remote sensing

MDPI

Article

CALCNet: A Novel Cross-Module Attention Network for Efficient Land Cover Classification

Muhammad Fayaz [1,†], Hikmat Yar [2,3,†], Weiwei Jiang [4], Anwar Hassan Ibrahim [5], Muhammad Islam [5] and L. Minh Dang [6,7,8,*]

[1] Department of Computer Science and Engineering, Sejong University, Seoul 05006, Republic of Korea; muhammadfayaz@sju.ac.kr
[2] Department of Mechanical Robotics and Energy Engineering, Dongguk University-Seoul, 30 Pildong-ro 1-gil, Jung-gu, Seoul 04620, Republic of Korea; hikmatyar@kaist.ac.kr
[3] KAIST InnoCORE PRISM-AI Center, Korea Advanced Institute of Science and Technology (KAIST), Daejeon 34141, Republic of Korea
[4] School of Information and Communication Engineering, Beijing University of Posts and Telecommunications, Beijing 100876, China; jww@bupt.edu.cn
[5] Department of Electrical Engineering, College of Engineering, Qassim University, Buraydah 52571, Saudi Arabia
[6] The Institute of Research and Development, Duy Tan University, Da Nang 550000, Vietnam
[7] Faculty of Information Technology, Duy Tan University, Da Nang 550000, Vietnam
[8] Department of Information and Communication Engineering and Convergence Engineering for Intelligent Drone, Sejong University, Seoul 05006, Republic of Korea
* Correspondence: danglienminh@duytan.edu.vn
† These authors contributed equally to this work.

Highlights

What are the main findings?

- A novel land cover classification network (CALCNet) is developed from scratch, combining a contracting and restoration backbone with cross-module attention for enhanced spatial and multi-scale feature representation
- A differential evolution–based neuron pruning strategy is developed that produces a compressed CALCNet variant that maintains high classification accuracy with reduced computational cost.

What are the implications of the main findings?

- The proposed CALCNet architecture improves performance in complex remote sensing scenarios beyond fine-tuned pre-trained models.
- The efficient and lightweight design enables practical deployment of LCC models in real-world and large-scale remote sensing applications.

Abstract

Land cover classification (LCC) is a fundamental task in remote sensing, which enables effective environmental monitoring, agricultural planning, and disaster management. The existing approaches often rely on fine-tuning pre-trained models, which are not specifically designed for LCC, which lead to suboptimal performance in complex scenarios. To address these limitations, we propose the Cross-Module Attention Land Cover Network (CALCNet), a novel architecture developed from scratch. CALCNet follows a contracting and restoration backbone, where the contracting path extracts progressively abstract semantic features while reducing spatial resolution, and the restoration path recovers fine-grained spatial details through upsampling and skip connections. In addition, CALCNet integrates a cross-module attention mechanism that combines spatial attention and multi-scale feature selection to enhance feature representation. Furthermore, we applied a differential evolution-based neuron pruning strategy to create a compressed CALCNet variant, which

Academic Editors: Peixian Zhuang, Xueyang Fu and Xiangyong Cao

Received: 25 January 2026
Revised: 9 April 2026
Accepted: 13 April 2026
Published: 17 April 2026

retains high classification performance while reducing computational cost. The CALCNet is evaluated on four benchmark LCC datasets, AID, UCMerced_LandUse, NWPU_RESISC45, and EuroSAT, demonstrating strong performance across all benchmarks. Specifically, the model achieves classification accuracies of 98.09%, 99.47%, 99.19%, and 99.19%, respectively. The compressed CALCNet variant reduces computational cost to 78.55 million floating point operations (FLOPs) with a model size of 43 MB, while achieving improved inference speeds (38.32 frames/sec on CPU and 118.3 frames/sec on GPU), representing approximately 45–50% reduction in FLOPs and model storage. These results highlight that CALCNet is both highly accurate and computationally efficient, making it well suited for real-world LCC applications.

Keywords: remote sensing; satellite image analysis; land cover classification; land area images; cross attention

1. Introduction

The rapid development of remote sensing (RS) technologies has generated the large-scale availability of earth observation data, from high-resolution multispectral and hyperspectral satellite imagery to radar and LiDAR. These datasets make significant contributions to a wide range of applications, which include global environmental monitoring, agricultural management, disaster response, climate change analysis, and especially land cover classification (LCC) [1]. Accurate and robust LCC is essential for geo-observation, supporting agricultural zoning, biodiversity conservation, promoting urban expansion, and reliable natural resource management [2]. Accurate classification is the core of abstraction and rules at a higher level in many of these types of systems, and errors here can impact the entire pipeline, which lead to policies or interventions that are inadequate. For example, the misclassification of croplands as being barren land could mislead agricultural planning, and mixing urban and forested areas would affect both sustainable city development and environmental protection efforts. Diverse approaches to classifying LCC have been introduced in the literature, which include both the existing techniques and the advanced deep learning (DL) techniques.

Conventional LCC methods often rely on hand-crafted features such as color histograms, pixel intensities, spectral indices, and texture descriptors [3]. Conventional machine learning algorithms, such as Support Vector Machines (SVMs) [4], Random Forests (RFs) [5], and K-means clustering [6] have been effectively applied in various scenarios in definitions for traditional approaches, but their performance remains limited by the high complexity and heterogeneity of RS data in various applications [7,8]. These methods have an inherent inability to capture contextual and hierarchical spatial patterns, which restricts their performance in regards to challenging land cover categories with high intra-class variability and low inter-class variance.

With the development of Deep Learning (DL), LCC has undergone tremendous progress in terms of automatic and hierarchical feature abstraction. For example, Convolutional Neural Networks (CNNs), Recurrent Neural Networks (RNNs), and more recently, Transformer-based architectures are being used in order to capture the spatial and/or spectral as well as temporal dependencies in RS data. These models have shown significant advances in classification accuracy and robustness over traditional methods. The integration of LCC into more knowledge-based systems which include precision agriculture, as well as real-time and scalable classification outputs, are critical to applications, such as environmental monitoring, and urban planning, which have also been enabled by the

performance increases from DL. LCC has consequently evolved from a purely classification task into a critical enabler of intelligent decision-support systems in remote sensing. Despite these advancements, many challenges are associated with DL, which are elaborated on in the next section, along with their respective solutions. The compact Transformer and hybrid CNN–Transformer models, such as Swin-T and ConvNeXt variants have shown competitive performance on general scene classification tasks, so CALCNet is specifically designed in order to capture fine-grained spatial–spectral features in remote sensing imagery through a contracting–expanding backbone with cross-module attention and multi-scale feature selection, which makes it particularly effective for land cover classification as well as maintaining computational efficiency.

1.1. Main Challenges

The main challenges outlined in the literature are as follows:

1. Most of the existing approaches rely on fine-tuning pre-trained models that are not specifically designed for LCC, which result in suboptimal feature representation for remote sensing imagery.
2. The previous methods often use shallow architectures or attention mechanisms in isolation, but there is no comprehensive study that integrates both attention and multi-scale feature selection for robust feature extraction.
3. DL models for LCC are typically computationally expensive, and no effective pruning or compression strategies have been applied in order to reduce model size and inference latency.
4. The performance of the existing models is often insufficient for real-time deployment, limiting their practical applicability in resource-constrained and large-scale remote sensing scenarios.

1.2. Contributions

The contributions of this work are as follows:

- CALCNet is developed from scratch using an encoder–decoder architecture with contracting and expansive paths. The contracting path captures hierarchical semantic information while gradually reducing the spatial resolution, whereas the expansive path reconstructs fine-grained details through upsampling and skip connections. This architecture allows the network to effectively learn the spatial and spectral variations that are inherent in remote sensing imagery.
- We integrate a cross-module attention mechanism that combines channel attention with multi-scale feature selection (MSFS). Channel attention strengthens informative feature maps, whereas MSFS gathers features across multiple scales, which help the network emphasize dominant patterns and maintain robustness under diverse environmental conditions.
- A neuron pruning strategy based on differential evolution is used to obtain a compressed variant of CALCNet. Despite compression, the algorithm maintains high classification performance, which allows for deployment in near-real-time on resource-limited platforms.
- CALCNet achieves a significantly higher classification accuracy performance than existing baseline methods over different datasets while being computationally efficient. These properties allow it to be suitable for practical applications, such as especially large-scale remote sensing scenarios.

1.3. Study Outline

The rest of this paper is organized as follows. Section 2 discusses related work on LCC, the details of the proposed CALCNet in Section 3, and datasets, experimental setup, and evaluation metrics followed by results and discussion in Section 4. Section 5 concludes the paper and outlines avenues for future work.

2. Related Work

Land cover classification is considered one of the major domains in remote sensing. For land cover classification, early methods relying on the manually defined rules called traditional machine learning (ML) were applied to images [9]. These traditional ML methods employing hand-crafted features are mostly constrained to low-level descriptors, such as spectral indices, color, and texture. Consequently, they fail to adequately capture complex land structures and interactions, which leads to limitations in their overall classification ability.

DL in particular has recently gained significant popularity as a powerful tool for LCC, because it automatically learns multi-scale and hierarchical features from large-scale remote sensing images (RSIs), which result in enhanced detection and classification accuracy and robustness [10]. DL-based techniques can generally be defined in terms of the spatial representation of their labels, which are patch-level and pixel-level methods. Patch-level approaches are commonly applied to medium-resolution RSIs, where fine structural details are limited [11]. Li et al. [12] introduced a patch-based recursive neural network (RNN), and Lv et al. [13] introduced a lightweight CNN model in land cover mapping. In contrast, pixel-based methods seek to label a pixel, similar to semantic segmentation in natural images. Pixel-level LCC baseline models generally assume the encoder–decoder network architecture to encode local and global contextual clues with multiple receptive fields [14,15]. McDonnell et al. [16] extended a stacked U-Net for semantic segmentation of RGB RSIs, and Liu et al. [17] utilized a compact dilated CNN to integrate contextual information at multiple scales.

Multisource data fusion is another promising research direction to improve classification performance, particularly given the limitations of individual RS sensors in providing high temporal, spatial, or spectral resolutions. Multisource fusion strategies usually consist of two stages that include (i) fusing heterogeneous data, and (ii) ML-based classification. For instance, in the studies by Iervolino et al. [18] and Kulkarni et al. [19], ML-based classifiers, such as genetic algorithms, SVM [20], or Markov random field-based models [21], are employed after the fusion process. Nonetheless, traditional hand-crafted feature methods are still restricted due to their low expressive power, which makes it difficult to accurately represent high-level semantic information.

The recent advancements in multimodal DL models have largely supplanted traditional approaches. Chen et al. [22] developed a two-branch DL approach that leverages two CNNs for feature extraction from multisource data, such as multi/hyperspectral and LiDAR data, which is followed by a fully connected Deep Neural Network (DNN) for feature fusion. Hughes et al. [23] proposed a pseudo-Siamese CNN in order to identify the corresponding patches in high-resolution optical and Synthetic Aperture Radar (SAR) images, by integrating information through concatenation and 1×1 convolutions at the decision stage. These types of advances have been made, but patch-based multimodal models do not typically have the granularity that is needed in order to perform pixel-level classification in high-resolution RSIs. To address this issue, Xu and Li et al. [24] developed a two-branch CNN architecture to classify the pixels at the pixel level, which combines both hyperspectral imagery and multisensors. Others have studied the concept of early and late fusion in the urban segmentation process. In particular, Audebert et al. [25] tried a

combination approach, where the late-fusion strategy learns to encapsulate the errors at the harder pixels, and where early-fusion approaches, like V-FesuNet, learn to generate strong multimodal features, and are less sensitive to missing or noise-influenced input. A fusion-fully convolutional network (FCN) model and Data , Hyperspectral + LiDAR + high-resolution imagery was used [26]. Capilez et al. [27] designed an M3 fusion network, which is a combination of CNNs and RNNs with the ability to utilize both spatial and temporal data.

