021*, Virtual, Apr. 2021.

[61] A. Bokhovkin, E. Burnaev, "Boundary Loss for Remote Sensing Imagery Semantic Segmentation," in *Adv. in Neural Net. – ISNN 2019*, Moscow, Russia, pp. 388–401, Jul. 2019.

[62] S.S.M. Salehi, D. Erdogmus, and A. Gholipour, "Tversky loss Function for Image Segmentation Using 3D Fully Convolutional Deep Networks," in *MLMI 2017*, pp. 379-387, Sep. 2017.

[63] I. Loshchilov, F. Hutter, "Decoupled Weight Decay Regularization," in *ICLR 2019*, New Orleans, LA, USA, Nov. 2017.

[64] I. Loshchilov, F. Hutter, "SGDR: Stochastic Gradient Descent with Warm Restarts," in *ICLR 2017*, Toulon, France, Aug. 2016.

[65] M. Tan, Q. V. Le, "EfficientNet: Rethinking Model Scaling for Convolutional Neural Networks," in *ICML 2019*, Long Beach, CA, USA, May 2019.

[66] N. Ma, X. Zhang, H. Zheng, and J. Sun, "ShuffleNet V2: Practical Guidelines for Efficient CNN Architecture Design," in *Comp. Vis. – ECCV 2018*, Munich, Germany, pp. 122–138, Oct. 2018.

[67] I. Radosavovic, R. P. Kosaraju, R. Girshick, K. He, and P. Dollár, "Designing Network Design Spaces," in *2020 IEEE/CVF Conf. on Comp. Vis., Patt. Recog. (CVPR)*, Seattle, WA, USA, pp. 10425–10433, Aug. 2020.

[68] K. He, X. Zhang, S. Ren, J. Sun, "Deep Residual Learning for Image Recognition," in *2016 IEEE Conf. on Comp. Vis., Patt. Recog. (CVPR)*, Las Vegas, NV, USA, pp. 770–778, Jun. 2016.

[69] Tu, Zhengzhong, et al. "Maxvit: Multi-axis vision transformer." in *European Conf. on Com. Vis.*. 2022.

[70] Liu, Ze, et al. "Swin transformer v2: Scaling up capacity and resolution." in *CVPR2022*. 2022.

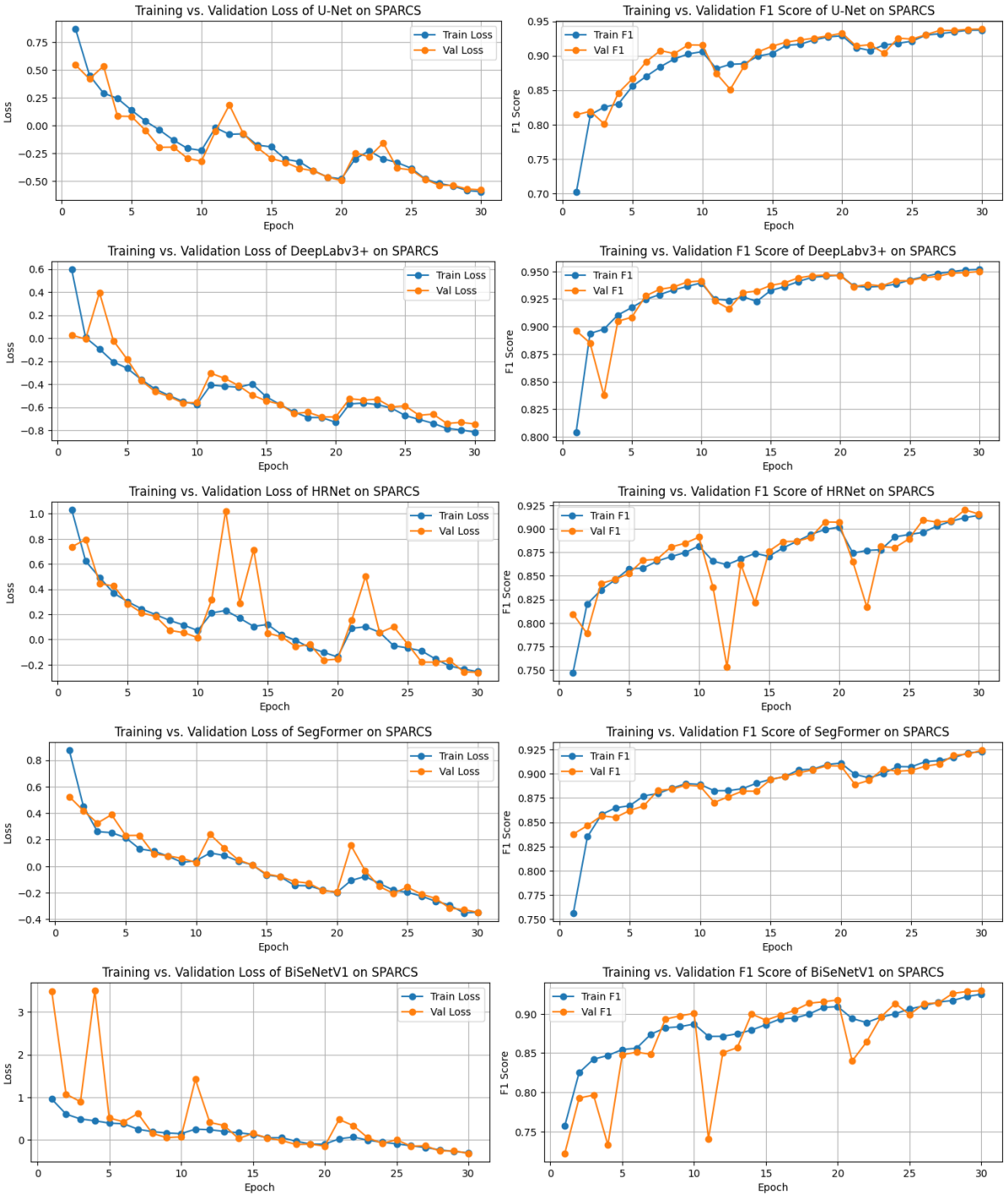
BAGUS SETYAWAN WIJAYA received the bachelor's degree in statistical computing from Politeknik Statistika STIS, Indonesia, in 2009, and the master's degree in chief information officer from Institut Teknologi Bandung (ITB), Indonesia, in 2016. He is currently pursuing the Ph.D. degree with the School of Electrical Engineering and Informatics, ITB. His research interests include computer vision and remote sensing.

