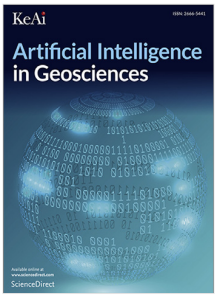
Journal Pre-proof



Enhancing land cover semantic segmentation with convolutional block attention modules and deep supervision

Sara Mobsite, Renaud Hostache, Laure Berti-Équille, Emmanuel Roux, Thibault Catry, Joris Guérin

PII: S2666-5441(26)00038-9
DOI: https://doi.org/10.1016/j.aiig.2026.100222
Reference: AIIG 100222

To appear in: *Artificial Intelligence in Geosciences*

Received date : 23 January 2026
Revised date : 25 March 2026
Accepted date : 2 May 2026

Please cite this article as: S. Mobsite, R. Hostache, L. Berti-Équille et al., Enhancing land cover semantic segmentation with convolutional block attention modules and deep supervision. *Artificial Intelligence in Geosciences* (2026), doi: https://doi.org/10.1016/j.aiig.2026.100222.

Enhancing Land Cover Semantic Segmentation with Convolutional Block Attention Modules and Deep Supervision

Sara Mobsite *[a], Renaud Hostache[a], Laure Berti-Équille[a], Emmanuel Roux[a], Thibault Catry[a] and Joris Guérin[a]

[a]*ESPACE-DEV, French National Research Institute for Sustainable Development (IRD), France*

ARTICLE INFO

Keywords:
Land Cover
Radar
Optical
Semantic Segmentation
Deep Supervision

ABSTRACT

High-resolution land cover semantic segmentation is challenged by strong class imbalance, spatial fragmentation of minority classes, and the presence of fine-scale textures and sensor noise that can dominate early feature learning. In addition, producing high-resolution labeled maps is time-consuming and requires expert annotation, while low-resolution maps are easier to obtain but lack spatial precision. To address these challenges, we propose MUSCLE-Net, a Multi-Scale Land Cover Network that explicitly enforces semantic consistency across spatial resolutions through deep supervision. By introducing an auxiliary low-resolution segmentation task during early decoding, the network is constrained to learn semantically meaningful regional representations before recovering fine spatial details, promoting a coarse-to-fine decoding process that mitigates overfitting to high-frequency noise. Convolutional Block Attention Modules are incorporated in the decoder to further refine spatial and channel-wise feature selection. For the DynamicEarthNet dataset, MUSCLE-Net achieves an overall accuracy of 66.48%, outperforming UNet by 1.05%, DeepLabV3 by 6.21%, and PSPNet by 7.70%. For the DFC2020 dataset, MUSCLE-Net reaches an overall accuracy of 70.10%, improving upon UNet by 2.86%, PSPNet by 4.98%, and DeepLabV3 by 6.10%, and consistently shows lower variability between runs, reflecting enhanced robustness in minority land-cover classes.

1. Introduction

High-resolution land cover (LC) mapping is commonly performed using deep learningbased semantic segmentation trained in a supervised manner on annotated satellite imagery [52]. LC semantic segmentation involves assigning a distinct label to each pixel from the input data. The applications of LC semantic segmentation vary widely, ranging from deforestation [9] and land degradation monitoring [42, 40] to sea ice recognition [45]. Depending on the task, data from different satellites can be used: some recent work uses radar data (Sentinel-1) for LC segmentation [23, 32, 54], while others prefer to utilize optical data (Sentinel-2) [5, 3]. Radar is often used due to its ability to operate under cloud cover. In contrast, optical imagery provides a richer spectral representation across the visible, near-infrared (NIR), and shortwave infrared (SWIR) wavelengths, enabling the extraction of detailed spectral and textural information. However, optical data are sensitive to cloud interference and atmospheric effects. As a result, many studies integrate both radar and optical images to achieve more reliable pixel-level LC identification [43].

One of the fundamental challenges in the classification of LC at the pixel-level is the imbalance between classes that represent large spatial extents and those that occupy much smaller areas [38]. For instance, forests and agricultural fields typically cover large regions, while features such as buildings and roads have relatively limited spatial extent. Another significant issue arises from the high semantic similarity between certain LC classes, which often exhibit overlapping spectral and spatial characteristics. Distinguishing between barren land and shrubland, for example, can be particularly difficult, as sparse vegetation on rocky or sandy terrain may appear nearly identical to bare soil in remote sensing imagery. These challenges are further intensified by the limited availability of high-resolution, expert-annotated LC datasets. The combination of underrepresented classes and limited high-quality annotations often leads deep learning models to overfit dominant classes and irrelevant fine-scale patterns, thereby limiting their generalization capability.

*Corresponding author: Sara Mobsite.
E-mail address: firstname.lastname@ird.fr

ORCID:0000-0002-5784-5018

In encoderdecoder architectures commonly used for land-cover semantic segmentation, the decoder progressively restores spatial resolution from compact, high-level feature representations. During the early stages of decoding, spatial resolution is increased while global semantic context remains incomplete, requiring the reconstruction of fine-grained details from semantically limited features. For high-resolution remote sensing imagery, early decoder features may emphasize local textures and noise rather than meaningful land-cover semantics, especially for fragmented and minority classes. As a result, constraining early decoder representations to encode semantically consistent information at coarser spatial scales is critical for achieving robust and generalizable land cover segmentation.

To address the previously discussed limitations, we introduce MUSCLE-Net (Multi-Scale Land Cover Network), an encoderdecoder architecture for land-cover semantic segmentation. The proposed model explicitly leverages multi-resolution semantic deep supervision to enhance the decoding process and improve robustness under class imbalance and spatial fragmentation. Deep supervision is implemented by introducing auxiliary segmentation outputs at early decoder stages, where global semantic information is enforced before fine spatial details are recovered. This strategy constrains the early decoding stages to learn semantically meaningful and scale-consistent representations, promoting a coarse-to-fine refinement of land-cover regions. The features learned through auxiliary supervision are subsequently reintegrated into the main decoding pathway, reinforcing the primary segmentation objective. In addition, skip connections between the encoder and decoder are employed to preserve both fine-scale spatial details and large-scale contextual information.

The decoder of MUSCLE-Net is composed of sequential convolutional layers integrated with Convolutional Block Attention Modules (CBAM) [47]. CBAM enhances the feature representation by adaptively focusing on the most informative channel-wise and spatial components extracted by the encoder. This attention mechanism enables the model to capture complex landscape structures more effectively.

The proposed MUSCLE-Net architecture is encoder-agnostic and can be combined with different convolutional or transformer-based backbones. To assess the impact of encoder choice and select a practical trade-off between performance and computational cost, we evaluated several encoders initialized with publicly available weights pre-trained on the reBEN dataset [18, 11]. Based on this analysis, ResNet-50 [19] is selected for the remainder of the study.

This paper presents the following primary contributions:

- To effectively leverage both high- and low-resolution labels, a deep supervision mechanism is integrated into the early decoding stages to promote semantic consistency, guide feature learning, and improve class discrimination in the initial layers through supervision from low-resolution maps.
- A decoder enhanced with convolutional block attention modules is introduced to selectively emphasize relevant spatial and channel-wise features that are critical for accurately distinguishing between different land cover types.
- Both transformer-based and convolution-based encoders are systematically evaluated within the proposed MUSCLE-Net framework under configurations with and without deep supervision, to determine the encoder architecture that offers the best trade-off between segmentation performance and computational efficiency.
- Features learned through the deep supervision task are re-injected into the main segmentation pathway, and skip connections from deep encoder layers are used to strengthen the representation of both fine-grained and large-scale LC features.

The remainder of this paper is organized as follows. The Related Work section reviews prior studies on LC semantic segmentation, with particular attention to approaches based on weak supervision and multi-task learning. The Methodology section describes the proposed MUSCLE-Net architecture, including the attention-enhanced decoder based on CBAM and the integration of multi-resolution deep supervision. The Results and Discussion section presents ablation and optimization studies that analyze the impact of encoder architectures, CBAM integration, and auxiliary supervision strategies. This section also provides a comparative evaluation of the proposed model against several semantic segmentation baselines using the same encoder configuration to ensure a fair comparison. Furthermore, additional comparisons are conducted with other architectures employing different encoder backbones and decoding strategies.

2. Related Work

LC segmentation using deep learning can be formulated either as a single-task problem, where the model directly predicts LC maps corresponding to predefined classes, or as a multi-task learning framework, where additional tasks are incorporated to refine the learning process and improve segmentation performance. In [33], several single-task-based networks, including FCN [29], UNet [35], SegNet [4], and DeepLabv3 [6], were trained and evaluated using Optical (OPT) data to segment green urban areas. To assess model performance, the study conducted a comparative analysis using various input configurations, including Red-Green-Blue (RGB) bands and the Normalized Difference Vegetation Index (NDVI). In [12], a transformer-based autoencoder was employed to capture temporal and spectral relationships in OPT data using positional encoding across spatial, temporal, and spectral dimensions to improve LC analyses. In [23], a UNet-based model was also proposed for flood mapping using synthetic aperture radar (SAR) data. This approach incorporated task-specific features through a wave-vision module and a vision multi-layer perceptron mechanism. Residual connections were used to enhance spatial information transfer from the encoder to the decoder. Similarly, SAR was used in [54] for flood mapping in open urban environments. Additionally, a new dataset named UrbanSARFloods was introduced, and multiple semantic segmentation models such as UNet++ [55], MANet[14], and PAN [26] were trained and evaluated on it. In [17], UNet, SF-Net [27], and LSTM [20] architectures were modified by incorporating attention blocks as well as residual and dilated connections. The goal of the study was to use SAR coherent features to improve land-use and deforestation mapping across different terrains in the southern Amazon, with the enhanced UNet architecture achieving the highest accuracy.