In summary, the conventional multimodal DL approaches to LCC typically extract features of both modalities and combine the features before making a classification. In contrast, the proposed CALCNet employs a cross-module attention mechanism that jointly integrates spatial and multi-scale discriminative features. The design not only enables the efficient combination of multimodal information but also captures details in each modality, improving the network's ability to distinguish between diverse, complex land cover types. As a result, CALCNet achieves superior classification performance across diverse remote sensing datasets while maintaining computational efficiency.

3. Proposed Method

The proposed CALCNet framework is illustrated in Figure 1, which consist of three primary modules that include feature extraction, cross-module attention, and multi-scale feature selection, and it is followed by classification. Each module of the CALCNet is briefly discussed in the subsequent sections.

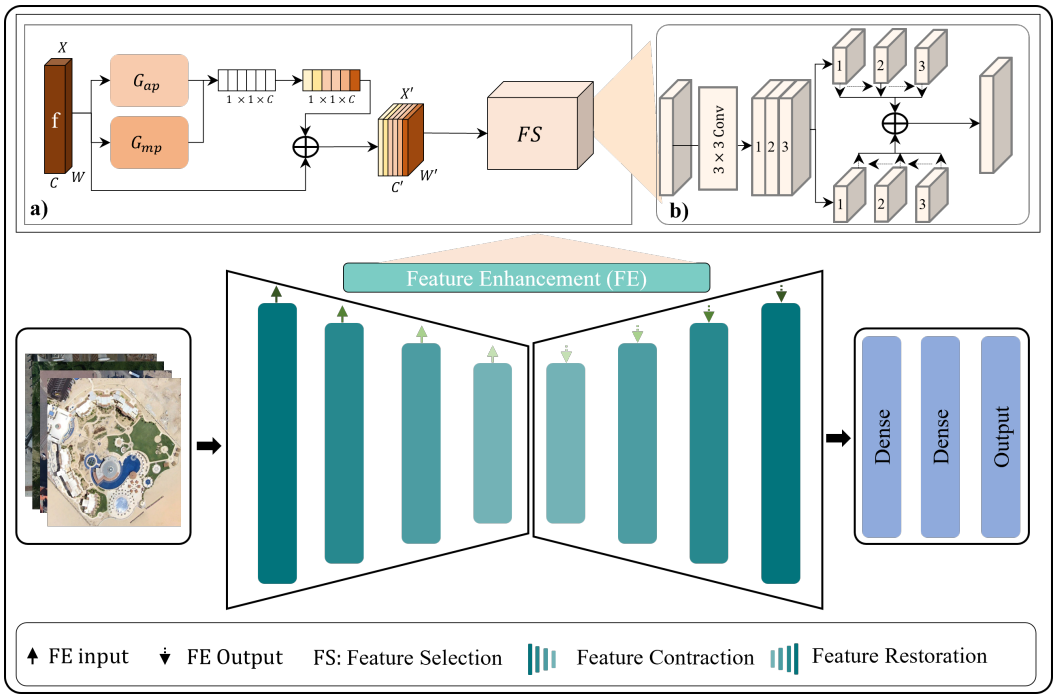

Figure 1. High-level framework for the CALCNet for land cover classification, where (**a**) represents the attention module and (**b**) is the feature selection module.

3.1. Backbone Feature Extraction

Let $X \in R^{H \times W \times C}$ be the input image. Each Feature Extraction Module (FEM) applies two convolutional layers followed by batch normalization (BN) and pooling. For the l-th layer in FEM, the convolutional operation is defined as:

$$F^{(l)} = \text{Conv}(F^{(l-1)}, W^{(l)}) + b^{(l)}, \quad (1)$$

where $W^{(l)}$ and $b^{(l)}$ are the weights and biases, respectively, and $F^{(0)} = X$. The output of BN is given by

$$\hat{F}^{(l)} = \frac{F^{(l)} - \mu^{(l)}}{\sqrt{(\sigma^{(l)})^2 + \epsilon}} \gamma^{(l)} + \beta^{(l)}, \tag{2}$$

where $\mu^{(l)}$ and $\sigma^{(l)}$ are the batch mean and variance, and $\gamma^{(l)}, \beta^{(l)}$ are learnable parameters. After activation

$$F_{\text{act}}^{(l)} = \text{ReLU}(\hat{F}^{(l)}), \tag{3}$$

a max-pooling layer reduces spatial dimensions:

$$F_{\text{pool}}^{(l)} = \text{MaxPool}(F_{\text{act}}^{(l)}, k_s, s), \tag{4}$$

where k_s is the kernel size and s the stride.

3.2. Contracting Path

The contracting path follows the encoder-style structure, where the main goal is to capture high-level semantic information as well as reduce the spatial dimensions of the input. At each stage, the spatial resolution decreases due to pooling, while the feature channels are doubled:

$$C_{l+1} = 2C_l, \quad H_{l+1} = \frac{H_l}{2}, \quad W_{l+1} = \frac{W_l}{2}. \tag{5}$$

This architecture is characterized by a gradual reduction in the input image into a smaller latent representation. The higher channels enable the network to represent finer hierarchies of the features, such as textures, edges, and region-level data, and the reduced spatial resolution promotes invariance to local deformations. The contracting path provides insight into the global context and develops strong discriminative elements.

3.3. Expansive Path

In contrast, the expansive path aims to reconstruct the spatial resolution and refine the localization of the features. It performs a symmetric decoding process, where each upsampling step doubles the resolution while halving the number of channels:

$$C_{l+1} = \frac{C_l}{2}, \quad H_{l+1} = 2H_l, \quad W_{l+1} = 2W_l. \tag{6}$$

To obtain larger feature maps, interpolation or transposed convolutional operations are used to upsample the features. Notably, skip connections from the contracting path are concatenated with the upsampled features to regain fine-grained details lost during down-sampling. By combining low-level spatial details with high-level semantics, the network maintains both localization accuracy and contextual understanding. Thus, the expansive path does not oppose the contracting path. it refines spatial information through the learned hierarchical features.

3.4. Cross-Module Attention

Cross-module attention enhances feature maps from FEMs using both channel and spatial attention. Let F be the input feature map of size $H \times W \times C$.

3.4.1. Channel Attention

Channel attention is designed to weigh the feature maps channel-wise, which allow the network to pay more attention to informative feature responses. This is done by consolidating both average pooling and maximum pooling over the entire channel, which

provides insights into the general distribution of activations as well as emphasizing salient features. A shared multi-layer perceptron (MLP) takes these pooled features as input to output the channel weights. Formally,

$$F_{\text{avg}} = \frac{1}{H \times W} \sum_{i=1}^{H} \sum_{j=1}^{W} F_{i,j,:}, \tag{7}$$

$$F_{\max} = \max_{i,j} F_{i,j,:}, \tag{8}$$

$$z = \sigma(W_2 \delta(W_1(F_{\text{avg}} + F_{\max}))), \tag{9}$$

where δ denotes the ReLU activation and σ denotes the sigmoid activation. The refined channel-attended feature representation is obtained by reweighting the original feature map:

$$F_{\text{ch}} = F \odot z. \tag{10}$$

3.4.2. Multi-Scale Feature Selection Module (MSFSM)

The objects in complex visual scenes, such as LCC regions are heterogeneous and widely vary in size, shape, and appearance. A single receptive field is often insufficient in order to capture these types of variations. We introduce a Multi-Scale Feature Selection Module (MSFSM) that captures informative features at various scales and fuses them to enhance the discriminative ability. This design preserves not only the fine-grained details captured by small kernels, but also larger contextual structures. By doing so, this design increases the robustness of the network under scaling variations and strengthens its capacity to distinguish land areas.

Formally, the MSFSM begins by splitting the input feature map into n subsets along the channel dimension, the feature map obtained from the Class Activation Map (CAM) module:

$$F_{\text{split}} = \text{Split}(F_{\text{CAM}}, n), \tag{11}$$

where n is the number of subsets of channels. All the subsets are then convolved with a set of kernels of size k_i that are intended to capture features at a specific scale:

$$F_i' = \text{Conv}_{k_i}(F_{\text{split},i}), \quad i = 1, \ldots, n. \tag{12}$$

Finally, the outputs from all scales are concatenated to form a unified multi-scale feature representation:

$$F_{\text{MSFSM}} = \text{Concat}(F_1', F_2', \ldots, F_n'). \tag{13}$$

The MSFSM combines complementary information across several receptive fields, which thereby enriches the feature space and improves the model's robustness in classification and detection. In our model deployment, the MSFSM splits the input feature map into a channel subset of 4, denoted as n. Conventional kernels of sizes 1×1, 3×3, 5×5, and 7×7 are applied to each subset, and thus, the network is able to extract features at different receptive fields. In the channel attention mechanism, a common multi-layer perceptron (MLP) is used with a reduction ratio of 16. The design minimizes the complexity of computation and still provides good channel-wise feature recalibration.

3.5. Contracting and Expanding Path Fusion

One of the main concepts of the proposed model is feature fusion between the contracting which is the encoder path, and expansive which s the decoder path. The contracting path offers abundant semantic content that has been extracted at successively coarser scales, and the expansive path restores spatial resolution in order to further refine localization.

The direct use of features from one path is not very effective at capturing fine detail or lacks semantic abstraction. To address this, a skip-connection approach, which combines information at both levels, was used in the proposed model. In particular, cross-module attention initially enhances features of the contracting path to maintain scale-conscious representations. The fused enriched features are then further combined with corresponding expansive path features to build a balanced formulation which maintains world semantics and fine-grained spatial detail.

Let F_l^{cont} be the output from the contracting path at level l and F_l^{exp} from the expansive path. The fused feature representation is given by

$$F_l^{\text{fused}} = F_l^{\text{exp}} + F_l^{\text{CAM}}. \tag{14}$$

The fused features are then upsampled to increase the spatial resolution, preparing them for subsequent layers of the expansive path:

$$F_l^{\text{up}} = \text{UpSample}(F_l^{\text{fused}}, 2). \tag{15}$$

This combined approach enables the network to effectively leverage the complementary strengths of both paths, which include semantic abstraction from the contracting path and spatial refinement from the expansive path, which results in more effective feature representations for the final classification.

3.6. Proposed Network Architecture

The proposed CALCNet network follows a contracting–expanding path structure with skip connections, which enable it to simultaneously capture high-level semantic features and preserve spatial localization. The input image of size $64 \times 64 \times C$ is initially fed into a series of four Feature Extraction Modules (FEMs). Each FEM comprises two convolutional layers with a kernel size of 3×3, which is followed by batch normalization and ReLU activation. The pooling layers reduce the spatial resolution by a factor of 2 at each stage, whereas the number of convolutional filters progressively doubles (16, 32, 64, 128) in order to extract increasingly abstract representations of the land cover types. The contracting path focuses on capturing discriminative features from the input data as well as gradually reduce the spatial dimensions.

After each FEM in the contracting path, the extracted feature maps are passed to the cross-module attention, which incorporates both channel and MSFS mechanisms. The channel attention module enhances the most informative features as well as suppresses irrelevant or noisy information, which ensure that the network focuses on key land cover structures whereas the MSFS further refines the features by splitting the attention-enhanced feature maps into multiple channel subsets. Each subset undergoes convolutional operations with different receptive fields, which capture multi-scale contextual information that is essential for distinguishing complex land cover patterns. The outputs of the MSFSM are concatenated along the channel dimension to generate a rich and diverse feature representation. In our implementation, the cross-module attention, including both channel attention and MSFSM, is applied after each Feature Extraction Module (FEM) in the contracting path. This results in four attention-enhanced stages, enabling progressive refinement of feature representations at different levels of abstraction.

The expansive path mirrors the contracting path in reverse, using upsampling layers to gradually restore spatial resolution while halving the number of channels at each stage. Skip connections fuse features from the corresponding contracting path FEMs, including the outputs from CAM, preserving fine-grained spatial details for pixel-level classification. Global max pooling aggregates the fused features across spatial dimensions, reducing

dimensionality while retaining the most critical information followed by a classification head with two fully connected layers. The final classification layer employs a SoftMax activation function to generate probability distributions over the land cover classes.

