International Journal of Geo-Information

MDPI

Article

DINOv3-MSFNet: A Multi-Scale Feature Fusion Network for Land Cover Semantic Segmentation in Mountainous Forest Ecosystems

Xu Zhao [1], Lizhi Liu [2,3] and Qiuliang Zhang [1,*]

[1] College of Forestry, Inner Mongolia Agricultural University, Hohhot 010019, China; zhaoxu20261212@emails.imau.edu.cn
[2] School of Electrical Engineering, Liaoning University of Technology, Jinzhou 121001, China; liulizhi@emails.imau.edu.cn
[3] Engineering Consulting Design and Research Institute Co., Ltd., Northeast Forestry University, Harbin 150040, China
* Correspondence: zqlemail@imau.edu.cn

Abstract

Mountainous forest ecosystems are characterized by spatially interwoven land-cover types, substantial scale variation, interclass spectral similarity, and mixed boundaries, posing challenges to both global semantic modeling and local spatial-detail recovery. This study proposes DINOv3-MSFNet, a land-cover semantic segmentation network built upon a frozen vision foundation model. Hierarchical features are extracted from different Transformer depths of DINOv3-ViT-L/16. A Feature Enhancement Module (FEM) performs channel projection and spatial reorganization, a Cross-Scale Feature Fusion Module (CSFM) enables bidirectional full-scale interaction among hierarchical representations, and a Multi-Scale Fusion Decoder (MSFD) progressively reconstructs fine-grained land-cover structures. Experiments were conducted on two Sentinel-2 RGB datasets with Dynamic World pseudo-labels and the TALANDCOVER benchmark, using spatially independent splits and separate within-dataset training for TALANDCOVER. DINOv3-MSFNet achieved OA, mF1, and mIoU values of 92.34%, 82.59%, and 72.68%, respectively, on the Greater Khingan Mountains dataset; 92.34%, 86.13%, and 77.20% on the Lesser Khingan Mountains-related dataset; and 75.61%, 65.85%, and 51.54% on TALANDCOVER, yielding the best results among the evaluated methods. Four independent random runs further showed that the improvements in mF1 and mIoU were repeatable. Ablation experiments indicated that the MSFD provided the principal performance gain, while the FEM and CSFM played complementary roles through hierarchical feature adaptation and cross-scale interaction. Regional validation using manually interpreted reference samples showed that, after limited fine-tuning, the model effectively preserved the spatial continuity of extensive forest regions, although wetland and grassland remained the principal sources of uncertainty. These evaluations assess agreement with Dynamic World pseudo-labels, within-dataset TALANDCOVER performance, and regional performance after limited fine-tuning. They support robustness within datasets, but not direct cross-regional generalization. Overall, DINOv3-MSFNet effectively integrates global semantic modeling with local boundary refinement, offering a solution for fine-grained land-cover semantic segmentation in complex mountainous forest ecosystems.

Keywords: land use and land cover; semantic segmentation; DINOv3; visual basic model; multi-scale feature fusion; mountain forest ecological region

Check for updates

Academic Editors: Wolfgang Kainz, Lirong Yin, Shan Liu and Kenan Li

Received: 17 August 2026
Revised: 12 November 2026
Accepted: 17 November 2026
Published: 20 November 2026

Copyright: © 2026 by the authors. Published by MDPI on behalf of the International Society for Photogrammetry and Remote Sensing. Licensee MDPI, Basel, Switzerland. This article is an open access article distributed under the terms and conditions of the Creative Commons Attribution (CC BY) license.

1. Introduction

In recent years, changes in land-cover patterns and surface disturbances, such as large-scale deforestation, have further highlighted the importance of land use/land cover (LULC) research [1]. Land-cover information is widely used in environmental monitoring [2], disaster management [3], agricultural production [4], and land planning [5]. With advances in remote-sensing technology, remotely sensed imagery has become an important data source for land-cover mapping because of its extensive coverage, short revisit cycles, and continuous observation capabilities [6]. Accordingly, related research has gradually shifted from conventional methods based on human experience and handcrafted features toward data-driven deep-learning approaches [7,8].

Early LULC studies relied primarily on visual interpretation, which was time-consuming and could not efficiently meet the requirements of large-area and frequently updated mapping. Subsequently, unsupervised methods, such as K-means clustering [9] and the expectation–maximization algorithm [10], as well as supervised methods, including support vector machines [11], decision trees [12], and maximum-likelihood classification [13], were widely applied. However, these methods generally rely on manually designed spectral, textural, and spatial features and remain limited in feature representation and boundary recognition in complex land-cover scenes [14].

Deep learning can automatically learn task-relevant features and has demonstrated strong representation capabilities in LULC classification and semantic segmentation [15,16]. Convolutional neural networks (CNNs) exploit local receptive fields and parameter sharing to extract textures, edges, and spatial structures. For example, Pushpalatha et al. [17] developed a CNN comprising four convolutional layers for the classification of built-up land, agricultural land, water bodies, and forest. However, the inherent locality of convolution limits the direct modeling of long-range dependencies and global contextual information [18]. Transformers enhance global relational modeling through self-attention. Zhang et al. [19] combined spectral and textural features and employed a Swin Transformer for land-cover recognition. Nevertheless, simultaneously maintaining global semantic consistency and fine spatial boundaries remains challenging in mountainous forest scenes [20].

Mountainous forest ecosystems are commonly characterized by pronounced topographic variation, fragmented land-cover patches, spatially interwoven classes, and complex transition boundaries. Classes such as forest, grassland, wetland, and cropland may exhibit similar visible-spectrum responses during particular seasons, whereas narrow objects and small-scale patches can easily be lost during downsampling. Consequently, models must not only recognize broad land-cover patterns but also accurately recover local structures and complex boundaries. Moreover, producing geographically extensive, class-consistent, and spatially independent pixel-level annotations is costly, which further limits methods that rely entirely on task-specific supervised training.

Vision foundation models (VFMs) have acquired strong general-purpose visual representations through large-scale pretraining and are increasingly being applied to remote-sensing image understanding [21–23]. Through self-supervised pretraining, DINOv3 learns spatially coherent dense visual representations and therefore provides a promising feature basis for land-cover semantic segmentation [24]. However, unlike hierarchical CNNs and conventional U-Net architectures, which rely on multi-resolution features for hierarchical representation learning and spatial detail recovery, outputs from different DINOv3 Transformer blocks generally retain the same patch-grid resolution but differ in their levels of abstraction: shallow features preserve more textural, boundary, and positional information, whereas deep features encode stronger global context and class semantics. Therefore, the principal scientific question addressed in this study is how to transform frozen features at

different semantic depths into hierarchical representations suitable for dense prediction and then enable effective cross-scale interaction and spatial reconstruction.

On this basis, we formulate and test the following hypothesis: under a frozen DINOv3 backbone, task-oriented reorganization, full-scale interaction, and progressive decoding of features extracted from different Transformer depths can improve the segmentation of narrow objects, fragmented patches, and complex class boundaries. Progressive decoding is expected to perform the principal dense-prediction function, whereas feature reorganization and cross-scale interaction are expected to provide complementary gains by improving the compatibility and complementarity of representations at different levels. This hypothesis is evaluated through comparisons with a linear segmentation baseline and representative semantic segmentation methods, module ablations, and multiple independent randomized runs.

To test this hypothesis, we propose DINOv3-MSFNet. The model extracts features from different Transformer depths of a frozen DINOv3-ViT-L/16 backbone. The Feature Enhancement Module (FEM) performs channel projection and spatial reorganization to construct hierarchical representations. The Cross-Scale Feature Fusion Module (CSFM) maps each source level to all target scales, enabling bidirectional interaction between shallow spatial information and deep semantic information. The Multi-Scale Fusion Decoder (MSFD) progressively upsamples and fuses these features to recover pixel-level predictions. The methodological innovation does not lie in basic operations such as convolution, interpolation, or skip connections themselves, but in the coordinated feature reorganization–full-scale, interaction–progressive and decoding pathway developed for frozen DINOv3 features.

Experiments were conducted under three complementary evaluation settings with spatially independent data partitions. First, two Sentinel-2 RGB datasets covering the Greater Khingan Mountains and the Lesser Khingan Mountains-related region were used to evaluate agreement between model predictions and reclassified Dynamic World pseudo-labels. Second, TALANDCOVER served as an additional benchmark, with the models trained and evaluated separately using its own data partitions. Third, manually interpreted regional reference samples were used to assess segmentation performance after limited fine-tuning, with spatially separated subsets for fine-tuning, model selection, and testing. Together, these evaluations provide complementary evidence of the effectiveness of DINOv3-MSFNet under different supervision and training settings rather than direct cross-regional transfer without retraining. The main contributions of this study are as follows:

(1) We propose DINOv3-MSFNet for frozen hierarchical DINOv3 features, coordinating the FEM, CSFM, and MSFD to establish a land-cover dense-prediction framework based on feature reorganization, full-scale interaction, and progressive decoding.
(2) We construct two Sentinel-2 RGB semantic segmentation datasets for the Greater Khingan Mountains and the Lesser Khingan Mountains-related region and employ geographically based spatially independent partitions and buffer zones to reduce spatial information leakage caused by adjacent or overlapping image patches.
(3) We systematically evaluate the model's segmentation performance, module contributions, run-to-run stability, computational cost, and the scope of result interpretation through spatially independent comparisons, module ablations, four independent randomized runs, computational-complexity analysis, and regional validation.

2. Related Works

2.1. Convolutional Neural Network

CNNs were among the earliest deep learning methods applied to LULC classification. By automatically learning hierarchical feature representations from remote sensing images through multiple convolutional layers, CNNs can enhance the discriminative capacity

of features for classification tasks. Owing to their local receptive fields, weight sharing, and spatial translation invariance, CNNs exhibit strong advantages in delineating object boundaries, extracting texture patterns, and modeling local spatial structures. Therefore, they have been widely used in land-cover classification tasks.

Kumar [25] constructed a neural network consisting of three convolutional layers and two pooling layers based on optical remote sensing images, thereby generating distribution maps for five land-use categories in mining areas. Li [26] developed the MCANet architecture based on the DeepLabv3+ model, in which optical and SAR images were fed into a pseudo-Siamese feature extraction module for feature extraction and interaction, enabling the recognition of urban areas, villages, and roads. Dong [27] employed lightweight depthwise separable convolutions and sequentially embedded position attention and channel attention modules to enhance the perception of boundaries and small objects, thereby improving land-cover classification accuracy. Reddy [28] performed principal component analysis on the original images and then constructed a 3D convolutional neural network comprising multiple Conv3D and MaxPooling3D layers for hyperspectral image classification. Albarakati [29] extracted image features using two parallel subnetworks, IBNR 65 and DenseNet 64, and optimized the network hyperparameters through Bayesian optimization to achieve land-use classification.

The expansion of the receptive field in CNNs typically relies on increasing network depth or applying pooling operations, which leads to certain limitations in modeling long-range dependencies. For LULC classification tasks characterized by complex spatial distributions, subtle inter-class differences, or significant cross-regional scene variations, relying solely on local convolutional features is often insufficient to fully capture global semantic relationships among land-cover objects. This limitation has motivated researchers to develop network architectures with stronger global modeling capabilities, thereby promoting the application of Transformer-based methods in LULC classification.

2.2. Transformer Method

Transformer-based methods have gradually become an important technical pathway in LULC classification research due to their superior global modeling capability. Unlike CNNs, which rely on local receptive fields for feature extraction, Transformers directly establish long-range dependencies among different image regions through the self-attention mechanism. This enables them to model spatial relationships and semantic associations among land-cover objects from a global perspective. Such characteristics provide Transformers with strong feature representation capability in complex surface scenes, making them particularly suitable for LULC classification tasks involving complex category distributions, significant scale variations, and pronounced spatial heterogeneity.

Xue et al. [30] first embedded multimodal data and then employed a deep Transformer for feature encoding. Subsequently, cross-attention operations were performed between class tokens and patch tokens from different modalities, and the resulting features were fed into an SVM to achieve high-accuracy classification. Shailaja et al. [31] first applied ESRGAN for super-resolution reconstruction to enhance the spatial details of remote sensing images, and then adopted Swin Transformer for classification. Fan et al. [32] combined multi-scale dilated convolutions with global–local attention, progressively fusing the global features from the Swin Transformer branch with the local features from the deformable convolution branch, thereby enhancing contextual information capture. Xu et al. [33] adaptively fused multi-scale receptive-field features using pointwise and depthwise convolutions, while selectively retaining the most relevant spatial–spectral tokens, effectively addressing the problems of fixed receptive fields and redundant information interference in land-cover classification. Han et al. [34] exploited the global self-attention mechanism of Vision Trans-

former (ViT) to capture long-range dependencies among cross-band probability maps, and introduced the Kolmogorov–Arnold Network as a classifier. Through the adaptive combination of spline basis functions, the method maintained high classification accuracy while significantly reducing the number of parameters. Recent studies have further extended advanced representation learning to weakly supervised, incomplete-supervision, and multimodal remote sensing classification. Yang et al. [35] proposed ITER, an image-to-pixel representation framework that derives dense pixel-level HSI predictions from image-level supervision, thereby reducing the dependence on refined pixel-wise annotations. To address the incomplete modeling of local details, global semantics, and information reuse in conventional HSI classifiers, Yang et al. [36] developed a full HSI classification model that combines recurrent Transformer-based representation learning with a feature reuse strategy. In addition, THSGR constructs heterogeneous graph representations for multimodal remote sensing data and employs a self-attention-free multi-convolutional modulator to model cross-modal relationships and long-range dependencies [37]. These methods demonstrate the importance of annotation-efficient learning, comprehensive local-to-global representation, and heterogeneous information interaction. Nevertheless, they mainly focus on HSI classification or multimodal classification, whereas the present study investigates fully supervised RGB semantic segmentation and emphasizes cross-scale interaction among hierarchical features extracted by a frozen vision foundation model.