Diversifying input data sources is one of the strategies explored to improve LC semantic segmentation. For example, the presence of clouds and atmospheric disturbances often limits the usability of OPT imagery. To address this limitation, [22] proposed the Optical and SAR Images Combined Mangrove Index (OSCMI), which integrates optical data with SAR observations. In this approach, the verticalvertical (VV) polarization from SAR data is used to complement optical information and enhance mangrove detection. Similarly, [48] introduced a multi-head network that incorporates long skip connections to preserve spatial details and a symmetric attention mechanism to progressively fuse SAR and optical features. These studies demonstrate the potential of combining complementary input modalities to improve segmentation performance under challenging observation conditions.

However, another strategy was adopted [1], where a multi-task learning fully convolutional network was proposed for simultaneous segmentation and boundary detection to support early-season agricultural mapping of field boundaries. The model relied exclusively on multispectral OPT imagery as input. Since segmentation alone was insufficient for this specific task, the authors enhanced the learning process by introducing boundary detection, along with auxiliary tasks such as distance-to-boundary estimation and classification confidence. A plastic-mulched land mapping was carried out in [30] using optical satellite data and the Index-Feature-Spatial-Attention Fused Deep Learning Model (IFSA-DLM). The IFSA-DLM model was designed with a dual-branch architecture: one branch extracted multi-scale index features, while the other processed raw spectral features directly from OPT imagery. The outputs of both branches were then integrated through a spatial attention mechanism to enhance feature discrimination. In [2], the authors improved the UNet architecture by incorporating channel-wise attention to the encoder features before forwarding them through the skip connections to the decoder. The proposed model leveraged OPT imagery to detect and map forest tree dieback. Instead of relying solely on the final decoder output, the authors introduced a teacher-student learning strategy, in which an auxiliary output layer from intermediate decoder layers (student) learned from the primary soft label (teacher). In [15], self-supervised learning was employed to leverage unlabeled OPT and SAR data. The proposed model uses contrastive learning between paired SAR and optical images, combined with an additional reconstruction task to enhance cross-modality feature representation. The resulting pretrained model can then be transferred to downstream applications and lightly fine-tuned for LC semantic segmentation tasks.

To improve LC semantic segmentation, weak supervision was adopted in [7]. The proposed method used a two-stage framework that first transferred reliable samples from low-resolution land cover maps and then trained a segmentation network using partially labeled superpixels with dynamic pseudo-label propagation. In [13], a weakly supervised framework was proposed to improve high-resolution land cover classification using low-resolution land cover products as guidance. The method first extracted coarse semantic information from low-resolution labels through a superpixel-based training strategy to reduce inconsistencies with high-resolution imagery. It then refined predictions by dynamically selecting high-confidence point labels, allowing the model to progressively learn more detailed features, while consistency regularization integrated coarse and refined knowledge to improve segmentation performance. Similarly, in [8], the authors employed superpixels as training units to reduce errors caused by resolution differences and small misclassified regions. They further introduced a dual-expert learning strategy that evaluated prediction credibility and adaptively corrected large-scale noisy labels during training, enabling the model to better handle label noise and generate more reliable high-resolution land cover maps.

Despite the progress achieved by these approaches (supervised and unsupervised), the imbalanced spatial coverage of LC classes remains a major challenge in LC segmentation tasks. Most datasets offer limited samples for certain classes, causing significant class imbalance and segmentation errors. Moreover, the high semantic similarity between some LC classes can cause models to overfit to dominant classes that span large spatial areas [38].

3. Method

This section describes the design and core components of the proposed MUSCLE-Net architecture for LC semantic segmentation. It presents the encoderdecoder framework, the attention-enhanced decoder based on CBAM modules, and the integration of deep supervision to improve multi-scale feature learning. This design enables effective exploitation of both high- and low-resolution information throughout training and provides better control over the learning process in the early stages of decoding.

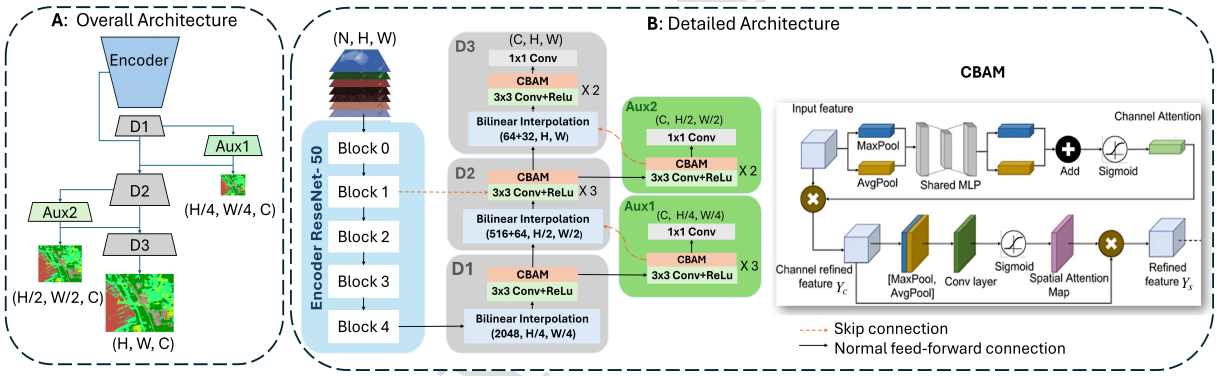


Figure 1: Overview of the proposed MUSCLE-Net architecture. The model processes multi-channel radar and/or optical input of size $H \times W \times N$ and outputs a C-channel segmentation mask. D1–D3 represent the decoder blocks, while *Aux* denotes the auxiliary output used for deep supervision.

3.1. MUSCLE-Net Overview

The proposed MUSCLE-Net follows an autoencoder architecture. The encoder can be any pretrained backbone, including convolutional networks, vision transformerbased models, or hybrid architectures. In practice, we conducted a comparative experimental study and selected a ResNet-50 backbone pretrained on LC classification using the reBEN dataset [18, 11]. As illustrated in Figure 1, the main contribution of MUSCLE-Net lies in its decoder design. The decoder is specifically tailored to progressively fuse multi-scale features and enable accurate spatial reconstruction, which distinguishes MUSCLE-Net from standard autoencoder architectures.

As depicted in Figure 1, the decoder architecture is composed of three sequential processing blocks with increasing depth. Each block integrates a convolutional operation (Conv), whether traditional or depthwise separable [10], followed by the CBAM [47]. The decoder incorporates CBAM to model both channel-wise and spatial dependencies in the extracted feature maps. To facilitate more effective gradient flow and feature refinement, deep supervision is

introduced in the decoding stage. This setup enables the extraction of highly discriminative and task-specific features, which can lead to improved segmentation performance. The placement and weighting of the deep supervision are empirically determined through comparative analysis.

3.2. Convolutional Block Attention Module

CBAM is integrated after the convolutional layers in the MUSCLE-Net decoder to refine feature representations. This mechanism consists of two sequential components: channel attention and spatial attention.

In the channel attention module, the input feature map $X \in \mathbb{R}^{D\times H\times W}$ is first subjected to global average pooling (AvgPool) and global max pooling (MaxPool) along the spatial dimensions, producing two descriptors of shape $\mathbb{R}^{D\times 1\times 1}$. Average pooling captures the general activation distribution, while max pooling highlights the most prominent features. These descriptors are flattened and passed through a shared multilayer perceptron (MLP) composed of two fully connected layers with a Rectified Linear Unit (ReLU) activation in between. The first layer reduces the dimensionality from D to D/r, and the second restores it to D, enabling efficient and expressive feature transformation.

The channel attention mechanism is described in Equation 1, where Θ_1 and Θ_2 denote the learnable weights of the MLP. The resulting attention map $M_c(X) \in \mathbb{R}^{D\times 1\times 1}$ is obtained by summing the MLP outputs of the two pooling paths and applying a sigmoid activation σ. The refined output Y_c is produced by reweighting the input feature map X through element-wise multiplication denoted as $\odot$ with $M_c(X)$.

$$\begin{aligned} \text{Avg} &= \text{AvgPool}(X), \\ \text{Max} &= \text{MaxPool}(X), \\ F_{\text{avg}} &= \Theta_2(\text{ReLU}(\Theta_1(\text{Avg}))), \\ F_{\text{max}} &= \Theta_2(\text{ReLU}(\Theta_1(\text{Max}))), \\ M_c(X) &= \sigma(F_{\text{avg}} + F_{\text{max}}), \\ Y_c &= M_c(X) \odot X. \end{aligned} \tag{1}$$

Subsequently, the spatial attention module operates on the output of the channel attention. It applies average pooling and max pooling along the channel axis, generating two spatial maps of size $\mathbb{R}^{1\times H\times W}$. These maps are concatenated to form a two-channel spatial descriptor $\mathbb{R}^{2\times H\times W}$, which is passed through a convolutional layer with a 7×7 kernel. A sigmoid activation then produces the spatial attention map, which is multiplied element-wise with the channel-refined feature map to emphasize informative spatial locations, resulting in the output Y_s. This process is expressed in Equation 2.

$$Y_s = \sigma\left(\text{Conv}_{7\times 7}\left(\text{Concat}\left(\text{AvgPool}(Y_c), \text{MaxPool}(Y_c)\right)\right)\right) \odot Y_c \tag{2}$$

The sequential application of channel and spatial attention in CBAM enables the network to emphasize informative features along both channel and spatial dimensions. In the original CBAM design, a residual connection with element-wise addition between the input feature map X and the output attention-refined feature maps from CBAM Y_s is employed to recover potentially lost information (Figure 2a). However, reintroducing the original features through this skip connection may attenuate the effect of the attention mechanism, as it allows features to bypass the learned attention weights.