Although the proposed CALCNet is designed for scene classification, the inclusion of the expansive (decoder) path plays a crucial role in preserving and refining spatial information. The complexity of spatial structures, textures, and relationships between objects are critical for differentiating similar land cover types in remote sensing images. Skip connections with the decoder path help to combine high-level semantic features and low-level spatial information together, helping to form more discriminative feature graphs. To further investigate this design choice, we also evaluated an encoder-only version of the model lacking the expansive path. These results suggest that the removal of the decoder degrades classification performance, especially on datasets with spatial variability. This indicates that the decoder enhances feature representation, even in a classification setting.

For the training, we ran the network for 100 epochs using an Adam optimizer with a learning rate of 0.001 and a batch size of 64. A categorical cross-entropy loss function is used to optimize the pixel-level predictions. Some data augmentation techniques, such as random rotations, flipping, and scaling were used in order to improve model generalization. This combination of hierarchical feature extraction, attention-guided multi-scale feature selection, and contracting-expanding path fusion, enables CALCNet to achieve a robust and accurate land cover classification. A summary of the training hyperparameters used in this study is provided in Table 1.

Table 1. Training hyperparameters used for the CALCNet and baseline models.

Parameter	Value/Configuration
Optimizer	Adam
Initial learning rate	0.001
Learning rate schedule	Fixed (no decay)
Batch size	32
Number of epochs	100
Loss function	Categorical cross-entropy
Input image size	$64 \times 64 \times C$
Weight initialization	Default Keras initialization
Activation function	ReLU (hidden layers), SoftMax (output layer)
Data augmentation	Random rotation, horizontal/vertical flipping, scaling
Normalization	Pixel values scaled to [0, 1]
Random seed	42
Hardware	NVIDIA RTX 3090 Ti GPU, Intel i9-14900K CPU
Framework	Keras 2.10.0 with TensorFlow 2.10.0 backend

3.7. CALCNet Compression via Differential Evolution

For practical deployment of LCC, the proposed model remains efficient in terms of both computation and memory usage. Deep networks perform well, but they come with a high number of parameters that leads to inference latency and demand higher storage space, which are unsuitable for edge devices. We applied a compression strategy in order to mitigate this, which was based on a differential evolution(DE)-driven neuron pruning method.

Binary Encoding of Neurons

Consider a hidden layer with N_h neurons. We represent a pruning configuration as a binary vector $v \in \{0,1\}^{N_h}$, where

$$v_i = \begin{cases} 0 & \text{neuron } i \text{ is removed,} \\ 1 & \text{neuron } i \text{ is retained.} \end{cases} \tag{16}$$

Given three distinct candidate solutions v_a, v_b, v_c, a donor vector u is generated as

$$u = v_a + F \cdot (v_b - v_c), \tag{17}$$

where $F \in (0,1)$ is the mutation factor (we use $F = 0.45$). Each component is binarized as

$$u_i = \begin{cases} 1 & \text{if } u_i \geq \tau, \\ 0 & \text{otherwise.} \end{cases} \tag{18}$$

A trial vector z is constructed by mixing the donor u with a target solution v_t:

$$z_i = \begin{cases} u_i & \text{if } r_i < C_r, \\ (v_t)_i & \text{otherwise,} \end{cases} \tag{19}$$

where $r_i \sim \mathcal{U}(0,1)$ and C_r is the crossover rate, which is fixed at 0.7.

Each candidate is evaluated using a fitness function that balances compression and accuracy:

$$\mathcal{F}(z) = \alpha \cdot \left(1 - \frac{N_{\text{red}}}{N_{\text{orig}}}\right) + \beta \cdot \text{Acc}(z), \tag{20}$$

where N_{red} is the reduced neuron count, N_{orig} is the original count, $\text{Acc}(z)$ is the accuracy achieved under configuration z, and α, β are balancing weights.

The better candidate for the next iteration is chosen as

$$v_{t+1} = \begin{cases} z & \text{if } \mathcal{F}(z) \geq \mathcal{F}(v_t), \\ v_t & \text{otherwise.} \end{cases} \tag{21}$$

This pruning approach removes redundant neurons and minimizes parameters as well as preserve model accuracy. The compressed CALCNet consequentlt yields a higher efficiency (lower computational cost and smaller memory footprint) and a faster inference while outperforming conventional baseline models. The C-CALCNet is the lightweight variant of CALCNet, designed to retain all essential features of CALCNet while being more feasible for both real-time and resource-constrained deployments.

4. Experimental Setup

In this section, we describe the experimental setup including our dataset, evaluation metrics, and comparison against baselines. The experiments were performed on a workstation Intel(R) Core(TM) i9–14900K CPU at 3.20 GHz, NVIDIA GeForce RTX 3090 Ti GPU, and Samsung 990 PRO 2 TB SSD. The DL algorithm was implemented using Keras over TensorFlow.

4.1. Datasets

We tested our method CALCNet on four widely used benchmarked datasets, which included AID [28], UCMerced_LandUse [29], NWPU [30], and EuroSAT [31]. These diverse datasets were selected in order to exhibit complementary aspects of scene diversity,

scale variation, spatial resolution, and spectral content for a holistic assessment of the proposed model.

The AID dataset is a large-scale aerial scene classification benchmark, which provides 30 classes with large inter-class variation and a complex scene structure. AID is appropriate for the evaluation of the robustness of classification models under practical remote sensing situations. The UCMerced_LandUse dataset contains 21 land use classes, and each class has 100 images of 256 × 256 pixels. The dataset is one of the most popular aerial datasets due to its careful curation, diverse notation per class with a balanced sample set, and comprehensive coverage of representative aerial scene categories, which render it as an important benchmark for assessing classification accuracy under controlled settings.

The NWPU dataset is a more challenging large-scale benchmark that contains more scene categories and significant variations in the scales, viewpoints, background, and illumination conditions of objects. It also makes it especially convenient to evaluate the generalization power of DL models in a complex and diverse space. In contrast, EuroSAT is based on Sentinel-2 satellite imagery and contains 10 classes with approximately 3000 images per class. Compared with the aerial image datasets, EuroSAT provides richer spectral–spatial information and is therefore useful for evaluating the ability of the model to learn discriminative representations from satellite-based land cover imagery.

These datasets share the common goal of land cover and land use scene classification, and they also present different strengths and limitations. AID and NWPU provide higher scene diversity and more challenging class variability, but they also introduce stronger inter-class similarity and intra-class complexity. UCMerced_LandUse is balanced and carefully curated, which is beneficial for standard benchmarking, but its relatively smaller size limits the diversity of real-world variations. EuroSAT benefits from multispectral satellite data and a relatively large number of samples per class, but it includes fewer categories than AID and NWPU. Therefore, each dataset captures different aspects of the land cover classification problem. A summary of the key characteristics of these datasets is presented in Table 2.

Table 2. Summary and comparison of benchmark datasets used in this study.

Dataset	Type	Classes	Images	Format	Strengths/Limitations
AID	Aerial imagery	30	10,000	JPG	The dataset is large, with a high degree of intra-class variability and complex scene composition, which makes it suitable for evaluating robustness and generalization. However, inter-class similarity is relatively high which complicates the classification.
UCMerced-LandUse	Aerial imagery	21	2100	TIFF	A well-curated and balanced benchmark dataset with clear scene categories that is commonly used for classification evaluation. Its major drawback is the limited dataset size, which could limit diversity and with it real-world variability.
NWPU-RESISC45	Aerial imagery	45	31,500	JPG	Massive and difficult dataset with significant variations in scale and viewpoint, in addition to background complexity and illumination. This makes it particularly useful for testing generalization, but the large volume of classes leads to increased confusion between classes.
EuroSAT	Satellite imagery	10	27,000	JPG	This is a dataset of satellite images from the Sentinel-1 satellite between 2016 and August 2020, and it can be utilized for spectral–spatial representation learning evaluations. It contains a high number of samples per class, which slowly impacts the diversity of scenes that outstrips AID and NWPU with its minimalistic category count.

We noticed in our experiments that each individual dataset can lack either enough categories or enough diversity, which limits the generalization of trained models in "real-world" conditions. Assessing CALCNet on these supplementary datasets allows us to perform a more rigorous evaluation of its robustness, classification performance, and generalizability across diverse remote sensing scenarios. A unified data splitting strategy was adopted across all experiments, with 70%, 20%, and 10% of the data used for training, validation, and testing, respectively. To ensure reproducibility, a fixed random seed (42) was used during dataset shuffling and splitting. Furthermore, care was taken to avoid any data leakage, ensuring that no overlapping images or scenes were present across the training, validation, and testing sets. Although alternative protocols such as 20% or 50% training splits are commonly used for datasets like AID and NWPU-RESISC45, we adopted a unified splitting strategy across all datasets to maintain consistency and enable fair comparative evaluation. Sample images from these datasets are shown in Figure 2.

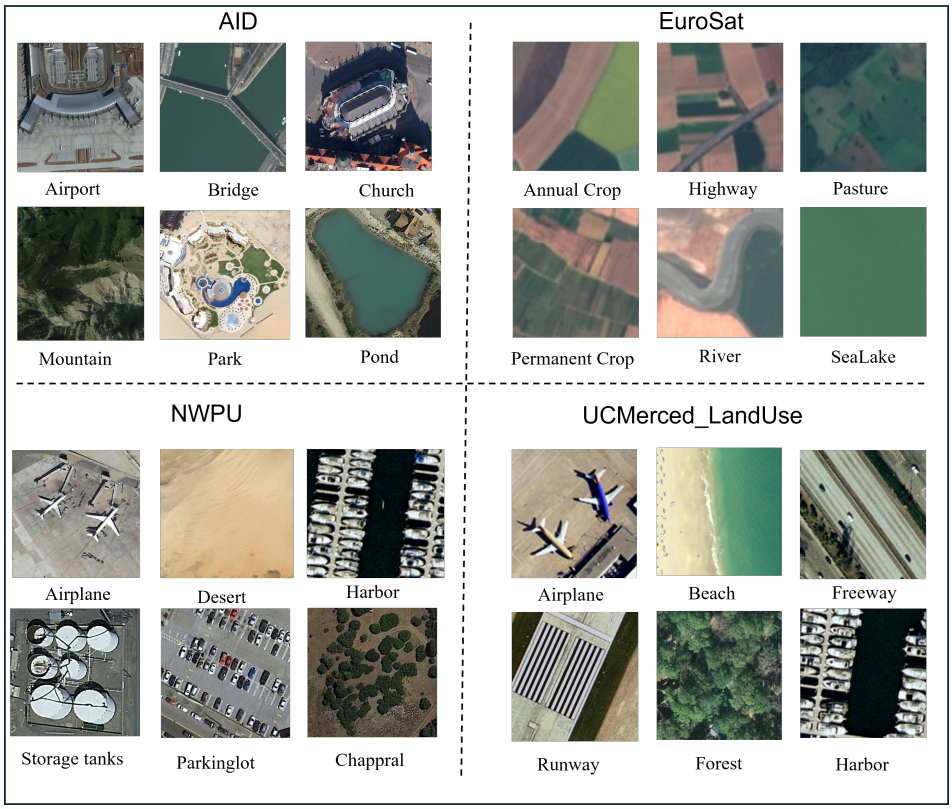

Figure 2. Sample images of different datasets.

4.2. Data Preprocessing

All input images from the AID, UCMerced_LandUse, NWPU-RESISC45, and EuroSAT datasets were resized to a fixed resolution of 64 × 64 pixels to ensure a consistent input size for the network. For the EuroSAT dataset, only the RGB channels were used instead of the full 13-band multispectral data, allowing uniform input representation across all datasets. Images were all normalized to the range of [0, 1] and then trained. All the models (including the proposed (CALCNet) and the other baseline architectures) had the same preprocessing pipeline, including resizing, channel selection, and normalization. This helps make a fair comparison by eliminating differences caused by input differences.