RINALDI MUNIR received the bachelor's degree in informatics engineering and the M.Sc. degree in digital image compression from Institut Teknologi Bandung (ITB), Bandung, Indonesia, in 1992 and 1999, respectively, and the Ph.D. degree in image watermarking from the School of Electrical Engineering and Informatics, ITB, in 2010. In 1993, he started his academic career as a Lecturer with the Department of Informat ics, ITB. He is currently an Associate Professor with the School of Electrical Engineering and Informatics, ITB, and the Informatics Research Group. His research interests include cryptography and steganography-related topics, digital image processing, fuzzy logic, and numerical computation.

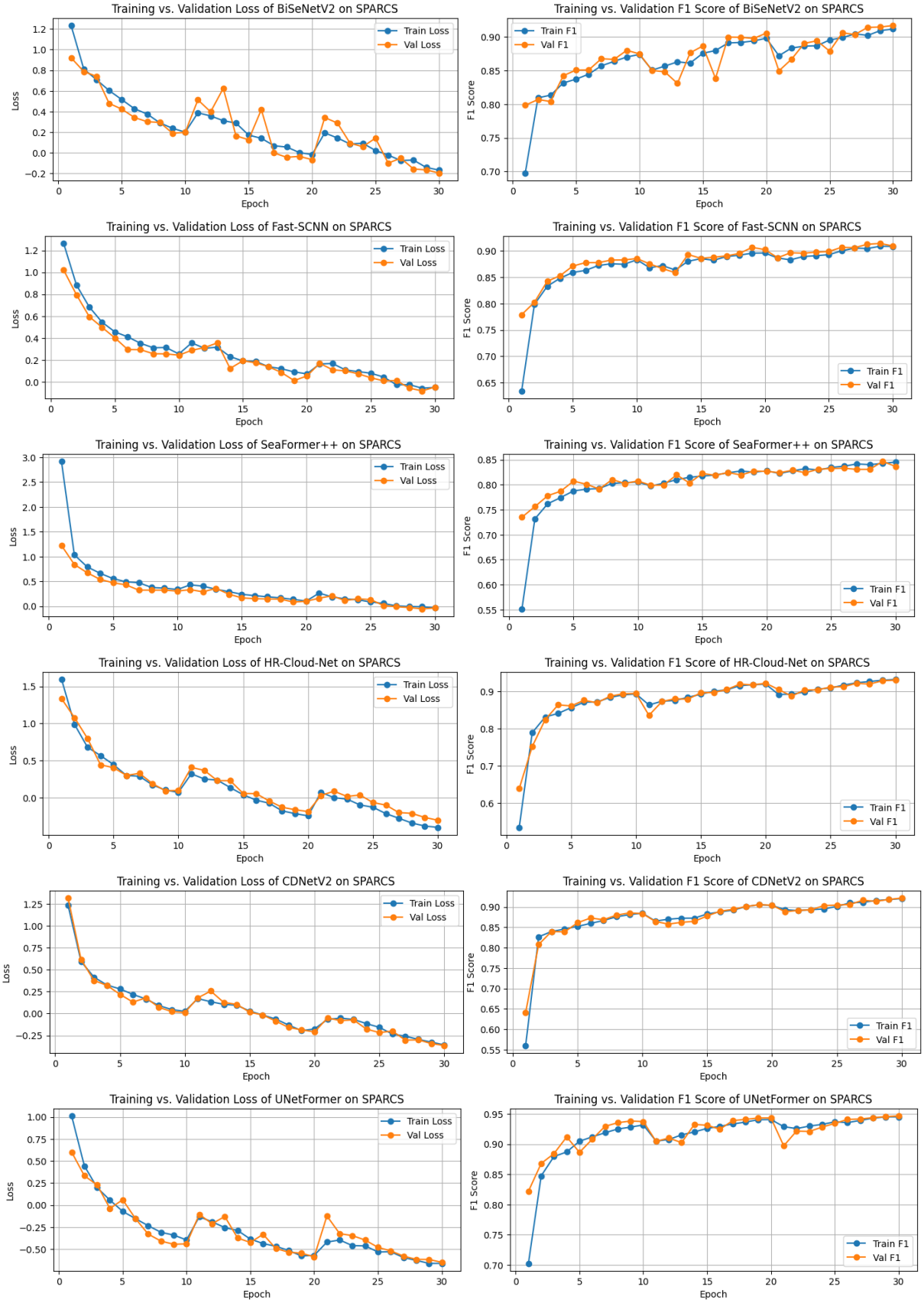
NUGRAHA PRIYA UTAMA (Member, IEEE) received the bachelor's degree in informatics from the Institut Teknologi Bandung (ITB), Bandung, Indonesia, in 2002, and the master's and Ph.D. degrees from the Tokyo Institute of Technology, in 2006 and 2009, respectively. His research interests include computer vision and neuroscience.