Transformer-based methods still face several challenges in LULC classification. In general, Transformers entail higher computational complexity and memory consumption than conventional CNNs, and their computational burden becomes particularly pronounced when processing high-spatial-resolution remote sensing images. To further exploit the representational advantages of Transformers, recent studies have begun to introduce large-scale pretraining strategies into remote sensing image analysis. By enhancing feature transferability and generalization through pretrained foundation models, these approaches provide a new direction for the development of LULC classification research.

2.3. Pretrained Model-Based Methods

In recent years, with the rapid development of large-scale pretraining techniques, vision foundation models (VFMs) have gradually become an important research direction in the field of computer vision and have shown considerable application potential in remote sensing image understanding tasks. Unlike traditional methods trained from scratch, pretrained vision foundation models typically learn general-purpose visual representations from massive image datasets and are then transferred to specific downstream tasks. This paradigm provides a new research perspective for LULC semantic segmentation in complex scenes.

FlexiSAM [38] adopts RGB imagery as the primary modality, refines spatial and channel information using multi-scale convolutions, and injects the refined features into the SAM encoder. By fine-tuning lightweight adapters, the model ultimately produces pixel-level segmentation predictions. RSAM-Seg [39] is built upon the SAM architecture and enhances feature adaptability through a lightweight bottleneck structure. It further extracts high-frequency image components using the fast Fourier transform as prior knowledge and combines them with image embeddings to automatically generate prompts, thereby achieving land-cover segmentation. Ma et al. [40] employed SAM as a feature extractor and designed an object-consistency loss to enforce consistent predicted categories for pixels within each object. Meanwhile, a boundary-preserving loss was introduced to improve the structural completeness and boundary clarity of regular objects. Xue et al. [41] fine-tuned the frozen SAM encoder using LoRA and adopted EfficientNet-B4 as a trainable

auxiliary encoder. Feature interaction between the dual encoders was achieved through a cross-attention module, and the final classification results were generated by UperNet.

Pretrained models are still at a developing stage in LULC semantic segmentation research and require further investigation. DINOv3 exhibits strong global semantic modeling capability and high-quality dense feature representation, and has demonstrated favorable adaptability across various downstream vision tasks. Therefore, this study adopts DINOv3 as the backbone feature extraction network to enhance the model's ability to represent complex land-cover categories and provide a more reliable feature basis for semantic segmentation.

3. Methodology

3.1. Overall Structure

DINOv3 is a self-supervised Vision Transformer capable of learning highly transferable visual representations from large-scale unlabeled data. Unlike hierarchical CNNs and conventional U-Net architectures, which rely on multi-resolution features for hierarchical representation learning and spatial detail recovery, DINOv3 typically produces feature representations with the same patch-grid resolution across different Transformer blocks, but at varying levels of abstraction. Shallow-layer features retain more texture, edge, and positional information, whereas deep-layer features capture richer global contextual information and higher-level semantic representations. Its architectural design and the introduced Gram-anchoring training objective effectively alleviate the degradation of dense features in local regions, enabling the generated features to be spatially clearer and more coherent. For tasks that require fine spatial detail, such as semantic segmentation, DINOv3 can simultaneously capture texture, shape, and contextual information, thereby providing a highly accurate and robust feature foundation for remote sensing image interpretation of complex land-cover structures.

As shown in Figure 1, this study designs a model named DINOv3-MSFNet for the semantic segmentation of land cover in complex mountainous forest ecosystems. The proposed model adopts a frozen DINOv3 as the backbone feature extraction network and further incorporates a Feature Enhancement Module (FEM), a Cross-Scale Feature Fusion Module (CSFM), and a Multi-Scale Fusion Decoder (MSFD). First, the FEM enhances the detailed information and semantic representation capability of multi-scale features through channel projection and spatial adjustment. Next, the CSFM improves the collaborative consistency among features at different hierarchical levels through cross-scale information interaction and fusion, thereby alleviating the problems of semantic imbalance and insufficient spatial representation in multi-scale features. Finally, the MSFD progressively fuses and decodes the enhanced multi-scale features to achieve fine-grained segmentation of land-cover categories in complex remote sensing scenes.

3.2. Feature Enhancement Module

The proposed Feature Enhancement Module (FEM) takes the outputs of the 4th, 11th, 17th, and 23rd Transformer blocks of DINOv3-ViT-L/16 as inputs. The features from the 4th layer primarily focus on fine-grained spatial details, while those from the 11th layer emphasize local structures and regional patterns. In contrast, the features from the 17th and 23rd layers are more oriented toward high-level semantic representation and global contextual relationships. Jointly leveraging features from these four hierarchical levels enables the model to capture progressive information ranging from boundary details and regional structures to semantic context. The FEM enhances the representational capacity of features at different scales from both the channel and spatial dimensions. Its design philosophy is to refine hierarchical features through multi-scale spatial adjustment, thereby

strengthening the edge and texture representation of shallow features, the local structural modeling capability of intermediate features, and the global semantic representation ability of deep features.

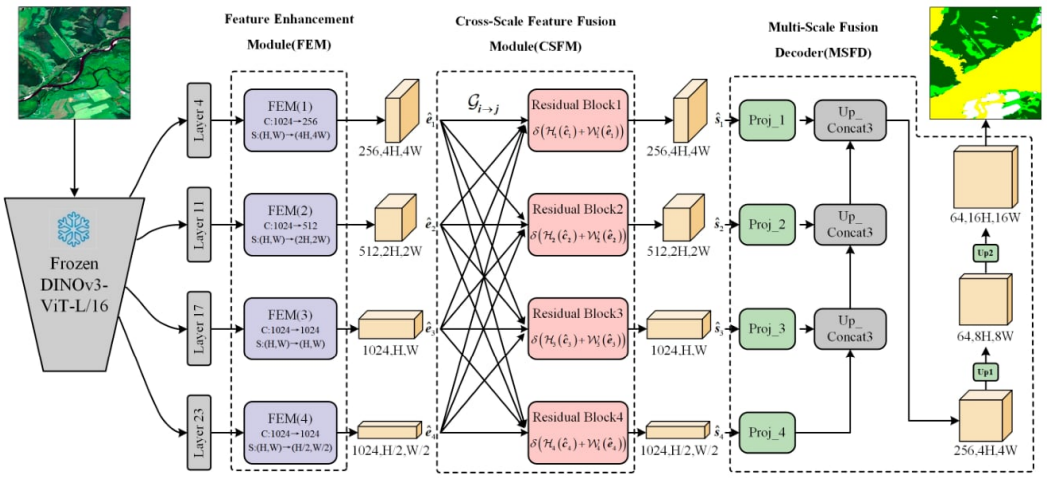


Figure 1. Overview of the proposed DINOv3-MSFNet.

Let the four feature maps output by the encoder be denoted $\{F_i\}_{i=1}^{4}$. For each feature map, the FEM first employs a lightweight Conv-BN-ReLU projection layer to perform channel-wise projection. This operation aligns the feature dimensions with the requirements of the decoder, while reducing parameter redundancy and improving the fusion compatibility of cross-layer features. Subsequently, the FEM performs spatial adjustment on each layer of features, including upsampling shallow features, moderately upsampling intermediate features, preserving the original spatial resolution of deep features, and downsampling the deepest features. In this way, the feature scales are unified or local information is further enhanced, enabling effective feature fusion in the subsequent decoding stage.

After the above channel projection and spatial adjustment operations, the feature set output by the FEM is denoted as $\{\hat{e}_i\}_{i=1}^{4}$, which serves as the input to the Multi-Scale Fusion Decoder (MSFD). This process enables a smooth transition from feature extraction to feature decoding. During forward propagation, the processing of the input feature F_i can be summarized by Equation (1), expressed as follows:

$$\hat{e}_i = \mathrm{Resize}_i\left(\mathrm{Proj}_{Li}(F_i)\right) \tag{1}$$

where Proj_{Li} denotes the Conv-BN-ReLU projection operation applied to the channel dimension of features at different hierarchical levels, and $\mathrm{Resize}_i(\cdot)$ represents the spatial adjustment operation for features at different levels, including upsampling, downsampling, or maintaining the original resolution. Through this process, the edge and texture information of shallow features is enhanced, the structural patterns of intermediate features are emphasized, and the global semantic and contextual information of deep features is preserved. Consequently, the decoder is provided with richer and more hierarchically organized multi-scale semantic representations.

Compared with directly feeding the raw features output by the backbone network into the decoder, the introduction of the FEM enhances the representational capacity and fusion compatibility of multi-scale features. This improves the model's segmentation performance for complex land-cover distributions, multi-scale objects, and class boundary regions.

3.3. Cross-Scale Feature Fusion Module

After feature extraction by the backbone network and processing by the Feature Enhancement Module (FEM), the features at different levels are enhanced to some extent at their respective scales. However, sufficient information interaction among different scales is still lacking, which may lead to inconsistency in semantic representation. To further improve the collaborative modeling capability of multi-scale features, this study introduces a Cross-Scale Feature Fusion Module (CSFM) between the FEM and the Multi-Scale Fusion Decoder (MSFD). The CSFM is designed to enable sufficient interaction and unified representation among features with different spatial resolutions.

The CSFM receives the four feature maps $\{\hat{e}_i\}_{i=1}^{4}$ from the FEM, which correspond to different spatial scales. For each input feature map, the CSFM does not perform single-scale processing alone. Instead, it constructs multi-scale transformation paths to map the feature to all target scales, thereby forming a full-scale interaction structure. Let the input feature be denoted as $\hat{e}_i$. Its representation at the target scale (j) is defined in Equation (2):

$$\hat{c}_i^j = \mathcal{G}_{i \to j}(\hat{e}_i) \tag{2}$$

where $\mathcal{G}_{i \to j}(\cdot)$ denotes the cross-scale feature transformation function. According to the relationship between the source scale and the target scale, this function adaptively selects different operation strategies. When the spatial resolution needs to be reduced, max pooling is adopted for downsampling. When the scale remains unchanged, a 3×3 convolution is used to enhance local information. When the scales are inconsistent, a 1×1 convolution is first applied for channel compression, followed by depthwise separable convolution to extract spatial information, thereby reducing computational complexity while maintaining representational capacity. In addition, for low-resolution features, bilinear interpolation is employed for upsampling to achieve spatial alignment with the target scale.

Through the above process, an aligned representation set from all input features can be obtained at each target scale j. These features are then concatenated along the channel dimension to produce the fused feature $\left\{\hat{c}_i^j\right\}_{i=1}^{4}$, as expressed in Equation (3):

$$\hat{c}_j = \mathrm{Cat}\left(\hat{c}_1^j, \hat{c}_2^j, \hat{c}_3^j, \hat{c}_4^j\right) \tag{3}$$

where $\mathrm{Cat}(\cdot)$ denotes the channel-wise concatenation operation. To further achieve efficient feature fusion, the CSFM introduces a Feature Fusion Module to model the concatenated features. This module first extracts cross-scale contextual information through a combination of pointwise convolution and depthwise convolution, then maps the features to the target channel dimension using a 1×1 convolution. A residual connection is further incorporated to enhance feature representation. The computation process can be expressed as Equation (4):

$$\hat{s}_j = \delta\left(\mathcal{H}_j\left(\hat{c}_j\right) + \mathcal{W}_j\left(\hat{e}_j\right)\right), j = 1, 2, 3, 4. \tag{4}$$

where $\mathcal{H}_j(\cdot)$ denotes the fusion branch composed of pointwise convolution and depthwise convolution, $\mathcal{W}_j(\cdot)$ represents the linear mapping of the original feature, and $\delta(\cdot)$ denotes the ReLU activation function. This residual structure effectively preserves the original feature representation while integrating cross-scale information, thereby improving the stability of network training.

The CSFM integrates feature information from all scales at each target scale. This enables shallow features to obtain stronger high-level semantic guidance, while allowing deep features to incorporate fine-grained spatial details, thereby achieving bidirectional enhancement of multi-scale features. Compared with traditional fusion strategies that rely solely on top-down or bottom-up pathways, the CSFM establishes a full-scale interaction mechanism, effectively improving the consistency and discriminative capability of feature representations.

It should be noted that the proposed CSFM differs fundamentally from the Transformer-based Heterogeneously Salient Graph Representation (THSGR) in both its research objective and feature interaction mechanism. THSGR is designed for multimodal remote sensing image semantic segmentation and constructs heterogeneous graph representations to model non-Euclidean relationships among observations from different sensor modalities. It further employs a self-attention-free multi-convolutional modulator to capture long-range dependencies. In contrast, the proposed CSFM operates on hierarchical features extracted from different depths of a single DINOv3 encoder. Rather than constructing graph nodes and inter-modal relations, CSFM maps each input feature to all target resolutions through convolutional projection, downsampling, and upsampling, followed by channel-wise concatenation and residual pointwise–depthwise convolutional fusion. Therefore, THSGR mainly addresses modality heterogeneity and cross-modal relational modeling, whereas CSFM focuses on bidirectional information exchange between shallow spatial details and deep semantic representations across network scales, thereby providing semantically consistent multi-scale features for dense land-cover prediction.

3.4. Multi-Scale Fusion Decoder

The Multi-Scale Fusion Decoder (MSFD) is designed to fully exploit the hierarchical representational capacity embedded in the enhanced multi-scale features, enabling progressive reconstruction and fusion of multi-level features. Considering the significant differences among the enhanced multi-scale features in terms of channel dimensions and semantic density, the MSFD first employs a channel projection module to align features at different scales. For an input feature, the mapping process can be expressed as follows:

$$\hat{F}_i = \delta(\mathrm{BN}(\mathrm{Conv}_{1\times1}(\hat{s}_i))) \tag{5}$$

where $\mathrm{Conv}_{1\times1}$ denotes the 1×1 convolution operation, BN denotes batch normalization, and $\delta(\cdot)$ represents the ReLU activation function. Through this mapping process, high-dimensional deep semantic features are compressed into a channel space more suitable for decoding and fusion, thereby reducing parameter redundancy while enhancing the compatibility among cross-level features. According to the network design, the four-level features are projected into 64-, 128-, 256-, and 512-channel representations, respectively, to construct a top-down decoding hierarchy.