4.3. Evaluation Metrics

We used several evaluation metrics to quantitatively evaluate the proposed CALCNet model's performance in land cover classification, including precision (P), recall (p),

F1-score (F1), and accuracy (Acc). These measures give a precise understanding of the model in the true classification of land cover classes as well as false positives and false negatives. Precision is the ratio of correctly predicted positive samples over all samples predicted to be positive:

$$\text{Precision} = \frac{\text{True Positives}}{\text{True Positives} + \text{False Positives}}, \tag{22}$$

Recall measures the proportion of correctly predicted positive samples among all actual positive samples:

$$\text{Recall} = \frac{\text{True Positives}}{\text{True Positives} + \text{False Negatives}} \tag{23}$$

The F1-score provides a harmonic mean of precision and recall, balancing both metrics:

$$\text{F1-score} = 2 \cdot \frac{\text{Precision} \cdot \text{Recall}}{\text{Precision} + \text{Recall}} \tag{24}$$

Accuracy indicates the proportion of correctly classified samples over the total number of samples:

$$\text{Accuracy} = \frac{\text{True Positives} + \text{True Negatives}}{\text{True Positives} + \text{True Negatives} + \text{False Positives} + \text{False Negatives}} \tag{25}$$

4.4. Ablation Study

We performed a highly thorough ablation study in order to thoroughly assess the contribution of each architectural component in the proposed CALCNet. The experiments are constructed to vary the influence of several essential components such as spatial attention and multi-scale feature extraction. More specifically, we evaluated four model variants that include (1) the base network with no additional enhancement modules, (2) the base network plus spatial attention, (3) the base network plus multi-scale feature selection, and (4) the full CALCNet model combining both spatial attention and feature selection modules. Moreover, several advanced CNNs, including EfficientNetB3, ResNet50, ResNet101, InceptionV3, DenseNet121, VGG16, Xception, and MobileNetV2, were used as baseline models to enable a comprehensive comparative analysis. The ablation studies were performed over four commonly used RS scene classification datasets, namely AID, EuroSAT, NWPU, and UCMerced LandUse. All evaluations were performed under identical training configurations to maintain fairness, and the results are summarized in Tables 3 and 4.

The outcomes clearly indicate the efficiency of the suggested architectural improvements. Across all datasets, the base CALCNet model outperforms traditional CNN architectures, demonstrating the effectiveness of its underlying feature extraction backbone. The discriminative power is further enhanced by introducing the spatial attention mechanism that adaptively focuses on the relevant spatial areas in input imagery, leading to increased precision and recall. On the same note, the multi-scale feature selection block is added to enhance our understanding of the context by incorporating hierarchical information across multiple receptive fields, enabling the network to learn both small-scale and large-scale patterns of the scene.

Table 3. Ablation study and comparison of various CNN models and proposed CALCNet variants for AID and EuroSAT datasets.

Method	P	R	F1	ACC
AID Dataset				
EfficientNetB3	93.12	93.07	93.09	95.14
ResNet50	91.05	91.12	91.08	93.21
ResNet101	92.04	92.18	92.10	94.12
InceptionV3	90.13	90.05	90.09	92.23
DenseNet121	91.01	91.09	91.04	93.18
Xception	92.08	92.03	92.05	94.09
MobileNetV2	89.15	89.09	89.12	91.14
VGG16	88.12	88.09	88.11	94.10
Base CALCNet	94.08	94.13	94.10	96.18
Base + Spatial Attention	95.12	95.09	95.10	97.16
Base + Feature Selection	95.09	95.13	95.11	97.11
CALCNet	**96.07**	**98.06**	**97.04**	**98.09**
EuroSAT Dataset				
EfficientNetB3	96.12	96.07	96.09	97.15
ResNet50	95.09	95.13	95.11	96.14
ResNet101	96.07	96.13	96.10	97.17
InceptionV3	95.08	95.11	95.09	96.13
DenseNet121	95.12	95.08	96.10	96.09
Xception	96.04	96.07	96.05	97.12
MobileNetV2	94.09	94.11	94.10	95.14
VGG16	93.11	93.07	93.09	94.12
Base CALCNet	97.06	97.09	97.07	98.13
Base + Spatial Attention	98.03	98.09	98.05	98.14
Base + Feature Selection	98.08	98.12	98.10	98.09
CALCNet	**99.07**	**99.11**	**99.09**	**99.19**

Table 4. Ablation study and comparison of various CNN models and proposed CALCNet variants for NWPU and UCMerced LandUse datasets.

Method	P	R	F1	ACC
NWPU Dataset				
EfficientNetB3	88.07	87.12	87.59	89.14
ResNet50	85.09	84.13	84.60	86.18
ResNet101	86.11	86.08	86.09	87.17
InceptionV3	84.09	83.12	83.60	85.15
DenseNet121	91.12	92.14	91.62	91.17
Xception	86.15	86.09	86.11	87.13
MobileNetV2	83.13	82.11	82.62	84.18
VGG16	82.08	81.12	81.59	83.16
Base CALCNet	89.12	88.09	88.60	90.19
Base + Spatial Attention	96.11	96.08	96.09	95.16
Base + Feature Selection	96.07	96.13	96.10	96.17
CALCNet	**98.10**	**98.44**	**98.26**	**98.53**
UCMerced LandUse Dataset				
EfficientNetB3	96.08	96.13	96.10	97.12
ResNet50	94.11	94.08	94.09	95.09
ResNet101	95.09	95.13	95.11	96.15
InceptionV3	94.13	94.08	94.10	95.11
DenseNet121	95.07	95.12	95.09	96.09
Xception	95.08	95.09	95.08	96.14
MobileNetV2	93.09	93.11	93.10	94.12
VGG16	92.07	92.13	92.09	93.15
Base CALCNet	96.09	96.11	96.10	97.12
Base + Spatial Attention	97.07	97.09	97.08	98.11
Base + Feature Selection	97.11	97.08	97.09	98.13
CALCNet	**99.09**	**98.96**	**99.02**	**99.47**

The proposed model that integrates spatial attention and multi-scale features selection obtained a higher performance in terms of all evaluation metrics over the targeted datasets. The results highlight the complementary roles of the two modules. Spatial attention im-

proves local discriminability, whereas feature selection enhances global robustness and cross-scale representation. Notable improvements on complex datasets such as NWPU and AID demonstrate that CALCNet reduces feature redundancy and overfitting as well as maintains strong generalization. The training and validation accuracy and loss curves further confirm the model's stability and robustness, which are shown in Figure 3.

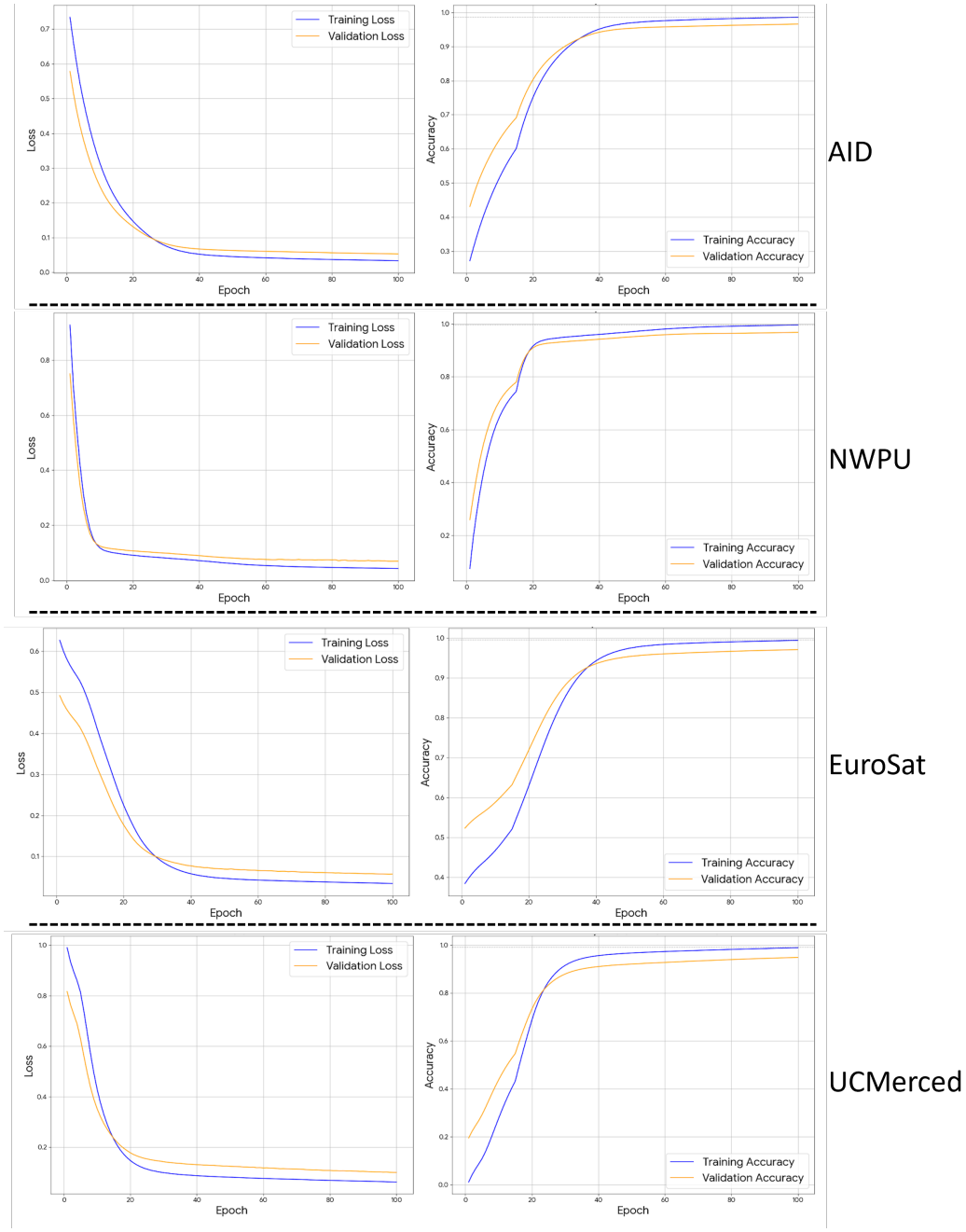

Figure 3. Training and validation accuracy and loss of the CALCNet for all the datasets.

CALCNet exhibits smooth and consistent convergence behavior with minimal oscillations, reflecting effective feature learning and well-regularized optimization. These findings collectively demonstrate that each component of CALCNet plays a vital role in enhancing classification performance, and their integration results in a powerful, efficient, and generalizable model for large-scale remote sensing applications.

The proposed attention module and MSFSM have a relatively small overhead in terms of computational complexity. The increment in the number of additional parameters and

floating point operations (FLOPs) is projected to be about 3–5 percent more than the base architecture. This is a lightweight design that makes sure that the enhanced performance of CALCNet is realized without the need to heavily generalize the computational load.

5. Comparison with State-of-the-Art Methods

To further support the effectiveness and generalization capacity of the proposed CALCNet, we have compared its results with a variety of state-of-the-art (SOTA) methods on four benchmark datasets. The comparative analysis shows that CALCNet has been performing better than traditional CNN-based and Transformer-based architectures, demonstrating that it is better at extracting discriminative, spatially significant features. This improvement in performance can be explained by CALCNet's integration of spatial attention and multi-scale feature selection mechanisms, which enable the system to dynamically highlight the most informative regions whilst maintaining global contextual relationships. In contrast to traditional models, which use fixed convolutional filters, CALCNet is an adaptive modulator of feature activations at scale and can be more robust to changes in texture, illumination, and spatial resolution.

UCMerced LandUse: Previously used CNNs such as ResNet or InceptionV3 could not perform well on this dataset, with a range of accuracy of 92% to 94%, which are shown in Table 5. Models that were more dense such as DenseNet121 and such hybrid models as OG-WOA–Bi-LSTM were relatively more successful, with an approximate accuracy of 98.99 percent. Nevertheless, these models continued to be susceptible to poor generalization across similar classes in appearance. The proposed CALCNet achieved the best accuracy of **99.47%**, with precision, recall, and F1-score of 99.09%, 98.96%, and 99.02%, respectively. These scores are a relative improvement of about 1–3% percent over the second-best model, which verifies that the dual attention and feature selection approach of CALCNet greatly increases the separability of the classes and minimizes confusion among similar patterns of land use.