(a) CBAM with residual connection

(b) Proposed CBAM with direct connection

Figure 2: CBAM integration strategies: (a) the original design employing a residual connection between the input feature map and the CBAM output, and (b) the proposed variant that directly feeds the CBAM output into the subsequent convolutional operation without a residual connection.

In this work, we evaluate the effectiveness of the residual connection with CBAM through ablation studies conducted with and without it (Figure 2). In the proposed network, CBAM is not applied to the encoder, as the backbone architectures are kept intact to allow weight initialization from the reBEN dataset.

3.3. Enhancing Land Cover Segmentation with Deep Supervision

To improve and guide learning from the early to the late stages of the decoder, auxiliary supervision with skip connections is applied during the initial steps of decoding. This task guides the learning process by promoting the extraction of features that are beneficial for the primary segmentation objective and re-injecting them into the decoder. It also facilitates the learning of intermediate representations that are semantically aligned with the resolution of each decoding stage. Unlike traditional auxiliary tasks, which often target related but distinct objectives, the auxiliary outputs in our approach are derived from the same primary segmentation task. They are therefore down-sampled using nearest neighbor interpolation to match the spatial dimensions of the feature maps within the corresponding decoder blocks. This alignment enables the extraction of relevant spatial semantics in the early decoding layers and improves the flow of meaningful information through the decoder hierarchy.

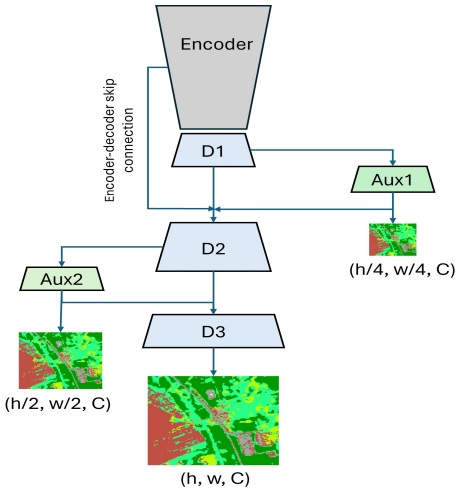


Figure 3: The proposed MUSCLE-Net incorporates deep supervision through two auxiliary outputs, Aux 1 and Aux 2. Their effectiveness is assessed through ablation and comparative analyses, where different weight configurations are assigned to Aux 1 and Aux 2 in the global loss function.

As shown in Figure 3, we evaluate the impact of auxiliary supervision strategies, namely an early supervision (Aux 1), a late supervision near the final output (Aux 2), and their combination with the final prediction (Final). During the decoding phase, convolutional operations combined with CBAM are employed to transform low-resolution features into high-resolution segmentation masks, effectively capturing both channel-wise and spatial dependencies. Upsampling is performed using bilinear interpolation instead of transposed convolution, reducing the number of trainable parameters. The spatial dimension is progressively restored in three stages: from $H/4 \times W/4$ to $H/2 \times W/2$, and finally to the original $H \times W$ size. Model training is guided by a weighted global loss function (Equation 3), where $\mathcal{L}_{\text{total}}$ represents the overall loss minimized during optimization. The term $\mathcal{L}_{\text{Final}}$ corresponds to the loss computed from the final high-resolution segmentation output, while $\mathcal{L}_{\text{Aux1}}$ and $\mathcal{L}_{\text{Aux2}}$ denote the auxiliary losses associated with the early and late supervision branches, respectively. The weighting coefficients λ_{Final}, λ_{Aux1}, and λ_{Aux2} regulate the relative contribution of each loss component.

$$\begin{aligned} \mathcal{L}_{\text{total}} &= \lambda_{\text{Final}}\mathcal{L}_{\text{Final}} + \lambda_{\text{Aux1}}\mathcal{L}_{\text{Aux1}} + \lambda_{\text{Aux2}}\mathcal{L}_{\text{Aux2}} \\ \text{subject to} \quad & \lambda_{\text{Final}} \geq \lambda_{\text{Aux1}} + \lambda_{\text{aux2}}, \\ & \lambda_{\text{Final}} + \lambda_{\text{Aux1}} + \lambda_{\text{Aux2}} = 1 \end{aligned} \tag{3}$$

In real-world scenarios, auxiliary outputs can correspond to products available at multiple resolutions for the same geographic region. For example, several land use land cover mapping services provide both high-resolution products, which capture fine spatial details but are computationally expensive to acquire and process, and low-resolution products, which are more widely available but contain coarser information. Instead of discarding the low-resolution data, it can be used as complementary supervision during training. By aligning auxiliary outputs with these different resolution levels, the model is encouraged to learn multi-scale feature representations, leveraging both the broad contextual information from low-resolution data and the detailed spatial patterns from high-resolution data. This multi-resolution supervision can ultimately improve model robustness and predictive performance.

4. Datasets and Experimental design

This section describes the datasets used in our study and the overall experimental protocol adopted to evaluate MUSCLE-Net. We first present the characteristics and preprocessing of the employed datasets. We then detail the experimental design, including the comparison strategies and ablation studies, followed by the evaluation metrics and implementation details used to ensure fair and reproducible results.

4.1. Datasets

4.1.1. DFC2020

This dataset, developed for the IEEE GRSS Data Fusion Contest (DFC) 2020 [36, 34] as an extension of the SEN12MS dataset [37], includes 8 LC classes along with Sentinel-1 and Sentinel-2 imagery and corresponding annotations at a spatial resolution of 10 meters per pixel. Following previous works [15, 44, 12], we used the original test set (5,128 samples) for training (85% train, 15% validation) and the original validation set (986 samples) for evaluation.

4.1.2. DynamicEarthNet

The DynamicEarthNet dataset [41] consists of Sentinel-2 imagery from 75 regions worldwide, with pixel-level annotations for 7 LC classes. Because our model employs a pre-trained encoder that expects inputs at a 10 m/pixel resolution, we downsampled the Sentinel-2 bands using bilinear interpolation and the corresponding labels using nearest-neighbor interpolation. Following the original study, the ice and snow class was excluded, as it does not appear in the validation or test sets. A key limitation for our task is that the dataset focuses on LC segmentation and change detection, with data collected monthly from 2018 to 2019. Many classes, such as forests and impervious surfaces, show little change, which makes temporal samples at each location often very similar. To introduce variability and help the model learn meaningful representations, we applied random unique crops and rotations to each monthly sample within each region. All optical data and labels were resized to 256×256 pixels with a 10 m spatial resolution. We used all samples from 10 regions for testing and 11 regions for validation, while the remaining regions were used for training. This strategy ensures that data from the same regions, even if collected in close or distant periods, do not appear in both training and validation sets simultaneously.

4.2. Experimental Design

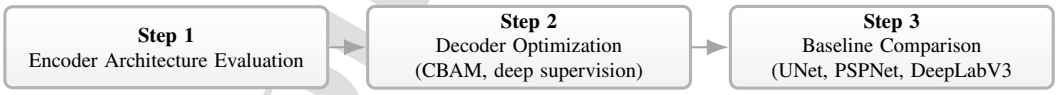


Figure 4: Experimental workflow illustrating the three main analysis stages: encoder and input evaluation, decoder ablation, CBAM and deep supervision analysis, and baseline comparison.

Figure 4 summarizes the experimental workflow used in our study. In Step 1, we conducted a comparative evaluation of encoder architectures without deep supervision. This step was designed to evaluate the performance of several encoders pre-trained on the reBEN LC classification dataset and to establish a fair comparison. The models used in this analysis were ResNet-50, MobileViT-S [31], ConvNeXt V2 Base [46], and Inception NeXt Base [51]. ResNet-50 incorporates skip connections throughout the network architecture to mitigate the vanishing gradient problem and preserve information across deep encoding blocks. Inception NeXt Base was designed to efficiently extract features at varying receptive field levels. Inspired by ConvNeXt [28] and InceptionNet [39], this architecture

applies depthwise separable convolutions instead of standard convolutions, along with residual skip connections between blocks. Moreover, similar to the MetaFormer block [49, 50], Inception NeXt employs block-to-block residual connections; however, unlike MetaFormer, where residual connections are added in the middle or after the MLP operation, Inception NeXt performs the residual connection strictly at the block level after the MLP. ConvNeXt V2 Base extends ConvNeXt by introducing Global Response Normalization (GRN), which improves representation learning and training stability. Finally, MobileViT-S combines convolutional operations with lightweight vision transformers, enabling efficient spatial context modeling while maintaining computational efficiency.

In Step 2, we performed decoder component analysis to isolate the contribution of each element, including CBAM, deep supervision, and the residual addition operation between the convolutional and CBAM outputs. These experiments were based on ResNet-50, which showed a good trade-off between performance and model complexity, particularly under low batch-size conditions, compared to the other encoders (Table C1 in the Appendix section). Moreover, we further analyzed the effects of CBAM and deep supervision using the optimal parameters obtained from these experiments across different encoders, to verify that the observed improvements generalize regardless of the encoder.