During the feature fusion stage, the MSFD generally adopts a progressive upsampling and skip-connection strategy. However, unlike conventional encoder–decoder architectures, the features fused in this study are not classical pyramid features generated by convolutional neural networks, but hierarchical semantic representations derived from different depths of DINOv3 Transformer blocks after feature enhancement. For two adjacent scales, the deeper feature is first upsampled by a factor of 2 and then concatenated with the corresponding shallow feature along the channel dimension. The concatenated feature is subsequently integrated through two consecutive 3×3 convolutional layers. This process can be formulated as Equation (6):

$$U_i = \psi_i\big(\mathrm{Cat}\big(\hat{F}_i, \mathrm{Up}(U_{i+1})\big)\big) \tag{6}$$

where $\mathrm{Up}(\cdot)$ denotes the bilinear interpolation-based upsampling operation, $\mathrm{Cat}(\cdot)$ represents channel-wise concatenation, and $\psi_i(\cdot)$ denotes the feature transformation function composed of two convolutional layers and nonlinear activations. With this top-down progressive fusion mechanism, the global semantic information embedded in deep features can be continuously propagated to high-resolution feature layers, while the boundary and texture details contained in shallow features can be effectively preserved during fusion. Consequently, both the spatial refinement and semantic consistency of the segmentation results are improved.

After three stages of backbone-scale fusion, the MSFD further restores the feature map to the original prediction resolution through two consecutive upsampling modules. Let the final fused feature be denoted as U_1. The generation process of the output prediction map (Y) is formulated in Equation (7):

$$Y = \mathrm{Conv}_{1\times1}(\mathrm{Up}_2(\mathrm{Up}_1(U_1))) \quad (7)$$

where $\mathrm{Up}_1(\cdot)$ and $\mathrm{Up}_2(\cdot)$ denote the upsampling and convolutional refinement processes that restore the feature map from 4H × 4W to 8H × 8W and then to 16H × 16W, respectively. $\mathrm{Conv}_{1\times1}(\cdot)$ is used to project the refined features into the category space, ultimately producing the multi-class semantic segmentation result.

4. Experiment Setup

4.1. Datasets

This study used three datasets with spatially independent training, validation, and test subsets: two Sentinel-2 datasets constructed for the Greater Khingan Mountains and the Lesser Khingan Mountains-related region (Figure 2a), and the publicly available TALANDCOVER dataset [42] (Figure 2b). The models were trained, validated, and tested separately on each dataset. Therefore, these experiments constitute independent within-dataset evaluations and are not interpreted as cross-regional transfer tests.

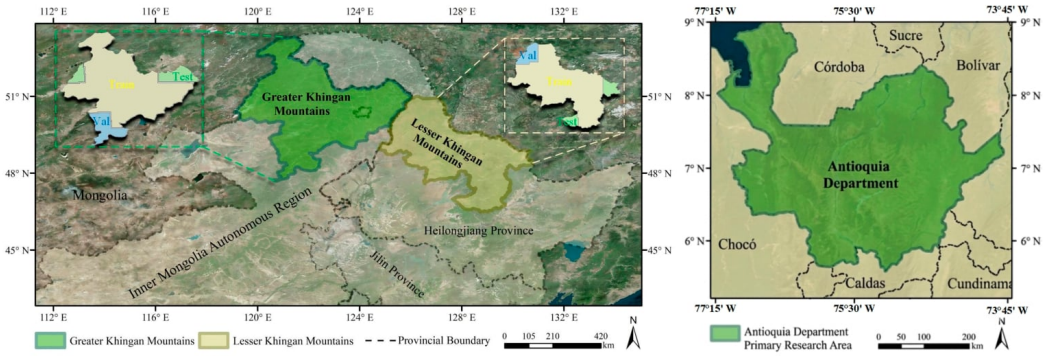


Figure 2. Locations and spatial partitions of the study areas. (**a**) Study areas in the Greater Khingan Mountains and the Lesser Khingan Mountains-related region, showing spatially independent training, validation, and test regions separated by 1 km buffer zones; (**b**) the Antioquia study area. The official spatially independent partition of the TALANDCOVER dataset was retained.

As shown in Table 1, Datasets #1 and #2 were constructed on the Google Earth Engine platform (Google LLC, Mountain View, CA, USA) using Sentinel-2 RGB imagery acquired during the peak growing season from May to September, thereby reducing the influence of seasonal differences on land-cover recognition. The imagery has a spatial resolution of 10 m, and each image-label pair has a size of 512 × 512 pixels. The labels were derived from the 10 m Dynamic World land-cover product. Dataset #1 covers the Greater Khingan Mountains and contains 8671 image-label pairs categorized as forest, grassland, wetland, and background. Dataset #2 covers the Lesser Khingan Mountains-related region and contains 7521 image-label pairs categorized as forest, grassland, cropland, and background. According to Reference [43], single-date Dynamic World classifications achieved an overall agreement of 73.8% with expert-consensus annotations across 409 global validation sites and performed relatively well in temperate and tree-dominated biomes. These findings indicate that Dynamic World can provide reasonably reliable supervisory information for this study. Nevertheless, because its accuracy varies among land-cover classes and

regional conditions, the reclassified labels were treated as pseudo-labels rather than fully independent ground truth.

Table 1. Basic information of three remote sensing semantic segmentation datasets.

Dataset	Area	Source	Label Source	Spatial Resolution	Size	Land-Cover Classes	Sample Number
#1	Greater Khingan Mountains	Sentinel-2 RGB Bands	Dynamic World pseudo-labels	10 m	512 × 512	Forest, grassland, wetland, and background	8671
#2	Lesser Khingan Mountains	Sentinel-2 RGB Bands	Dynamic World pseudo-labels	10 m	512 × 512	Forest, grassland, cropland, and background	7521
#3	Antioquia	NICFI RGB Bands	TALAND-COVER	4.77 m	128 × 128	Dense forest, grassland, bare-degraded land, Heterogeneous agricultural areas, Water bodies, Built-up areas	5000

To avoid spatial information leakage caused by assigning adjacent or overlapping image patches to different subsets, Datasets #1 and #2 were partitioned using a spatially independent strategy based on their actual geographic locations. First, the spatial footprint of each image patch was obtained from the coordinate reference system and affine transformation of its label GeoTIFF and then transformed into a UTM projected coordinate system with distances measured in meters. The image patches were subsequently grouped into spatial blocks according to their geographic locations, with all samples in the same block assigned exclusively to the training, validation, or test subset. During block allocation, an 8:1:1 ratio was used as the target for the training, validation, and test subsets, while the number of samples, class-wise pixel distributions, and spatial relationships among neighboring blocks were jointly considered. Multiple initializations were performed to obtain an optimized block-level allocation. Because image patches from adjacent blocks could still overlap or be located too close to one another, a 1 km spatial buffer was further imposed between the actual footprints of different subsets. Boundary samples that overlapped across subsets or were separated by less than 1 km were assigned to the buffer zone using a deterministic conflict-resolution strategy and excluded from subsequent experiments. Samples in the buffer zone were not used for model training, model selection, or final testing. The resulting spatial partitions were fixed before model training and used unchanged in all comparative, ablation, and repeated-run experiments.

Dataset #3 is the TALANDCOVER dataset, which contains 5000 image-label pairs with a spatial resolution of 4.77 m and a size of 128 × 128 pixels. It provides ground-truth annotations for six classes: bare and degraded land, grassland, heterogeneous agricultural areas, dense forest, water bodies, and built-up areas. Because its original partition already satisfies the requirement of spatial independence, the existing split was retained without repartitioning.

To preliminarily investigate the spatial distribution of the major land-cover types in the Greater Khingan Mountains and the Lesser Khingan Mountains-related region, 400 sampling locations were selected in a spatially uniform manner based on geographic coordinates within each study area, and a 512 × 512-pixel Sentinel-2 image was extracted at each location. Because spatially comprehensive field-reference data were unavailable and the Dynamic World pseudo-labels may contain class- and region-dependent uncertainty, pixel-level reference labels were manually delineated from Sentinel-2 imagery by one interpreter

and subsequently reviewed by a second interpreter. High-resolution Google Earth imagery and, for the Greater Khingan Mountains, the national wetland map of China [44] were used for auxiliary interpretation and cross-checking. The second interpreter reviewed the initial delineations rather than independently re-annotating the complete dataset; therefore, inter-interpreter agreement was not quantified. These labels were produced independently of the Dynamic World pseudo-labels. However, since the imagery of the same study area is used in both stages, these sample positions are not required to be spatially independent from Dataset #1 and Dataset #2. Instead, the 400 images from each study area were divided according to geographic location into spatially disjoint training, validation, and test subsets at a ratio of 3:1:1, comprising 240, 80, and 80 images, respectively. The training subset was used for limited regional fine-tuning, the validation subset was used for model selection, and the test subset was used only to evaluate agreement between model predictions and the manually interpreted reference labels; it was not involved in parameter updating or model selection. Because these validation data were not collected using a probability-based sampling design, the resulting land-cover proportions are used only to describe the spatial distribution patterns identified by the model and should not be interpreted as error-adjusted estimates of true class areas.

4.2. Experimental Configuration

In comparative and ablation experiments, the training configuration was set as follows. The training subset was used for parameter optimization, the validation subset for model selection, and the test subset exclusively for final evaluation. The input image size of #1 and #2 was 512 × 512, while that of #3 was 128 × 128, the batch size was set to 16, and the number of training epochs was set to 50. AdamW was adopted as the optimizer, with the weight decay coefficient set to 0.01 to balance convergence efficiency and parameter regularization. A cosine annealing learning rate scheduling strategy was used, with an initial learning rate of 1×10^{-4} and a minimum learning rate of 1×10^{-7}. The loss function was formulated as a joint loss consisting of Soft Cross-Entropy Loss with label smoothing and Dice Loss, with both loss weights set to 1.0. The smoothing coefficient was set to 0.05, and the ignore index was set to the number of classes, aiming to alleviate class imbalance and improve the robustness of segmentation boundaries. The backbone network was initialized with pretrained DINOv3-ViT-L/16 weights and frozen during training. Only the decoder and task-specific modules were optimized, thereby fully exploiting the representational capability of the pretrained model while reducing the parameter update cost during downstream training. The overall framework was implemented using PyTorch (version 2.8.0+cu126), and all experiments, including the comparative methods, were conducted on an NVIDIA RTX 3090 Ti GPU with 24 GB of memory.

To evaluate the sensitivity of the results to stochastic variation across different runs, four independent training runs were conducted. In each run, the complete training set was randomly shuffled without replacement, ensuring that every training sample was used. After loading, the training images were augmented on the fly using random cropping, random scaling, and random flipping, whereas no stochastic data augmentation was applied during validation or testing. The spatial data partitions, source images, pretrained backbone weights, optimizer configuration, training hyperparameters, and evaluation protocol were kept identical across runs. The reported means and standard deviations therefore reflect the overall run-to-run variability jointly arising from data-loading order, stochastic data augmentation, and initialization of the trainable modules. For each dataset, repeated experiments were conducted for the proposed method and the best-performing comparison method, and the means and standard deviations of OA, mF1, and mIoU were reported.

Model complexity was evaluated using a standardized input tensor of 3 × 512 × 512. The total parameter count includes both frozen and trainable parameters, and FLOPs were calculated for a single input image. Inference time was measured on the same NVIDIA RTX 3090 Ti GPU with a batch size of 1. Each model was set to evaluation mode with gradient computation disabled. Following two GPU warm-up iterations, 10 consecutive forward passes were performed. CUDA synchronization was applied to obtain the average inference time per image. Data loading, preprocessing, metric calculation, and result writing were excluded from the timing.

For the separate regional analysis, the models were initialized using the optimal weights obtained from Datasets #1 and #2, respectively. This weight reuse served only as the technical starting point for limited fine-tuning; the experiments on Datasets #1 and #2 themselves evaluated model agreement with the shared pseudo-label systems. During fine-tuning, the DINOv3 backbone remained frozen, and only the FEM, CSFM, and MSFD were updated. The training subsets were used for limited fine-tuning, the validation subsets for model selection, and the spatially independent test subsets exclusively for final evaluation. OA, PA, UA, F1, and IoU were calculated from the original pixel-count confusion matrices, from which mF1 and mIoU were further derived. These practices were consistent with the principles emphasized by Foody [45], Olofsson et al. [46], and Stehman and Foody [47], including test-data independence, reference-data quality control, response design, and error matrix analysis.

4.3. Evaluation Indicators

This study adopts Total Accuracy, mean F1-score (mF1), and mean Intersection over Union (mIoU) as evaluation metrics for the semantic segmentation task. These metrics are used to comprehensively assess model performance from three perspectives: overall pixel-level accuracy, class-level recognition capability, and regional overlap consistency.

Total Accuracy is used to measure the overall correctness of pixel-level prediction across all classes. Let the confusion matrix be denoted as CM, where CMij represents the number of pixels whose true class is class i but are predicted as class j. Total Accuracy is defined as the ratio of the sum of the main diagonal elements of the confusion matrix to the total number of pixels. The calculation is given in Equation (8).

$$\mathrm{CM} = \frac{\sum_{i=1}^{C} \mathrm{CM}_{ii}}{\sum_{i=1}^{C} \sum_{j=1}^{C} \mathrm{CM}_{ij}} \tag{8}$$

where C denotes the total number of classes, and CM_{ij} represents the number of correctly predicted pixels for class *i*. This metric provides an intuitive measure of the model's overall pixel-level prediction capability. However, when the class distribution is imbalanced, it can be easily dominated by classes with a large proportion of samples.

The mF1 represents the mean value of the F1-scores across all classes and is used to comprehensively evaluate the model's ability to balance precision and recall for different categories. The mean F1-score is obtained by averaging the F1-scores over all classes, as defined in Equation (9):

$$\mathrm{mF1} = \frac{1}{C} \sum_{i=1}^{C} \frac{2\mathrm{CM}_{ii}}{\sum_{j=1}^{C} \mathrm{CM}_{ij} + \sum_{j=1}^{C} \mathrm{CM}_{ji}} \tag{9}$$

where $\sum_{j=1}^{C} \mathrm{CM}_{ij}$ denotes the total number of pixels whose true class is class *i*, and $\sum_{j=1}^{C} \mathrm{CM}_{ji}$ denotes the total number of pixels predicted as class i. The mF1 metric can effectively reflect the comprehensive recognition capability of the model across different classes and is particularly suitable for performance evaluation under class-imbalanced scenarios.