Table 5. Comparison of SOTA methods and the proposed CALCNet on the UCMerced LandUse dataset.

Method	ACC (%)	P (%)	R (%)	F1 (%)
ResNet [32]	92.00	88.00	86.00	86.00
Inception V3 [33]	94.00	88.00	87.00	87.00
Hybrid model [34]	98.00	96.00	96.00	96.00
ViT [35]	98.00	–	–	–
DAS-RHDIS [36]	56.00	65.00	70.00	67.00
DenseNet121 [37]	98.00	98.00	98.00	98.00
AlexNet [38]	94.00	–	–	–
GoogLeNet [28]	97.00	–	–	–
Transformer based [39]	96.00	92.00	90.00	94.00
Inception V3 [10]	92.00	93.00	92.00	92.00
KL [40]	98.06	–	–	–
GCN [41]	89.00	–	–	–
EfficientNet-B3-Attn-2 [42]	98.00	–	–	–
OG-WOA–Bi-LSTM [43]	97.09	–	–	–
DenseNet201 [44]	94.00	95.00	94.00	94.00
TEX-Net [45]	97.00	–	–	–
Proposed CALCNet	**99.47**	**99.09**	**98.96**	**99.02**

Moreover, Figure 4 indicates the classification performance of the proposed model in various categories of land use. The average precision, recall, and F1-score of the model were 99.09%, 98.96%, 99.02%, which demonstrated a high level of accuracy and reliability of the model in terms of classification. Based on the analysis by classes, it was possible to find

several categories, airplane, beach, and tennis court, which obtained a perfect classification result of 100 percent precision, recall, and F1-score, indicating that the model is capable of successfully defining specific scene patterns. Likewise, classes such as harbor, runway, and storage tanks achieved very high F1-scores of over 99.8 per cent, indicating an excellent performance of the model in identifying well-structured objects in aerial imagery. Other classes like agricultural, parkinglot, overpass and chaparral scored F1-scores of above 99 percent, and residential classes like dense residential, medium residential, mobile home park, and sparse residential scored high with above 98 percent F1-scores. Forest and golf course slightly underperformed, and this could be because of similar patterns of vegetation and complicated textures. In general, the findings indicate that the proposed model offers strong, consistent classification performance across various aerial scene categories.

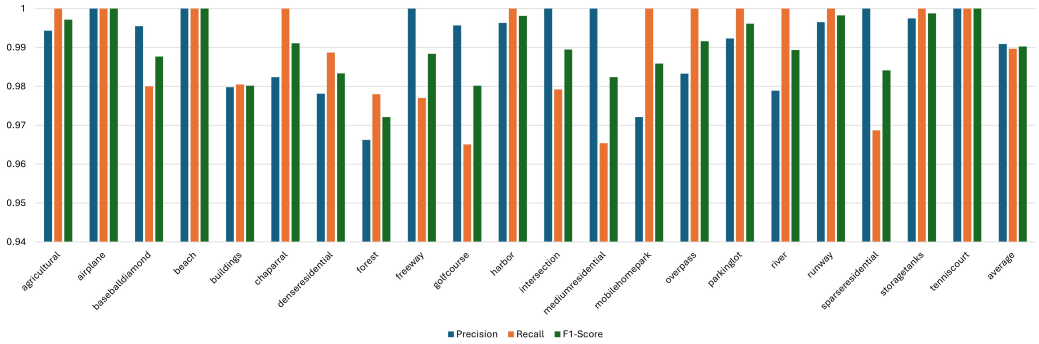

Figure 4. Classification report of the proposed model using UCMerced LandUse dataset.

AID: The AID dataset is more difficult due to increased intra-class variability and complex spatial layouts. Table 6 illustrates that more complex hybrid models, such as DNNE-SWA and OG-WOA -Bi-LSTM are able to achieve 97% accuracy because of increased generalization (ACC = 89%) in earlier DL models such as SceneNet and DenseNet201. CALCNet surpassed these methods by achieving a **98.09%** accuracy, along with precision = 96.07%, recall = 98.06%, and F1-score = 97.04%. This enhancement demonstrates that CALCNet can learn both global appearance semantics and localized discriminative features, which can be used to classify images accurately, even in visually ambiguous categories such as residential/dense residential or industrial/storage tanks.

Table 6. Comparative analysis of the proposed method with baselines over the AID dataset.

Method	ACC (%)	P (%)	R (%)	F1 (%)
DNNE-SWA [46]	97.00	97.00	97.00	97.00
CNN [47]	96.00	97.00	97.00	97.00
TEX-Net [45]	95.00	–	–	–
CaffeNet [28]	89.00	–	–	–
KL [40]	96.00	–	–	–
GCN [41]	95.00	–	–	–
ViT [35]	94.00	–	–	–
EfficientNet-B3-Attn-2 [42]	96.00	–	–	–
SceneNet [48]	89.00	–	–	–
OG-WOA–Bi-LSTM [43]	97.00	–	–	–
DenseNet201 [44]	89.00	90.00	89.00	89.00
Proposed CALCNet	**98.09**	**96.07**	**98.06**	**97.04**

Figure 5 shows the classification results of the proposed model in the 30 scene categories. The model was found to have an average precision of 96.07%, recall of 98.06% and an F1-score of 97.04%, which means that it can classify various remote sensing scenes with high and consistent precision. Based on the analysis by classes, a few class categories like

BareLand, Desert, Farmland, River, and Port also performed remarkably well with an F1-score of over 98%, indicating the effectiveness of the model in classifying the unique natural and geographical features. Correspondingly, other classes such as Airport, DenseResidential, Parking, RailwayStation, and SparseResidential also performed well with an F1-score of over 97%, indicating the capability of the model to predict urban infrastructure and residential patterns. Moderate scores were noticed under the categories like Industrial, Pond, School, and Playground with F1-scores of about 95–96%, meaning that there was slight confusion between man-made structures, which were visually similar. The F1-scores of Church, Commercial, Resort, and Square were relatively lower and can be explained by structural similarities and complicated spatial layouts on urban scenes. The findings as a whole indicate that the given model will yield a strong, and stable classification performance across natural and urban land use types, which makes it effective for large-scale remote sensing scene classification problems.

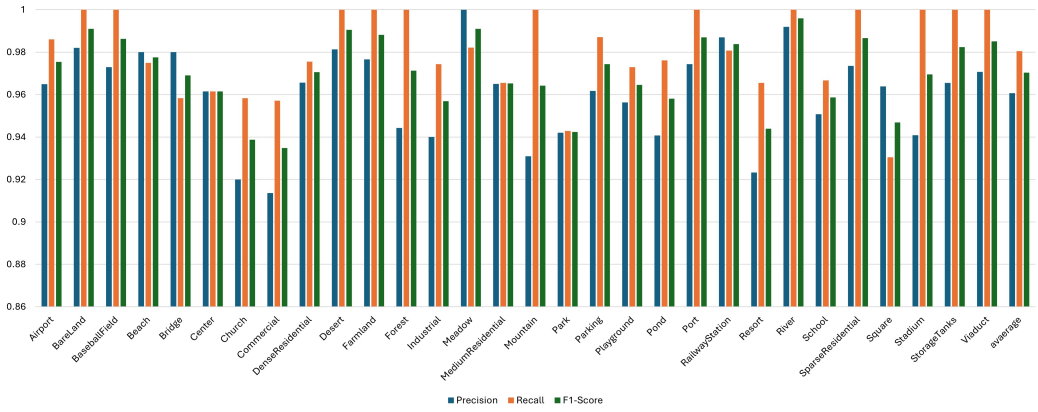

Figure 5. Classification report of the proposed model using the AID dataset.

EuroSAT: The EuroSAT dataset includes Sentinel-2 imagery with varying spectral and spatial resolutions. Table 7, shows that the earlier approaches, such as self-attention-fused CNNs and Scale-Invariant Feature Transform (SIFT)-based models achieved relatively low accuracies around 89–92%, which was primarily due to the limited cross-channel interaction and contextual reasoning. Transformer-based models and deep transfer learning approaches achieved up to 98% accuracy. The proposed CALCNet further advanced performance to **99.19% accuracy**, with a precision of 99.07%, recall of 99.11%, and F1-score of 99.09%. The model's superior generalization across heterogeneous land cover types confirms its effectiveness in multispectral scene interpretation and robust spectral–spatial feature encoding.

Table 7. Comparison on the EuroSAT dataset.

Method	ACC (%)	P (%)	R (%)	F1 (%)
InceptionV3 [37]	98.00	98.00	98.00	98.00
TinyViT [49]	96.00	–	–	–
ViT [50]	95.00	–	–	–
Self-attention-fused CNN [51]	89.00	–	–	–
SIFT-based with CNN [52]	92.00	–	–	–
ResNet152+ViT [53]	94.00	–	–	–
Multi-branch DL [54]	96.00	–	–	–
BiLSTM UNet [55]	97.00	–	–	–
EfficientNetV2B0+ResNet152 [32]	97.00	–	–	–
PSO+CNN+SVM [56]	95.00	–	–	–
Deep transfer learning [57]	98.00	–	–	–
Transformer-based [39]	98.00	98.00	98.00	98.00
Proposed CALCNet	**99.19**	**99.07**	**99.11**	**99.09**

Figure 6 illustrates the classification performance of the proposed model across the EuroSAT scene categories. The model achieved an average precision of 99.07%, a recall of 99.11%, and an F1-score of 99.09%, which indicates a highly accurate and reliable classification performance. Several categories, such as Industrial Buildings, Residential Buildings, Forest, and Sea & Lake achieved extremely high performance with F1-scores above 99.5%, from the class-wise analysis, which demonstrate the model's strong capability in in order to recognize well-defined land cover patterns. Similarly, classes including Highway, Permanent Crop, River, and Herbaceous Vegetation also showed strong performance with F1-scores close to 99%, highlighting the model's effectiveness in distinguishing different natural and man-made environments. Other categories such as AnnualCrop and Pasture achieved slightly lower yet strong performance with F1-scores above 98%, which may be attributed to similarities in vegetation textures and seasonal variations between agricultural land types. Overall, the results demonstrate that the proposed model provides robust and consistent classification performance across diverse land cover categories, confirming its effectiveness for remote sensing scene classification tasks.

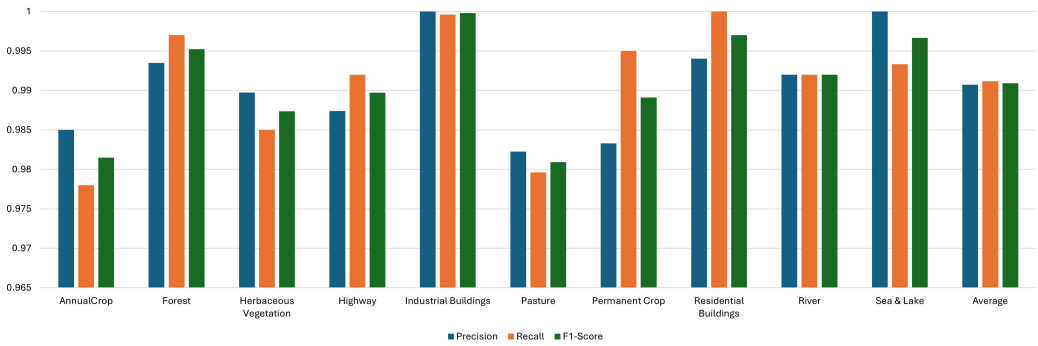

Figure 6. Classification report of the proposed model using the EuroSAT dataset.