Finally, in Step 3, we benchmarked the optimized MUSCLE-Net configuration against established semantic segmentation baselines, including UNet, PSPNet [53], and DeepLabV3, using the same encoder architecture and identical initial weights pre-trained on the reBEN dataset to ensure a fair comparison. UNet was selected as a widely adopted encoder–decoder architecture that leverages skip connections to preserve spatial details. PSPNet was included to represent pyramid poolingbased approaches that capture multi-scale contextual information. DeepLabV3 was chosen as a representative of atrous convolutionbased models, which enhance receptive fields while maintaining spatial resolution. This selection enables a comprehensive comparison across different decoder design philosophies.

4.3. Evaluation metrics

During the analyses, the Mean Intersection over Union (MIoU) (Equation 5), computed as the average of class-wise IoU values (Equation 4), together with the overall accuracy (OA) (Equation 6), were used as evaluation metrics, where K denotes the number of classes. For each class i, TP_i, FP_i, FN_i, and TN_i represent the number of true positives, false positives, false negatives, and true negatives, respectively. True positives correspond to pixels correctly classified as class i, false positives to pixels incorrectly assigned to class i, false negatives to pixels of class i incorrectly classified as another class, and true negatives to pixels correctly classified as not belonging to class i.

$$IoU_i = \frac{TP_i}{TP_i + FP_i + FN_i}, \quad i = 1, 2, \ldots, K \tag{4}$$

$$MIoU = \frac{1}{K} \sum_{i=1}^{K} IoU_i \tag{5}$$

$$\mathrm{OA} = \frac{\sum_{i=1}^{K} TP_i}{\sum_{i=1}^{K} (TP_i + FN_i)} \tag{6}$$

4.4. Implementation Details

The input size of the Sentinel images was 256×256 pixels for both datasets. Each input was composed of 10 OPT bands and 2 SAR channels; specifically, 10 Sentinel-2 bands and 2 Sentinel-1 channels for DFC2020, and only optical bands for DynamicEarthNet (Table 1). When combining SAR and OPT inputs, the channel order was as follows: 'VV', 'VH', 'B02', 'B03', 'B04', 'B05', 'B06', 'B07', 'B08', 'B8A', 'B11', and 'B12'.

To maintain consistent initialization across all experiments, the encoders (ResNet-50, MobileViT-S, ConvNeXt V2 Base, and Inception NeXt Base) were initialized with pre-trained weights from the reBEN dataset during the comparative analyses. For comparative analyses, the encoder was fixed to the ResNet-50 architecture and applied uniformly across all models (UNet, PSPNet, DeepLabV3, and MUSCLE-Net) to ensure fair evaluation. For DFC2020, the encoder weights were trained, whereas for DynamicEarthNet, the encoder weights remained frozen across all models in order to systematically assess both training strategies.

For all experiments, cross-entropy loss [16] was used as the training objective. Models were trained for a maximum of 200 epochs, with early stopping applied using a patience of 20 epochs for DFC2020 and 10 epochs

Table 1
Satellite sources and channels used from each dataset

Satellite	Bands used	DFC2020	DynamicEarthNet
Sentinel-1	VV, VH	✓	–
Sentinel-2	B02, B03, B04, B05, B06, B07, B08, B8A, B11, B12	✓	✓

for DynamicEarthNet, due to its smaller number of training samples. The Adam optimizer [25] was employed with an initial learning rate of 0.001, and batch sizes ranged from 16 to 64.

5. Results and Discussion

This section presents a comprehensive evaluation of the proposed MUSCLE-Net architecture and discusses the impact of its design choices on LC segmentation performance. We first analyze the influence of different encoder architectures to identify the most suitable backbone. Subsequently, we investigate the effects of key decoder components, including CBAM and deep supervision, through extensive ablation and comparative studies. Finally, the optimized MUSCLE-Net configuration is compared against state-of-the-art segmentation models on the DFC2020 and DynamicEarthNet datasets to demonstrate its effectiveness and generalization capability.

5.1. Encoder Architecture Evaluation

To identify the most suitable encoder architecture for the proposed MUSCLE-Net model, we conducted experiments with several networks, evaluating their performance using combined SAR and OPT inputs at the input level. The choice of using combined SAR and OPT data, rather than either modality alone, was motivated by a comparative analysis presented in Appendix A (Table A1), which demonstrated superior performance for the OPT+SAR configuration on the DFC2020 dataset. Consequently, all experiments reported in this section were conducted using the OPT+SAR input setting to enable a fair comparison across encoder architectures.

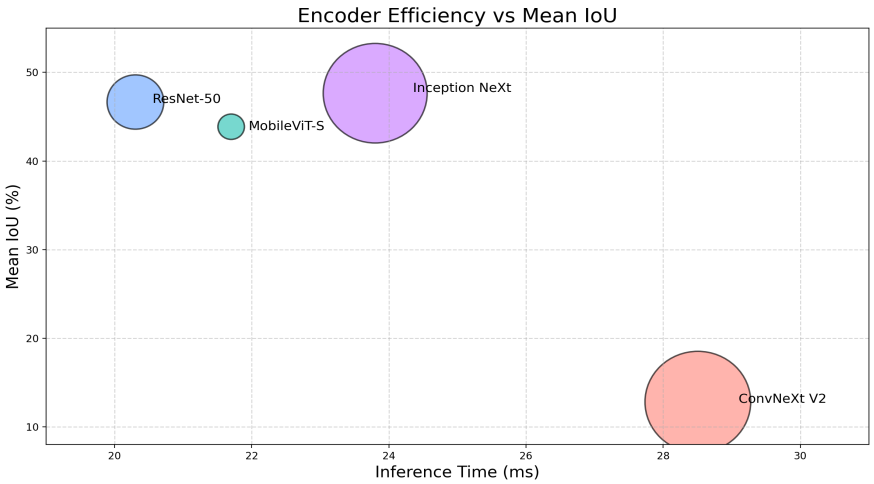


Figure 5: Inference time, number of parameters in relation to MIoU for different encoders in the MUSCLE-Net framework. Each bubbles size represents the number of model parameters.

Table 2 reports the per-class IoU and MIoU achieved by the proposed MUSCLE-Net using different encoder architectures initialized with reBEN weights. All experiments were conducted with a batch size of 64, and the comparative analysis was performed without deep supervision. Although Inception NeXt achieves the highest MIoU, ResNet-50 attains comparable segmentation performance, with only a marginal difference in MIoU. As shown in Figure 5, ResNet-50 exhibits a more favorable balance between segmentation accuracy and computational efficiency,

Table 2
Per-class IoU (%) and MIoU (%) of the proposed model on the DFC2020 test set using different encoder architectures with initial weights from the reBEN dataset. Best MIoU is shown in blue and second-best MIoU in green.

Encoder Model	Forest	Shrubland	Savanna	Grassland	Croplands	Urban	Barren	Water	MIoU
Inception NeXt	67.34	42.08	18.73	0.96	30.92	64.91	59.22	97.24	**47.68**
MobileViT-S	35.59	38.27	19.45	0.01	**51.65**	56.65	55.92	93.60	43.89
ConvNeXt V2	13.29	0.00	0.00	0.00	0.50	0.00	0.02	88.88	12.84
ResNet-50	67.38	40.27	13.67	0.76	33.90	63.33	58.14	95.77	**46.65**
								MIoU gap (best − 2nd):	**1.03**

requiring fewer parameters and lower computational cost while achieving shorter inference time compared to the best-performing model. Considering this trade-off between performance and resource consumption, ResNet-50 was selected as the encoder for the proposed MUSCLE-Net and was used for the comparative analysis and the remainder of the study.

5.2. Optimization of MUSCLE-Net Parameters

To analyze the impact of deep supervision placement within the decoder, several weight configurations for Aux1 ($\lambda_{\mathbf{Aux1}}$) and Aux2 ($\lambda_{\mathbf{Aux2}}$) were evaluated. In this comparative analysis, the sum of weights assigned to the auxiliary supervision branches was fixed at 0.10 to prevent them from dominating the optimization process and maintaining the primary focus on the main segmentation task ($\lambda_{\mathbf{Final}}$).

These auxiliary tasks were intended to guide feature learning in the early decoding layers rather than serving as the primary objective of the study. As shown in Table 3, the best performance was achieved when a low weight of 0.1 was assigned exclusively to Aux1, with the remaining 0.9 allocated to the main task, resulting in a MIoU of 53.82 and an OA of 70.10 on the DFC2020 dataset. Notably, all evaluated configurations incorporating Aux1 and/or Aux2 outperformed the baseline model without auxiliary deep supervision, highlighting the effectiveness of this strategy irrespective of the specific supervision placement. The reported results (Table 3) were obtained by training each model five times with a batch size of 32, and reporting the mean and standard deviation across runs.