Mean Intersection over Union (mIoU) is one of the most commonly used evaluation metrics in semantic segmentation tasks. It is used to measure the degree of overlap between the predicted regions and the ground-truth regions. The mIoU is obtained by averaging the Intersection over Union values across all classes, as shown in Equation (10):

$$\mathrm{mIoU} = \frac{1}{C}\sum_{i=1}^{C}\frac{\mathrm{CM}_{ii}}{\sum_{j=1}^{C}\mathrm{CM}_{ij} + \sum_{j=1}^{C}\mathrm{CM}_{ji} - \mathrm{CM}_{ii}} \tag{10}$$

where the numerator of the IoU represents the number of correctly predicted pixels for class i, and the denominator represents the number of pixels in the union of the predicted and ground-truth regions for that class. Compared with Total Accuracy, mIoU is more sensitive to both false positives and false negatives, and therefore can more accurately reflect the segmentation quality of the model across different categories, particularly in boundary regions and small-object areas.

5. Experiment Results

5.1. Comparative Experimental Analysis

In this section, DINOv3-MSFNet is compared with several state-of-the-art semantic segmentation methods developed in recent years, including SegFormer [48], UNetFormer [49], MultiSenseSeg [50], SegDINO [51], DINO U-Net [52], and AFENet [53]. The backbone architectures and optimization strategies of these methods are summarized in Table 2. Qualitative and quantitative evaluations are conducted on the three datasets. SegDINO and DINO U-Net were originally developed for medical image segmentation; therefore, their input layers, segmentation heads, and numbers of output classes were adapted to accept 3 × H × W remote-sensing images and generate predictions consistent with the land-cover taxonomy of each dataset. For all comparative methods, the officially released implementations were used, and the experiments were conducted under the same data splits and experimental settings. The evidence provided by the three datasets is interpreted separately. Metrics on Datasets #1 and #2 quantify agreement with Dynamic World-derived pseudo-labels, whereas those on TALANDCOVER quantify performance against its reference labels after within-dataset retraining.

Table 2. Backbone architectures and optimization strategies of the compared methods.

Model	Backbone	Backbone Optimization
Segformer (2021)	Mix Transformer (MiT-B0)	Trainable
UNetFormer (2022)	ResNet-18	Trainable
MultiSenseSeg (2024)	Custom Hierarchical Transformer	Trainable
SegDINO (2025)	DINOv3 ViT-L/16	Frozen
Dino U-Net (2025)	DINOv3 ViT-L/16	Frozen
AFENet (2025)	SWSL-ResNet-18	Trainable
Ours	DINOv3 ViT-L/16	Frozen

5.1.1. Greater Khingan Mountains Dataset

Table 3 shows that DINOv3-MSFNet attains OA, mF1, and mIoU values of 92.34%, 82.59%, and 72.68%, respectively. SegDINO is the second-best method for all three metrics, with 92.24% OA, 81.51% mF1, and 70.85% mIoU. The corresponding margins of the proposed method are therefore 0.10, 1.08, and 1.83 percentage points. Although these improvements

are moderate, their consistency across the three complementary metrics indicates a stable advantage under the shared experimental setting.

Table 3. Quantitative comparison results on the Greater Khingan Mountains dataset.

#1	OA	mF1	mIoU
Segformer (2021)	83.32	60.60	49.61
UNetFormer (2022)	91.66	80.09	69.82
MultiSenseSeg (2024)	90.98	78.81	68.03
SegDINO (2025)	92.24	81.51	70.85
Dino U-Net (2025)	88.73	71.73	60.29
AFENet (2025)	78.05	48.74	39.08
Ours	**92.34**	**82.59**	**72.68**

Note: Bold values indicate the best performance for each evaluation metric.

Meanwhile, SegDINO delivers the strongest overall performance among the compared methods, confirming the effectiveness of DINO-based feature representations. Nevertheless, the consistent improvements achieved by DINOv3-MSFNet suggest that a task-specific decoding and multi-scale feature-fusion mechanism can more fully exploit hierarchical DINOv3 features, particularly in terms of class-balanced recognition and spatial overlap quality.

The qualitative results in Figure 3 show that the Greater Khingan Mountains dataset primarily comprises forest, grassland, wetland, and background. Overall, SegFormer, DINO U-Net, and AFENet exhibit more pronounced misclassification, whereas SegDINO and the proposed method achieve comparatively better segmentation performance. Nevertheless, SegDINO still shows varying degrees of deviation in narrow structures, fragmented regions, and complex class-transition areas.

Specifically, Figure 3a,b contain elongated and branching wetland regions surrounded by extensive forest, placing considerable demands on boundary continuity and small-target preservation. SegFormer produces relatively fragmented predictions and isolated noisy pixels, while DINO U-Net omits substantial portions of the wetland regions. Although UNetFormer, MultiSenseSeg, SegDINO, and AFENet recover the principal wetland structures, local omissions and boundary deformation remain visible within the red boxes. The proposed method reconstructs both the major regions and small wetland targets more consistently with the reference pseudo-labels, indicating that the complete DINOv3-MSFNet more effectively preserves narrow boundaries and small-scale structures in this scene.

Figure 3c–e contain highly interwoven distributions of forest, grassland, wetland, and background, imposing greater demands on local discrimination and multiscale semantic consistency. SegFormer generates substantial noisy predictions and exhibits confusion between background and either grassland or wetland, while DINO U-Net and AFENet produce more pronounced regional misclassification. UNetFormer, MultiSenseSeg, and SegDINO recover the major regions relatively well, but still exhibit varying degrees of regional erosion, boundary displacement, and class confusion within the red boxes. In contrast, the proposed method reduces isolated misclassification while more completely reconstructing the shapes and spatial relationships of different land-cover regions.

Figure 3f,g contain large homogeneous regions together with small and fragmented targets. In Figure 3f, SegFormer produces severely fragmented predictions, while some comparison methods fail to preserve the narrow grassland structures within the extensive wetland region. In Figure 3g, SegFormer and AFENet generate noticeable false wetland predictions, whereas the other methods show varying degrees of deviation along the boundaries among forest, grassland, and background. The proposed method maintains greater class consistency across large regions while more accurately recovering the small targets

and fine boundaries highlighted by the red boxes. Consequently, its overall predictions are the most consistent with the reference pseudo-labels. These results indicate that the complete DINOv3-MSFNet more effectively balances class consistency across large regions with the recovery of small targets and fine boundaries.

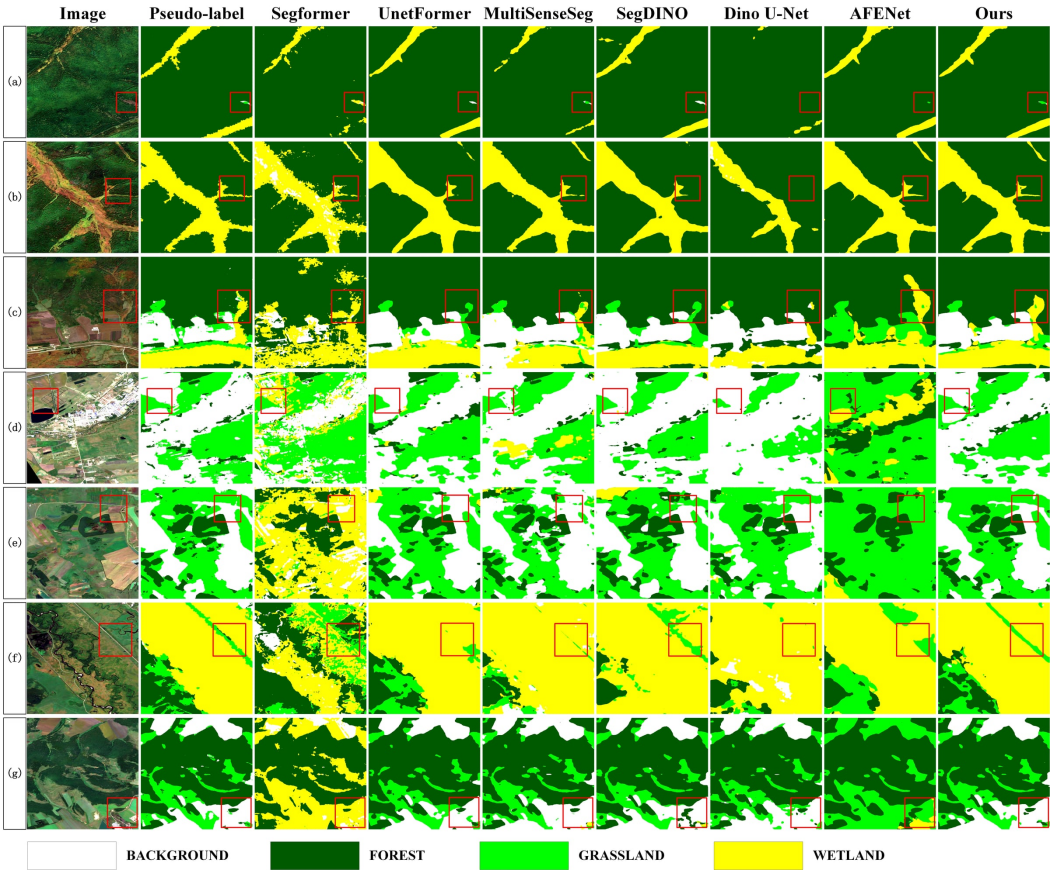


Figure 3. Qualitative comparison on the Greater Khingan Mountains dataset. (**a**,**b**) elongated and branching wetlands; (**c**–**e**) interwoven land-cover areas; (**f**) wetlands with narrow grassland structures; (**g**) complex land-cover areas with fragmented patches. Pseudo-label denotes the reclassified Dynamic World reference. Class colours follow the legend, and red boxes highlight representative local differences among the methods.

5.1.2. Lesser Khingan Mountains Dataset

To further assess the robustness of the model, quantitative and qualitative comparisons were conducted for three major land-cover classes in the Lesser Khingan Mountains region—forest, grassland, and cropland—as reported in Table 4 and Figure 4. DINOv3-MSFNet achieved an OA of 92.34%, an mF1 score of 86.13%, and an mIoU of 77.20%, outperforming the second-best method, MultiSenseSeg, by 0.56, 1.07, and 1.25 percentage points, respectively.

These results demonstrate that the proposed model performs consistently under the dataset-specific training and testing setting. The largest improvement is observed in mIoU, indicating that the model more effectively recovers spatial overlap in interwoven land-cover regions and along complex boundaries. SegFormer and DINO U-Net perform substantially worse in this scenario, while SegDINO also trails the leading methods, suggesting that pretrained representations alone cannot replace task-specific multiscale decoding and feature-interaction mechanisms. Although the Greater and Lesser Khingan Mountains are geographically and ecologically related, their class taxonomies are not directly equivalent. DINOv3-MSFNet nevertheless achieved the highest OA, mF1, and mIoU among the evalu-

ated methods on both datasets, indicating effective segmentation performance under two distinct Dynamic World pseudo-label-supervised settings.

Table 4. Quantitative comparison results on the Lesser Khingan Mountains dataset.

#2	OA	mF1	mIoU
Segformer (2021)	80.13	54.07	45.62
UNetFormer (2022)	91.61	84.62	75.05
MultiSenseSeg (2024)	91.78	85.06	75.95
SegDINO (2025)	90.69	82.91	73.15
Dino U-Net (2025)	81.04	55.76	47.08
AFENet (2025)	89.01	80.80	70.00
Ours	**92.34**	**86.13**	**77.20**

Note: Bold values indicate the best performance for each evaluation metric.

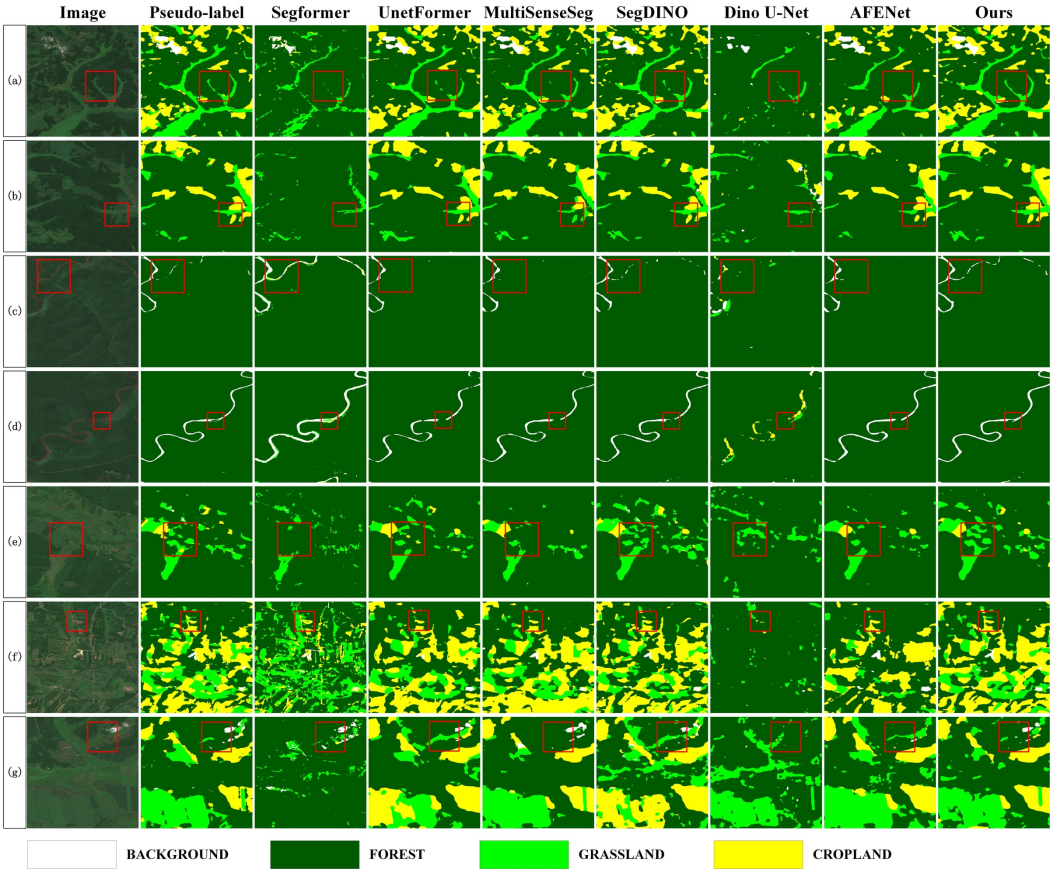


Figure 4. Qualitative comparison on the Lesser Khingan Mountains-related dataset. (**a**,**b**) forest-dominated areas with elongated grassland and cropland patches; (**c**,**d**) areas with curved linear features of the background class; (**e**) an area with interwoven grassland and cropland patches; (**f**,**g**) highly fragmented land-cover areas with grassland and cropland patches of varying sizes. Pseudo-label denotes the reclassified Dynamic World reference. Class colours follow the legend, and red boxes highlight representative local differences among the methods.