NWPU: The NWPU-RESISC45 dataset poses a higher degree of complexity due to the large number of classes and varying intra-class spatial scales. Table 8 shows that conventional networks such as GCN and ViT struggled with large-scale generalization (ACC $\leq$ 93%), whereas DenseNet121 and DNNE-SWA achieved accuracies of up to 97–98%. The proposed CALCNet achieved a remarkable **98.53% accuracy**, with precision of 98.10%, recall of 98.44%, and an F1-score of 98.26%, and it estalished a new state-of-the-art performance on this challenging dataset. The superior results indicate CALCNet's ability to model complex spatial dependencies and maintain discriminative capacity across wide geographical diversity. Figure 7 depicts the classification performance of the proposed model across 45 remote sensing scene categories. The model achieved an average precision of 98.10%, a recall of 98.44%, and an F1-score of 98.26%, which demonstrates strong and consistent classification capability across diverse aerial scenes. Several categories such as cloud and harbor achieved perfect classification performances with 100% precision, recall, and F1-score from the class-wise analysis, which indicates that these classes contain highly distinctive visual patterns. Similarly, classes including chaparral, island, runway, sea-ice, and thermal-power-station achieved F1-scores above 99%, highlighting the model's effectiveness in identifying both natural landscapes and structured infrastructure. Many other categories such as airplane, beach, bridge, circular-farmland, parking-lot, and tennis-court also achieved strong performance with F1-scores above 98%, demonstrating the model's robustness in recognizing diverse land use and object categories. Moderate performance was observed for church, commercial-area, industrial-area, and medium-residential, with F1-scores around 95–97%, which may be attributed to structural similarities and complex spatial layouts within urban environments. Overall, the results indicate that the proposed

model provides robust and reliable classification performance across a wide range of remote sensing scene categories, effectively capturing both natural and man-made features in aerial imagery.

The balanced precision–recall trade-off and accuracy gains (2–4%) confirm that there is a high level of generalization and resistance of the model to class imbalance. This discussion affirms that our method, which relies on CALCNet, establishes a new state of the art in remote sensing scene classification by effectively combining attention-based spatial refinement with multi-scale contextual learning, providing both the highest performance and interpretability for real-world remote sensing tasks. The performance of some baseline models was similar across a few datasets, but CALCNet demonstrates better stability across the entire set of benchmarks. This means that there are better generalization abilities in the presence of varying data complexities and distributions within a scene. CALCNet also offers a better trade-off between accuracy, recall, and computational efficiency, which is more practical in real-world deployment settings.

Table 8. Comparison on the NWPU dataset.

Method	ACC (%)	P (%)	R (%)	F1 (%)
DenseNet121 [37]	98.00	98.00	98.00	98.00
DNNE-SWA [46]	97.00	97.00	97.00	97.00
Self-attention fused CNN [51]	92.00	–	–	–
KL [40]	94.00	–	–	–
WNN [58]	92.00	91.00	91.00	91.00
ViT [35]	93.00	–	–	–
GCN [41]	89.00	–	–	–
SceneNet [48]	95.00	–	–	–
OG-WOA–Bi-LSTM [43]	97.00	–	–	–
BestC [59]	95.00	–	–	95.00
DenseNet201 [44]	93.00	93.00	93.00	93.00
Proposed CALCNet	**98.53**	**98.10**	**98.44**	**98.26**

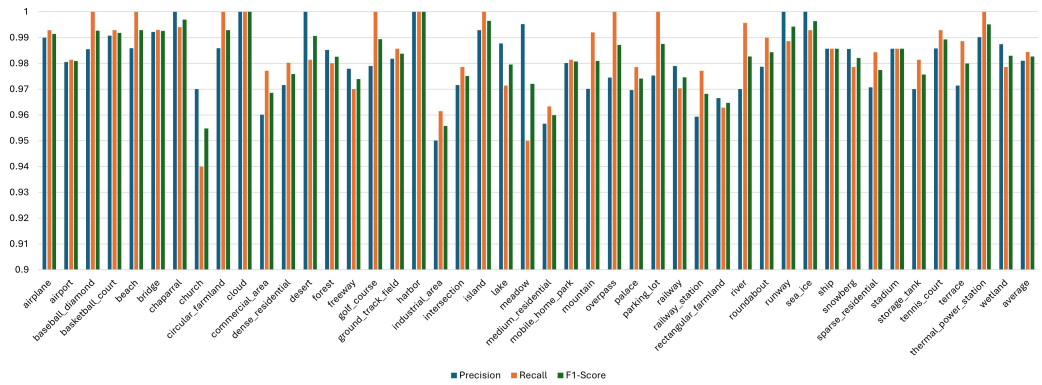

Figure 7. Classification report of the proposed model using the NWPU RESISC45 dataset.

5.1. Model Compression Analysis

We conducted a thorough model compression analysis using neuron pruning in order to further assess the robustness, scalability, and deployment efficiency of the proposed CALCNet. In this experiment, we used neuron pruning with compression ratios of 0% (the uncompressed model) to 50%. After each compression step, redundant neurons that contributed less to overall network activation were pruned, and the resulting models were fine-tuned to restore representational balance. Table 9 presents the results of compression, showing the changes in precision (P), recall (R), and accuracy (ACC) at different compression levels.

Table 9. Performance of CALCNet under different neuron compression ratios for AID, EuroSAT, NWPU, and UCMerced LandUse datasets. P = Precision, and R = recall, ACC = accuracy.

Dataset	Compression (%)	P (%)	R (%)	ACC (%)
AID	0	96.07	98.06	98.09
	10	95.80	95.85	97.95
	20	95.40	95.45	97.70
	30	94.90	94.95	97.35
	40	94.10	94.05	96.85
	50	93.50	93.45	96.10
EuroSAT	0	99.07	99.11	99.19
	10	98.80	98.85	98.90
	20	98.40	98.45	98.60
	30	98.00	98.05	98.20
	40	97.50	97.55	97.80
	50	97.00	97.05	97.20
NWPU	0	98.10	98.44	98.53
	10	98.04	98.35	98.40
	20	97.80	97.75	98.10
	30	96.90	96.85	97.60
	40	95.70	95.65	96.90
	50	94.80	94.75	96.10
UCMerced LandUse	0	99.09	98.96	99.47
	10	97.80	97.85	99.10
	20	97.40	97.45	98.70
	30	96.90	96.95	98.20
	40	96.20	96.25	97.70
	50	95.70	95.75	97.10

From the results, it is evident that CALCNet demonstrates strong resilience to compression. At 10% and 20% pruning ratios, the performance degradation is minimal across all datasets, with accuracy drops typically below 0.3%. This confirms that the network architecture contains a certain degree of parameter redundancy, allowing moderate pruning without affecting the discriminative power of the model. Even at 30% compression, CALCNet sustains competitive accuracy, maintaining above 97% on AID and EuroSAT, and above 98% on UCMerced LandUse. The ability to preserve such performance with significantly fewer parameters suggests that the proposed cross-attentional feature learning mechanism successfully captures essential spatial–spectral correlations within a compact representation space. However, at higher compression ratios of 40% and 50%, a more noticeable decline in performance is observed. The reduction in precision and recall becomes more prominent, particularly in datasets with higher inter-class similarity such as NWPU and AID. Nevertheless, even at 50% compression, CALCNet retains satisfactory accuracy levels above 96% for AID and EuroSAT, and around 97% for UCMerced LandUse, indicating the robustness of its learned feature representations. This trend implies that although aggressive pruning removes some fine-grained discriminative neurons, the network still maintains its global scene understanding capability through multi-level attention integration.

5.2. Visual Analysis

In Figure 8, we demonstrate a comparative visual analysis of the ensemble model, DLAN, CALCNet (Original), and CALCNet (Compressed), using attention map activations. The ensemble model from the baseline gives us broad and diffuse activations with low localization accuracy, which suggest an overall lack of ability to capture meaningful

structures in the land cover data. In practice, DLAN generates only a marginal increase in the quality of the (sampled) activation maps but is still inconsistent and insufficient for detailed structural representation. The CALCNet (Original) and CALCNet (Compressed) versions show clearer and more concentrated activations compared to the earlier models, demonstrating the effectiveness of CALCNet in refining feature emphasis. However, even CALCNet still fails to fully resolve complex contours and multi-scale patterns present in the scene. These visual results qualitatively confirm that CALCNet (Original) and CALCNet (Compressed) outperform all baseline variants, aligning with their superior quantitative performance across remote sensing datasets.

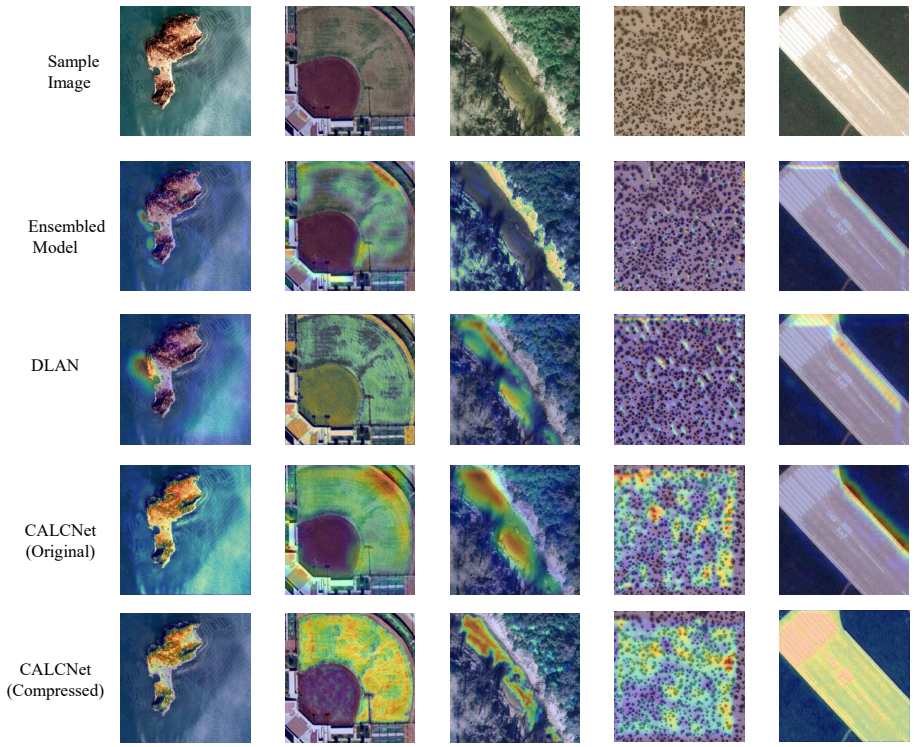

Figure 8. Visual results of CALCNet and other models. CALCNet achieves higher accuracy than other models.

5.3. Statistical Analysis Test

Table 10 illustrates the performance of the proposed model across five independent experimental runs on four benchmark remote sensing datasets, which include AID, EuroSAT, NWPU, and UCMerced LandUse. The proposed model achieved consistently high performance across all datasets. The model obtained an average accuracy of 98.09% for the AID dataset, and the best accuracy reached 98.76%. The mean accuracy with standard deviation was 98.09 ± 0.48 which indcates, stable performance with minimal variation across the five runs. The model achieved an average accuracy of 99.19% for the EuroSAT dataset, and the highest recorded accuracy was 99.31%. The mean ± SD value of 99.19 ± 0.10 demonstrates an extremely consistent performance and very low variability between experimental runs. Similarly, on the NWPU dataset, the proposed approach achieved an average accuracy of 98.53%, while the best performance reached 99.19%. The standard deviation of 1.17 indicates slightly higher variability compared to the other datasets, but the results still confirm strong and reliable classification capability. For the UCMerced LandUse dataset, the proposed model produced an average accuracy of 99.47%, with a peak accuracy of 99.62% across the five runs. The mean ± SD value of 99.47 ± 0.11 reflects a highly stable and consistent classification performance.

Table 10. Statistical analysis test with different runs.