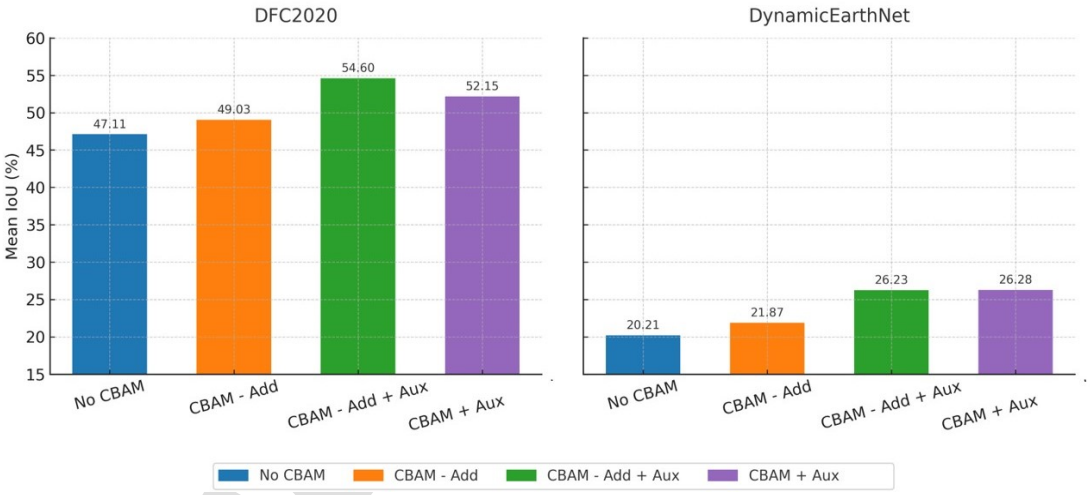


Figure 6: Ablation study on the DFC2020 and DynamicEarthNet test sets, assessing the effects of CBAM, deep supervision (Aux1 = 0.10), skip connections, and the removal of addition operations between convolutional and CBAM blocks. Experiments used a batch size of 32 for DFC2020 and 16 for DynamicEarthNet, with ResNet-50 as the backbone across all analyses.

The results reported in Table 4 further confirm the positive impact of deep supervision (Aux1 = 0.10), as incorporating deep supervision consistently improved the performance of all variants of the proposed model across different encoder architectures. Further analyses were conducted on the DFC2020 and DynamicEarthNet datasets to evaluate the impact of integrating CBAM, skip connections, and the addition operation between convolutional layer

Table 3
Ablation study of deep supervision weights on DFC2020, reported as mean ś Std over 5 runs with a ResNet-50 backbone.

λ_{Final}	λ_{Aux1}	λ_{Aux2}	Mean mIoU ± StD	Mean OA ± StD
0.9	0.05	0.05	50.89 ± 3.83	68.57 ± 2.77
0.9	0	0.10	52.47 ± 2.23	69.17 ± 1.8
0.9	0.10	0	53.82 ± 0.71	70.10 ± 0.8
1	0	0	46.28 ± 2.28	65.20 ± 2.32

Table 4
Per-class IoU scores of the proposed model under different encoder configurations, with and without deep supervision, on the DFC2020 dataset. The final column shows the gain in MIoU (Δ mIoU) compared to the baseline without deep supervision. All experiments used a batch size of 64 and were initialized with reBEN-pretrained weights.

Encoder Model	Decoding Technique	Forest	Shrubland	Savanna	Grassland	Croplands	Urban	Barren	Water	Mean IoU (%)	Δ Mean IoU (%)
Inception NeXt	CBAM - Add	67.34	42.08	18.73	0.96	30.92	64.91	59.22	97.24	47.68	
	CBAM - Add + Aux	66.09	44.69	16.88	2.41	32.37	63.97	56.94	98.31	47.71	+0.03
MobileViT-S	CBAM - Add	35.59	38.27	19.45	0.01	51.65	56.65	55.92	93.60	43.89	
	CBAM - Add + Aux	61.92	44.27	23.78	3.16	47.95	75.07	62.88	98.79	52.23	+8.34
ConvNeXt V2	CBAM - Add	13.29	0.00	0.00	0.00	0.50	0.00	0.02	88.88	12.84	
	CBAM - Add + Aux	13.41	46.67	9.11	2.86	23.91	72.31	60.29	99.03	40.95	+28.11
ResNet-50	CBAM - Add	67.38	40.27	13.67	0.76	33.90	63.33	58.14	95.77	46.65	
	CBAM - Add + Aux	72.87	45.51	8.25	4.32	35.91	69.42	60.65	98.78	49.46	+2.81

outputs and CBAM. As shown in Figure 6, applying CBAM without a residual addition connection (CBAM−Add) resulted in noticeable performance gains (MIoU +1.92 on DFC2020), highlighting its effectiveness in enhancing both spatial and channel-wise attention. When combined with auxiliary supervision from Aux1, this configuration achieved the highest MIoU, indicating that early-stage guidance promotes more effective semantic feature extraction. In contrast, introducing an addition-based residual connection between the CBAM and convolutional blocks (CBAM+Add) did not yield further improvements on the DFC2020 dataset. Although a slight benefit was observed on DynamicEarthNet, performance decreased on DFC2020, likely due to the reintroduction of redundant features that had previously been suppressed by CBAM.

5.3. Comparison of MUSCLE-NET with Baseline Architectures

Based on the previous optimization and ablation analyses, the proposed MUSCLE-Net incorporates deep supervision through an auxiliary loss weighted at 10%, without applying the skip-with-addition operation between the convolutional blocks and the CBAM module. The model was compared with baseline architectures, including UNet, PSPNet, and DeepLabV3, using the same encoder backbone (ResNet-50) and identical initialization to ensure a fair evaluation. This configuration isolated decoder performance, as the main architectural differences lay in how features were upsampled and fused. PSPNet employed pyramid pooling to capture features at multiple spatial scales, DeepLabV3 utilized Atrous Spatial Pyramid Pooling (ASPP) to enhance contextual understanding, and UNet leveraged skip connections to combine encoder and decoder features for precise segmentation.

Table 5 presents semantic segmentation results (mIoU) on the DFC2020 test set, comparing our approach with state-of-the-art (SOTA) methods. SatMAE [12] relied on an OPT reconstruction-based pretraining strategy and was subsequently fine-tuned on the DFC2020 dataset. In contrast, CROMA [15], in both its base and large Vision Transformer variants (ViT-B/ViT-L), was pretrained using a combination of contrastive and reconstruction objectives, and was then evaluated using linear probing on DFC2020. Another recent model included in this comparison was FG-MAE [44], which was based on feature-guided masked reconstruction and was later fine-tuned on the DFC2020 dataset. Compared to these approaches, MUSCLE-Net demonstrated competitive and often superior performance across different backbone architectures. Notably, MUSCLE-Net with a ResNet50 backbone achieved the best performance, reaching 54.60% mIoU on the same test set.

Table 5
Semantic segmentation results (mIoU) on the same test set from the DFC2020 dataset of our MUSCLE-Net compared to SOTA with different encoders.

Model	mIoU (%) on DFC2020
SatMAE (ViT-L) (2022) [12]	44.13
Croma (ViT-B) (2023) [15]	51.58
Croma (ViT-L) (2023) [15]	53.24
FG-MAE (2024) [44]	51.8
MUSCLE-Net (MobileViT-S)	49.53
MUSCLE-Net (ConvNeXt V2)	53.56
MUSCLE-Net (ResNet50)	**54.60**

As shown in Table 6, a further comparison using the same encoder architecture and identical initial weights was performed. Five independent training and testing runs were conducted for MUSCLE-Net and the baseline networks on the DFC2020 dataset, and the average mIoU and OA were reported. Incorporating the CBAM module after each convolutional block, along with the use of deep supervision, resulted in an improvement of 2.89 in mIoU and 2.86 in OA compared to UNet.
The performance gains were particularly pronounced for challenging and underrepresented classes such as Savanna, Barren, and Grassland, indicating improved robustness to class imbalance. While all models achieved high accuracy on dominant classes such as Water and Forest, MUSCLE-Net demonstrated superior consistency, as reflected by lower standard deviations across runs. Moreover, MUSCLE-Net outperformed the baseline models on the highly imbalanced DynamicEarthNet dataset (Figure 7). It achieved the highest overall accuracy of 66.48 $\pm$ 0.68%, surpassing UNet (65.43 $\pm$ 2.42%), DeepLabV3 (60.27 $\pm$ 1.21%), and PSPNet (58.78 $\pm$ 0.71%). Notably, MUSCLE-Net also exhibited the lowest variability across runs, indicating more stable and consistent performance on the DynamicEarthNet dataset.

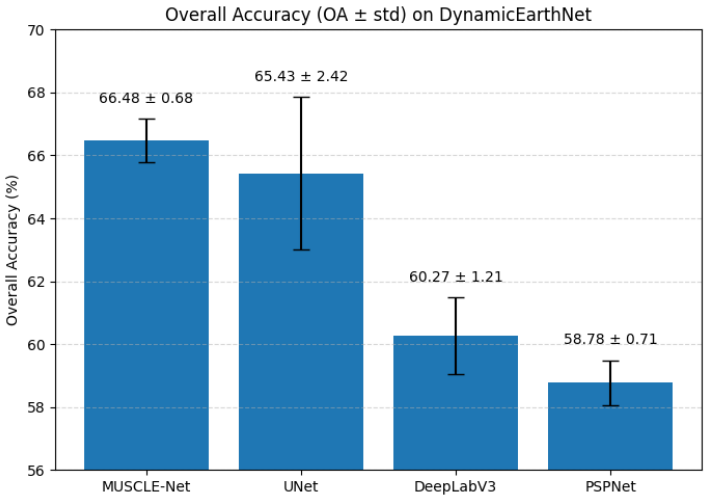


Figure 7: Overall Accuracy of MUSCLE-Net, UNet, PSPNet, and DeepLabV3 on the DynamicEarthNet test set, reported as OA ś standard deviation over five runs.