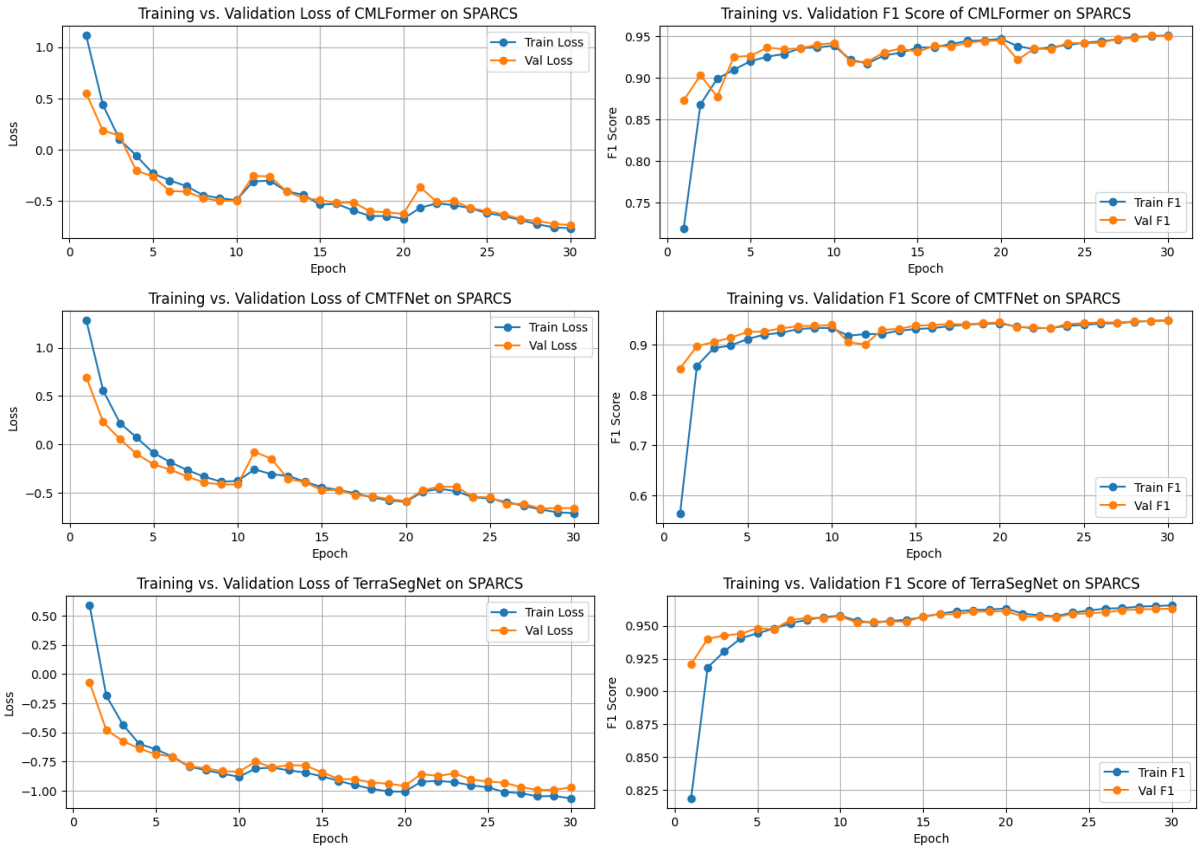
APPENDIX A
TRAINING CURVE

A. SPARCS DATASET

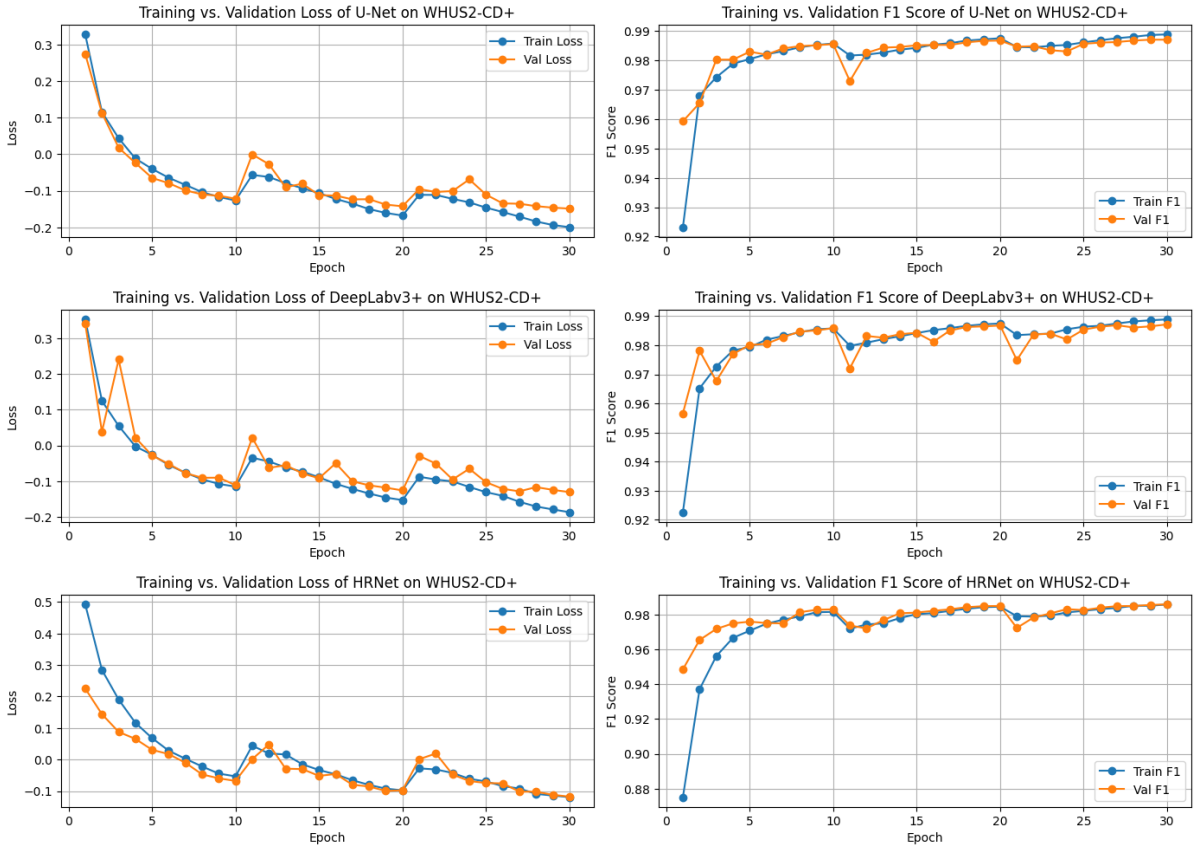


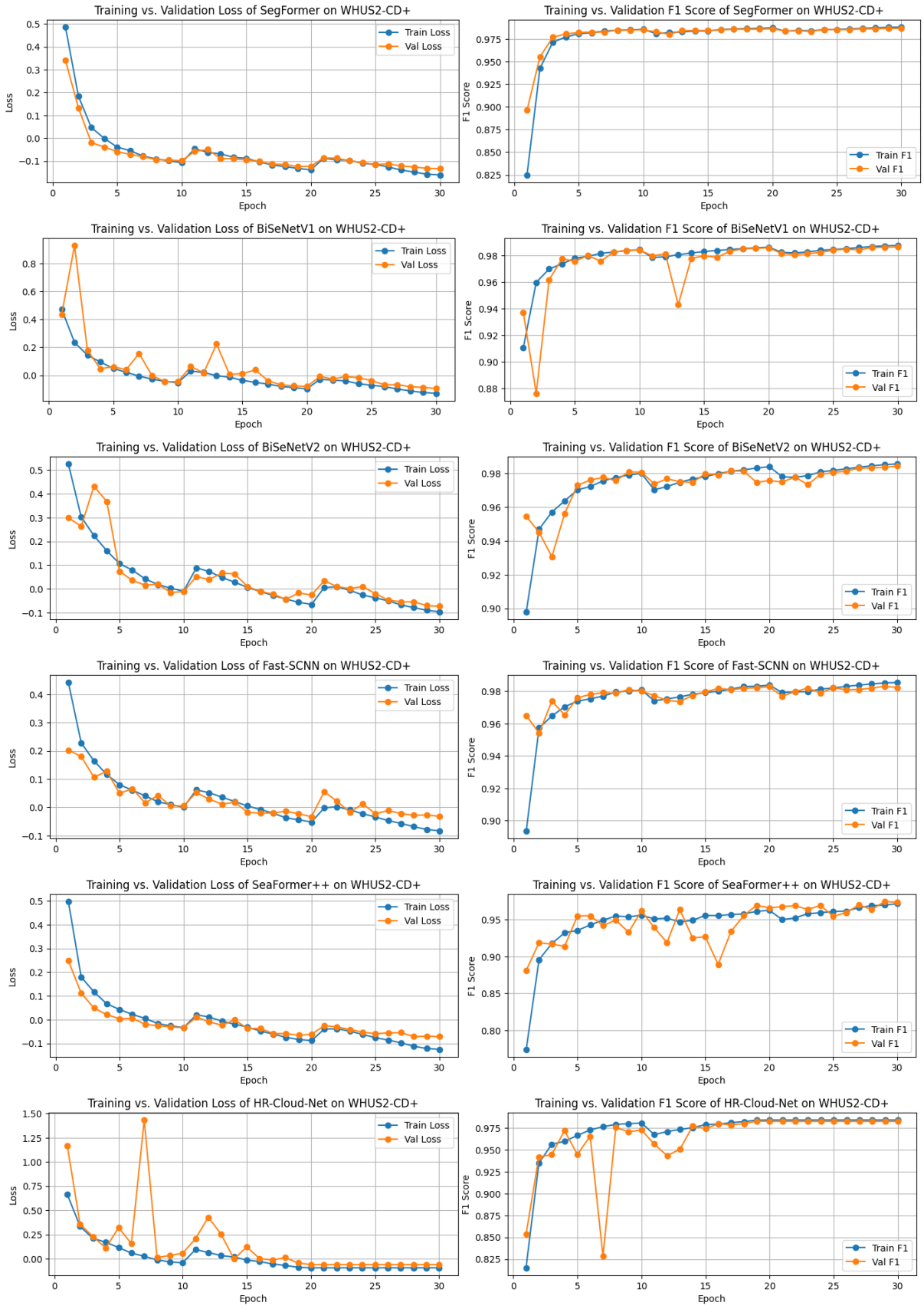
This article has been accepted for publication in IEEE Access. This is the author's version which has not been fully edited and content may change prior to final publication. Citation information: DOI 10.1109/ACCESS.2025.3633712

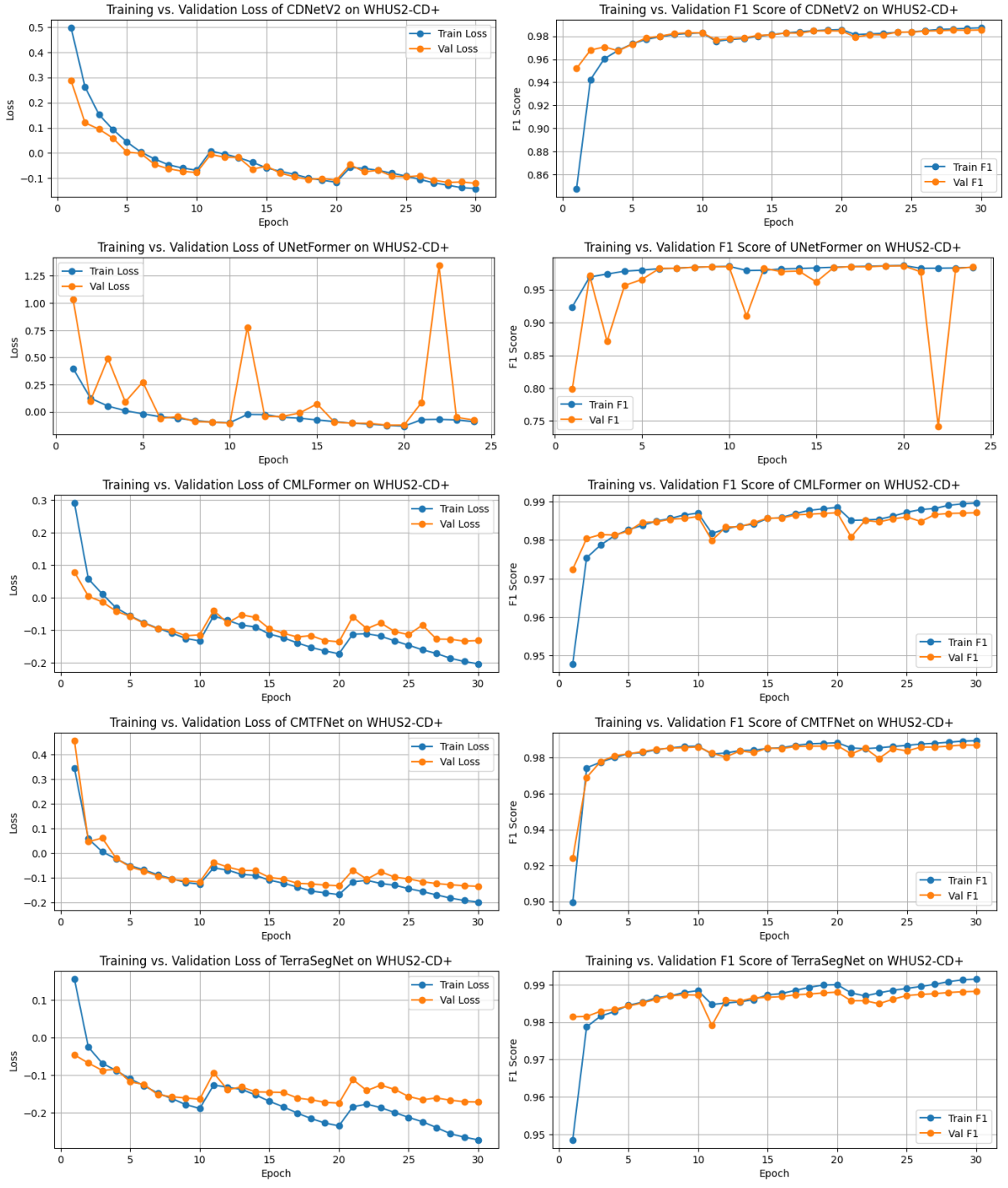




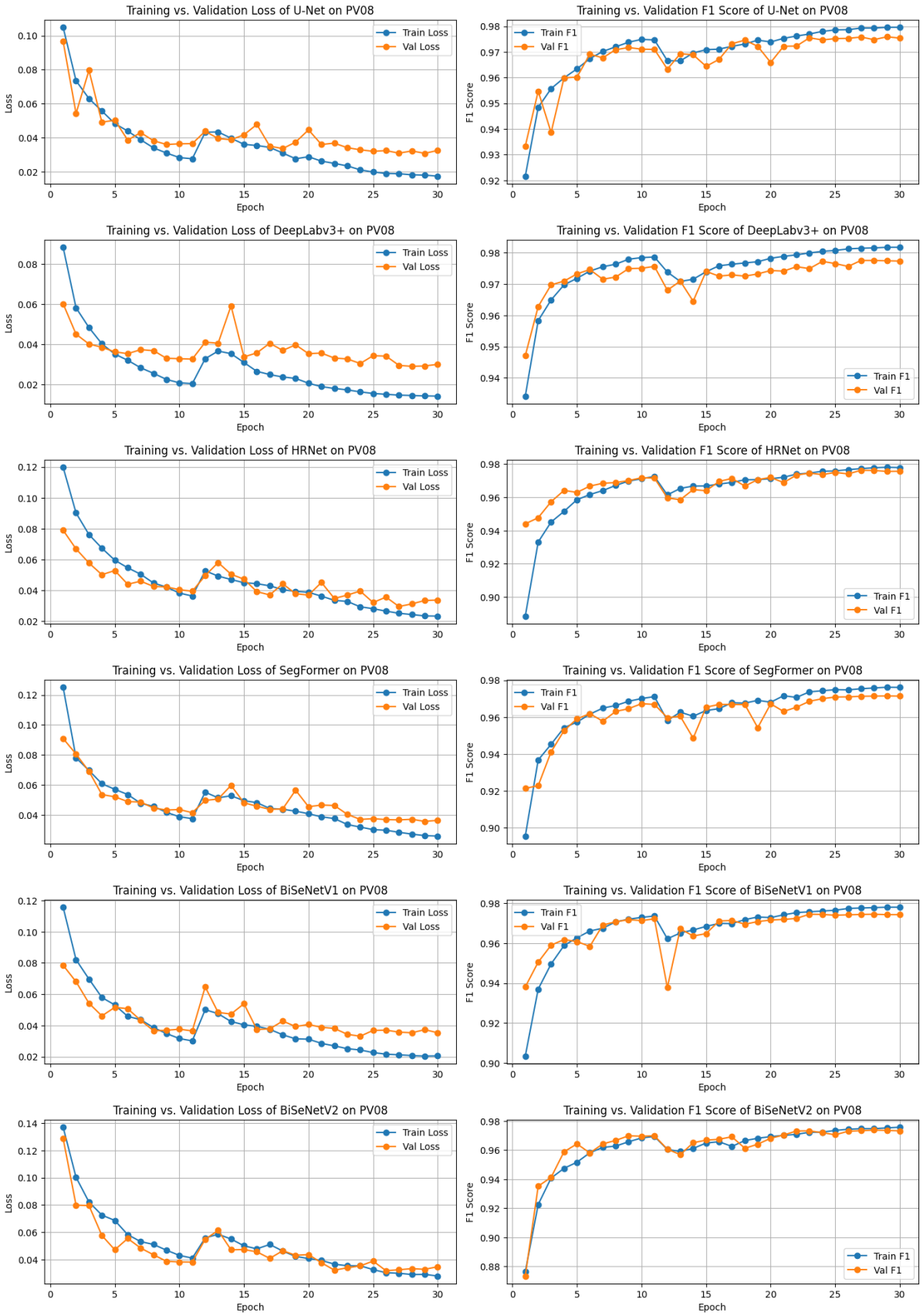
B. *WHUS2-CD+ DATASET*

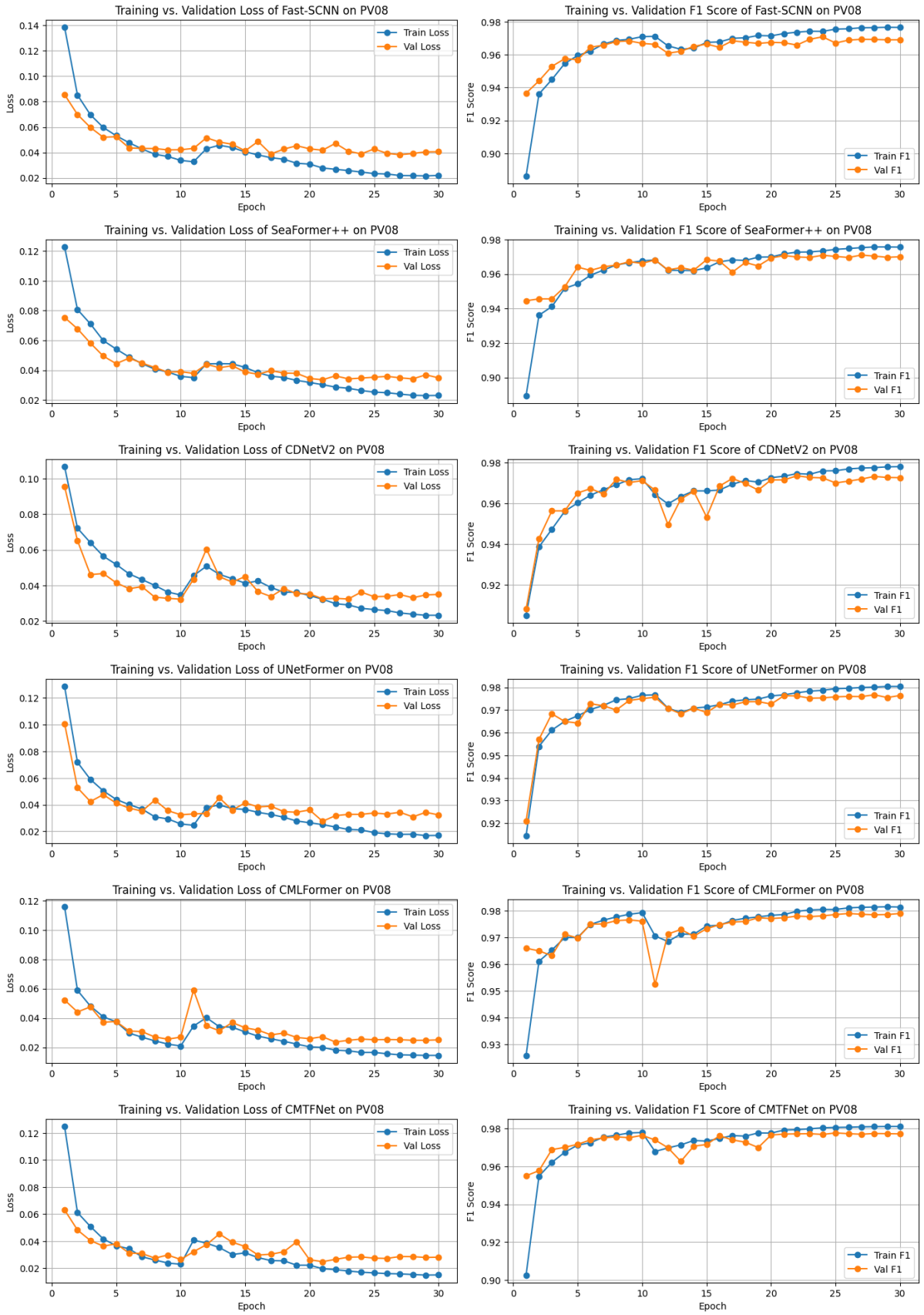


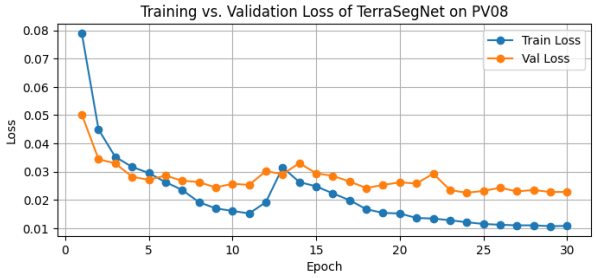


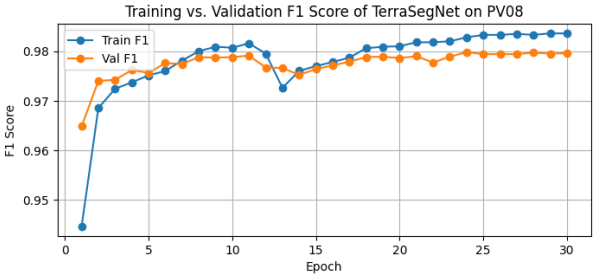


C. PV08 DATASET

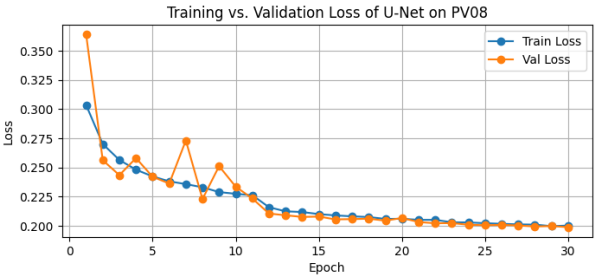


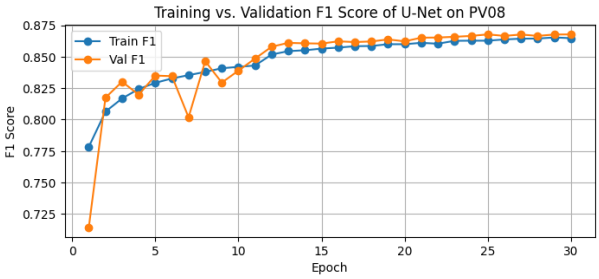


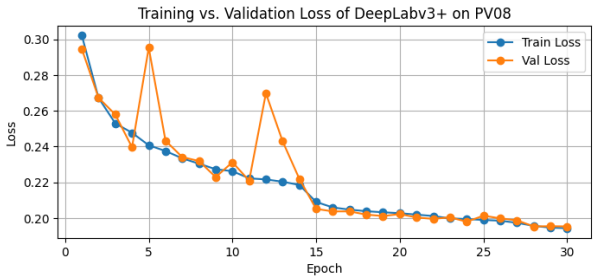


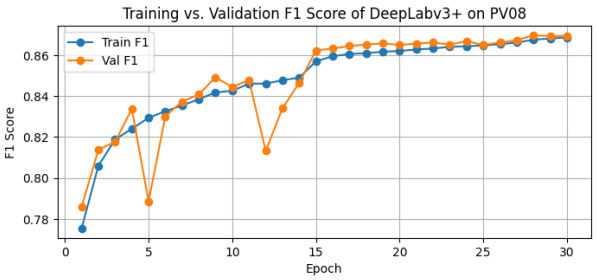


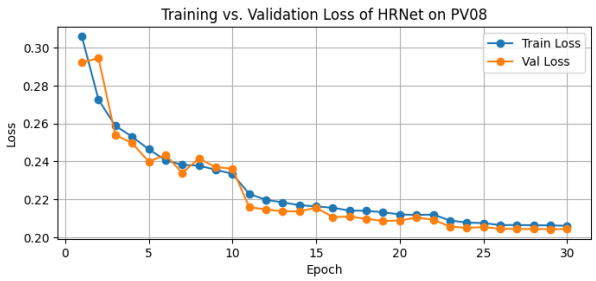
D. PLOT-RICE V1.0 DATASET

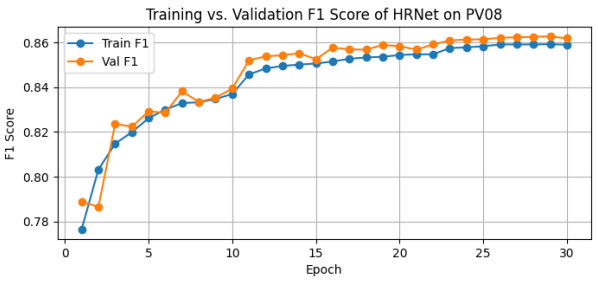


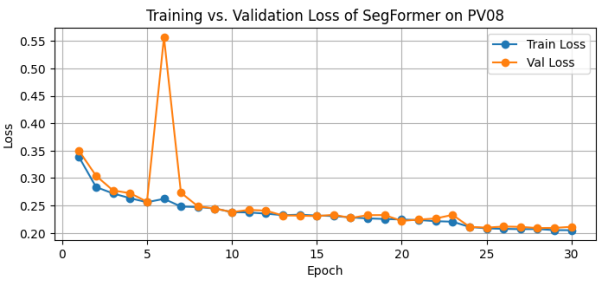


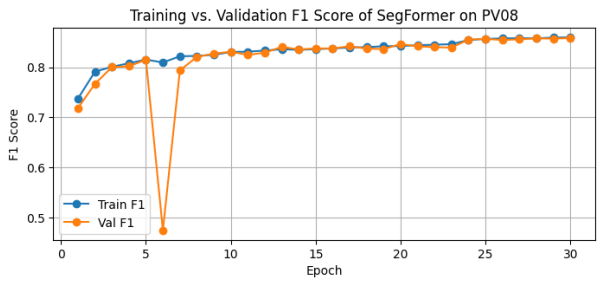


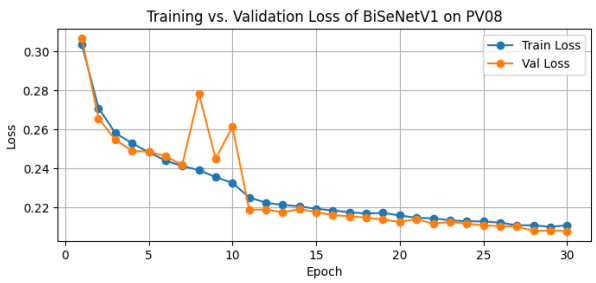


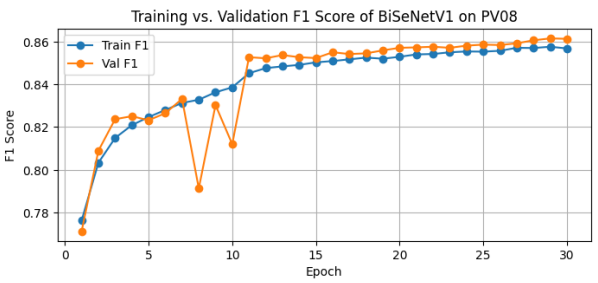




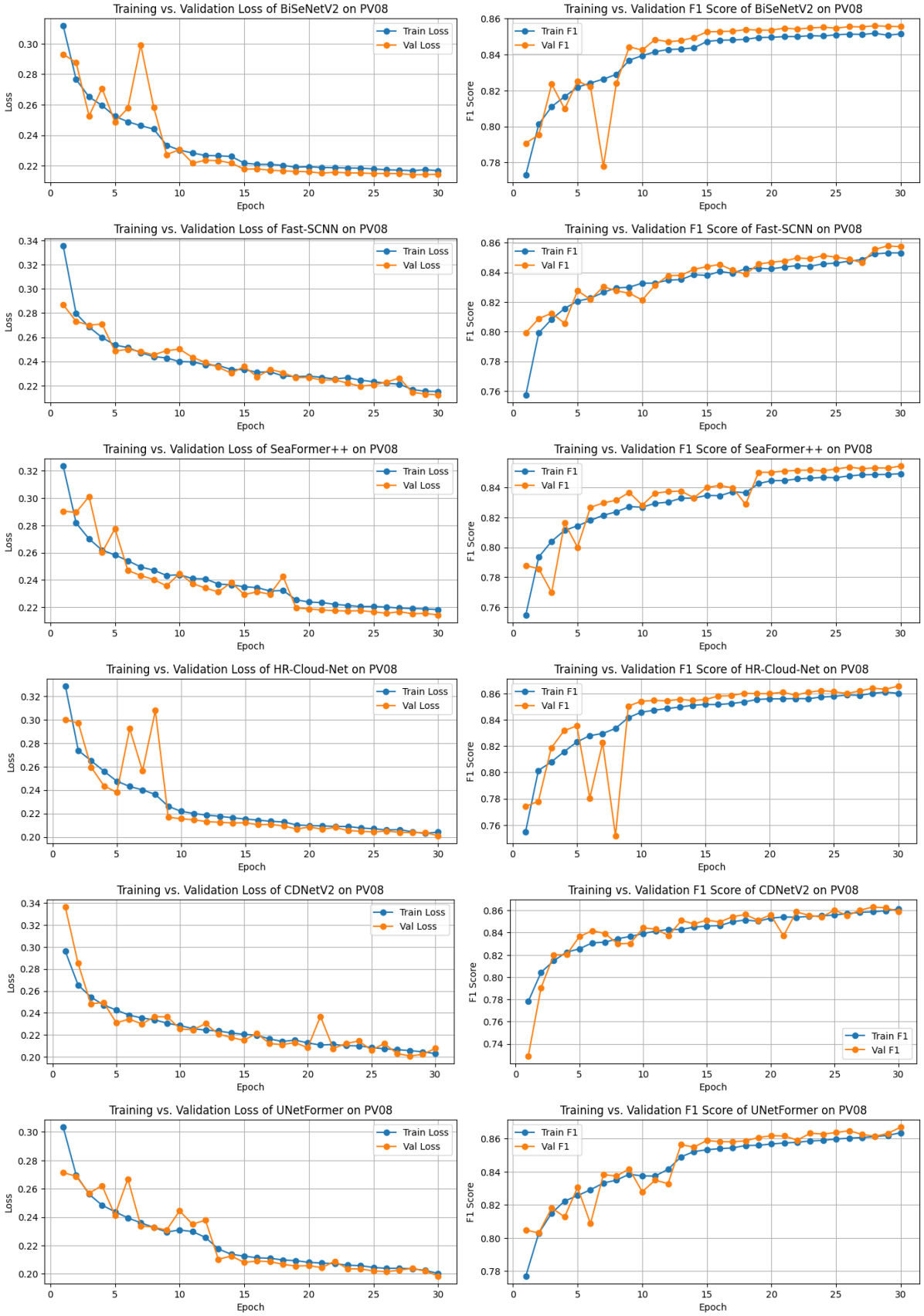


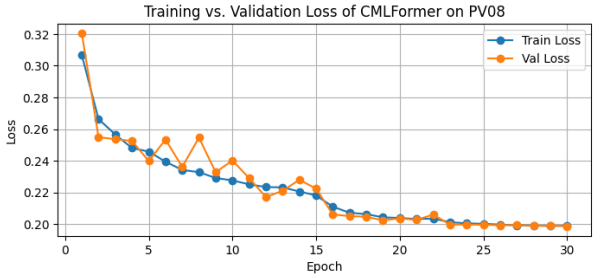


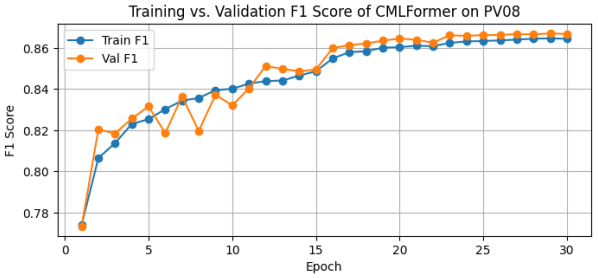


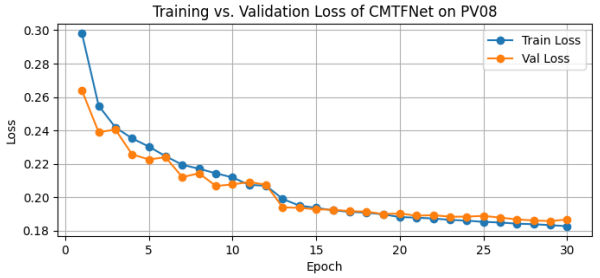


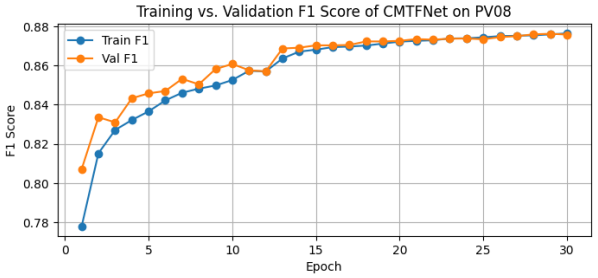
This article has been accepted for publication in IEEE Access. This is the author's version which has not been fully edited and content may change prior to final publication. Citation information: DOI 10.1109/ACCESS.2025.3633712

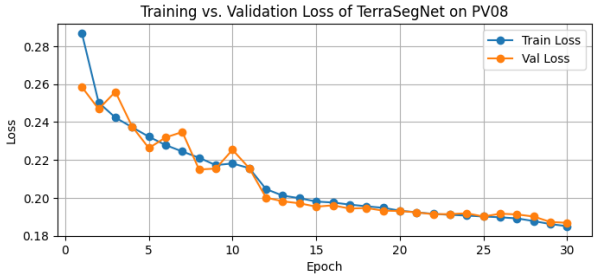


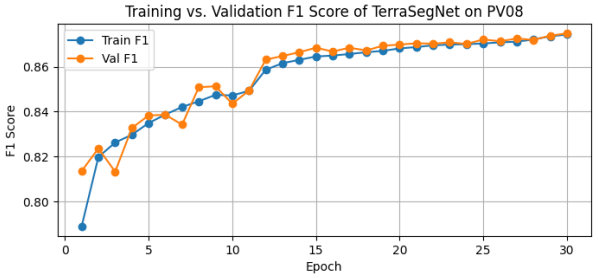












E. AIR-POLSAR-SEG-2.0 DATASET