Experiment	AID	EuroSAT	NWPU	UCMerced LandUse
1	97.35	99.13	96.50	99.61
2	97.92	99.06	98.80	99.29
3	98.44	99.29	99.19	99.38
4	98.76	99.31	99.00	99.49
5	97.98	99.16	99.16	99.56
Average	98.09	99.19	98.53	99.48
Best	98.76	99.31	99.19	99.62
Mean ± SD	**98.09 ± 0.48**	**99.19 ± 0.10**	**98.53 ± 1.17**	**99.47 ± 0.11**

5.4. Time Complexity

We compare the computational efficiency of the proposed CALCNet in this section, which include both in its original and compressed forms, with a number of state-of-the-art (SOTA) models. The model size and the overall number of model parameters are the determinants to inference time. Table 11 shows a comparative study of various models on the basis of parameters (millions), model size (MB), latency (seconds), and accuracy (%). ResNet-101 and VGG16 have high parameter counts (44 M and 138 M), and the larger model sizes, which translate to inference times of 90 seconds and 93 seconds, respectively, and moderate accuracy. The latency of lightweight models, such as MobileNet-V2 and EfficientNet-B0 is lower, but their accuracy decreases compared to the accuracy of models based on ensembles and transformers. The original CALCNet has a high precision 99%, a moderate latency of 72 seconds, and a model size 85 MB that can be handled. Notably, the compressed form of CALCNet can substantially reduce latency to 41 s and model size to 43 MB, while still achieving a high accuracy of 98%, indicating a good trade-off between model size, computational efficiency, and performance. This discussion points out that CALCNet provides a trade-off between practical implementation and high accuracy, and the compressed implementation further increases its usefulness in resource-limited systems.

Table 11. Comparison of computational time and model efficiency of various SOTA models with original and compressed CALCNet.

Methods	Parameters	Size	Latency
ResNet-101 [60]	44	171	90
Xception [61]	22	88	90
Inception-V3 [62]	23.9	92	67
MobileNet-V2 [63]	3.5	14	87
DenseNet-121 [64]	20.2	80	92
VGG16 [65]	138	528	93
Deep Ensembled Model [34]	64.7	247	32
DVIT [66]	6.6	–	–
Swin Transformer [67]	4.6	–	–
PlantXViT [68]	4.6	–	–
DLAN [69]	23	84	71
CALCNet (Original)	**23**	**85**	**72**
CALCNet (Compressed)	**12**	**43**	**41**

Furthermore, Figure 9 shows the proposed model's computational complexity in terms of Flops, model size, and frames per second. The original CALCNet requires 150.53 M FLOPs with a model size of 85 MB. After compression, the CALCNet-Compressed model significantly reduces the computational cost to 78.55 M FLOPs and decreases the model size to 43 MB, resulting in approximately 45–50% reduction in computational operations and model storage. In terms of inference performance, the compressed model shows an improved processing speed on both hardware platforms. The CPU inference speed increases

from 25.45 to 38.32, while the GPU performance improves from 75.4 to 118.3, demonstrating the efficiency of the compressed architecture. These improvements indicate that the compressed model reduces computational overhead and accelerates inference without significantly affecting the model's effectiveness. Overall, the compressed CALCNet-Compressed model provides a favorable balance between model efficiency and computational performance, making it more suitable for deployment in real-world applications with limited computational resources.

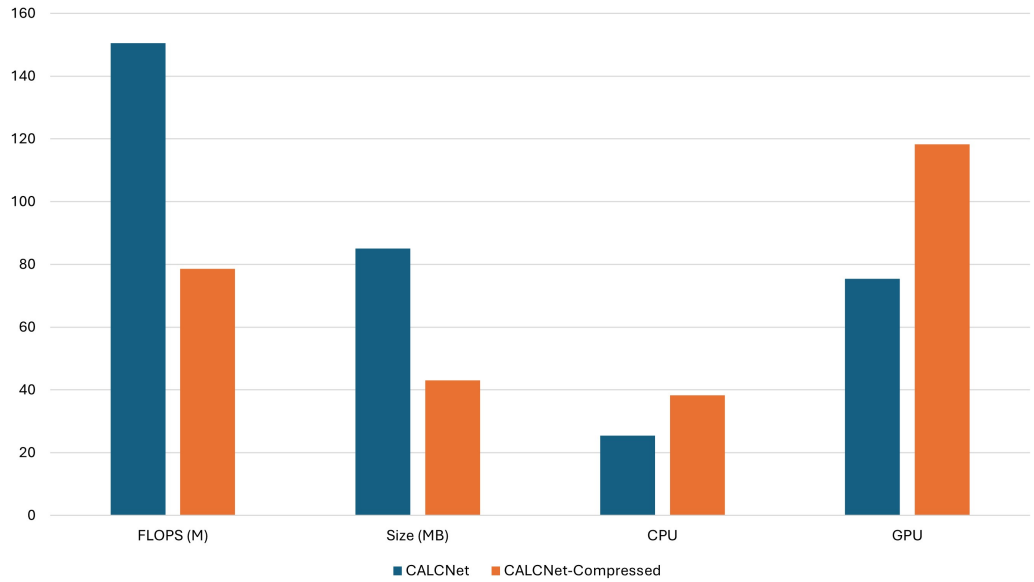

Figure 9. The proposed model's computational complexity in terms of Flops, model size, and frames per second.

6. Conclusions

This paper presents a new dual attention network called CALCNet that combines spatial attention and multi-scale feature selection towards land cover classification. Because of its special architecture, CALCNet can simultaneously obtain spatial dependencies and multi-scaled contextual information, and extract discriminative features that are important in making appropriate decisions on which complex land cover types to correctly differentiate.

Extensive experiments carried out in four benchmark datasets, AID, EuroSAT, NWPU or UCMerced LandUse, were used to test the robustness and generalization ability of the new model. CALCNet has continually been ahead of conventional CNN designs like ResNet, InceptionV3, EfficientNet, and others. Through these ablation findings, we were able to validate that the spatial attention module and Multi-Scale Feature Selection Module play an important role in improving the ability of models to capture relevant features and improve classification performance. In particular, CALCNet is able to achieve 98.09%, 99.47%, 98.53%, and 99.19% overall classification accuracies on AID, UCMerced LandUse, NWPU-RESISC45, and EuroSAT datasets, respectively, while the compressed variant (C-CALCNet) greatly reduces parameters and FLOPs but suffers less than 1% accuracy degradation.

Moreover, a compressed version of CALCNet was tested to assess computational speed. The findings showed that the model is highly accurate with minimal performance degradation, making it suitable for resource-constrained systems and real-time scenarios.

In future work, we will investigate integrating multimodal data, including optical and SAR data, to enhance classification under adverse conditions. Future research in lightweight model optimization and transfer learning in different parts of the world can

help make CALCNet a more general task in the large-scale remote sensing and land cover analysis processes.

Author Contributions: Conceptualization, M.F.; methodology, M.F.; software, A.H.I.; validation, H.Y. and M.I.; formal analysis, H.Y.; investigation, H.Y. and L.M.D.; resources, H.Y. and L.M.D.; data curation, M.F. and M.I.; writing—original draft preparation, M.F.; writing—review and editing, H.Y., W.J. and L.M.D.; visualization, M.F. and A.H.I.; supervision, L.M.D.; project administration, H.Y. and L.M.D.; funding acquisition, W.J. All authors have read and agreed to the published version of the manuscript.

Funding: This work was supported by the InnoCORE Program of the Ministry of Science and ICT (N10260002) and by Basic Science Research Program through the National Research Foundation of Korea (NRF) funded by the Ministry of Education (RS-2024-00461244). Furthermore, AI-powered tools have been incorporated to enhance the readability in specific sections of this article.

Data Availability Statement: The data will be available from the corresponding author upon reasonable request.

Conflicts of Interest: The authors declare that they have no known conflicting financial interests or personal ties that may have seemed to affect the work presented in this study.

References

1. Wang, X.; Jiang, W.; Deng, Y.; Yin, X.; Peng, K.; Rao, P.; Li, Z. Contribution of land cover classification results based on Sentinel-1 and 2 to the accreditation of wetland cities. *Remote Sens.* **2023**, *15*, 1275.
2. Yan, X.; Li, J.; Smith, A.R.; Yang, D.; Ma, T.; Su, Y. Rapid land cover classification using a 36-year time series of multi-source remote sensing data. *Land* **2023**, *12*, 2149.
3. Rizayeva, A.; Nita, M.D.; Radeloff, V.C. Large-area, 1964 land cover classifications of Corona spy satellite imagery for the Caucasus Mountains *Remote Sens. Environ.* **2023**, *284*, 113343. https://doi.org/10.1016/j.rse.2022.113343.
4. Melgani, F.; Bruzzone, L. Classification of hyperspectral remote sensing images with support vector machines. *IEEE Trans. Geosci. Remote Sens.* **2004**, *42*, 1778–1790.
5. Ayerdi, B.; Romay, M.G. Hyperspectral image analysis by spectral–spatial processing and anticipative hybrid extreme rotation forest classification. *IEEE Trans. Geosci. Remote Sens.* **2015**, *54*, 2627–2639.
6. Lin, T.H.; Li, H.T.; Tsai, K.C. Implementing the Fisher's Discriminant Ratio in ak-Means Clustering Algorithm for Feature Selection and Data Set Trimming. *J. Chem. Inf. Comput. Sci.* **2004**, *44*, 76–87.
7. He, D.; Shi, Q.; Xue, J.; Atkinson, P.M.; Liu, X. Very fine spatial resolution urban land cover mapping using an explicable sub-pixel mapping network based on learnable spatial correlation. *Remote Sens. Environ.* **2023**, *299*, 113884.
8. Shi, Q.; He, D.; Liu, Z.; Liu, X.; Xue, J. Globe230k: A benchmark dense-pixel annotation dataset for global land cover mapping. *J. Remote Sens.* **2023**, *3*, 0078.
9. Amare, M.T.; Demissie, S.T.; Beza, S.A.; Erena, S.H. Land cover change detection and prediction in the Fafan catchment of Ethiopia. *J. Geovisualization Spat. Anal.* **2023**, *7*, 19.
10. Fayaz, M.; Nam, J.; Dang, L.M.; Song, H.K.; Moon, H. Land-cover classification using deep learning with high-resolution remote-sensing imagery. *Appl. Sci.* **2024**, *14*, 1844.
11. Frimpong, B.F.; Koranteng, A.; Atta-Darkwa, T.; Junior, O.F.; Zawiła-Niedźwiecki, T. Land cover changes utilising landsat satellite imageries for the kumasi metropolis and its adjoining municipalities in ghana (1986–2022). *Sensors* **2023**, *23*, 2644.
12. Li, R.; Gao, X.; Shi, F.; Zhang, H. Scale effect of land cover classification from multi-resolution satellite remote sensing data. *Sensors* **2023**, *23*, 6136.
13. Lv, Z.; Zhang, P.; Sun, W.; Benediktsson, J.A.; Li, J.; Wang, W. Novel adaptive region spectral–spatial features for land cover classification with high spatial resolution remotely sensed imagery. *IEEE Trans. Geosci. Remote Sens.* **2023**, *61*, 5609412.
14. Moharram, M.A.; Sundaram, D.M. Land use and land cover classification with hyperspectral data: A comprehensive review of methods, challenges and future directions. *Neurocomputing* **2023**, *536*, 90–113.
15. Dash, P.; Sanders, S.L.; Parajuli, P.; Ouyang, Y. Improving the accuracy of land use and land cover classification of Landsat data in an agricultural watershed. *Remote Sens.* **2023**, *15*, 4020.
16. McDonnell, M.D. Training wide residual networks for deployment using a single bit for each weight. *arXiv* **2018**, arXiv:1802.08530.
17. Liu, Q.; Kampffmeyer, M.; Jenssen, R.; Salberg, A.B. Dense dilated convolutions' merging network for land cover classification. *IEEE Trans. Geosci. Remote Sens.* **2020**, *58*, 6309–6320.

18. Iervolino, P.; Guida, R.; Riccio, D.; Rea, R. A novel multispectral, panchromatic and SAR data fusion for land classification. *IEEE J. Sel. Top. Appl. Earth Obs. Remote Sens.* **2019**, *12*, 3966–3979.
19. Kulkarni, S.C.; Rege, P.P. Pixel level fusion techniques for SAR and optical images: A review. *Inf. Fusion* **2020**, *59*, 13–29.
20. Sukawattanavijit, C.; Chen, J.; Zhang, H. GA-SVM algorithm for improving land-cover classification using SAR and optical remote sensing data. *IEEE Geosci. Remote Sens. Lett.* **2017**, *14*, 284–288.
21. Qin, Y.; Xiao, X.; Dong, J.; Zhang, G.; Shimada, M.; Liu, J.; Li, C.; Kou, W.; Moore III, B. Forest cover maps of China in 2010 from multiple approaches and data sources: PALSAR, Landsat, MODIS, FRA, and NFI. *ISPRS J. Photogramm. Remote Sens.* **2015**, *109*, 1–16.
22. Chen, Y.; Li, C.; Ghamisi, P.; Jia, X.; Gu, Y. Deep fusion of remote sensing data for accurate classification. *IEEE Geosci. Remote Sens. Lett.* **2017**, *14*, 1253–1257.
23. Hughes, L.H.; Schmitt, M.; Mou, L.; Wang, Y.; Zhu, X.X. Identifying corresponding patches in SAR and optical images with a pseudo-siamese CNN. *IEEE Geosci. Remote Sens. Lett.* **2018**, *15*, 784–788.
24. Xu, X.; Li, W.; Ran, Q.; Du, Q.; Gao, L.; Zhang, B. Multisource remote sensing data classification based on convolutional neural network. *IEEE Trans. Geosci. Remote Sens.* **2017**, *56*, 937–949.
25. Audebert, N.; Le Saux, B.; Lefèvre, S. Beyond RGB: Very high resolution urban remote sensing with multimodal deep networks. *ISPRS J. Photogramm. Remote Sens.* **2018**, *140*, 20–32.
26. Xu, Y.; Du, B.; Zhang, L. Multi-source remote sensing data classification via fully convolutional networks and post-classification processing. In *Proceedings of the IGARSS 2018—2018 IEEE International Geoscience and Remote Sensing Symposium*; IEEE: Piscataway, NJ, USA, 2018; pp. 3852–3855.
27. Capliez, E.; Ienco, D.; Gaetano, R.; Baghdadi, N.; Salah, A.H.; Le Goff, M.; Chouteau, F. Multisensor temporal unsupervised domain adaptation for land cover mapping with spatial pseudo-labeling and adversarial learning. *IEEE Trans. Geosci. Remote Sens.* **2023**, *61*, 5405716.
28. Xia, G.S.; Hu, J.; Hu, F.; Shi, B.; Bai, X.; Zhong, Y.; Zhang, L.; Lu, X. AID: A benchmark data set for performance evaluation of aerial scene classification. *IEEE Trans. Geosci. Remote Sens.* **2017**, *55*, 3965–3981.
29. Yang, Y.; Newsam, S. Bag-of-visual-words and spatial extensions for land-use classification. In *Proceedings of the 18th SIGSPATIAL International Conference on Advances in Geographic Information Systems*; Association for Computing Machinery: New York, NY, USA, 2010; pp. 270–279.
30. Cheng, G.; Han, J.; Lu, X. Remote sensing image scene classification: Benchmark and state of the art. *Proc. IEEE* **2017**, *105*, 1865–1883.
31. Helber, P.; Bischke, B.; Dengel, A.; Borth, D. Eurosat: A novel dataset and deep learning benchmark for land use and land cover classification. *IEEE J. Sel. Top. Appl. Earth Obs. Remote Sens.* **2019**, *12*, 2217–2226.
32. Dastour, H.; Hassan, Q.K. A comparison of deep transfer learning methods for land use and land cover classification. *Sustainability* **2023**, *15*, 7854.
33. Alem, A.; Kumar, S. Transfer learning models for land cover and land use classification in remote sensing image. *Appl. Artif. Intell.* **2022**, *36*, 2014192.
34. Fayaz, M.; Dang, L.M.; Moon, H. Enhancing land cover classification via deep ensemble network. *Knowl.-Based Syst.* **2024**, *305*, 112611.
35. Bazi, Y.; Bashmal, L.; Rahhal, M.M.A.; Dayil, R.A.; Ajlan, N.A. Vision transformers for remote sensing image classification. *Remote Sens.* **2021**, *13*, 516.
36. Sumbul, G.; Ravanbakhsh, M.; Demir, B. Informative and representative triplet selection for multilabel remote sensing image retrieval. *IEEE Trans. Geosci. Remote Sens.* **2021**, *60*, 5405811.
37. Adegun, A.A.; Viriri, S.; Tapamo, J.R. Review of deep learning methods for remote sensing satellite images classification: Experimental survey and comparative analysis. *J. Big Data* **2023**, *10*, 93.
38. Shafaey, M.A.; Salem, M.A.M.; Ebeid, H.; Al-Berry, M.; Tolba, M.F. Comparison of CNNs for remote sensing scene classification. In *Proceedings of the 2018 13th International Conference on Computer Engineering and Systems (ICCES)*; IEEE: Piscataway, NJ, USA, 2018; pp. 27–32.
39. Adegun, A.; Viriri, S.; Tapamo, J.R. Automated classification of remote sensing satellite images using deep learning based vision transformer. *Appl. Intell.* **2024**, *54*, 13018–13037.
40. Xie, H.; Chen, Y.; Ghamisi, P. Remote sensing image scene classification via label augmentation and intra-class constraint. *Remote Sens.* **2021**, *13*, 2566.
41. Xu, K.; Huang, H.; Deng, P.; Li, Y. Deep feature aggregation framework driven by graph convolutional network for scene classification in remote sensing. *IEEE Trans. Neural Netw. Learn. Syst.* **2021**, *33*, 5751–5765.
42. Alhichri, H.; Alswayed, A.S.; Bazi, Y.; Ammour, N.; Alajlan, N.A. Classification of remote sensing images using EfficientNet-B3 CNN model with attention. *IEEE Access* **2021**, *9*, 14078–14094.

43. Vinaykumar, V.; Babu, J.A.; Frnda, J. Optimal guidance whale optimization algorithm and hybrid deep learning networks for land use land cover classification. *EURASIP J. Adv. Signal Process.* **2023**, *2023*, 13.
44. Gupta, N.; Mittal, A.; Singh, S. *Feature Extraction for Remote Sensing Image Classification Using Variants of Deep Learning Pre-Trained Models Densenet-169, Densenet-121 and Densenet-201*; Elsevier: Amsterdam, The Netherlands, 2023.
45. Anwer, R.M.; Khan, F.S.; Van De Weijer, J.; Molinier, M.; Laaksonen, J. Binary patterns encoded convolutional neural networks for texture recognition and remote sensing scene classification. *ISPRS J. Photogramm. Remote Sens.* **2018**, *138*, 74–85.
46. Ekim, B.; Sertel, E. Deep neural network ensembles for remote sensing land cover and land use classification. *Int. J. Digit. Earth* **2021**, *14*, 1868–1881.
47. Obianuju, N.L.; Agwu, N.; Ikechukwu, O. Medium resolution satellite image classification system for land cover mapping in Nigeria: A multi-phase deep learning approach. In *Proceedings of the Intelligent Computing: Proceedings of the 2021 Computing Conference*; Springer: Berlin/Heidelberg, Germany, 2021; Volume 2, pp. 1056–1072.
48. Ma, A.; Wan, Y.; Zhong, Y.; Wang, J.; Zhang, L. SceneNet: Remote sensing scene classification deep learning network using multi-objective neural evolution architecture search. *ISPRS J. Photogramm. Remote Sens.* **2021**, *172*, 171–188.
49. Wu, K.; Zhang, J.; Peng, H.; Liu, M.; Xiao, B.; Fu, J.; Yuan, L. Tinyvit: Fast pretraining distillation for small vision transformers. In *Proceedings of the European Conference on Computer Vision*; Springer: Berlin/Heidelberg, Germany, 2022; pp. 68–85.
50. Sukumar, A.; Anil, A.; Sajith, V.V.; Sowmya, V.; Krichen, M.; Ravi, V. Influence of spectral bands on satellite image classification using vision transformers. In *Proceedings of the International Conference on Frontiers of Intelligent Computing: Theory and Applications*; Springer: Berlin/Heidelberg, Germany, 2022; pp. 243–251.
51. 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.
52. Ahmed, V.A.; Jouini, K.; Tuama, A.; Korbaa, O. A fusion approach for enhanced remote sensing image classification. *Proc. Copyr.* **2024**, *554*, 561.
53. Nampally, T.; Wu, J.; Dev, S. Performance comparison of multispectral channels for land use classification. In *Proceedings of the IGARSS 2023—2023 IEEE International Geoscience and Remote Sensing Symposium*; IEEE: Piscataway, NJ, USA, 2023; pp. 6178–6181.
54. Khan, S.D.; Basalamah, S. Multi-branch deep learning framework for land scene classification in satellite imagery. *Remote Sens.* **2023**, *15*, 3408.
55. Yele, V.P.; Alegavi, S.; Sedamkar, R. Effective segmentation of land-use and land-cover from hyperspectral remote sensing image. *Int. J. Inf. Technol.* **2024**, *16*, 2395–2412.
56. Suganya, D.; Sugumar, R. PSO-Optimized CNN for feature extraction and accurate classification of satellite images using machine learning. In *Proceedings of the 2024 International Conference on Computing and Data Science (ICCDS)*; IEEE: Piscataway, NJ, USA, 2024; pp. 1–6.
57. Nagaraju, K.A.; Chaurasia, K. Identifying Land features from satellite images using deep learning. In *Proceedings of the 2023 16th International Conference on Developments in eSystems Engineering (DeSE)*; IEEE: Piscataway, NJ, USA, 2023; pp. 54–59.
58. Khan, M.A.; Hamza, A.; Ibrar, W.; Jamel, L.; Alasiry, A.; Marzougui, M.; Kumari, S.; Nam, Y. Coastal and Land Use Land Cover Area Recognition from High-Resolution Remote Sensing Images using a Novel Multimodal Attention Inception Residual Deep Network. *IEEE J. Sel. Top. Appl. Earth Obs. Remote Sens.* **2025**, *18*, 17460–17475.
59. Hu, W.; Lan, C.; Chen, T.; Liu, S.; Yin, L.; Wang, L. Scene Classification of Remote Sensing Image Based on Multi-Path Reconfigurable Neural Network. *Land* **2024**, *13*, 1718.
60. Jamali, A.; Mahdianpari, M.; Brisco, B.; Granger, J.; Mohammadimanesh, F.; Salehi, B. Comparing solo versus ensemble convolutional neural networks for wetland classification using multi-spectral satellite imagery. *Remote Sens.* **2021**, *13*, 2046.
61. Chollet, F. Xception: Deep learning with depthwise separable convolutions. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*; IEEE: Piscataway, NJ, USA, 2017; pp. 1251–1258.
62. Szegedy, C.; Vanhoucke, V.; Ioffe, S.; Shlens, J.; Wojna, Z. Rethinking the inception architecture for computer vision. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*; IEEE: Piscataway, NJ, USA, 2016; pp. 2818–2826.
63. Sandler, M.; Howard, A.; Zhu, M.; Zhmoginov, A.; Chen, L.C. Mobilenetv2: Inverted residuals and linear bottlenecks. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*; IEEE: Piscataway, NJ, USA, 2018; pp. 4510–4520.
64. Shafiq, M.; Gu, Z. Deep residual learning for image recognition: A survey. *Appl. Sci.* **2022**, *12*, 8972.
65. Simonyan, K.; Zisserman, A. Very deep convolutional networks for large-scale image recognition. *arXiv* **2014**, arXiv:1409.1556.
66. Bansal, K.; Tripathi, A.K. Dual level attention based lightweight vision transformer for streambed land use change classification using remote sensing. *Comput. Geosci.* **2024**, *191*, 105676.
67. Liu, Z.; Lin, Y.; Cao, Y.; Hu, H.; Wei, Y.; Zhang, Z.; Lin, S.; Guo, B. Swin transformer: Hierarchical vision transformer using shifted windows. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*; IEEE: Piscataway, NJ, USA, 2021; pp. 10012–10022.

68. Thakur, P.S.; Khanna, P.; Sheorey, T.; Ojha, A. Explainable vision transformer enabled convolutional neural network for plant disease identification: PlantXViT. *arXiv* **2022**, arXiv:2207.07919.
69. Fayaz, M.; Dang, L.M.; Moon, H. DLAN: A dual attention network for effective land cover classification in remote sensing. *Knowl.-Based Syst.* **2025**, *319*, 113620.