The results shown in Figure 8 demonstrate that the proposed model outperforms both UNet and PSPNet in pixel-level classification, particularly along LC class boundaries and in regions where certain classes are sparsely distributed. This improvement is attributed to the use of deep supervision, which guides the learning of global contextual features at early stages and progressively enhances the extraction of fine spatial details in later stages of the network. As a result, the proposed approach achieves more accurate edge delineation and class discrimination, even when compared with UNets encoderdecoder skip connections and PSPNets pyramid-based feature learning strategy.

Table 6
IoU per class and MIoU/OA for MUSCLE-Net, UNet, PSPNet, and DeepLabV3 on the DFC2020 dataset. Each value shows the mean ± Standard Deviation (%) over 5 runs. The best mean results are shown in bold.

Class	Class %	MUSCLE-Net	UNet	PSPNet	DeepLabV3
		IoU (%)	IoU (%)	IoU (%)	IoU (%)
Forest	25.64	74.0 ± 4.9	72.69 ± 2.03	64.80 ± 1.48	64.01 ± 1.19
Shrubland	5.93	45.4 ± 1.1	46.42 ± 1.60	40.40 ± 0.75	40.60 ± 1.06
Savanna	10.13	17.6 ± 3.4	13.13 ± 4.17	15.23 ± 6.95	12.80 ± 3.65
Grassland	2.09	10.4 ± 7.8	6.74 ± 4.11	2.44 ± 1.96	2.42 ± 2.24
Croplands	19.54	44.5 ± 4.4	33.14 ± 4.54	38.99 ± 5.66	30.88 ± 4.03
Urban/Built-up	10.73	73.9 ± 1.6	73.85 ± 1.85	61.95 ± 0.26	61.16 ± 2.34
Barren	2.61	64.2 ± 0.7	62.69 ± 0.92	57.07 ± 1.83	56.30 ± 2.37
Water	23.33	98.8 ± 0.2	98.99 ± 0.09	96.52 ± 0.30	95.08 ± 2.04
MIoU (%)	–	**53.82 ± 0.71**	50.93 ± 1.25	46.41 ± 2.11	45.41 ± 1.61
OA (%)	–	**70.10 ± 0.78**	67.24 ± 1.14	65.12 ± 1.83	64.00 ± 1.24

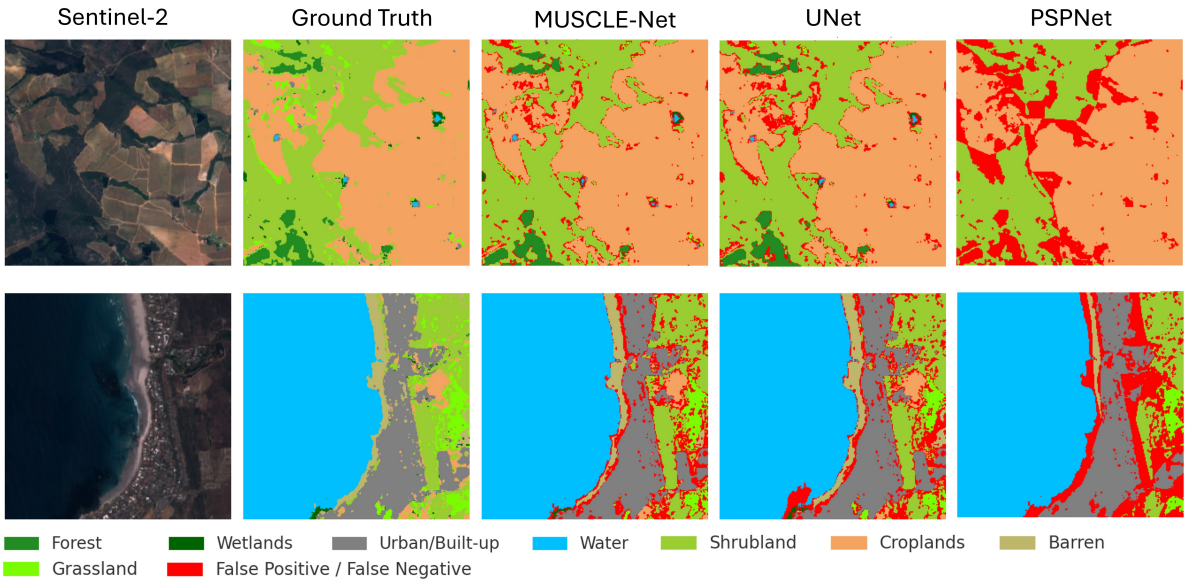


Figure 8: The predicted and ground truth masks generated using the MUSCLE-Net model on samples from the test set of the DFC2020 dataset.

6. Conclusion

Land cover segmentation remains a challenging problem due to pronounced class imbalance, spatial fragmentation of minority classes, and the limited availability of high-resolution annotated data. In this work, we demonstrate that constraining the decoding process with multi-resolution deep supervision is an effective strategy to address these challenges. By enforcing semantic consistency at coarse spatial scales during early decoding, the proposed approach promotes a coarse-to-fine refinement of land-cover regions and reduces overfitting to semantically misleading fine-scale patterns. Combined with spatialchannel attention in the decoder, this strategy leads to more robust and discriminative feature representations.

Experimental validation on the DFC2020 and DynamicEarthNet datasets confirms consistent performance improvements across different land cover classes. These results highlight the effectiveness of leveraging both low- and

high-resolution labels through deep supervision. Future work will focus on extending the framework to multi-temporal data and exploring weakly supervised learning strategies to further reduce dependence on densely annotated high-resolution ground truth.

Declaration of competing interest

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

CRediT authorship contribution statement

Sara Mobsite: Writing original draft, Visualization, Software, Methodology, Investigation, Formal analysis, Conceptualization. Renaud Hostache: Writing review & editing, Supervision, Methodology, Formal analysis, Conceptualization. Laure Berti-Équille: Writing review & editing, Supervision, Methodology, Formal analysis, Conceptualization. Joris Guérin: Writing review & editing, Supervision, Methodology, Formal analysis, Conceptualization. Emmanuel Roux: Writing review & editing, Funding acquisition, Conceptualization. Thibault Catry: Writing review & editing, Conceptualization.

Acknowledgments

This work has received funding from the European Unions Horizon Europe Research and Innovation program under Grant Agreement Nž 101137398.

Data and Code Availability

The code for MUSCLE-Net, along with the data preprocessing pipeline, is publicly available at: code.

Appendix

A. Assessment of OPT, SAR, and Their Integration

In this section, we evaluate the impact of using OPT data only, SAR data only, and their combination at the input level on model performance. All model weights reported in Table A1 were initialized using the reBEN dataset and kept frozen during the comparative analysis to ensure a fair evaluation of the different input configurations. All experiments were conducted with a batch size of 64. The decoder excluded deep supervision and consisted solely of convolutional operations followed by a CBAM block without an ADD residual connection.

The experimental results reveal several trends across encoder architectures and input modalities. Overall, multimodal input (SAR+OPT) yielded higher MIoU and OA values across all encoders, highlighting the benefit of fusing complementary information from both sensors. Among the evaluated architectures, Inception NeXt Base achieved the highest MIoU (47.21%) and OA (67.59%) in the multimodal setting, with noticeable improvements in challenging classes such as Savanna, Croplands, and Barren. ResNet-50 demonstrated competitive performance across all input configurations, particularly in the SAR-only and OPT-only settings, where it achieved the highest MIoU under SAR-only input (35.85%) and consistently strong OA values. MobileViT-S showed moderate performance, performing well on Croplands in the multimodal setting, but exhibiting lower overall consistency compared to ResNet-50 and Inception NeXt. ConvNeXt V2 Base showed limited transferability to the DFC2020 dataset, with lower MIoU and OA across all input modalities. Performance degradation was especially evident for minority classes such as Grassland and Barren, where near-zero IoU values were observed under OPT-only input.

Across all models and input configurations, Water consistently achieved the highest and most stable IoU values, indicating that it is the easiest class to segment. In contrast, Grassland and Savanna remained the most challenging classes, exhibiting low IoU across architectures and modalities, which can be attributed to their limited representation in the training data.

B. Trainable vs. Frozen Encoder

The performance of the proposed model was evaluated using a decoder composed solely of convolutional operations followed by a CBAM block, without an ADD residual connection, while assessing each encoder under two training

Table A1
IoU per class (%), MIoU(%), and OA (%) for MUSCLE-Net without skip connection and deep supervision using different encoder architectures on the DFC2020 dataset under different sensor modalities (SAR+OPT, SAR only, OPT only).