The qualitative results in Figure 4 show that the Lesser Khingan Mountains-related dataset primarily comprises forest, grassland, cropland, and background. Overall, SegFormer and DINO U-Net exhibit more pronounced class confusion and target omission, whereas UNetFormer, MultiSenseSeg, AFENet, and the proposed method achieve com-

paratively better segmentation performance. Nevertheless, all comparison methods show varying degrees of deviation along local boundaries, around small targets, and within complex interwoven regions.

Specifically, Figure 4a,b contain elongated grassland and cropland patches surrounded by extensive forest areas, placing considerable demands on local-structure recognition and boundary preservation. SegFormer and DINO U-Net omit substantial portions of the grassland and cropland regions, resulting in incomplete target structures. Although UNetFormer, MultiSenseSeg, SegDINO, and AFENet recover the major regions, boundary deformation and the omission of small targets remain evident within the red boxes. The proposed method reconstructs the narrow structures and local patches more consistently with the reference pseudo-labels, indicating that the complete DINOv3-MSFNet more accurately preserves spatial details and local structures.

Figure 4c,d contain curved, linear background targets, which place greater demands on local discrimination and multiscale semantic consistency. Most methods approximately identify their spatial locations, but local discontinuities, width deviations, and noisy regions remain. DINO U-Net, in particular, exhibits pronounced class confusion. In contrast, the proposed method more completely preserves the extension direction and boundary morphology of the targets while reducing isolated noise and local misclassification. Figure 4e contains interwoven grassland and cropland patches. SegFormer and several other methods exhibit patch omission, regional erosion, or boundary displacement, whereas the proposed method more completely reconstructs the target shapes and their spatial relationships. This result indicates that the complete DINOv3-MSFNet more accurately represents interwoven land-cover patterns and their spatial relationships.

Figure 4f,g contain highly fragmented land-cover patterns and grassland and cropland targets with substantial size variations. SegFormer and DINO U-Net show particularly severe omission of grassland and cropland information. Although SegDINO recovers the primary distribution patterns, discrepancies remain in the small patches, narrow boundaries, and class-transition regions highlighted by the red boxes. The proposed method maintains greater class consistency across large regions while more accurately recovering local contours and small targets. Its overall predictions are therefore the most consistent with the reference pseudo-labels, indicating that the complete DINOv3-MSFNet more effectively maintains large-area class consistency while recovering local contours and small-scale targets.

5.1.3. Tropical Andes Land Cover Dataset

Table 5 presents a quantitative comparison of different semantic segmentation methods on the Tropical Andes Land Cover Dataset (TALANDCOVER). The proposed DINOv3-MSFNet achieves the highest scores across all three evaluation metrics, with an OA of 75.61%, an mF1 score of 65.85%, and an mIoU of 51.54%. These results demonstrate that the proposed method provides consistent advantages in overall pixel-wise prediction accuracy, class-balanced recognition, and regional overlap quality.

Among all competing methods, SegDINO exhibits the strongest overall performance on small-sized images, achieving an OA of 74.69%, an mF1 score of 64.06%, and an mIoU of 49.73%. Compared with SegDINO, the proposed method improves these three metrics by 0.92, 1.79, and 1.81 percentage points, respectively. Although the improvement in OA is relatively modest, the more substantial gains in mF1 and mIoU indicate that the advantage of the proposed method is reflected not only in the overall number of correctly predicted pixels, but also in more balanced recognition across different land-cover classes and more accurate spatial overlap between predicted and ground-truth regions. These results further demonstrate the effectiveness of DINOv3-MSFNet on an additional mountainous land-cover

benchmark under dataset-specific training and evaluation, highlighting the applicability of the proposed architecture to different land-cover classification settings.

Table 5. Quantitative comparison results on the Tropical Andes Land Cover dataset.

#3	OA	mF1	mIoU
Segformer (2021)	63.51	33.60	23.89
UNetFormer (2022)	69.88	63.41	48.48
MultiSenseSeg (2024)	71.63	56.99	41.75
SegDINO (2025)	74.69	64.06	49.73
Dino U-Net (2025)	67.92	48.18	34.92
AFENet (2025)	70.54	59.85	45.70
Ours	**75.61**	**65.85**	**51.54**

Note: Bold values indicate the best performance for each evaluation metric.

Figure 5 provides a qualitative comparison of the segmentation results obtained by different methods on the TALANDCOVER dataset. The major land-cover classes in this dataset include bare and degraded land, grassland, heterogeneous agricultural areas, dense forest, water bodies, and built-up areas. Overall, SegFormer exhibits pronounced class confusion and spatial fragmentation, while DINO U-Net tends to miss small objects and produce overly smoothed boundaries in some complex scenes. UNetFormer, MultiSenseSeg, SegDINO, and AFENet recover the major land-cover regions relatively well, but still show varying degrees of boundary displacement or misclassification in small patches, elongated structures, and transition zones between different classes. By comparison, the predictions produced by the proposed method are more consistent with the ground truth in terms of regional integrity, boundary continuity, and local-detail recovery, while also containing fewer isolated noisy pixels.

Specifically, the scene in Figure 5a is dominated by grassland, with heterogeneous agricultural areas and irregularly shaped built-up regions distributed throughout the image. The red box highlights an elongated built-up structure with complex boundaries, posing a considerable challenge to the discrimination of built-up areas from the surrounding grassland and agricultural land. SegFormer incorrectly classifies extensive regions as built-up areas and dense forest, resulting in severe class confusion. Although MultiSenseSeg and DINO U-Net approximately locate the built-up area, their predictions contain erroneous bare-land and water classes, as well as prominent internal holes. UNetFormer and AFENet recover the main built-up structure, but noticeably overestimate its spatial extent. In contrast, the proposed method more completely reconstructs the principal outline of the built-up area within the red box and effectively reduces the intrusion of surrounding classes into the target region. Its predicted morphology is therefore more consistent with the ground truth, demonstrating stronger recognition of complex built-up boundaries and local spatial structures.

In Figure 5b, heterogeneous agricultural areas, grassland, and dense forest are spatially interwoven. SegFormer substantially overestimates the forest extent, while MultiSenseSeg, SegDINO, and DINO U-Net exhibit boundary displacement and spatial fragmentation at the interfaces between agricultural land and forest. The proposed method better preserves the overall spatial relationships between these two classes. Nevertheless, some extremely small objects within the red box remain undetected, indicating that the recognition of very small targets remains challenging.

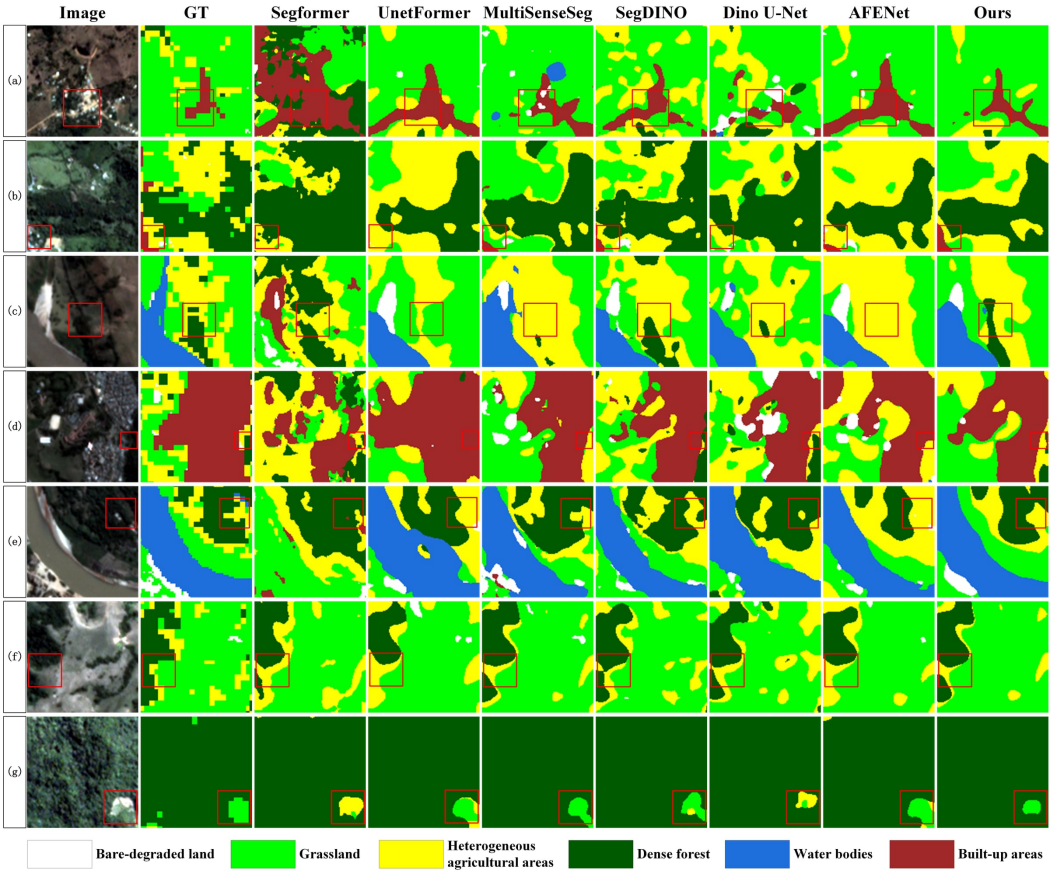


Figure 5. Qualitative comparison on the TALANDCOVER dataset. (**a**) grassland-dominated area with an elongated built-up structure; (**b**,**d**) interwoven land-cover areas; (**c**,**e**) areas with water bodies and complex land-cover boundaries; (**f**) transition zones among grassland, dense forest, and heterogeneous agricultural areas; (**g**) dense forest area with a small grassland patch. GT denotes the ground-truth reference. Class colours follow the legend, and red boxes highlight representative local differences among the methods.

Figure 5c,e both contain water bodies and complex class boundaries. In Figure 5c, the competing methods exhibit missing water regions, displaced boundaries, or omission of the elongated forest structure within the red box. The proposed method more accurately reconstructs the spatial relationships among water bodies, agricultural areas, and forest. Figure 5e contains water bodies, dense forest, grassland, and heterogeneous agricultural areas, with a curved shoreline and a small agricultural patch embedded within the forest. SegFormer exhibits substantial class confusion, while UNetFormer overestimates the agricultural patch. SegDINO and AFENet recover the main water and forest regions but show local boundary displacement, whereas DINO U-Net incompletely identifies the embedded target. In contrast, the proposed method better preserves the shoreline, interclass spatial relationships, and the morphology of the agricultural patch. In contrast, the proposed method better preserves the shoreline, interclass spatial relationships, and the morphology of the agricultural patch. These results indicate that the complete DINOv3-MSFNet achieves more accurate delineation of small embedded targets and complex boundaries in this scene.

Figure 5d is dominated by extensive built-up areas interspersed with grassland, agricultural land, and a small amount of bare land. SegFormer produces highly confused predictions, whereas MultiSenseSeg, SegDINO, and DINO U-Net exhibit obvious internal holes and fragmentation. The proposed method better preserves the continuity of the main built-up region. Although discrepancies from the ground truth remain in some agricultural

areas, the red-boxed region shows that the proposed method more effectively reconstructs the internal land-cover structure.

In Figure 5f, grassland, forest, and heterogeneous agricultural areas form complex transition zones. Several methods either overestimate or underestimate the forest boundary or generate scattered spurious patches. The proposed method more accurately reconstructs the curved boundary and elongated structure within the red box, indicating that it effectively exploits shallow spatial details to improve boundary-localization accuracy.

Figure 5g primarily evaluates the recognition of a small grassland patch embedded within a dense-forest background. SegFormer almost completely misses this target, while several other methods either overestimate its extent or misclassify it as agricultural land. The proposed method more accurately reconstructs the location, extent, and contour of the target, indicating that DINOv3-MSFNet is more effective in preserving small-scale land-cover objects within complex backgrounds.

Overall, DINOv3-MSFNet achieves a better balance between global semantic consistency and local spatial-detail preservation when segmenting elongated objects, curved shorelines, small patches, and complex class-transition regions. Consequently, it reduces class confusion, boundary displacement, and target omission. These qualitative observations are consistent with the quantitative results reported in Table 5 and demonstrate the overall effectiveness of the complete architecture in complex land-cover scenes.

5.2. Ablation Study Analysis

To evaluate the effectiveness of the proposed modules, ablation experiments were conducted on the Greater Khingan Mountains dataset, and the results are presented in Table 6. A frozen DINOv3 was used as the backbone, and eight variants were compared: the baseline with a linear segmentation head, DINOv3 + FEM, DINOv3 + CSFM, DINOv3 + MSFD, DINOv3 + FEM + CSFM, DINOv3 + FEM + MSFD, DINOv3 + CSFM + MSFD, and the complete DINOv3-MSFNet.

Table 6. Ablation study results on the Greater Khingan Mountains dataset.

No.	FEM	CSFM	MSFD	OA	mF1	mIoU
No. 1	×	×	×	90.28	76.05	65.48
No. 2	√	×	×	90.55	78.63	67.94
No. 3	×	√	×	90.95	79.11	68.57
No. 4	×	×	√	91.94	82.03	71.95
No. 5	√	√	×	90.97	78.51	68.02
No. 6	√	×	√	92.08	82.23	72.20
No. 7	×	√	√	92.34	82.33	72.40
No. 8	√	√	√	**92.34**	**82.59**	**72.68**

Note: √ indicates that the module is included, and × indicates that it is excluded. Bold values indicate the best performance for each evaluation metric. FEM: Feature Enhancement Module; CSFM: Cross-Scale Feature Fusion Module (CSFM); MSFD: Multi-Scale Fusion Decoder.

The baseline achieves an OA of 90.28%, an mF1 score of 76.05%, and an mIoU of 65.48%. Introducing the FEM alone improves these metrics by 0.27, 2.58, and 2.46 percentage points, respectively, indicating that channel projection and spatial adjustment enhance the quality and fusion compatibility of hierarchical features. The CSFM alone produces gains of 0.67, 3.06, and 3.09 percentage points, demonstrating that full-scale interaction improves information exchange between shallow spatial details and deep semantic representations.

Among the individual modules, the MSFD provides the largest improvements, increasing OA, mF1, and mIoU by 1.66, 5.98, and 6.47 percentage points, respectively. This result

confirms that task-specific progressive decoding is essential for effectively transforming hierarchical DINOv3 features into dense land-cover predictions.

Combining the FEM and CSFM without the MSFD yields an OA of 90.97%, an mF1 score of 78.51%, and an mIoU of 68.02%. The absence of an additive improvement indicates that enhanced and cross-scale-interacted features still require an effective decoder to translate their representational advantages into accurate segmentation results. When the FEM or CSFM is added to the MSFD, performance improves further. Relative to the MSFD-only variant, DINOv3 + FEM + MSFD increases OA, mF1, and mIoU by 0.14, 0.20, and 0.25 percentage points, whereas DINOv3 + CSFM + MSFD produces gains of 0.40, 0.30, and 0.45 percentage points.

The complete DINOv3-MSFNet achieves the highest mF1 and mIoU values of 82.59% and 72.68%, respectively, together with an OA of 92.34%. Compared with the baseline, the complete model improves OA, mF1, and mIoU by 2.06, 6.54, and 7.20 percentage points. Although its OA is equal to that of DINOv3 + CSFM + MSFD, adding the FEM further increases mF1 and mIoU by 0.26 and 0.28 percentage points, indicating improved class-balanced recognition and spatial overlap. Overall, the MSFD contributes the principal performance gain, while the FEM refines hierarchical spatial representations and the CSFM strengthens bidirectional cross-scale interaction. Their coordinated use enables the complete model to achieve the best overall segmentation performance.

5.3. Computational Efficiency and Repeated-Run Results

Table 7 compares the parameter counts, computational complexity, and inference time of the evaluated methods using a unified 3 × 512 × 512 input. DINOv3-MSFNet contains 255.38 M total parameters, of which 27.85 M, approximately 10.9%, are trainable because the DINOv3 backbone remains frozen. The proposed model has the highest total parameter count and computational complexity, requiring 316.67 G FLOPs. Nevertheless, its measured inference time is 0.29 s per image, which is slightly higher than that of SegDINO (0.24 s) but lower than those of the other evaluated methods (0.40–3.97 s). Its trainable parameter count is also lower than that of MultiSenseSeg (57.66 M), although it exceeds those of the remaining methods. Overall, these results characterize DINOv3-MSFNet as a computationally intensive architecture with a relatively limited proportion of trainable parameters and the second-shortest inference time under the specified experimental environment.

Table 7. Comparison of parameter counts, computational complexity, and inference time for an input tensor of 3 × 512 × 512.

Model	Total Params (M)	Trainable Params (M)	FLOPs (G)	Inference Time (s)
SegFormer	3.72	3.72	6.79	0.40
UNetFormer	11.67	11.67	11.73	1.74
MultiSenseSeg	57.66	57.66	259.37	0.41
SegDINO	230.93	2.21	241.64	0.24
DINO U-Net	246.25	18.14	102.90	0.67
AFENet	20.23	20.23	25.59	3.97
Ours	255.38	27.85	316.67	0.29

Table 8 further reports the means and standard deviations from four independent training runs with randomized data-loading orders, data-augmentation realizations, and trainable-module initialization. On Datasets #1, #2, and #3, the proposed method improves the average OA over the corresponding best-performing comparison methods by 0.15, 0.65,

and 2.50 percentage points, respectively. The corresponding improvements in mF1 are 1.46, 1.17, and 1.23 percentage points, while those in mIoU are 2.25, 1.30, and 1.57 percentage points. The standard deviations of all reported metrics do not exceed 0.76 percentage points.

Table 8. Robustness comparison across four independent random runs (mean ± standard deviation).

Dataset	Runs	Model	OA (%)	mF1 (%)	mIoU (%)
#1	4	SegDINO	92.09 ± 0.21	81.02 ± 0.29	70.13 ± 0.30
		Ours	92.24 ± 0.19	82.48 ± 0.22	72.38 ± 0.25
#2	4	MultiSenseSeg	91.58 ± 0.26	84.85 ± 0.31	75.71 ± 0.38
		Ours	92.23 ± 0.18	86.02 ± 0.27	77.01 ± 0.22
#3	4	SegDINO	73.71 ± 0.69	63.65 ± 0.36	49.19 ± 0.58
		Ours	76.21 ± 0.25	64.88 ± 0.76	50.76 ± 0.61

Note: All four training runs randomly shuffled the complete training set without replacement. The reported variation reflects the combined effects of data-loading order, stochastic data augmentation, and initialization of the trainable modules, while the spatial data partitions and training configuration remained unchanged. Not to conduct inferential tests of statistical significance.

5.4. Regional Validation Results

To further evaluate the performance of DINOv3-MSFNet in the two study areas, regional validation was conducted using manually interpreted reference samples after limited fine-tuning. As described in Section 4.1, 400 reference images were prepared for each study area and spatially divided into training, validation, and test subsets, comprising 240, 80, and 80 images, respectively. The models were fine-tuned using the training subsets, while the spatially separated test subsets were used for final evaluation. Table 9 presents the class-level accuracy metrics, and Figure 6 shows the corresponding confusion matrices.

Table 9. Class-level accuracy metrics calculated from the original confusion matrices (%).

Study Area	Class	PA	UA	F1	IoU	OA	mF1	mIoU
Greater Khingan Mountains	Background	97.20	96.75	96.97	94.12	93.41	88.33	80.04
	Forest	95.94	95.35	95.64	91.65			
	Grassland	80.43	82.89	81.64	68.98			
	Wetland	77.92	80.27	79.08	65.40			
Lesser Khingan Mountains-related area	Background	94.84	96.89	95.85	92.04	92.47	89.31	81.75
	Forest	94.14	95.63	94.88	90.26			
	Grassland	74.77	73.68	74.22	59.01			
	Cropland	94.03	90.62	92.29	85.69			

Note: PA denotes producer's accuracy, and UA denotes user's accuracy. The test set for each study area contains 80 spatially separated 512 × 512-pixel image patches.

As shown in Table 9, DINOv3-MSFNet achieved an overall accuracy (OA) of 93.41%, a mean F1-score (mF1) of 88.33%, and a mean Intersection over Union (mIoU) of 80.04% in the Greater Khingan Mountains study area. The background and forest classes exhibited relatively high classification performance, with IoU values of 94.12% and 91.65%, respectively. In contrast, wetland showed the lowest IoU of 65.40%, with misclassifications occurring mainly between wetland and forest. In the Lesser Khingan Mountains-related study area, the model achieved an OA of 92.47%, an mF1 of 89.31%, and an mIoU of 81.75%. Background, forest, and cropland achieved IoU values of 92.04%, 90.26%, and 85.69%, respectively, whereas grassland exhibited the lowest IoU of 59.01%, with confusion pri-

marily involving forest and cropland. The confusion matrices in Figure 6 further illustrate these class-dependent differences in prediction–reference agreement. The reported metrics quantify agreement with image-interpreted reference labels rather than accuracy against independent field observations.

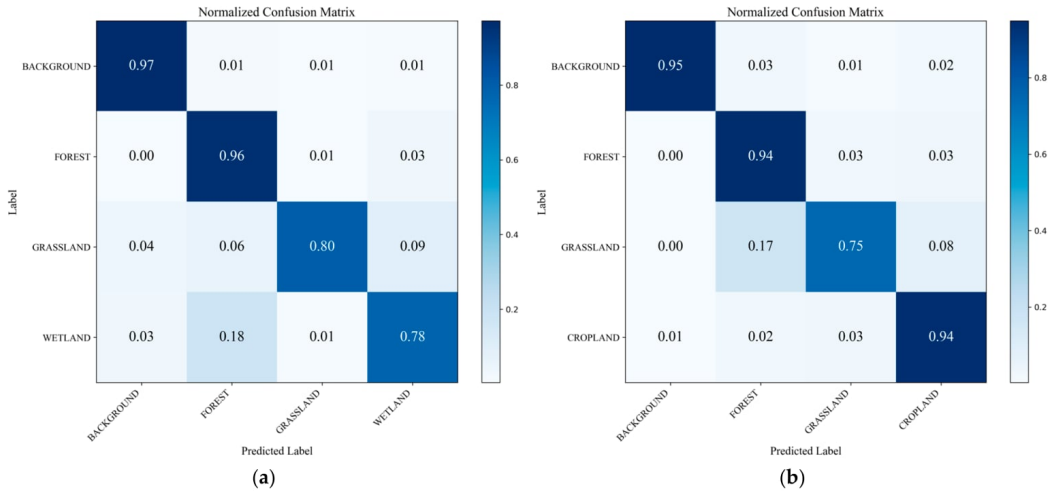


Figure 6. Row-normalized confusion matrices for the held-out, manually interpreted regional test subsets: (**a**) Greater Khingan Mountains study area; (**b**) Lesser Khingan Mountains-related study area. Rows represent reference labels, columns represent DINOv3-MSFNet predictions, and darker colours indicate higher proportions.

Figure 7 presents the land-cover segmentation maps generated by DINOv3-MSFNet for the two study areas. In the Greater Khingan Mountains study area, forest constitutes the dominant predicted land-cover class, with grassland and wetland distributed across different parts of the region. Similarly, the Lesser Khingan Mountains-related study area is predominantly classified as forest, while grassland and cropland exhibit distinct spatial distributions. These maps illustrate the regional spatial patterns of the predicted land-cover classes after fine-tuning.

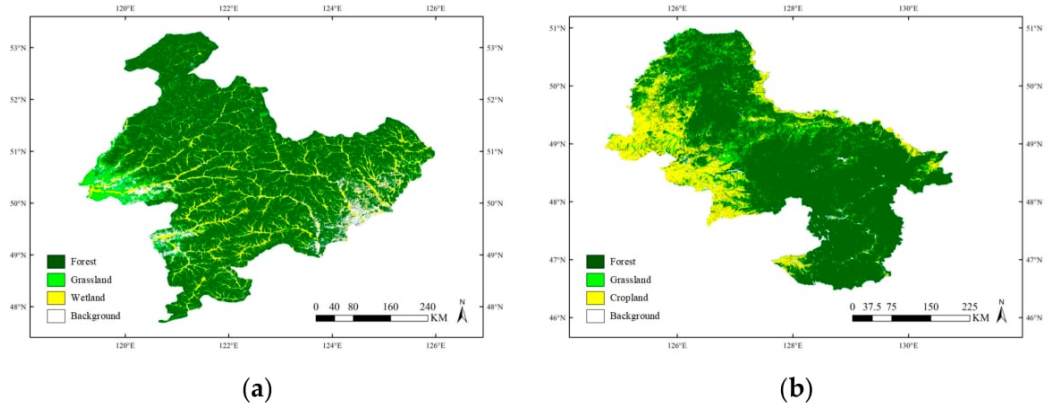


Figure 7. Regional land-cover predictions generated by DINOv3-MSFNet after limited fine-tuning: (**a**) Greater Khingan Mountains study area; (**b**) Lesser Khingan Mountains-related study area. Class colours follow the legend. The maps represent model predictions rather than independently observed land-cover maps.

Figure 8 summarizes the land-cover class proportions derived from the regional segmentation maps. Among the three reported non-background classes, forest accounts for 80.21% of the mapped area in the Greater Khingan Mountains study area, followed by

wetland (15.73%) and grassland (4.06%). In the Lesser Khingan Mountains-related study area, forest accounts for 75.47%, followed by cropland (19.93%) and grassland (4.60%). These results provide a quantitative description of the relative distributions of the predicted land-cover classes in the two study areas.

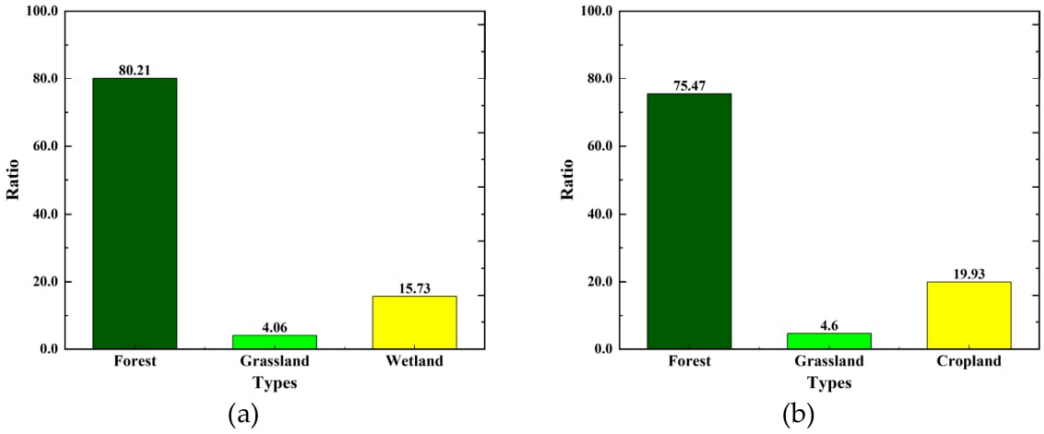


Figure 8. Predicted proportions of the non-background land-cover classes derived from the results in Figure 7: (**a**) Greater Khingan Mountains study area; (**b**) Lesser Khingan Mountains-related study area. The values are descriptive model outputs rather than independently observed or error-adjusted area estimates.

6. Discussion

6.1. Comparison with Existing Methods and Methodological Innovations

The use of DINOv3 as the encoder alone does not constitute the methodological novelty of this study. Both SegDINO [51] and DINO U-Net [52] employ representations derived from DINO-family foundation models, while the frozen-DINOv3 linear baseline in Table 6 isolates the contribution of pretrained representations without task-specific hierarchical interaction or multiscale decoding. Therefore, the principal distinction of DINOv3-MSFNet lies in how the hierarchical features extracted from different Transformer depths are reorganized, interacted, and progressively decoded for dense land-cover prediction.

Compared with conventional U-Net-like decoders and Feature Pyramid Networks, the proposed framework is not limited to adjacent-level skip connections or a single top-down feature-fusion pathway. FEM first projects the features extracted from different DINOv3 layers into appropriate channel dimensions and spatial resolutions, improving their compatibility before fusion. CSFM subsequently maps every source feature level to all target resolutions and performs residual pointwise–depthwise convolutional fusion. This design enables bidirectional full-scale interaction among shallow spatial details, intermediate local structures, and deep semantic representations before decoding. MSFD then progressively reconstructs the segmentation result through top-down upsampling and adjacent skip fusion. Thus, although MSFD adopts a U-shaped progressive decoding strategy, it operates on features that have already undergone hierarchical reorganization and full-scale interaction rather than directly decoding the original encoder outputs.

The differences from the compared methods can also be understood from their respective design objectives. SegFormer [48] combines a hierarchical Transformer encoder with a lightweight all-MLP decoder, whereas UNetFormer [49] employs a convolutional encoder and a global–local Transformer decoder. MultiSenseSeg [50] primarily focuses on multimodal feature adaptation and shared representation learning, while AFENet [53] enhances remote-sensing features through spatial–frequency interaction. Among the DINO-based methods, SegDINO [51] aligns multilevel features to a common resolution and applies a

lightweight MLP decoder, whereas DINO U-Net [52] incorporates DINOv3 representations into a U-shaped encoder–decoder architecture using specialized feature adapters. In contrast, DINOv3-MSFNet explicitly establishes full-scale, bidirectional interactions among features extracted from different Transformer depths before progressively reconstructing the prediction. This coordinated pathway is intended to reconcile high-level semantic consistency with the fine spatial details required for narrow, fragmented, and boundary-sensitive objects in mountainous land-cover scenes.

The qualitative comparisons characterize the overall segmentation behavior of the complete DINOv3-MSFNet, whereas the contributions of its individual modules are determined from the controlled ablation experiments in Table 6. Relative to the frozen-DINOv3 linear baseline, FEM, CSFM, and MSFD individually improve mIoU by 2.46, 3.09, and 6.47 percentage points, respectively, indicating that MSFD provides the largest individual contribution. When FEM or CSFM is combined separately with MSFD, mIoU increases by a further 0.25 and 0.45 percentage points, respectively. Incorporating both modules increases mIoU from 71.95% for DINOv3 + MSFD to 72.68% for the complete model, corresponding to an additional gain of 0.73 percentage points. Overall, the complete model improves OA, mF1, and mIoU over the linear baseline by 2.06, 6.54, and 7.20 percentage points, respectively.

These results suggest that FEM and CSFM should not be regarded as independent modules producing large additive improvements. Instead, they play complementary roles around the principal decoding contribution of MSFD: FEM improves the quality and compatibility of the hierarchical representations, CSFM facilitates full-scale information exchange before decoding, and MSFD converts the resulting representations into spatially precise predictions. Accordingly, the methodological contribution of DINOv3-MSFNet lies in the coordinated organization of these feature-transformation, cross-scale interaction, and decoding pathways rather than in claiming novelty for the individual convolution, interpolation, or skip-connection operations.

6.2. Model Performance and Robustness Analysis

The repeated-run results indicate that DINOv3-MSFNet improves mF1 and mIoU over the strongest comparison methods by 1.17–1.46 and 1.30–2.25 percentage points, respectively. These moderate but relatively consistent improvements should be considered alongside the model's computational requirements. As reported in Table 7, DINOv3-MSFNet has the highest total parameter count (255.38 M) and FLOPs (316.67 G) among the evaluated methods. Although freezing the DINOv3 backbone limits the number of trainable parameters to 27.85 M, it does not reduce the total model footprint or the computations required during inference. The proposed model therefore represents an accuracy–efficiency trade-off, prioritizing class-balanced recognition and spatial-overlap quality over lightweight deployment. It may be more appropriate for offline or server-side applications in which sufficient computational resources are available, and the observed accuracy improvements justify the additional cost, whereas lighter methods may remain preferable for real-time or resource-constrained applications. The measured inference time of 0.29 s was obtained using a batch size of 1 on an RTX 3090 Ti GPU and should be interpreted only as platform-specific execution performance. Because actual latency also depends on hardware parallelism, kernel implementation, memory access, numerical precision, and computational-graph structure, this benchmark does not establish an inference-efficiency advantage on edge devices. Moreover, GPU memory consumption, energy use, and latency on deployment-oriented hardware were not evaluated.

The repeated-run results also indicate that the performance advantage of DINOv3-MSFNet is more consistent in mF1 and mIoU than in OA. In particular, the OA advantage

on Dataset #1 is of a similar magnitude to the variation across runs and should not be interpreted as a decisive improvement in overall pixel-wise accuracy. By contrast, the repeated improvements in mF1 and mIoU across the three datasets support the reproducibility of the model's advantages in class-balanced recognition and spatial-overlap quality. Because no inferential significance test was conducted, this evidence supports performance repeatability and stability rather than statistical significance.

The interpretation of the results on Datasets #1 and #2 is further constrained by the use of reclassified Dynamic World pseudo-labels for both training and evaluation. Class-dependent errors, ambiguity along mixed boundaries, and regional biases in Dynamic World may be inherited by the models during training. During evaluation, predictions that reproduce erroneous pseudo-labels may be treated as correct, whereas predictions that better represent the actual land cover but disagree with the pseudo-labels may be penalized. Consequently, the reported metrics primarily quantify agreement with the shared pseudo-label system rather than absolute accuracy against independent ground truth. Nevertheless, because all methods were trained and evaluated using the same spatially independent partitions and annotations, these datasets still provide a controlled basis for relative model comparison. The repeated-run results quantify sensitivity to random training variation but do not account for systematic uncertainty in the pseudo-labels.

The three datasets were trained, validated, and tested independently. Results on Datasets #1 and #2 quantify agreement with the shared pseudo-label systems, whereas TALANDCOVER provides evidence from an additional benchmark after within-dataset retraining. These experiments support within-dataset robustness but do not demonstrate direct cross-regional or cross-resolution generalization without retraining.

6.3. Regional Mapping Uncertainty and Limitations

The regional validation results indicate that spectrally similar and spatially fragmented land-cover classes remain challenging to distinguish. Wetlands and forests may exhibit similar colors and textures in RGB imagery, while grasslands may share visible-spectrum characteristics with forests and croplands during the growing season. Mixed pixels, spatial heterogeneity, and phenological variation may further contribute to classification ambiguity. Although DINOv3-MSFNet is designed to integrate hierarchical semantic information and spatial details, these architectural features cannot fully compensate for insufficient spectral information.

The reliability of the regional analysis also depends on the quality of the reference labels. In the absence of systematic field observations, the labels were manually interpreted from remote-sensing imagery independently of Dynamic World and reviewed by a second interpreter. However, because their image locations may overlap with Datasets #1 and #2, the regional results assess agreement with manually interpreted references after limited fine-tuning rather than generalization to previously unseen areas. Independent parallel annotations were unavailable, and uncertainties may remain for spectrally similar classes and mixed-boundary regions; therefore, reference-label uncertainty cannot be quantitatively separated from model prediction errors.

Furthermore, the land-cover proportions derived from model predictions should not be interpreted as unbiased estimates of actual class areas, as probability-based sampling and area-error adjustment were not performed. Future research will incorporate additional spectral information, such as near-infrared and shortwave-infrared bands, together with topographic data and independent reference observations, to improve land-cover discrimination and regional mapping reliability.

7. Conclusions

This study proposes DINOv3-MSFNet, a semantic segmentation model built upon a frozen DINOv3 backbone, to address spatially interwoven land-cover types, pronounced scale variation, and complex boundaries in mountainous forest ecosystems. The FEM projects and spatially reorganizes features at different semantic depths, the CSFM enables full-scale interaction between shallow spatial information and deep semantic information, and the MSFD progressively reconstructs pixel-level predictions. The methodological contribution lies in establishing a coordinated feature reorganization–full-scale, interaction–progressive and decoding pathway, rather than treating DINOv3 or conventional operations such as convolution, interpolation, and skip connections as innovations.

On Datasets #1 and #2, DINOv3-MSFNet achieved the strongest agreement with the Dynamic World-derived pseudo-labels under spatially independent within-dataset partitions. After independent retraining on TALANDCOVER, the model also achieved the best comparative performance on this additional benchmark. Regional evaluation against manually interpreted references showed that, after limited target-region fine-tuning, the model effectively preserved the spatial continuity of the major land-cover types, although wetland and grassland remained subject to greater uncertainty. Multiple randomized runs supported the repeatability of the within-dataset results, while the ablation experiments identified the MSFD as the principal source of improvement and the FEM and CSFM as complementary components. Collectively, the evidence supports within-dataset robustness and performance after target-region fine-tuning, but does not demonstrate direct geographic transfer without target-region training.

This study remains limited by pseudo-label supervision, manually interpreted reference data, RGB-only input, and relatively high computational cost. Future research will incorporate field observations and additional information from NIR, SWIR, SAR, and topographic data; further evaluate reference-label uncertainty and model performance across regions and seasons; and investigate more efficient feature-fusion strategies and lightweight decoders to improve the reliability and computational efficiency of land-cover semantic segmentation.

Author Contributions: Conceptualization, Xu Zhao and Lizhi Liu; methodology, Lizhi Liu; software, Qiuliang Zhang; validation, Xu Zhao, Lizhi Liu and Qiuliang Zhang; formal analysis, Lizhi Liu; investigation, Qiuliang Zhang; resources, Qiuliang Zhang; data curation, Lizhi Liu; writing—original draft preparation, Xu Zhao; writing—review and editing, Xu Zhao; visualization, Xu Zhao; supervision, Qiuliang Zhang; project administration, Qiuliang Zhang; funding acquisition, Qiuliang Zhang. All authors have read and agreed to the published version of the manuscript.

Funding: This research was supported by the Inner Mongolia Science and Technology Program under Grant Number 2025KYPT017801, the Inner Mongolia Natural Science Foundation for Young Scientists under Grant Number 2024QN03077, the Inner Mongolia Science and Technology Program under Grant Number 2023YFDZ0026, the Heilongjiang Post-doctoral Foundation under Grant Number LBH-Z25064, and the Natural Science Foundation of Xinjiang Uygur Autonomous Region under Grant Number 2026D01C597, Inner Mongolia Autonomous Region Science and Technology "Breakthrough" Project (Unveiling and Commanding) under Grant Number 2025KJTW0023.

Data Availability Statement: The code and data required to reproduce this study are available at https://github.com/zhaoxu-doctor/DINOv3-MSFNet, accessed on 16 November 2026. The original Sentinel-2 imagery, Dynamic World labels, and TALANDCOVER dataset were obtained from https://dataspace.copernicus.eu/, https://dynamicworld.app/, and https://zenodo.org/records/8219553, accessed on 16 November 2026, respectively.

Acknowledgments: Thank you to the editors and reviewers for their hard work.

Conflicts of Interest: The authors declare no conflicts of interest.

References

1. Ologunde, O.H.; Kelani, M.O.; Biru, M.K.; Olayemi, A.B.; Nunes, M.R. Land use and land cover changes: A case study in Nigeria. *Land* **2025**, *14*, 389. [CrossRef]
2. Irfan, A.; Li, Y.; E, X.; Sun, G. Land Use and Land Cover classification with deep learning-based fusion of SAR and optical data. *Remote Sens.* **2025**, *17*, 1298. [CrossRef]
3. Benitez, F.L.; Mena, C.F.; Laso, F.; Zapata, M.B.; Rivas-Torres, G.; Gobin, A. Trajectories of farming systems reshape land use/land cover dynamics in the agricultural zones of the Galapagos Islands. *Agric. Syst.* **2026**, *231*, 104506. [CrossRef]
4. Patel, S.; Indraganti, M.; Jawarneh, R.N. A comprehensive systematic review: Impact of Land Use/Land Cover (LULC) on Land Surface Temperatures (LST) and outdoor thermal comfort. *Build. Environ.* **2024**, *249*, 111130. [CrossRef]
5. Liu, R.; Ling, J.; Zhang, H. SoftFormer: SAR-optical fusion transformer for urban land use and land cover classification. *ISPRS J. Photogramm. Remote Sens.* **2024**, *218*, 277–293. [CrossRef]
6. Chen, J.; Chen, G.; Zhou, P.; He, Y.; Yue, L.; Ding, M.; Lin, H. CTSeg: CNN and ViT collaborated segmentation framework for efficient land-use/land-cover mapping with high-resolution remote sensing images. *Int. J. Appl. Earth Obs. Geoinf.* **2025**, *139*, 104546. [CrossRef]
7. Nandi, D.; Thakur, R.R.; Bera, D.K.; Pati, A.K.; Mishra, P.; Alshehri, F.; Zhran, M. A Deep Learning-Geospatial Analytics Fusion Framework for Predicting Land Use Land Cover Dynamics in Eastern Region of India. *Results Eng.* **2026**, *29*, 109201. [CrossRef]
8. Li, W.; Liang, S.; Chen, K.; Chen, Y.; Ma, H.; Xu, J.; Ma, Y.; Zhang, Y.; Guan, S.; Fang, H. AgriFM: A multi-source temporal remote sensing foundation model for Agriculture mapping. *Remote Sens. Environ.* **2026**, *334*, 115234. [CrossRef]
9. Lemenkova, P.; Debeir, O. R Libraries for remote sensing data classification by k-means clustering and NDVI computation in Congo River Basin, DRC. *Appl. Sci.* **2022**, *12*, 12554. [CrossRef]
10. Taghavi-Shahri, S.M.; Fassò, A.; Mahaki, B.; Amini, H. Concurrent spatiotemporal daily land use regression modeling and missing data imputation of fine particulate matter using distributed space-time expectation maximization. *Atmos. Environ.* **2020**, *224*, 117202. [CrossRef]
11. Jozdani, S.E.; Johnson, B.A.; Chen, D. Comparing deep neural networks, ensemble classifiers, and support vector machine algorithms for object-based urban land use/land cover classification. *Remote Sens.* **2019**, *11*, 1713. [CrossRef]
12. Phiri, D.; Simwanda, M.; Nyirenda, V.; Murayama, Y.; Ranagalage, M. Decision tree algorithms for developing rulesets for object-based land cover classification. *ISPRS Int. J. Geo-Inf.* **2020**, *9*, 329. [CrossRef]
13. Agumagu, O.O.; Marchant, R.; Stringer, L.C. Land use and land cover change dynamics in the Niger Delta region of Nigeria from 1986 to 2024. *Land* **2025**, *14*, 765. [CrossRef]
14. Subbarayan, S.; Youssef, Y.M.; Singh, L.; Dąbrowska, D.; Alarifi, N.; Ramsankaran, R.; Visweshwaran, R.; Saqr, A.M. Soil and water assessment Tool-Based prediction of runoff under scenarios of land use/land cover and climate change across Indian Agro-Climatic zones: Implications for sustainable development goals. *Water* **2025**, *17*, 458. [CrossRef]
15. Li, J.; Yan, Y.; Stouffs, R. Simulating land surface temperature impacts of proposed land use and land cover plans using an integrated deep neural network approach. *Energy Build.* **2025**, *332*, 115437. [CrossRef]
16. Li, X.; Xu, F.; Li, J.; Su, Y.; Li, L.; Lyu, X.; Xu, Z.; Kaup, A. Frequency Domain-Enhanced Spectral-Spatial Fusion Transformer for Semantic Segmentation of Remote Sensing Images. *Inf. Fusion* **2026**, *132*, 104248. [CrossRef]
17. Pushpalatha, V.; Mallikarjuna, P.; Mahendra, H.; Subramoniam, S.R.; Mallikarjunaswamy, S. Land use and land cover classification for change detection studies using convolutional neural network. *Appl. Comput. Geosci.* **2025**, *25*, 100227. [CrossRef]
18. Shu, T.; Shen, L.; Wang, Y.; Zhang, P. FADConv: A frequency-aware dynamic convolution for cropland non-agriculturalization identification and segmentation. *Int. J. Appl. Earth Obs. Geoinf.* **2026**, *146*, 105014. [CrossRef]
19. Zhang, Y.; Huang, M.; Chen, Y.; Xiao, X.; Li, H. Land cover classification in high-resolution remote sensing: Using Swin Transformer deep learning with texture features. *J. Spat. Sci.* **2025**, *70*, 205–229. [CrossRef]
20. Zhou, R.; Li, J.; Zhang, W.; Li, S.; Chen, J.; Li, C.; Tong, B.; Zhang, W.; Sun, X.; Fu, K. CAMformer: A Single-Stage CNN-Transformer Hybrid for Weakly Supervised Semantic Segmentation in Aerial Imagery. *IEEE Trans. Geosci. Remote Sens.* **2026**, *64*, 5605020. [CrossRef]
21. Zhu, Y.; Huang, B.; Zhou, F.; Liu, Y. Causal-SAM: Enhancing Segment Anything for Remote Sensing Instance Segmentation via Causal Representation Learning. *Knowl.-Based Syst.* **2026**, *343*, 115915. [CrossRef]
22. Xiong, S.; Wu, L.; Zhang, B.; Zhang, D.; Tao, Y.; Tang, Y. HL-SAM-Seg: Complementary High-and Low-Resolution Features Based on SAM for Remote Sensing Image Semantic Segmentation. *IEEE Trans. Geosci. Remote Sens.* **2026**, *64*, 4402214. [CrossRef]
23. Jiang, Y.; Yuan, Z.; Zhou, T.; Chen, J.; Xie, F.; Jiang, Z.; Zhang, H. SGCT-Net: SAM-Guided Cross-Teaching Network for Weakly Supervised Semantic Segmentation for Generating High-Quality CAMs in High-Resolution Remote Sensing Imagery. *IEEE J. Sel. Top. Appl. Earth Obs. Remote Sens.* **2026**, *19*, 13395–13407. [CrossRef]
24. Siméoni, O.; Vo, H.V.; Seitzer, M.; Baldassarre, F.; Oquab, M.; Jose, C.; Khalidov, V.; Szafraniec, M.; Yi, S.; Ramamonjisoa, M. Dinov3. *arXiv* **2025**, arXiv:2508.10104. [CrossRef]

25. Kumar, A.; Gorai, A.K. A comparative evaluation of deep convolutional neural network and deep neural network-based land use/land cover classifications of mining regions using fused multi-sensor satellite data. *Adv. Space Res.* **2023**, *72*, 4663–4676. [CrossRef]
26. Li, X.; Zhang, G.; Cui, H.; Hou, S.; Wang, S.; Li, X.; Chen, Y.; Li, Z.; Zhang, L. MCANet: A joint semantic segmentation framework of optical and SAR images for land use classification. *Int. J. Appl. Earth Obs. Geoinf.* **2022**, *106*, 102638. [CrossRef]
27. Dong, Y.; Liu, Q.; Du, B.; Zhang, L. Weighted feature fusion of convolutional neural network and graph attention network for hyperspectral image classification. *IEEE Trans. Image Process.* **2022**, *31*, 1559–1572. [CrossRef]
28. Reddy, C.K.K.; Daduvy, A.; Mohana, R.M.; Assiri, B.; Shuaib, M.; Alam, S.; Sheneamer, M.A. Enhancing precision agriculture and land cover classification: A self-attention 3D convolutional neural network approach for hyperspectral image analysis. *IEEE Access* **2024**, *12*, 125592–125608. [CrossRef]
29. Albarakati, H.M.; Khan, M.A.; Hamza, A.; Khan, F.; Kraiem, N.; Jamel, L.; Almuqren, L.; Alroobaea, R. A novel deep learning architecture for agriculture land cover and land use classification from remote sensing images based on network-level fusion of self-attention architecture. *IEEE J. Sel. Top. Appl. Earth Obs. Remote Sens.* **2024**, *17*, 6338–6353. [CrossRef]
30. Xue, Z.; Yang, G.; Yu, X.; Yu, A.; Guo, Y.; Liu, B.; Zhou, J. Multimodal self-supervised learning for remote sensing data land cover classification. *Pattern Recognit.* **2025**, *157*, 110959. [CrossRef]
31. Shailaja, P.; Kumar, P.M.; Nikhitha, N.; Reddy, K.N.K.; Reddy, E.M.; Reddy, G.G.; Indu, V. LCC-NET: SWIN Transformer-CNN Hybrid for enhanced land cover classification in natural disaster monitoring. *Syst. Soft Comput.* **2025**, *7*, 200303. [CrossRef]
32. Fan, Y.; Zhang, D.; Li, J.; Xiao, J. A focusing-attention deformable convolution and transformer network with multi-scale contour-render for land cover classification in high-resolution remote-sensing images. *Eng. Appl. Artif. Intell.* **2025**, *160*, 111949. [CrossRef]
33. Xu, Y.; Wang, D.; Zhang, L.; Zhang, L. Dual selective fusion transformer network for hyperspectral image classification. *Neural Netw.* **2025**, *187*, 107311. [CrossRef]
34. Han, S.; Ren, D.; Gao, F.; Yang, J.; Ma, H. ViT–KAN synergistic fusion: A novel framework for parameter-efficient multi-band PolSAR land cover classification. *Remote Sens.* **2025**, *17*, 1470. [CrossRef]
35. Yang, J.; Du, B.; Wang, D.; Zhang, L. ITER: Image-to-pixel representation for weakly supervised HSI classification. *IEEE Trans. Image Process.* **2024**, *33*, 257–272. [CrossRef]
36. Yang, J.; Du, B.; Zhang, L. Overcoming the barrier of incompleteness: A hyperspectral image classification full model. *IEEE Trans. Neural Netw. Learn. Syst.* **2024**, *35*, 14467–14481. [CrossRef]
37. Yang, J.; Du, B.; Liu, R.; Mao, Z.; Zhang, L. Boosting multimodal remote sensing image classification with Transformer-based heterogeneously salient graph representation. *IEEE Trans. Geosci. Remote Sens.* **2026**, *64*, 5512117. [CrossRef]
38. Zhang, Z.; Shu, D.; Liao, C.; Liu, C.; Zhao, Y.; Wang, R.; Huang, X.; Zhang, M.; Gong, J. FlexiSAM: A flexible SAM-based semantic segmentation model for land cover classification using high-resolution multimodal remote sensing imagery. *ISPRS J. Photogramm. Remote Sens.* **2025**, *227*, 594–612. [CrossRef]
39. Zhang, J.; Li, Y.; Yang, X.; Jiang, R.; Zhang, L. Rsam-seg: A SAM-based model with prior knowledge integration for remote sensing image semantic segmentation. *Remote Sens.* **2025**, *17*, 590. [CrossRef]
40. Ma, X.; Wu, Q.; Zhao, X.; Zhang, X.; Pun, M.-O.; Huang, B. SAM-assisted remote sensing imagery semantic segmentation with object and boundary constraints. *IEEE Trans. Geosci. Remote Sens.* **2024**, *62*, 5636916. [CrossRef]
41. Xue, B.; Cheng, H.; Yang, Q.; Wang, Y.; He, X. Adapting segment anything model to aerial land cover classification with low-rank adaptation. *IEEE Geosci. Remote Sens. Lett.* **2024**, *21*, 2502605. [CrossRef]
42. Gomez-Ossa, L.F.; Sanchez-Torres, G.; Branch-Bedoya, J.W. Land cover classification in the Antioquia region of the Tropical Andes using NICFI Satellite Data Program imagery and semantic segmentation techniques. *Data* **2023**, *8*, 185. [CrossRef]
43. Brown, C.F.; Brumby, S.P.; Guzder-Williams, B.; Birch, T.; Hyde, S.B.; Mazzariello, J.; Czerwinski, W.; Pasquarella, V.J.; Haertel, R.; Ilyushchenko, S.; et al. Dynamic World, Near Real-Time Global 10 m Land Use Land Cover Mapping. *Sci. Data* **2022**, *9*, 251. [CrossRef]
44. Mao, D.; Wang, Z.; Du, B.; Li, L.; Tian, Y.; Jia, M.; Zeng, Y.; Song, K.; Jiang, M.; Wang, Y. National Wetland Mapping in China: A New Product Resulting from Object-Based and Hierarchical Classification of Landsat 8 OLI Images. *ISPRS J. Photogramm. Remote Sens.* **2020**, *164*, 11–25. [CrossRef]
45. Foody, G.M. Status of Land Cover Classification Accuracy Assessment. *Remote Sens. Environ.* **2002**, *80*, 185–201. [CrossRef]
46. Olofsson, P.; Foody, G.M.; Herold, M.; Stehman, S.V.; Woodcock, C.E.; Wulder, M.A. Good Practices for Estimating Area and Assessing Accuracy of Land Change. *Remote Sens. Environ.* **2014**, *148*, 42–57. [CrossRef]
47. Stehman, S.V.; Foody, G.M. Key Issues in Rigorous Accuracy Assessment of Land Cover Products. *Remote Sens. Environ.* **2019**, *231*, 111199. [CrossRef]
48. Xie, E.; Wang, W.; Yu, Z.; Anandkumar, A.; Alvarez, J.M.; Luo, P. SegFormer: Simple and Efficient Design for Semantic Segmentation with Transformers. In Proceedings of the 35th Conference on Neural Information Processing Systems (NeurIPS 2021), Online Conference, 6–14 December 2021.

49. Wang, L.; Li, R.; Zhang, C.; Fang, S.; Duan, C.; Meng, X.; Atkinson, P.M. UNetFormer: A UNet-like Transformer for Efficient Semantic Segmentation of Remote Sensing Urban Scene Imagery. *ISPRS J. Photogramm. Remote Sens.* **2022**, *190*, 196–214. [CrossRef]
50. Wang, Q.; Chen, W.; Huang, Z.; Tang, H.; Yang, L. MultiSenseSeg: A cost-effective unified multimodal semantic segmentation model for remote sensing. *IEEE Trans. Geosci. Remote Sens.* **2024**, *62*, 4703724. [CrossRef]
51. Yang, S.; Wang, H.; Xing, Z.; Chen, S.; Zhu, L. Segdino: An efficient design for medical and natural image segmentation with dino-v3. *arXiv* **2025**, arXiv:2509.00833. [CrossRef]
52. Gao, Y.; Li, H.; Yuan, F.; Wang, X.; Gao, X. Dino u-net: Exploiting high-fidelity dense features from foundation models for medical image segmentation. *arXiv* **2025**, arXiv:2508.20909. [CrossRef]
53. Gao, F.; Fu, M.; Cao, J.; Dong, J.; Du, Q. Adaptive Frequency Enhancement Network for Remote Sensing Image Semantic Segmentation. *IEEE Trans. Geosci. Remote Sens.* **2025**, *63*, 5619415. [CrossRef]

Disclaimer/Publisher's Note: The statements, opinions and data contained in all publications are solely those of the individual author(s) and contributor(s) and not of MDPI and/or the editor(s). MDPI and/or the editor(s) disclaim responsibility for any injury to people or property resulting from any ideas, methods, instructions or products referred to in the content.