Class / Metric	SAR+OPT				SAR only				OPT only			
	ResNet-50	MobileViT-S	ConvNeXt V2	Inception NeXt	ResNet-50	MobileViT-S	ConvNeXt V2	Inception NeXt	ResNet-50	MobileViT-S	ConvNeXt V2	Inception NeXt
Forest	57.41	44.28	23.92	57.73	50.47	37.88	43.76	53.62	31.22	31.99	20.26	57.12
Shrubland	38.15	28.65	22.36	37.80	12.92	14.89	10.97	17.74	23.33	27.56	16.37	15.41
Savanna	8.20	15.14	5.73	26.12	1.89	2.36	3.16	2.44	11.46	18.46	0.21	8.96
Grassland	1.55	0.09	0.00	0.64	0.19	0.78	0.07	0.53	0.19	0.05	0.00	0.42
Croplands	27.62	55.21	38.62	49.53	33.49	31.10	40.21	31.99	42.26	39.19	36.77	45.31
Urban/Built-up	59.11	28.24	39.22	59.40	58.51	48.51	56.12	57.67	60.19	48.10	42.03	60.67
Barren	44.64	29.87	27.93	55.52	35.28	35.58	25.11	21.88	54.92	32.00	0.01	18.99
Water	93.08	81.88	84.25	90.94	94.03	84.48	94.68	94.36	94.24	82.99	81.90	94.04
MIoU (%)	41.22	35.42	30.25	47.21	35.85	31.95	34.26	35.03	39.72	35.04	24.69	37.62
OA (%)	61.51	59.18	54.54	67.59	59.18	57.64	56.88	60.08	62.14	59.13	51.42	61.38

Table B1
Per-class IoU (%) and MIoU (%) of the proposed model on the DFC2020 test set using different encoder architectures under two training strategies: with frozen encoder weights and with trainable encoder weights. The last column reports the improvement (Δ MIoU) relative to the frozen-weight baseline.

Encoder Model	Training Strategy	Forest	Shrubland	Savanna	Grassland	Croplands	Urban	Barren	Water	MIoU	Δ MIoU
Inception NeXt	Frozen	57.73	37.80	26.12	0.64	49.53	59.40	55.52	90.94	47.21	
	Trainable	67.34	42.08	18.73	0.96	30.92	64.91	59.22	97.24	47.68	+0.47
MobileViT-S	Frozen	44.28	28.65	15.14	0.09	55.21	28.24	29.87	81.88	35.42	
	Trainable	35.59	38.27	19.45	0.01	51.65	56.65	55.92	93.60	43.89	+8.47
ConvNeXt V2	Frozen	23.92	22.36	5.73	0.00	38.62	39.22	27.93	84.25	30.25	
	Trainable	13.29	0.00	0.00	0.00	0.50	0.00	0.02	88.88	12.84	17.41
ResNet-50	Frozen	57.41	38.15	8.20	1.55	27.62	59.11	44.64	93.08	41.22	
	Trainable	67.38	40.27	13.67	0.76	33.90	63.33	58.14	95.77	46.65	+5.34

strategies: frozen encoder weights and fully trainable encoder weights. A fixed batch size of 64 was used across all experiments to ensure consistent and fair comparisons. As reported in Table B1, enabling encoder fine-tuning generally led to performance improvements for most architectures, with the exception of ConvNeXt V2. Inception NeXt exhibited a marginal increase in MIoU (+0.47), indicating that its pretrained representations learned from reBEN were already well aligned with the DFC2020 segmentation task and benefited only slightly from further adaptation. In contrast, MobileViT-S and ResNet-50 showed substantial gains when trained end-to-end, achieving MIoU improvements of +8.47 and +5.34, respectively. These gains were driven by notable improvements across several classes, particularly Urban, Barren, and Water, suggesting that these architectures were able to effectively refine their pretrained features to better capture task-specific characteristics.

Conversely, ConvNeXt V2 experienced a significant degradation in performance when trained with unfrozen weights, with MIoU dropping from 30.25% to 12.84%. This degradation was accompanied by near-zero IoU values for most land-cover classes, indicating instability during fine-tuning on the relatively small DFC2020 dataset. This behavior suggests that the features learned during pretraining on the larger reBEN dataset were more suitable for this task when kept frozen, whereas full fine-tuning led to overfitting or loss of useful representations.

C. Impact of Batch Size on Model Performance

The performance of the proposed models was evaluated with two different batch sizes (Table C1). ResNet-50 demonstrated the best performance among the tested encoders within the MUSCLE-Net framework when a batch size of 32 was used. The superior performance with the smaller batch size was attributed to its ability to provide more frequent weight updates and introduce beneficial gradient noise, which often improved generalization of neural networks [21, 24].

Table C1
Performance comparison of encoders (ResNet50, Inception, MobileViT, and ConvNeXtV2) on the DFC2020 test set with batch sizes 32 and 64. Reported values include IoU for each class (%), along with mIoU (%) and OA (%). All encoder weights were initialized from pre-trained reBEN models and kept trainable during training.

Class	Batch Size 32				Batch Size 64			
	ResNet-50	Inception NeXt	MobileViT-S	ConvNeXt V2	ResNet-50	Inception NeXt	MobileViT-S	ConvNeXt V2
Water	98.96	98.46	98.77	98.18	98.78	98.31	98.79	99.03
Barren	64.79	43.60	60.04	63.51	60.65	56.94	62.88	60.29
Urban	75.58	67.13	74.29	69.66	69.42	63.97	75.07	72.31
Croplands	41.98	29.22	33.29	41.84	35.91	32.37	47.95	23.91
Grassland	16.64	0.37	4.80	20.85	4.32	2.41	3.16	2.86
Savanna	14.62	12.18	20.03	22.13	8.25	16.88	23.78	9.11
Shrubland	45.43	34.66	40.69	45.85	45.51	44.69	44.27	46.67
Forest	78.82	61.98	64.34	66.44	72.87	66.09	61.92	13.41
MIoU (%)	54.60	43.45	49.53	53.56	49.46	47.71	52.23	40.95
OA (%)	70.27	63.26	67.48	70.59	66.48	65.30	70.00	59.15

References

[1] Amin, G., Oberlin, T., Demarez, V., 2025. Early-season delineation of agricultural fields using a fully convolutional multi-task network and satellite images. Science of Remote Sensing 12, 100256.

[2] Andresini, G., Appice, A., Malerba, D., 2024a. A deep semantic segmentation approach to map forest tree dieback in sentinel-2 data. IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing .

[3] Andresini, G., Appice, A., Malerba, D., 2024b. Leveraging sentinel-2 time series for bark beetle-induced forest dieback inventory, in: Proceedings of the 39th ACM/SIGAPP Symposium on Applied Computing, pp. 875–882.

[4] Badrinarayanan, V., Kendall, A., Cipolla, R., 2017. Segnet: A deep convolutional encoder-decoder architecture for image segmentation. IEEE transactions on pattern analysis and machine intelligence 39, 2481–2495.

[5] Buttar, P.K., Sachan, M.K., 2024. Generating land cover maps in semi-arid regions based on a 3d semantic segmentation architecture using multi-temporal sentinel-2 satellite images: a case study of ludhiana district in punjab, india. Journal of the Indian Society of Remote Sensing 52, 383–398.

[6] Chen, L.C., Papandreou, G., Schroff, F., Adam, H., 2017. Rethinking atrous convolution for semantic image segmentation. arXiv preprint arXiv:1706.05587 .

[7] Chen, Y., Zhang, G., Cui, H., Li, X., Hou, S., Ma, J., Li, Z., Li, H., Wang, H., 2023. A novel weakly supervised semantic segmentation framework to improve the resolution of land cover product. ISPRS Journal of Photogrammetry and Remote Sensing 196, 73–92.

[8] Chen, Y., Zhang, G., Cui, H., Li, X., Hou, S., Zhu, C., Xie, Z., Li, D., 2025. Superpixel-aware credible dual-expert learning for land cover mapping using historical land cover product. ISPRS Journal of Photogrammetry and Remote Sensing 223, 296–316.

[9] Choi, Y., Lee, D., Moon, S., 2024. Deforestation segmentation approach based on time of event occurrence using multi-temporal satellite data. IEEE Sensors Letters .

[10] Chollet, F., 2017. Xception: Deep learning with depthwise separable convolutions, in: Proceedings of the IEEE conference on computer vision and pattern recognition, pp. 1251–1258.

[11] Clasen, K.N., Hackel, L., Burgert, T., Sumbul, G., Demir, B., Markl, V., 2025. reBEN: Refined bigearthnet dataset for remote sensing image analysis, in: IEEE International Geoscience and Remote Sensing Symposium (IGARSS).

[12] Cong, Y., Khanna, S., Meng, C., Liu, P., Rozi, E., He, Y., Burke, M., Lobell, D., Ermon, S., 2022. Satmae: Pre-training transformers for temporal and multi-spectral satellite imagery. Advances in Neural Information Processing Systems 35, 197–211.

[13] Cui, H., Zhang, G., Chen, Y., Li, X., Hou, S., Li, H., Ma, X., Guan, N., Tang, X., 2024. Knowledge evolution learning: A cost-free weakly supervised semantic segmentation framework for high-resolution land cover classification. ISPRS Journal of Photogrammetry and Remote Sensing 207, 74–91.

[14] Fan, T., Wang, G., Li, Y., Wang, H., 2020. Ma-net: A multi-scale attention network for liver and tumor segmentation. IEEE Access 8, 179656–179665.

[15] Fuller, A., Millard, K., Green, J., 2023. Croma: Remote sensing representations with contrastive radar-optical masked autoencoders. Advances in Neural Information Processing Systems 36, 5506–5538.

[16] Goodfellow, I., Bengio, Y., Courville, A., Bengio, Y., 2016. Deep learning. volume 1. MIT press Cambridge.

[17] Guimarães, U.S., Rodrigues, T.B., Vieira, A.C., Hung, E.M., Soja, M.J., Eriksson, L.E., Ulander, L., . Deep learning models to map deforestation based on sentinel 1 coherent features in the southern border of amazon. Available at SSRN 5258317 .

[18] Hackel, L., Clasen, K.N., Demir, B., 2024. Configilm: A general purpose configurable library for combining image and language models for visual question answering. SoftwareX 26, 101731.

[19] He, K., Zhang, X., Ren, S., Sun, J., 2016. Deep residual learning for image recognition, in: Proceedings of the IEEE conference on computer vision and pattern recognition, pp. 770–778.

[20] Hochreiter, S., Schmidhuber, J., 1997. Long short-term memory. Neural computation 9, 1735–1780.

[21] Hoffer, E., Hubara, I., Soudry, D., 2017. Train longer, generalize better: closing the generalization gap in large batch training of neural networks. Advances in neural information processing systems 30.

[22] Huang, K., Yang, G., Yuan, Y., Sun, W., Meng, X., Ge, Y., 2022. Optical and sar images combined mangrove index based on multi-feature fusion. Science of Remote Sensing 5, 100040.

[23] Jamali, A., Roy, S.K., Beni, L.H., Pradhan, B., Li, J., Ghamisi, P., 2024. Residual wave vision u-net for flood mapping using dual polarization sentinel-1 sar imagery. International Journal of Applied Earth Observation and Geoinformation 127, 103662.

[24] Keskar, N.S., Mudigere, D., Nocedal, J., Smelyanskiy, M., Tang, P.T.P., 2016. On large-batch training for deep learning: Generalization gap and sharp minima. arXiv preprint arXiv:1609.04836 .

[25] Kingma, D.P., 2014. Adam: A method for stochastic optimization. arXiv preprint arXiv:1412.6980 .

[26] Li, H., Xiong, P., An, J., Wang, L., 2018. Pyramid attention network for semantic segmentation. arXiv preprint arXiv:1805.10180 .

[27] Li, X., Zhang, J., Yang, Y., Cheng, G., Yang, K., Tong, Y., Tao, D., 2024. Sfnet: Faster and accurate semantic segmentation via semantic flow. International Journal of Computer Vision 132, 466–489.

[28] Liu, Z., Mao, H., Wu, C.Y., Feichtenhofer, C., Darrell, T., Xie, S., 2022. A convnet for the 2020s, in: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition, pp. 11976–11986.

[29] Long, J., Shelhamer, E., Darrell, T., 2015. Fully convolutional networks for semantic segmentation, in: Proceedings of the IEEE conference on computer vision and pattern recognition, pp. 3431–3440.

[30] Lu, L., Xu, Y., Huang, X., Zhang, H.K., Du, Y., 2025. Large-scale mapping of plastic-mulched land from sentinel-2 using an index-feature-spatial-attention fused deep learning model. Science of Remote Sensing 11, 100188.

[31] Mehta, S., Rastegari, M., 2021. Mobilevit: light-weight, general-purpose, and mobile-friendly vision transformer. arXiv preprint arXiv:2110.02178 .

[32] Pech-May, F., Aquino-Santos, R., Álvarez-Cárdenas, O., Arandia, J.L., Rios-Toledo, G., 2024. Segmentation and visualization of flooded areas through sentinel-1 images and u-net. IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing .

[33] Pešek, O., Brodskỳ, L., Halounová, L., Landa, M., Bouček, T., 2024. Convolutional neural networks for urban green areas semantic segmentation on sentinel-2 data. Remote Sensing Applications: Society and Environment 36, 101238.

[34] Robinson, C., Malkin, K., Jojic, N., Chen, H., Qin, R., Xiao, C., Schmitt, M., Ghamisi, P., Hänsch, R., Yokoya, N., 2021. Global land-cover mapping with weak supervision: Outcome of the 2020 ieee grss data fusion contest. IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing 14, 3185–3199.

[35] Ronneberger, O., Fischer, P., Brox, T., 2015. U-net: Convolutional networks for biomedical image segmentation, in: Medical image computing and computer-assisted intervention–MICCAI 2015: 18th international conference, Munich, Germany, October 5-9, 2015, proceedings, part III 18, Springer. pp. 234–241.

[36] Schmitt, M., Hughes, L., Ghamisi, P., Yokoya, N., Hansch, R., 2019a. 2020 ieee grss data fusion contest. URL: https://dx.doi.org/10.21227/rha7-m332, doi:10.21227/rha7-m332.

[37] Schmitt, M., Hughes, L.H., Qiu, C., Zhu, X.X., 2019b. Sen12ms–a curated dataset of georeferenced multi-spectral sentinel-1/2 imagery for deep learning and data fusion. arXiv preprint arXiv:1906.07789 .

[38] Sharma, S., Gosain, A., 2025. Addressing class imbalance in remote sensing using deep learning approaches: a systematic literature review. Evolutionary Intelligence 18, 1–28.

[39] Szegedy, C., Liu, W., Jia, Y., Sermanet, P., Reed, S., Anguelov, D., Erhan, D., Vanhoucke, V., Rabinovich, A., 2015. Going deeper with convolutions, in: Proceedings of the IEEE conference on computer vision and pattern recognition, pp. 1–9.

[40] Tian, Y., Zhao, F., Meng, R., Sun, R., Zhang, Y., Shen, Y., Wang, B., Liu, J., Li, M., 2025. A vision foundation model-based method for large-scale forest disturbance mapping using time series sentinel-1 sar data. Remote Sensing of Environment 325, 114775.

[41] Toker, A., Kondmann, L., Weber, M., Eisenberger, M., Camero, A., Hu, J., Hoderlein, A.P., Senaras, C., Davis, T., Cremers, D., et al., 2022. Dynamicearthnet: Daily multi-spectral satellite dataset for semantic change segmentation, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 21158–21167.

[42] Vo Quang, A., Delbart, N., Jaffrain, G., Pinet, C., 2025. Detection of degraded forests in guinea, west africa, using convolutional neural networks and sentinel-2 time series. Frontiers in Remote Sensing 6, 1538808.

[43] Wang, S., Cai, B., Hou, D., Ding, Q., Wang, J., Shao, Z., 2024a. Mf-bhnet: A hybrid multimodal fusion network for building height estimation using sentinel-1 and sentinel-2 imagery. IEEE Transactions on Geoscience and Remote Sensing .

[44] Wang, Y., Hernández, H.H., Albrecht, C.M., Zhu, X.X., 2024b. Feature guided masked autoencoder for self-supervised learning in remote sensing. IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing .

[45] Wang, Z., Li, Y., Wang, J., Liu, B., 2025. Lh-unet: A lighted histoformer-encoded u-net for sea ice recognition with high-resolution remote sensing images. IEEE Geoscience and Remote Sensing Letters .

[46] Woo, S., Debnath, S., Hu, R., Chen, X., Liu, Z., Kweon, I.S., Xie, S., 2023. Convnext v2: Co-designing and scaling convnets with masked autoencoders, in: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition, pp. 16133–16142.

[47] Woo, S., Park, J., Lee, J.Y., Kweon, I.S., 2018. Cbam: Convolutional block attention module, in: Proceedings of the European conference on computer vision (ECCV), pp. 3–19.

[48] Xu, D., Li, Z., Feng, H., Wu, F., Wang, Y., 2024. Multi-scale feature fusion network with symmetric attention for land cover classification using sar and optical images. Remote Sensing 16, 957.

[49] Yu, W., Luo, M., Zhou, P., Si, C., Zhou, Y., Wang, X., Feng, J., Yan, S., 2022. Metaformer is actually what you need for vision, in: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition, pp. 10819–10829.

[50] Yu, W., Si, C., Zhou, P., Luo, M., Zhou, Y., Feng, J., Yan, S., Wang, X., 2023. Metaformer baselines for vision. IEEE Transactions on Pattern Analysis and Machine Intelligence 46, 896–912.

[51] Yu, W., Zhou, P., Yan, S., Wang, X., 2024. Inceptionnext: When inception meets convnext, in: Proceedings of the IEEE/cvf conference on computer vision and pattern recognition, pp. 5672–5683.

[52] Yuan, X., Shi, J., Gu, L., 2021. A review of deep learning methods for semantic segmentation of remote sensing imagery. Expert Systems with Applications 169, 114417.

[53] Zhao, H., Shi, J., Qi, X., Wang, X., Jia, J., 2017. Pyramid scene parsing network, in: Proceedings of the IEEE conference on computer vision and pattern recognition, pp. 2881–2890.

[54] Zhao, J., Xiong, Z., Zhu, X.X., 2024. Urbansarfloods: Sentinel-1 slc-based benchmark dataset for urban and open-area flood mapping, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 419–429.

[55] Zhou, Z., Rahman Siddiquee, M.M., Tajbakhsh, N., Liang, J., 2018. Unet++: A nested u-net architecture for medical image segmentation, in: Deep learning in medical image analysis and multimodal learning for clinical decision support: 4th international workshop, DLMIA 2018, and 8th international workshop, ML-CDS 2018, held in conjunction with MICCAI 2018, Granada, Spain, September 20, 2018, proceedings 4, Springer. pp. 3–11.

Highlights

- Introduces MUSCLE-Net with multi-resolution deep supervision in the decoder
- Enforces semantic consistency across scales to improve minority class learning
- Proposes an attention-enhanced decoder using CBAM without residual addition
- Re-injects auxiliary supervision features to guide early decoder representations
- Demonstrates consistent accuracy and stability gains on two benchmark datasets

Declaration of Interest Statement

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

☐ The author is an Editorial Board Member/Editor-in-Chief/Associate Editor/Guest Editor for this journal and was not involved in the editorial review or the decision to publish this article.

☐ The authors declare the following financial interests/personal relationships which may be considered as potential competing interests: