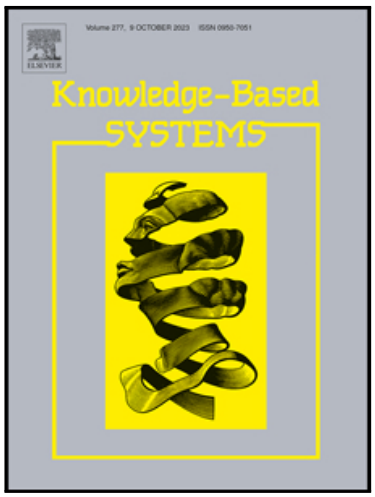

DLAN: A Dual Attention Network for Effective Land Cover Classification in Remote Sensing

Muhammad Fayaz , L. Minh Dang , Hyeonjoon Moon

PII: S0950-7051(25)00666-5
DOI: https://doi.org/10.1016/j.knosys.2025.113620
Reference: KNOSYS 113620

To appear in: *Knowledge-Based Systems*

Received date: 26 November 2024
Revised date: 31 March 2025
Accepted date: 21 April 2025

Please cite this article as: Muhammad Fayaz , L. Minh Dang , Hyeonjoon Moon , DLAN: A Dual Attention Network for Effective Land Cover Classification in Remote Sensing, *Knowledge-Based Systems* (2025), doi: https://doi.org/10.1016/j.knosys.2025.113620

DLAN: A Dual Attention Network for Effective Land Cover Classification in Remote Sensing

Muhammad Fayaz, L. Minh Dang, and Hyeonjoon Moon*

Sejong university, Seoul, South Korea

*Corresponding author: Hyeonjoon Moon

Abstract:

In the era of remote sensing (RS), the demand for accurate land cover classification (LCC) has intensified due to various environmental challenges such as deforestation and urbanization. Conventional approaches often rely on shallow features for classification, limiting their effectiveness in capturing spatial patterns and diverse land cover types. In response, this study introduces a novel LCC approach utilizing a convolutional neural network (CNN) equipped with a dual land cover attention segment. The proposed module integrates channel attention (CA) and spatial attention mechanisms (SA) to enhance the discriminative capabilities of deep models. Leveraging inter-channel and inter-spatial relationships, the dual attention module enables the identification of various land cover types, spatial patterns, and color variations. Through thorough experimentation, the InceptionV3 feature extractor was identified as the optimal backbone for the proposed network architecture. Furthermore, to address the challenge of diverse land cover types, highly curated datasets are utilized. Additionally, to optimize model efficiency and reduce size, an improved model compression approach is employed. The effectiveness of the proposed Dual Land Cover Attention Network (DLAN) was evaluated through extensive experimentation, demonstrating superior performance compared to conventional methods. The results indicate the potential of DLAN in advancing LCC tasks, facilitating detailed agricultural zoning, environmental monitoring, and urban planning at a regional scale.

Index Terms: land area images; Remote sensing; land use classification; land cover classification; Satellite imagery; Dual attention mechanism

1. Introduction:

In the present era, RS technology offers an extensive volume of earth observation data, encompassing satellite images and LiDAR. These datasets serve besides global-scale environmental monitoring but also find applications in regions for example change detection, LCC, and the observing and evaluation of natural disasters (X. Wang, et al., 2023). Furthermore, enhancing the precision of LCC holds paramount significance for Geo-observation, agricultural zoning, environmental protection at regional level, and urban development (Yan, et al., 2023). Recent incidents, such as changes in land cover patterns and extensive deforestation, have highlighted the urgent need for effective LCC. Therefore, numerous methods are developed for LCC primarily rely on shallow features like pixel color and texture for classification (Rizayeva, Nita, & Radeloff, 2023). Moreover, machine learning-driven algorithms are applied to remote sensing (RS) image classification, encompassing methods such as

SVM (Melgani & Bruzzone, 2004), random forests (RF) (Ayerdi & Romay, 2015), K-means clustering (Lin, Li, & Tsai, 2004), etc. However, changes in land cover have been observed in regions beyond natural disasters, particularly due to urbanization, agriculture, and other human activities, requiring more advanced classification methods. While several methods have been developed for LCC, they often face challenges. These methods typically rely on shallow architectures that stack multiple layers without considering optimal feature selection. Some studies have explored attention mechanisms in the spatial or channel dimensions, but these methods are insufficient to effectively capture dominant spatial and channel details. Moreover, the inference time of existing methods makes them computationally expensive (D. He, Shi, Xue, Atkinson, & Liu, 2023; Shi, He, Liu, Liu, & Xue, 2023).

However, satellite images often suffer from interference due to factors such as cloud cover, varying light conditions, and atmospheric disturbances. To address these challenges, the proposed method integrates robust feature extraction and attention mechanisms, such as the InceptionV3 model and the DAM, which enhance the model's resilience to such interferences. The CA and modified SA modules allow the model to focus on the most informative features, mitigating the negative effects of noise while ensuring effective land cover classification. Furthermore, the model can be fine-tuned on specific datasets affected by environmental interferences, making it adaptable to diverse real-world conditions. Future enhancements will include additional preprocessing strategies to improve the model's robustness against cloud cover and lighting variations.

Therefore, this study developed an efficient LCC method by integrating DLAN module with backbone feature extractor. The DLAN module enhances the discriminatory capabilities of the backbones, enabling them to identify various land cover types, nuanced patterns and shade variations. The main contributions to our work are described below:

- *Introduction of Dual Land Cover Attention Module:* We propose a novel DLAN featuring an innovative dual attention module that integrates CA and SA mechanisms. This module enhances the network's ability to discern and classify various land cover types by effectively capturing both channel-wise and spatial relationships within remote sensing data.
- *Optimal Backbone Selection and Comprehensive Evaluation:* Through extensive experimentation, we identify InceptionV3 as the most effective backbone for our network architecture. We rigorously evaluate the proposed DLAN using a highly curated dataset of 21 distinct land cover classes and openly available datasets, indicating that they provide higher performance in LCC than existing state-of-the-art approaches.
- *Model Compression and Efficiency Improvement:* To address the challenge of model efficiency and size, we incorporate a refined model compression approach. This optimization ensures that DLAN remains effective while being resource-efficient, making it suitable for deployment in diverse environments with varying computational resources.
- *Empirical Validation and Qualitative Analysis:* We provide a thorough empirical validation of the DLAN's performance, including qualitative analyses that highlight its advantages over

conventional methods. Our results underscore the potential of DLAN for advancing LCC tasks, with implications for agricultural zoning, environmental monitoring, and urban planning.

The following sections of the article are organized as shown below: Section 2 encompasses a related study, offering a concise overview of the literature, and Hybrid methods. A brief exposition of the proposed method is presented in Section 3. Section 4 delves into discussions regarding the dataset, performing assessment, parameter settings, and outcomes. Lastly, Section 5 serves as the conclusion, addressing the shortcomings of the study and proposing directions for the future.

2. Literature Review

Traditional approaches for LCC in RSIs often rely on a limited set of rules applied to distinct spatial units, like pixels and objects (Amare, Demissie, Beza, & Erena, 2023). However, these traditional Machine Learning (ML) approaches with artificial feature descriptors typically involve low-level features, making it challenging to efficiently identify complicated land structures. Deep Learning (DL) has gained widespread application in LCC due to its improvements in multiscale and multilevel feature extraction, yielding optimal outcomes (Fayaz, Nam, Dang, Song, & Moon, 2024). DL-based land cover categorization approaches can be largely categorized into two groups based on the spatial representation level of the labels: patches and pixels. Patch-level algorithms are proper for medium-resolution RSIs, lacking fine structural information (Frimpong, Koranteng, Atta-Darkwa, Junior, & Zawiła-Niedźwiecki, 2023). For instance, Li et al. (R. Li, Gao, Shi, & Zhang, 2023) introduced a patch-based recurrent neural network (RNN), and Lv et al. (Lv, et al., 2023) developed a lightweight CNN for land cover mapping. On the other hand, Pixel-level algorithms seek to classify each pixel in remote sensing images (RSIs) with specific land cover labels using end-to-end deep learning models, akin to the process of semantic segmentation in natural images. State SOTA semantic segmentation architectures for RSIs often employ encoder–decoder architecture to capture detailed multilevel contextual information across extensive receptive fields (Dash, Sanders, Parajuli, & Ouyang, 2023; Moharram & Sundaram, 2023). Examples include McDonnell et al. (McDonnell, 2018) expanded stacked U-Net for semantic segmentation of RGB RSIs and Liu et al. (Q. Liu, Kampffmeyer, Jenssen, & Salberg, 2020) Compact dilated convolution integration network for LCC utilizing combined contextual information at both local and global levels.

Given the limitations of RSI's sensors in meeting high temporal, spatial, and spectral resolution requirements, the merging of corresponding information from multisource RSIs has emerged as a promising approach to enhance accuracy (Dang, et al., 2024). The fusion of information from different sources contributes to overcoming sensor limitations and improving the overall effectiveness of LCC techniques in RSIs.

Traditional approaches involve two phases: multisource RSI fusion and ML-based classification, with examples such as Iervolino et al. (Iervolino, Guida, Riccio, & Rea, 2019),and Kulkarni et al. (Kulkarni & Rege, 2020). ML-based classification methods include genetic algorithms and SVM by Sukhavatanavit et al. (Sukawattanavijit, Chen, & Zhang, 2017) and a general model by Qin et al. (Qin, et al., 2015) based on the Markov random field. However, these artificial feature descriptor techniques possess limited expressive power, restricting their capability to completely represent intricate high-level semantic details.

In the last few years, SOTA multisource LCC techniques predominantly rely on DL. Chen et al. (Chen, Li, Ghamisi, Jia, & Gu, 2017) developed a deep neural network (DNN) with two CNNs for extracting features from multi/hyperspectral and LiDAR data, and a fully connected DNN for combining heterogeneous features. Hughes et al.(Hughes, Schmitt, Mou, Wang, & Zhu, 2018) presented a pseudo-Siamese structure (PSCNN) for recognizing corresponding patches in high-resolution optical and SAR imagery, where information is combined in a final fully connected decision layer through concatenation and a 1 × 1 convolutional operation.

Despite these advancements, Patch-based multimodal depth models are unable to carry out pixel-level classification of high-resolution multisource RSIs. Xu and li et al. (X. Xu, et al., 2017) explored a dual-branch CNN for pixel-level classification of multisource remote sensing data (MRSDC) by combining hyperspectral imagery (HSI) with data from various sensors. Audebert et al. (Audebert, Le Saux, & Lefèvre, 2018) discussed Strategies for urban segmentation include both early and late fusion approaches, with experiments showing late fusion recovering critical errors on hard pixels, while early fusion (V-FesuNet) facilitates more robust multimodal combined features but exhibits greater susceptibility to missing or noisy data. Xu et al. (Y. Xu, Du, & Zhang, 2018) proposed a fusion-FCN architecture for classification using LiDAR data, HSI data, and high-resolution images. Capilez et al. (Capliez, et al., 2023) developed an M3 fusion architecture integrating CNN and recurrent neural networks for spatial and time-series information.

In short, multimodal DL-based LCC models often retrieve features from individual modal networks and create fusion representations for categorization. However, existing research has mainly concentrated on directly deriving multimodal features through interconnected two-stream approaches networks, limiting the exploration of specific modality features. The proposed joint learning strategy aims to effectively extract modality-specific discriminative information by concurrently learning both fusion and individual modality branches.

3. Methodology

As analyzed in **Section 2**, The CNN-based models achieved significantly superior performances compared to Traditional Machine Learning based methods. However, a receptive field is used for

feature extraction in all these models, specifically in shallow layers giving us limited results and making it difficult to differentiate between the different scenes of different classes. To deal with this challenging issue, we utilize a dual attention module consisting of CA and modified SA mechanisms in DLAN to gather extra polished spatial feature information and useful channels for LCC. The overall structure of the DLAN is illustrated in **Figure 1**, and it is explained in the following parts.

3.1 Data Feature Extraction

The remarkable capacity of CNNs to extract valuable aspects or features from rough images and adapt these images to a variety of computer vision functional domains. Selecting a domain adapted CNN structure is a tough task that needs to be completed to produce consistent algorithmic complexity and reliable predictions for practical use. The advancement of CNNs for visual detection task has been the subject of remarkable research, analyst have employed pre-trained models used as core feature extractors and subsequently fine-tune advance models on datasets for land cover categorization and localization. The idea of finetune and optimizing a pre-trained model involves adjusting the weights and parameters that have been learned, which helps to learn the unique visual characteristics of the domain. Pre-trained

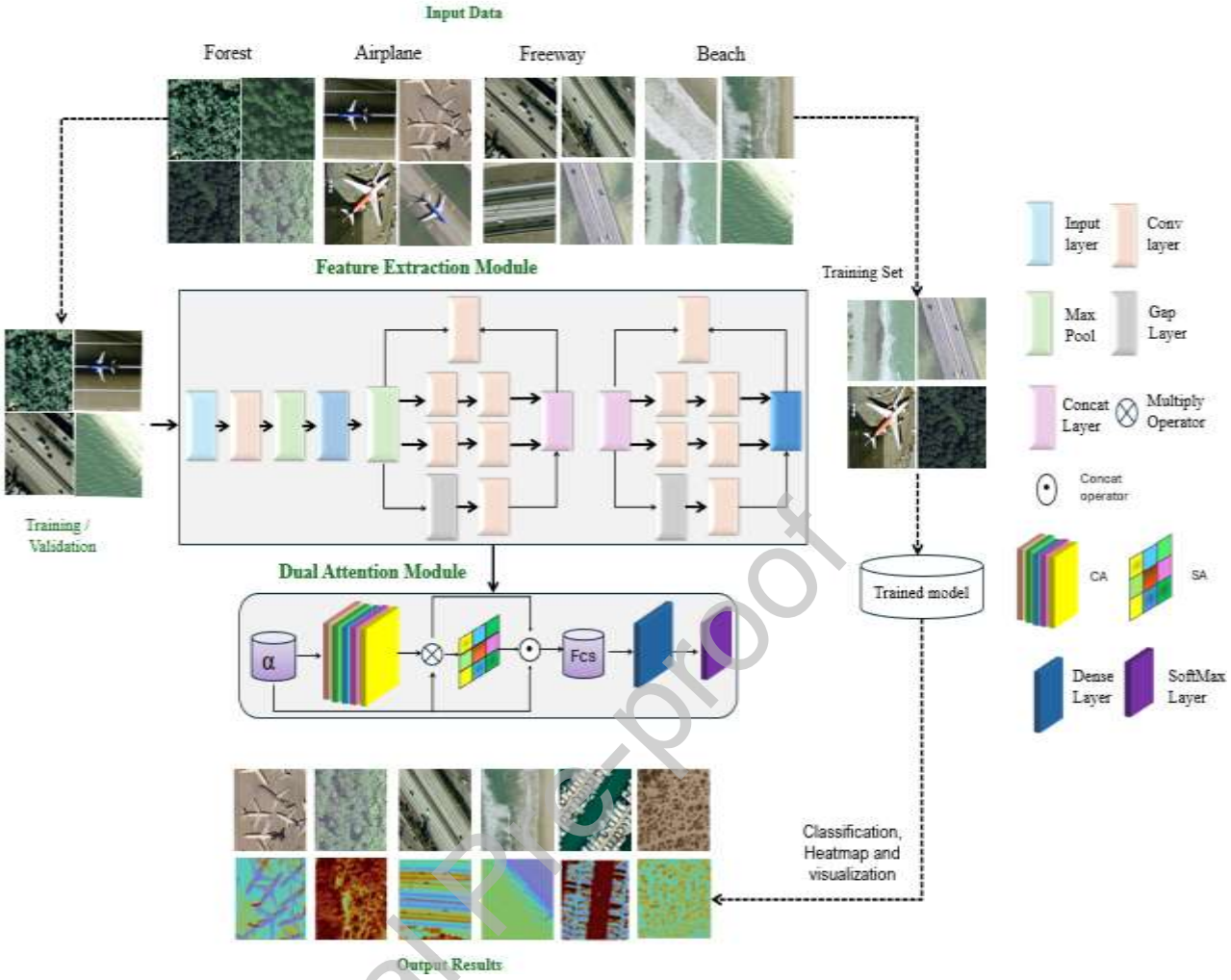

Figure 1 : Proposed land cover architecture.

networks include a powerful and diverse feature extraction process that may be efficiently utilized to adjust the system for any visual classification task.

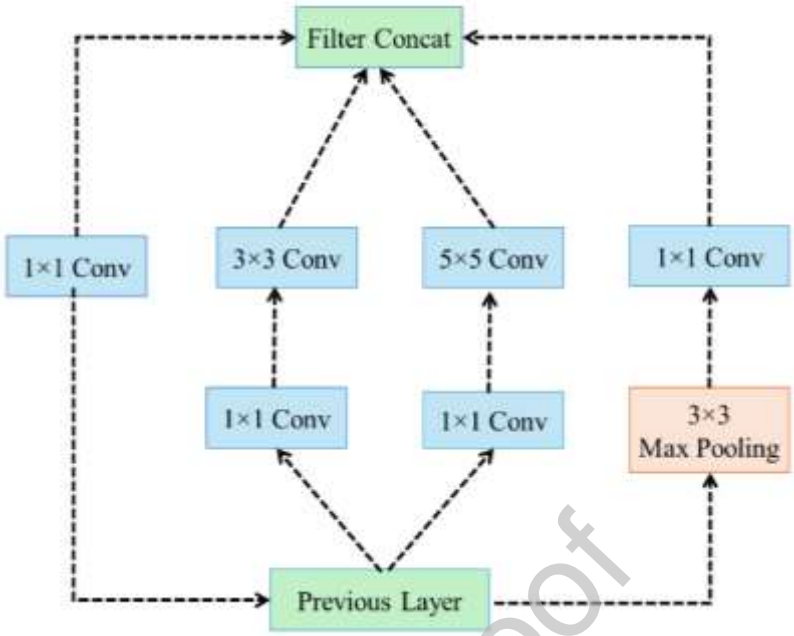

Figure 2 : Representation of inception V3 model for land Cover classification

We use numerous core feature extractors, like EffiecientNet, Xception, MobileNet, ResNet, InceptionV3, and efficientNetV2S to establish the best method for identifying key features for LCC in exceptionally difficult circumstances. Our approach is motivated by the progress ratio of active feature extraction approaches in various computer vision-based fields. In Section 4, the optimality of employing InceptionV3 features is empirically proven. Theoretically, InceptionV3 has been adjusted to produce better outcomes than its older version and has more Inception modules (Ahmad, Jan, Farman, Ahmad, & Ullah, 2020). Inception modules are multi-scale processing, which provides excellent results in a variety of tasks. The Inception (A), (B), and (C) modules are the three fundamental Inception modules found in InceptionV3. **Figure 2** illustrates the several convolutional and pooling layers that are included in each Inception module. These modules use 1×1, 1×3, and 3×1 and 3×3 tiny convolutional layers with small filter sizes to reduce the amount of training parameters. The channels for red, green, and blue are present in the 299 x 299 input size that is the default for InceptionV3. Five convolutional layers are employed to process the input images at first. Each convolutional layer applies multiple 3 × 3 kernels. After extracting the last few dense layers from InceptionV3, we obtain an 8x8 feature vector with 2048 channels for DLAN,

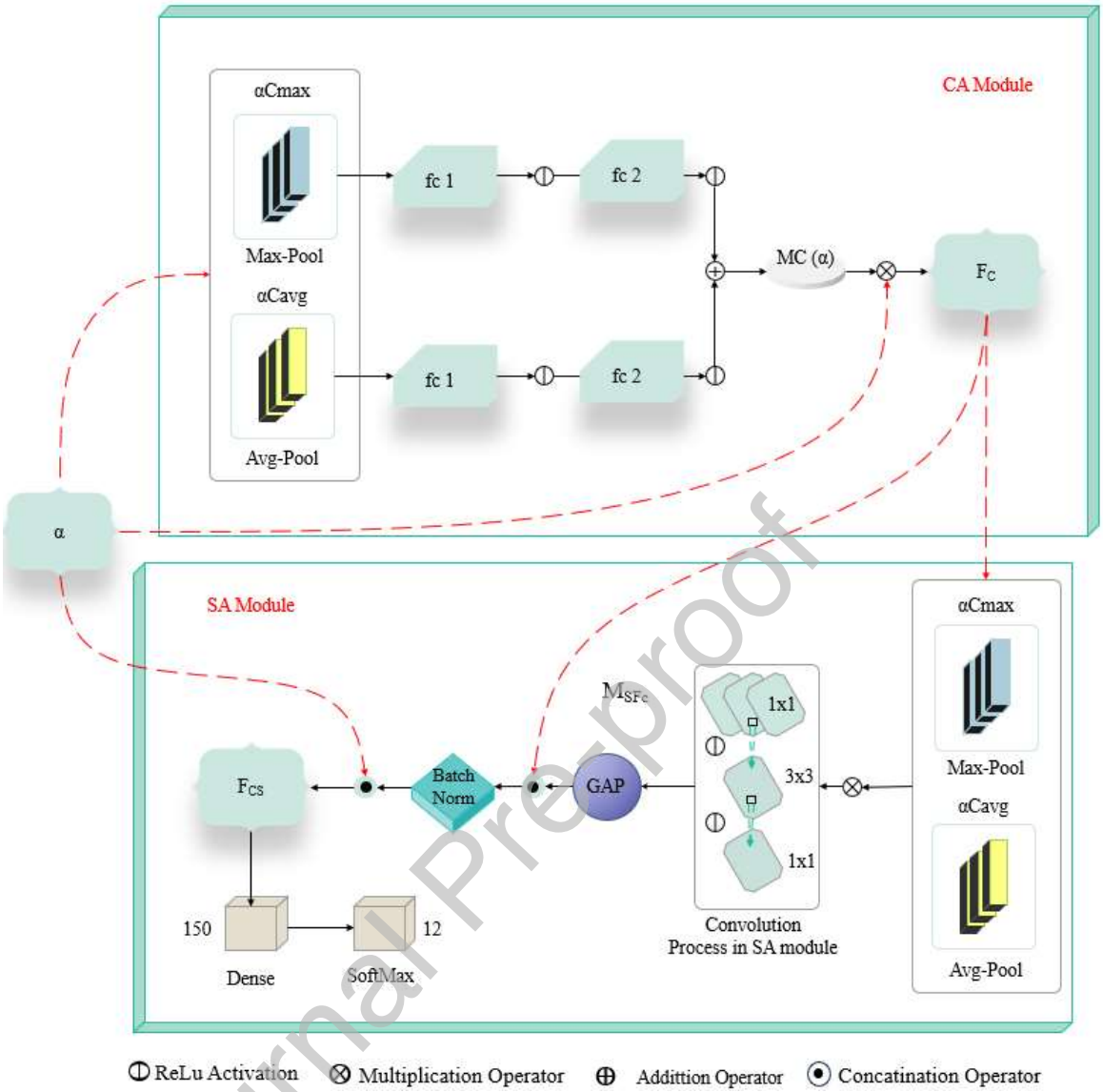

Figure 3 : Conceptual representation of dual attention module.

called α; The mathematical representation is expressed in **Eq 1**:

$$\alpha = \partial\left(\partial\left(\partial\left(\left(\mu\left(\chi\right)\right)\right)\right)\right) \qquad (1)$$

where μ denotes the first convolutional operation applied to the input $\boldsymbol{\chi}$ and each $\boldsymbol{\partial}$ stands for one of the three Inception modules used in the DLAN. **Equation 1** has given us a feature vector α that is very detailed and includes a lot of information about the object, such as its colors, forms, and data about its edges and structure. But they are coarse features and using them to classify visual sequences will inevitably lead to under-depiction of the localization data and incorrect predictions, particularly in complex scenarios. As explained in the subsections, our dual attention module collects the most significant channels and spatial information to further enhance the α features.

3.2 Dual Attention

Numerous application domains have been analyzed CNN-based networks with different attention modules (Khan, et al., 2025) as reported in this research (Fayaz, Dang, & Moon, 2024a; Ullah, Ullah, Hussain, Khan, & Baik, 2021). Due to the combination's remarkable frame resemblance with every sequence, which produces positive performance consequences, it is especially outstanding in video applications. However, because of the variety of information and the use of specific CA approach or SA approach, the attention-based mechanisms created in this study led to limited performance for image data. Additionally, for image-oriented scene recognition, several papers just included a CA module in CNN architecture. For simple scenarios like classifying land cover objects, combining a CA incorporating base model characteristics is an efficient strategy. However, in more complicated scenarios identifying and localizing the object on scene requires more than just a CA module, the results are Restricted. The attention portion of the dual attention network is represented by the CA and modified SA modules that we introduced in this paper. This allows the dual attention network to concentrate on the highest priority areas for LCC, scene localization, and the identification of land-based objects. Using the dual (Channel Attention + modified Spatial Attention) attention strategy, the attention module effectively extracts and localizes the most significant regions.

Channel Attention: We employ CA to utilize the inter-channel co-ordination amongst features. Each channel's feature map is regarded as a feature detector, as depicted in **Figure 3**, where the CA includes fully interconnected layers, a convolution process, a global average pooling (GAP) layer, and a maximum pooling layer, each channel feature map is considered as a feature detector. During the training process, different channels convolution feature maps contribute differently to the representation of an object; some of the channels exhibit model structures better as compared to others, and vice versa. To decrease the complexity of the calculated weights, many researchers, including shu-xiangbo et al. (Shu, Yang, Yan, & Song, 2022) and yar et al. (Yar, et al., 2022) employed attention models that made use of maximum or average pooling. To develop our CA module, we used both pooling strategies rather than just maximum or average pooling. The max-pooling strategy concentrates on the most refined aspects of an object, on the other hand avg-pooling presents the most extensive information about the feature map. Hence, the two individual pooling techniques are employed individually on the spatial aspect or dimension of the assigned feature map α to create two separate spatial background details: β_{CHavg} and β_{CHmax}, which are computed as given below:

$$\beta_{CHavg} = avg - pool\ (\beta) \tag{2}$$

$$\beta_{CHmax} = max - pool\ (\beta) \tag{3}$$

The output description of β_{CHavg} and β_{CHmax} are received by two FC layers, fc^1 and fc^2, which have mutual constraints. Then, a ReLU function is implemented in every FC layer to obtain M_{max} and M_{avg} . Furthermore, a summation operation is carried out on both feature maps to calculate the weight parameters, resulting in Mc(α):

$$M_{max} = \ominus\,(fc^2(\ominus\,(fc^1(\beta_{CHmax})))) \quad (4)$$

$$M_{avg} = \ominus\,(fc^2(\ominus\,(fc^1\alpha CH_{avg})))) \quad (5)$$

$$M_c(\alpha) = (M_{max} \oplus M_{avg}) \quad (6)$$

In the above equations, fc^1 & fc^2 layers utilize pointwise convolution filters to reduce and increase the features of every channel, in which $\ominus$ denotes ReLU function, and $\oplus$ signifies the summation process. Lastly, we obtain the $Mc^{(\alpha)}$ features by using a summing operation on M_{max} and M_{avg}. Next, we apply a skip residual connection, $\emptyset$, to retain the feature transfer from the input and generate the CA map F_c:

$$F_C = M_C(\alpha)\ \emptyset\ \alpha \quad (7)$$

3.3. Modified Spatial Attention:

This segment uses spatial interaction of features. The SA technique is distinct from the CA technique; its goal is to identify the most significant area, which matches the actions of the CA section. To determine the SA, we employ maximum and avg pooling techniques within the channel, combining them with effective feature descriptors. To effectively highlight informational locations, pooling operations are applied along the channel axis. The next two-dimensional feature maps are created by utilizing the two pooling operations:

$$\alpha S_{avg} = Avg - pool\,(F_C) \quad (8)$$

$$\alpha S_{max} = max - pool\,(F_C) \quad (9)$$

Then, to create a two-dimensional SA feature map, the resulting feature maps are convolved by a convolutional layer and concatenated using an addition operation. We employed three convolutional layers in the modified SA module, preceded by the ReLU function. Each layer has 64 different filters: 1 × 1 convolution is used in the first layer, 3 × 3 filter is used in the second layer, and 1 × 1 filter is used in the final layer. Instead of using dilated convolution operations, we use standard convolution method employed in every layer:

$$MsF_c = (\ominus\,(f^{1\times1}(\ominus\,(f^{3\times3}(\ominus\,(f^{1\times1}(\alpha s_{avg} \oplus \alpha s_{max}))))))) \quad (10)$$

In ***eq (10)***, f portrays the filter dimensions utilized within the convolutional layers of the SA module. Employing GAP with the feature maps of Ms_{Fc} and combining their outcome with F_c, will yield the SA map $M_{SF_{cGAP}}$ as shown below.

$$M_{SF_{cGAP}} = GAP(F_c) \tag{11}$$

$$F_s = \odot\left[M_{SF_{cGAP}}, F_c\right] \tag{12}$$

Thereafter the combination procedure, we attained F_S; subsequently, batch normalization (BN) was carried out on F_S. In the end, we derived the feature maps of BN and α to obtain F_{CS}:

$$F_{CS} = \odot\left[Batchnorm, \alpha\right] \tag{13}$$

Then, the F_{CS} were fed into a fully linked layer with 150 neurons, at end, the input images were classified to their corresponding classes by using SoftMax layer. In this study, the two units, the Channel and Spatial attention modules, were employed to determine corresponding details that focus on "what" evidence is significant and "where" this detail info is located.

3.4. DLAN Compression Module

The timely change and intricate characteristics of land cover demand swift and timely detection, underscoring the necessity for low latency, rapid inference time, and instantaneous decision-making. To address this imperative in the context of land cover applications, leveraging edge devices becomes obtainable. However, the restricted computational power and storage capacity inherent to these devices necessitates the deployment of efficient CNN models. In this research endeavor, we introduce a pioneering model compression technique tailored to the specific requirements of land cover analysis. This technique focuses on eliminating redundant neurons, effectively reducing the number of learning parameters while preserving the high performance of the model, ensuring a seamless balance between computational efficiency & accurate LCC.

Image compression plays a crucial role in optimizing the performance of deep learning models, particularly when dealing with large datasets and the need for real-time processing. In the context of land cover classification, image compression helps reduce the computational load and enhances inference speed without sacrificing accuracy. The compression process typically involves several key steps, such as reducing pixel resolution, applying quantization techniques, and using model-based compression methods that eliminate redundant parameters. Herein, we employ a model compression technique based on differential evolution (DE), which optimizes the network by identifying and removing redundant neurons while retaining critical features for accurate classification. This ensures that the model remains efficient and suitable for deployment in resource-constrained environments,

such as drones, where low latency and quick decision-making are essential. Future work will include a more detailed explanation of the compression steps and their impact on the overall model efficiency.

In this research endeavor, we explore a meta-heuristic strategy based on differential evolution (DE) to optimize the Dual Land Cover Attention Network (DLAN) for land cover analysis. Recognizing the parallels with biological evolution, DE serves as a meta-heuristic technique to enhance the efficiency of the DLAN model. By strategically decreasing learning parameters count through DE, we aim to streamline the model for improved performance in land cover analysis, aligning with the specific demands and nuances of this dynamic environmental domain.

V1	V2	V3	V_res	V_donor
0	1	0	0.5	1
1	1	0	1.5	1
1	0	1	0.5	1
0	0	1	-0.5	0
1	0	1	0.5	1
…	…	…	…	…
0	1	1	0	0
1	1	0	1.5	1
1	0	0	1	1
Recombination (second block)				
V_target	V_donor	Random numbers	V_final	- -
1	0	0.4	0	- -
0	0	0.8	0	- -
1	1	0.3	1	- -
1	0	0.4	1	- -
0	1	0.8	0	- -
…	…	…	…	- -
1	0	0.9	1	- -
1	1	0.1	1	- -
0	1	0.4	0	- -
			0	- -

Table 1 : process of mutation (first block) and recombination (second block).

The optimization process begins with a population pool of vectors, where each vector corresponds to the number of neurons in a hidden layer. Each element in the vector is assigned a value of either 0 or

1, where 0 indicates that the corresponding neuron will be discarded, and 1 indicates that the neuron will be retained. Over multiple iterations, a series of steps mutation, recombination, and selection are applied to identify and eliminate redundant neurons, ultimately optimizing the model's efficiency. In the mutation step, three randomly selected vectors are used to generate a donor vector, which is then calculated using Equation 14:.

$$V_{donor} = v1 + F \times (v_2 - v_3) \quad (14)$$

Here, F represents the mutation factor (set to 0.5 in this study), and $v1, v2$ and $v3$ are randomly selected vectors from the population pool. The resulting values from the mutation process, which may not initially fall within the set 0 and 1 are re-scaled so that values below 0.5 are set to 0, and values greater than or equal to 0.5 are set to 1. This ensures that the values are consistent with the binary nature required for neuron retention or discarding. The recombination step follows, where each vector element is assigned a random value as shown in **Table 1**. If this random value is below a remerging factor (set to 0.7 in this study), the resulting vector element is taken from the target vector; otherwise, the element is derived from the donor vector. This process helps introduce variation and explore new configurations that may lead to improved performance.

The overall objective of this optimization approach is to balance the trade-off between maintaining classification accuracy and reducing model complexity. The fitness function used during the process incorporates both the F1-score (a measure of classification accuracy) and the compression ratio (a measure of model size reduction). This dual-objective strategy ensures that the selected neurons contribute to accurate classification while minimizing the model's size, making it suitable for deployment in resource-constrained environments, such as drones for real-time land cover classification. By applying this DE-based optimization mechanism, the model becomes more efficient without compromising its ability to accurately classify land cover data, ultimately resulting in a more practical and deployable solution for real-world applications. The model fitness function is represented by **Eq. 15 and 16** listed below:

$$Z = g \times \left(1 - \frac{\omega i}{\alpha i}\right) + (1 - g) \times F_1(k) \quad (15)$$

$$(\omega i \leq \alpha i, \forall\, 1 \leq i \leq M) \quad (16)$$

Here, ωi denotes the reduction factor applied to the count of hidden neurons in the i^{th} hidden unit layer, and αi represents the original count of neurons in that specific hidden unit layer, both customized to accommodate the intricacies of land cover analysis. The parameter g signifies the weight assigned to the first objective, which involves minimizing the number of neurons, while 1−g corresponds to the weight assigned to the second objective, focusing on optimizing model

performance. This dual-objective optimization strategy is crucial for tailoring the model to the specific demands of land cover analysis, balancing the trade-off between model complexity and classification accuracy.

4. Experimental Setup and Results

This section provides an overview of our experimental approach, focusing on the tailored aspects for land cover analysis. It covers dataset selection, training procedures, and specific evaluation metrics. Following this, a thorough comparison is presented, both quantitatively and qualitatively, between our proposed method and SOTA techniques within the realm of LCC. This comparison aims to highlight the effectiveness and advancements introduced by our model in addressing the unique challenges inherent in land cover analysis. Finally, to validate the dual attention network performance we executed an ablation study. We executed the experiments on an Intel(R) Core (TM) i9-14900K 3.20 GHz with an NVIDIA GeForec RTX 3090 Ti GPU and SAMSUNG 990 PRO 2TB SSD employing Keras for deep learning which utilizes TensorFlow as the backend. A detailed summary of hardware and software is given in **Table 2**.

Table 2 : Hardware and software specifications for the proposed system

Label Name	Description
Libraries	NumPy, TensorFlow, Keras, sklearn, matplotlib, OpenCV
Processor	3.20 GHz, NVIDIA GeForce RTX 3090
Development tools	Windows-10 ,64-bit, Python 3.8
Memory	24 GB

The DLAN model, along with the ablation studies, with the training period of the model is 100 epochs using the standard input dimension (224×224×3) as specified by the Inception-V3 model. The batch size 32 and SGD optimizer is employed utilizing a learning rate (lr) of 0.001 and momentum coefficient of 0.8.

4.1. Datasets & performance metrics.

In the experimental evaluation of the DLAN, we carried out experiments by utilizing various datasets, which includes UCMerced_LandUse (Yang & Newsam, 2010), NWPU (Q. Wang, Gao, Lin, & Li, 2020) dataset and EuroSAT (Helber, Bischke, Dengel, & Borth, 2019). EuroSat is a dataset with a small size for recognition for land use. Dataset has ten classes, and each class has about 3000 images.

The **UCMerced Land Use** dataset was carefully chosen for scholarly research, exhibiting a stringent selection procedure to maintain academic standards and data quality. Finally, they created a huge scale dataset for land cover and land use classification and detection comprising 21000 images; each class

has 1000 images. Throughout our experimentations, we observed that currently available datasets either lack diversity or are confined to a restricted number of classes.

Figure 4 : Sample images from datasets

Models trained in those types of datasets may struggle to perform effectively in real-time, complex conditions. Hence, we select a compact, imbalanced, and exceptionally diverse dataset, encompassing 21 distinct land cover classes. The dataset originated from (Yang & Newsam, 2010). Thus, for the training process, we included 70% of the dataset, for the validation 20%, and for the testing 10%. Sample images are shown in **Figure 4**.

We employed the evaluation metrics previously used to assess various SOTA LCC methods (Fayaz, Nam, et al., 2024; Hussain, Ul Amin, Fayaz, & Seo, 2023; Meng, Xie, Sun, Liu, & Han, 2023; Stanimirova, et al., 2023) comprising the accuracy, precision, F1-score, and recall, discriminated by the false-positive and false-negative rate. Detailed explanations regarding the arithmetic formulation of these metrics are provided in the references.

Ablation Study

We performed multiple studies to determine the most effective configuration for the Dual-Attention Fusion Network in the context of LCC. These studies addressed exploring different base core models

and assessing the efficiency of the proposed bilateral attention approach among many features. The outcomes of these ablation studies are presented in **Table 3** and are elaborated upon in the following subsections.

Table 3 : Effectiveness study of the proposed technique using the suggested UCmerced dataset for ablation study. Here blue implies the top performance and green conveys the second-best performance value whereas the line between all models differentiates the core feature from the performance of distinct dual attention ablation.

Backbone models	LAN	UCmercedLandUse Dataset			
		Precision	Recall	F1 score	Accuracy
MobileNet	×	91.00	92.00	92.00	87.00
Inception V3	×	93.00	90.00	91.00	89.00
Xception	×	89.00	91.00	90.00	88.00
ResNet50	×	88.00	88.00	89.00	89.00
EfficientNetV2S	×	90.00	90.00	91.00	90.00
MobileNet	✓	92.00	93.00	93.00	91.00
Xception	✓	91.00	92.00	92.00	89.60
ResNet50	✓	90.00	90.00	90.00	90.00
$DLAN_{Comp}$	✓	**97.00**	**96.00**	**96.00**	**97.00**
DLAN	✓	**97.00**	**97.00**	**98.00**	**98.00**

We employed various standard CNN models for core feature extraction. These models include EffiecientNet, ResNet101, MobileNet, DenseNet121, and Inception. Furthermore, we incorporated the dual attention module within these models to boost the precision of LCC and localization with our proposed dataset.

The standard CNN models incorporating a dual attention module performed better than approaches that only used deep features, as can be shown in the first block of **Table 3**. This superiority over simple ImageNet classification is explained by the greater difficulties involved in land categorization. The incorporation of attention modules serves to augment the extraction capabilities of distinctive features of objects, consequently enhancing the overall accuracy of LCC. This method has better feature extraction capabilities than the baseline models, the InceptionV3 features paired or when paired with a SoftMax classifier, achieved the best performance. Therefore, from **Table 3**, we can finalize that DLAN provides the top performances, while Xception and EffiecientNet yield the worst results, owing to the non-corrective nature of some of their features and also restricted functionality of some of the features in LCC.

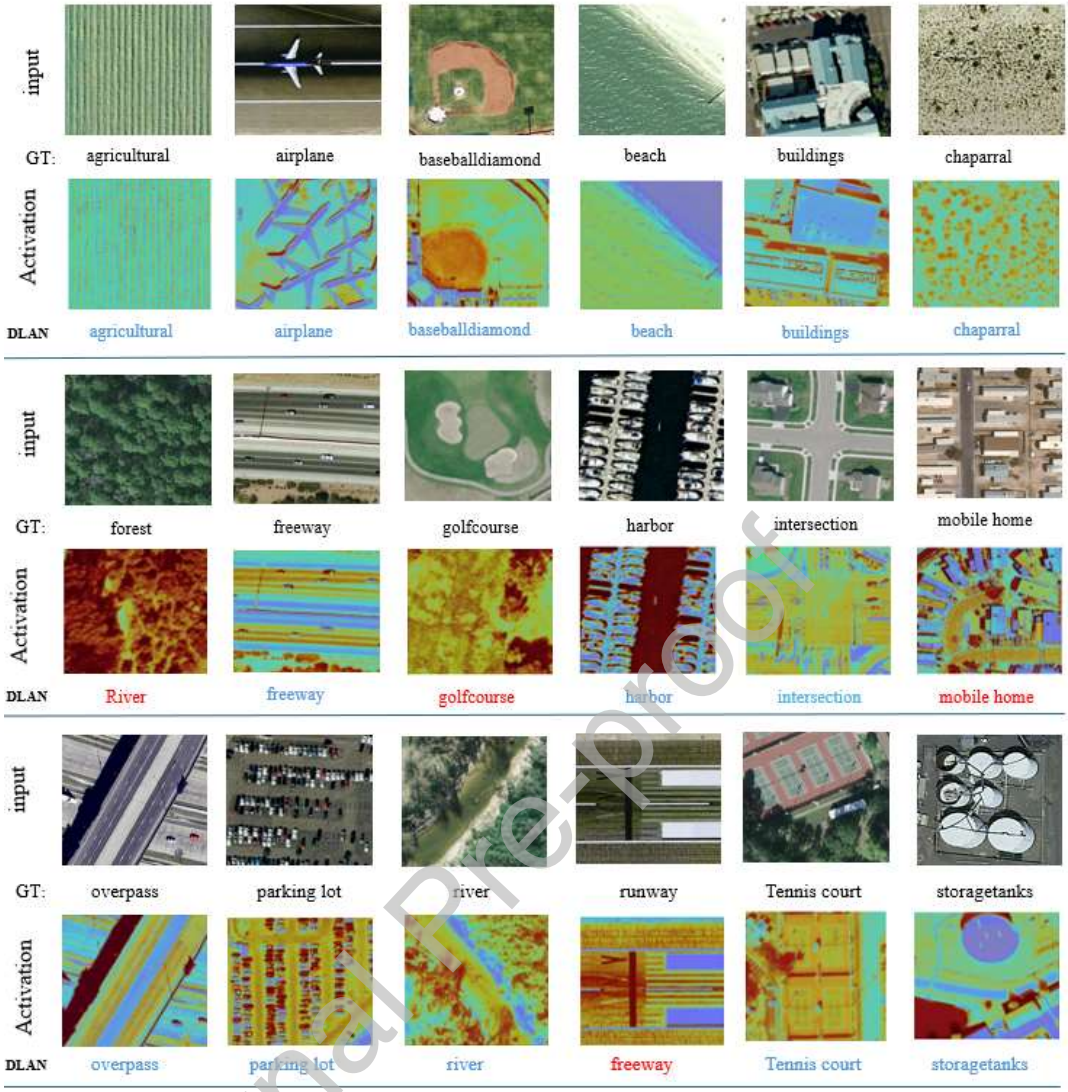

Figure 5 : Visual representation of our DLAN utilizing the proposed dataset. Blue and red colors show accurate and inaccurate classification results for each dataset.

The forest images are classified as rivers because of the strong visual resemblance among these classes. Furthermore, the runway images are wrongly classified as a freeway because the land seems like a freeway. **Figure 5** illustrates that the DLAN is a proficient model capable of perfectly identifying land cover in challenging conditions. In **Figure 5**, certain images are misidentified and not accurately localized, primarily attributed due to the visual resemblance between different land cover classes.

Furthermore, to validate the effectiveness of our dual attention module, we conducted a comparative analysis with other widely used attention mechanisms, including Channel Attention (CA), Spatial Attention (SA), Squeeze-and-Excitation Network (SENet), Multi-Head Attention (MHA), Convolutional Block Attention Module (CBAM) and proposed one. As shown in **Figure 6**, our proposed Dual Attention Network (DLAN) achieved the highest accuracy of 98.00%, outperforming

all other attention mechanisms over UCmercedLandUse dataset. While CBAM and MHA also demonstrated strong performance with accuracies of 97% and 96%, respectively, our approach effectively leverages both spatial and channel information synergistically, leading to superior results. This highlights the advantage of integrating complementary attention mechanisms to enhance LCC performance.

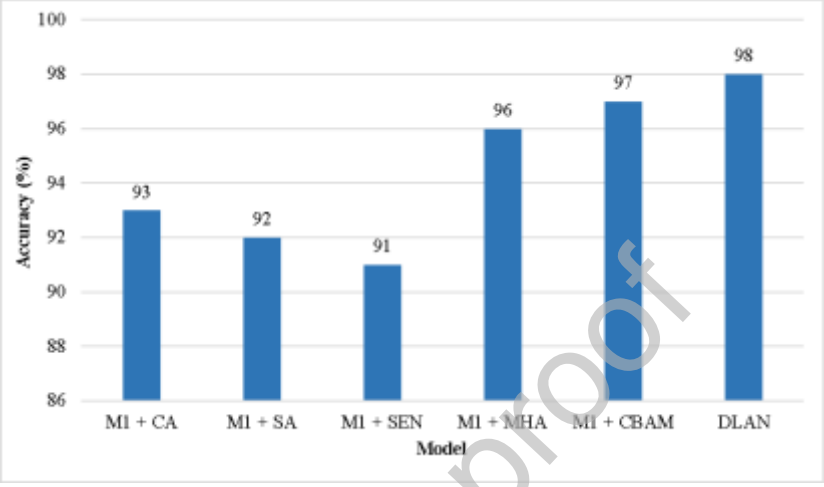

Figure 6. Comparative analysis of different attention modules integrated with inceptionv3 architecture over UCmercedLandUse dataset.

To further validate the robustness and generalization performance of the proposed model, we evaluated it on real-world images collected from YouTube videos. These images were captured from different angles and under varying lighting conditions across multiple scene categories, including beaches, airports, and buildings. Specifically, the dataset includes images from Incheon International Airport in Seoul, Haeundae Beach in Busan, and various buildings in Seoul. Since drone-based data collection at such heights is restricted in South Korea, YouTube video frames serve as a viable alternative for testing real-world applicability. The results are as given in **Figure 7**. demonstrate the model's ability to accurately identify targets despite variations in viewpoint and illumination, further supporting its robustness in diverse environments.

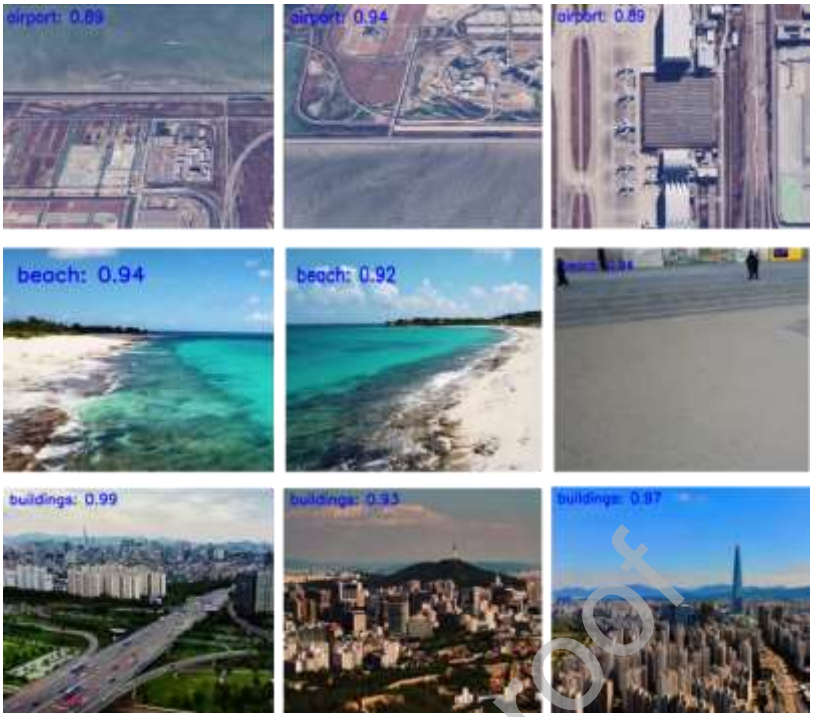

Figure 7. Generalization performance evaluation using real-world images from YouTube videos, captured from different angles and lighting conditions across various scenes.

Comparative Analysis:

We evaluated the efficacy of the dual land cover attention network (DLAN) when compared with SOTA methods based on CNNs, assessing their respective performances. We utilize various benchmark datasets, including UC Merced Land_Use dataset, to demonstrate the applicability of dual attention in LCC and localization. To ensure an unbiased assessment, we conducted an ablation study on the suggested approach, employing various baseline CNN models.

Table 3 provides the ablation study results for the proposed DLAN leveraging the suggested UC Merced Land_Use dataset. The outcomes indicate that the DLAN surpasses SOTA methods on all existing datasets, as evidenced by superior values regarding accuracy, precision, recall, and F1 score.

The evaluation of the dual attention network's effectiveness compared to Traditional Machine Learning approaches was conducted using the UC Merced Land cover, AID and NWPU datasets. The assessment and comparison were based on the evaluation parameters specified in references (S. Li, Yan, & Liu, 2020; Talukdar, et al., 2020). **Table 4** illustrates that the dual attention network demonstrated superior performance compared to existing Traditional Machine Learning methods on the specified datasets. Concerning the False Negative (FN) rate in AID dataset, the top-performing methods were (Ekim & Sertel, 2021). Additionally, our model enhanced the Overall Accuracy from 94% to 98.00%. Nevertheless, the DLAN attained the maximum values for Accuracy, Precision, and F1 score, as outlined in **Table 4**, underscoring the strength and resilience of the proposed DLAN.

Therefore, to evaluate the efficiency of land cover categorization and localization, we conducted an assessment of the DLAN and Compressed DLAN (DLANComp) performance on different standard datasets, as detailed in **Table 4**. This evaluation specifically focuses on the model ability to classify and localize land cover types, acknowledging the importance of accurate performance metrics in understanding the capability of these models in handling diverse land cover scenarios. The finest efficiency on the UCmerced LandUse is obtained by DLAN, and the next best outcomes are attained with DLANComp; the lowest outcomes are acquired by EuroSat dataset. In the experiments, our techniques, as illustrated in **Table 4**, according to the other metrics, the DLAN and DLANComp had the highest values, while the second-highest values for these metrics belong to the DLANComp, which means that the proposed models are more stable than the SOTA Methods.

Table 4: Comparative analysis our model and other approaches on the pre-existing Datasets. The blue and Green Illustrates the Top and second-Top Performances.

Models / Methods	**AID**				**EuroSAT**				**NWPU**				**UCmercedLandUse**			
	P	R	F1	ACC	P	R	F1	ACC	P	R	F1	ACC	P	R	F1	ACC
DNNE (Ekim & Sertel, 2021)	95	95	95	95	-	-	-	-	-	-	-	-	-	-	-	-
ResNet (Dastour & Hassan, 2023; Fayaz, Nam, et al., 2024)	-	-	-	94	97	97	97	95	-	-	-	-	88	86	86	92
Google Net (Helber, et al., 2019)	-	-	-	94	-	-	-	98	-	-	-	-	-	-	-	97
InceptionV3(A. A. Adegun, Viriri, & Tapamo, 2023; Alem & Kumar, 2022b)	-	-	-	-	75	75	75	75	-	-	-	-	87	87	87	94
VGG19	-	-	-	-	-	-	-	-	-	-	-	-	90	88	88	94.3
CNN (Obianuju, Agwu, & Ikechukwu, 2021; Yamashkin, Yamashkin, Zanozin, Radovanovic, & Barmin, 2020)	97	97	97	96	-	-	-	88	94	94	94	94	-	-	-	-
Hybird model (Fayaz, Dang, et al., 2024a)	-	-	-	-	91	91	91	92	-	-	-	-	96	96	96	98
TEX-Net (Anwer, Khan, Van De Weijer, Molinier, & Laaksonen, 2018)	-	-	-	95	-	-	-	-	-	-	-	-	-	-	-	97
Bi LSTM (Vinaykumar, Babu, & Frnda, 2023)	-	-	-	97	-	-	-		93	93	93	96	-	-	-	-
VGG16 (Dastour & Hassan, 2023) [35]	-	-	-	89	79	79	79	79	-	-	-	-	-	-	-	95
ViT (Bazi, Bashmal, Rahhal, Dayil, & Ajlan, 2021)	-	-	-	95	-	-	-	-	-	-	-	93	-	-	-	98
LAN_{CA}	-	-	-	-	91	92	92	93	-	-	-	-	94	93	93	93
LAN_{SA}	-	-	-	-	91	91	91	92	-	-	-	-	93	94	94	92
$DLAN_{Comp}$	**91**	**92**	**92**	**92**	**93**	**92**	**95**	**95**	**85**	86	87	**97**	**97**	**96**	**96**	**97**
DLAN	**92**	**92**	**93**	**94**	**95**	**93**	**96**	**94**	**56**	97	86	**96**	**97**	97	**98**	**98**

In the land classification domain, **AID** datasets are among the most commonly utilized publicly accessible datasets. Traditional methods achieve the optimal false-negative (FN) ratio on this dataset. However, DLAN and DLANComp attain superior False Positive and Accuracy values, surpassing the

SOTA approaches, as illustrated in **Table 4**. AID is recognized as a very difficult dataset in the field of LCC. On AID dataset, the maximum R value is attained because of the graphical resemblance between the land cover classes. However, DLAN surpasses the SOTA deep models in precision, F1-score, and accuracy; the F1 score represents a single metric that balances both Precision and Recall concerns. Among these models, the proposed DLANComp attained the second-highest performance, whereas the DLAN got the top performance, as illustrated in **Table 3 and Table 4**. Therefore, the complete statistical assessment reveals that the proposed model delivers the top performance in handling challenging LCC tasks.

Qualitative analysis: We also examined the qualitative implementation of the DLAN by differentiating with different images that contain and those that lack specific land cover features, relying on the localization results. The outcomes depicted in **Figure 5** demonstrate the strength of the DLAN; it effectively identifies and detects distinct land cover territories within challenging sights. For every test image, we integrated activation maps of DLAN to highlight the most useful section of an image that the system focuses on. The activation of the classes reveals that the model excels at pinpointing areas with a high probability of featuring specific land cover types.

Figure 5 presents the pictorial outcomes of the DLAN for intricate samples from UCmercedLandUse dataset, showcasing the model's performance in handling diverse land cover scenarios. The first row represents the real scenes of LCC, whereas the 2nd and 6th rows depict classification performance; the 3rd and 5th rows are the localization performance. In the 2nd and 4th rows, the DLAN correctly localizes and classifies all input samples with respect to land cover characteristics.

Table 4, we presented the accuracy of the proposed compressed model, which is designed to efficiently decrease the count of parameters and overall dimension, while maintaining the efficiency of the DLAN unaffected. **Table 4** reveals that the compressed DLAN attains the second-best results across all assessment metrics, excluding the P metric. The lower performance in the P metric is attributed to misclassifying some challenging conditions. Nevertheless, the compressed model consistently delivers the next-best outcomes in terms of F1 score, suggesting a stable performance in the classification.

Furthermore, we conducted additional experiments using the SIRI WHU dataset, which includes a wide range of image categories with varying complexities to further verify the robustness of the proposed model. As shown in **Figure 8**, we compared the performance of our method with different methods such as Alem et al. (Alem & Kumar, 2022a), Wenyi et al. (Hu, et al., 2024), Linjuan et al. (L. Li, Zhang, Xie, & Zhang, 2024), Adegun et al. (A. Adegun, Viriri, & Tapamo, 2024), Raju et al. (Raju, Natarajan, & Vasamsetty, 2022) , and Hussain et al. (Albarakati, et al., 2024), on SIRI WHU dataset. The proposed model achieved superior performance compared to other methods with 98% precision,

97% recall, and 98% accuracy. The highest accuracy is achieved by Hussain et al., while other methods show either lower overall performance (like Alem et al.'s 87% scores) or imbalanced metrics (such as Linjuan et al.'s 91% precision versus 86% recall). Notably, the proposed model maintains balance between precision and recall (98% vs 97%), indicating equally strong performance in both avoiding false positives and identifying all relevant cases. This balanced verify robustness of proposed architecture over a challenging benchmark compared to baseline methods.

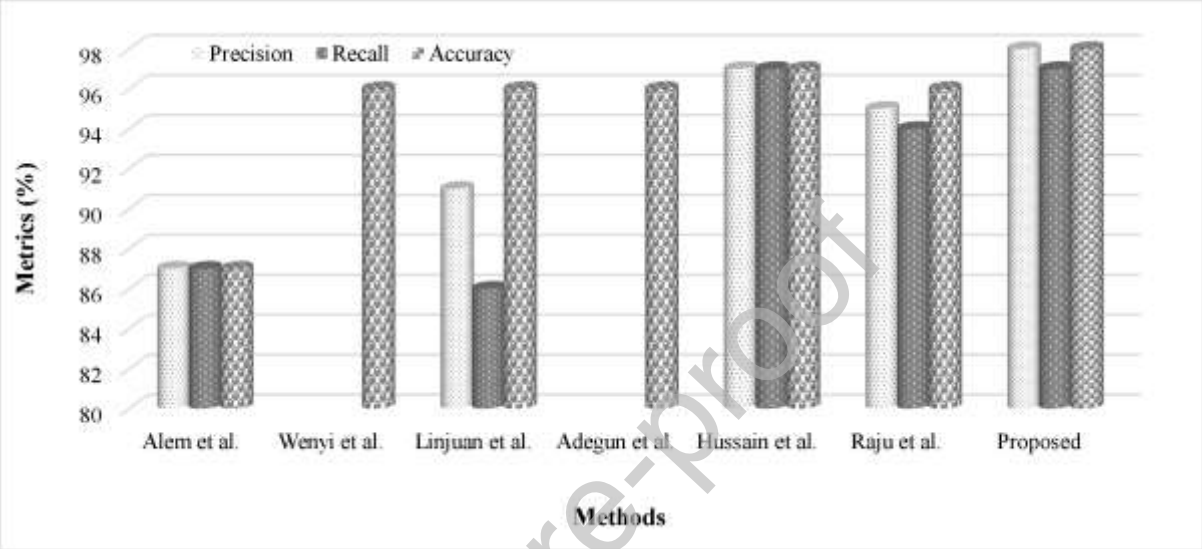

Figure 8: Comparative analysis our model and other approaches over SIRI WHU dataset.

Effect of Dual Attention:

In examining the resilience of different CNNs for LCC, this study primarily focused on assessing the effectiveness of the proposed dual attention module. The DLAN integrate CA and SA to address the challenges associated with LCC by effectively focusing on both the most informative channels and the most relevant spatial regions. Specifically, the SA helps the model identify critical spatial regions that are important for accurate classification while CA operates by analyzing inter-channel relationships, allowing the model to weigh the importance of different feature channels. The rationale behind combining these two attention mechanisms is to enhance the model's ability to capture both “what” (relevant features) and “where” (important regions) information, which significantly improves performance in more complex LCC tasks. The DLAN configuration proved to be the most effective, outdoing M+CA by an accurate margin of 2.00%. This highlights DLAN superior robustness and efficiency in LCC tasks compared to the other tested models. To further validate the effectiveness of this approach, additional comparative experiments will be conducted with other attention mechanisms, including CBAM, SENet, and Transformer-based attention. These experiments will provide a more comprehensive comparison and solidify the scientific justification for the proposed dual attention mechanism, demonstrating its superiority in LCC tasks.

Model Compression:

In this article we employed DE to reduce redundant neurons, thereby improving computational efficiency and enabling the model to operate more effectively in resource-constrained environments. DE is a meta-heuristic optimization process that iteratively adjusts the model's parameters by applying mutation, recombination, and selection to identify and remove unnecessary neurons. This optimization process is specifically designed to preserve the most essential features for classification while minimizing the number of parameters, resulting in a more efficient model. While model compression often presents a trade-off with classification accuracy, our approach ensures that the critical neurons and features necessary for LCC are retained. This allows us to maintain a high level of classification accuracy while reducing the model's complexity. To facilitate the implementation of the proposed LCC model in real-world settings with limited computational resources, we utilized DE to compress the DLAN model. We successfully reduced the model size from approximately 84 MB to 43 MB, and the total number of model parameters decreased from 23851874 to 13,385,649 with a slight decline from 98.00% to 97.00%. This demonstrates that our model compression performs a beneficial balance between computational cost and accuracy.

Table 5: Comparative analysis of inception V3 model filters before and after applying the compression in the convolutional

Layer	Original filters	Reduced filters
i	320	147
ii	448	217
Iii	384	186
Iv	384	160
v	320	139
vi	448	205
vii	284	184
viii	284	170

The specific layers affected were mixed 4 through mixed 7, which were refined, and mixed 9 and mixed 10, which underwent compression through the DE process. The details of the decreased filters in the layers are provided in **Table 5**. The visual representation demonstrates that the compressed model effectively focuses on land regions in a manner comparable to the original model.

Time Complexity: In this study, we evaluate the computational time and accuracy trade-off of our proposed model, DLAN, by comparing it against several state-of-the-art (SOTA) lightweight models. Computational complexity and model size are the two primary factors influencing deep learning model inference time. As illustrated in Table 6, We assess how the proposed DLAN and compressed DLAN models perform in terms of processing time and accuracy, and compare these results with other models like ResNet, InceptionV3, and EfficientNet. To further explore the trade-off between computation time and accuracy, we present a detailed comparison of computational time versus accuracy across different models, highlighting the balance between efficiency and performance. This analysis underscores the practical implications of our method, especially in real-world applications where both speed and accuracy are critical.

Table 6: Comparative analyses of different methods in terms of parameters(millions), model size (MB), and latency(sec).

Methods	Parameters (in million)	Model size (MB)	Latency (sec)	Accuracy (%)
ResNet-101 (K. He, Zhang, Ren, & Sun, 2016; Jamali, et al., 2021)	44	171	90	87
Xception (Chollet, 2017)	22	88	90	94
Inception-V3(C Szegedy, 2016)	23.9	92	67	--
MobileNet-V2 (M Sandler, 2018)	3.5	14	87	--
DenseNet-121(M Shafiq, 2022)	20.2	80	92	--
EfficientNet-B0 (C Szegedy, 2016)	5.3	29	84	--
VGG16 (Simonyan & Zisserman, 2014)	138	528	93	--
Deep Ensembled model (Fayaz, Dang, & Moon, 2024b)	64.7	247	32	97
DVIT (Bansal & Tripathi, 2024)	6.6	--	--	88
SwinTransformer (Z. Liu, et al., 2021)	4.6	--	--	87
PlantXViT (Thakur, Khanna, Sheorey, & Ojha)	4.6	--	--	90
DLAN	**23 M**	**84 Mb**	**71 sec**	**98**
$DLAN_{Comp}$	**13 M**	**43 Mb**	**36 sec**	**97**

Discussion:

The proposed method demonstrates significant potential for deployment in drone-based aerial surface image recognition. While this study primarily focuses on satellite imagery, drones equipped with

high-resolution cameras can offer real-time, detailed images of land surfaces. These drones can capture images at various altitudes, angles, and times, providing an advantage over satellite-based systems. The proposed model, with its dual attention mechanisms Channel Attention (CA) and Spatial Attention (SA) is well-suited for integration into drone systems. The CA mechanism helps analyze inter-channel relationships, while the SA mechanism allows the model to focus on critical spatial regions, making it particularly effective in aerial imagery where varying altitudes, angles, and conditions present unique challenges for classification.

Deploying this model on drone systems would enable efficient, on-the-fly land cover classification, a key feature for various real-world applications. These include urban monitoring, agricultural land assessment, and disaster response, where drones can offer high-resolution images in areas where satellite data may not be as effective. Real-time processing is crucial in these applications, and the proposed model's efficient feature extraction ensures that it can work within the constraints of drone systems. Furthermore, by focusing on both "what" (important features) and "where" (relevant regions) information, the model is robust enough to handle complex, noisy environments in aerial imagery, ensuring high classification accuracy even under variable conditions. However, several challenges need to be addressed to adapt this model for drone-based platforms, particularly in real-time processing and computational efficiency. Drones typically operate under dynamic conditions—such as varying altitudes, lighting, and weather which can affect image quality. Future work will focus on optimizing the model for real-time deployment, ensuring it handles these variations effectively. Additionally, improving the model's robustness to environmental interference, such as cloud cover and atmospheric disturbances, will be a key focus. By further fine-tuning the model for drone-specific imagery and optimizing it for real-time use, the method can be fully integrated into drone systems, enhancing land cover classification in a range of dynamic, real-world conditions.

5. Conclusion and future work

In the domain of computer vision, leveraging CNNs has markedly advanced the efficacy of land cover classification techniques. Despite considerable progress, prevalent CNN-based methodologies for detecting land cover have encountered notable challenges. These include the misclassification of land scenes within complex environments, alongside issues pertaining to the impractical time complexities and substantial model sizes that impede their applicability. In response to these limitations, our research introduces a novel approach DLAN. This method is rooted in the extraction of deep features coupled with the implementation of a newly devised dual land attention mechanism. Our approach further incorporates an innovative model compression strategy aimed at strengthening the model's efficiency while concurrently minimizing its size. The evaluation of DLAN was conducted using a uniquely assembled dataset, characterized by its imbalance, diversity, and the high level of challenge it presents, thereby establishing a new benchmark for LCC. Through rigorous experimenting on

standard datasets and comparative analysis against SOTA methods, our findings reveal that DLAN achieves an optimal balance in terms of accuracy, processing speed, and model compactness. Such a balance underscores the utility of our vision-based method for land cover classification across a wide array of application domains, including but not limited to forested regions, roadways, demanding outdoor environments, and industrial areas. This versatility is attributed to the comprehensive and varied nature of our training dataset.

Looking forward, the research aims to extend the capabilities of DLAN by integrating multi-modal remote sensing data, including LiDAR, SAR, and hyperspectral imagery, to enhance classification performance in diverse landscapes. Additionally, the adoption of semi-supervised learning approaches will be explored to improve model generalization by leveraging both labeled and unlabeled data, making DLAN more adaptable to real-world scenarios with limited annotated datasets.

Furthermore, there is an ambition to enhance the precision of the model in delineating land regions within images through the inclusion of object detection or semantic segmentation models a facet not currently addressed by DLAN. These advancements will not only refine the accuracy of land cover detection but also expand the model's applicability in resource-constrained and challenging surveillance contexts, significantly contributing to the field of computer vision and land cover analysis.

Acknowledgement

This work was supported by Basic Science Research Program through the National Research Foundation of Korea (NRF) funded by the Ministry of Education (2020R1A6A1A03038540).

Refrences

Adegun, A., Viriri, S., & Tapamo, J.-R. 2024. Automated classification of remote sensing satellite images using deep learning based vision transformer. *Applied Intelligence*, 1-20.

Adegun, A.A., Viriri, S., & Tapamo, J.-R. 2023. Review of deep learning methods for remote sensing satellite images classification: experimental survey and comparative analysis. *Journal of Big Data, 10*, 93.

Ahmad, J., Jan, B., Farman, H., Ahmad, W., & Ullah, A. 2020. Disease detection in plum using convolutional neural network under true field conditions. *Sensors, 20*, 5569.

Albarakati, H.M., Khan, M.A., Hamza, A., Khan, F., Kraiem, N., Jamel, L., Almuqren, L., & Alroobaea, R. 2024. 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 Journal of Selected Topics in Applied Earth Observations and Remote Sensing*.

Alem, A., & Kumar, S., 2022a. Deep Learning Models for Remote Sensed Hyperspectral Image Classification, 2022 13th International Conference on Computing Communication and Networking Technologies (ICCCNT). Publishing, pp. 1-7.

Alem, A., & Kumar, S. 2022b. Transfer learning models for land cover and land use classification in remote sensing image. *Applied Artificial Intelligence, 36*, 2014192.

Amare, M.T., Demissie, S.T., Beza, S.A., & Erena, S.H. 2023. Land cover change detection and prediction in the Fafan catchment of Ethiopia. *Journal of Geovisualization and Spatial Analysis, 7*, 19.

Anwer, R.M., Khan, F.S., Van De Weijer, J., Molinier, M., & Laaksonen, J. 2018. Binary patterns encoded convolutional neural networks for texture recognition and remote sensing scene classification. *ISPRS journal of photogrammetry and remote sensing, 138*, 74-85.

Audebert, N., Le Saux, B., & Lefèvre, S. 2018. Beyond RGB: Very high resolution urban remote sensing with multimodal deep networks. *ISPRS journal of photogrammetry and remote sensing, 140*, 20-32.

Ayerdi, B., & Romay, M.G. 2015. Hyperspectral image analysis by spectral–spatial processing and anticipative hybrid extreme rotation forest classification. *IEEE Transactions on Geoscience and Remote Sensing, 54*, 2627-2639.

Bansal, K., & Tripathi, A.K. 2024. Dual level attention based lightweight vision transformer for streambed land use change classification using remote sensing. *Computers & Geosciences, 191*, 105676.

Bazi, Y., Bashmal, L., Rahhal, M.M.A., Dayil, R.A., & Ajlan, N.A. 2021. Vision transformers for remote sensing image classification. *Remote Sensing, 13*, 516.

C Szegedy, V.V., S Ioffe, J Shlens, Z Wojna. 2016. Rethinking the inception architecture for computer vision. *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2016*, 2818-2826.

Capliez, E., Ienco, D., Gaetano, R., Baghdadi, N., Salah, A.H., Le Goff, M., & Chouteau, F. 2023. Multi-Sensor Temporal Unsupervised Domain Adaptation for Land Cover Mapping with spatial pseudo labelling and adversarial learning. *IEEE Transactions on Geoscience and Remote Sensing*.

Chen, Y., Li, C., Ghamisi, P., Jia, X., & Gu, Y. 2017. Deep fusion of remote sensing data for accurate classification. *IEEE Geoscience and Remote Sensing Letters, 14*, 1253-1257.

Chollet, F. 2017. Xception: Deep Learning with Depthwise Separable Convolutions. *In Proceedings of the IEEE conference on computer vision and pattern recognition*, 1251-1258.

Dang, L.M., Danish, S., Khan, A., Alam, N., Fayaz, M., Nguyen, D.K., Song, H.-K., & Moon, H. 2024. An efficient zero-labeling segmentation approach for pest monitoring on smartphone-based images. *European Journal of Agronomy, 160*, 127331.

Dash, P., Sanders, S.L., Parajuli, P., & Ouyang, Y. 2023. Improving the accuracy of land use and land cover classification of landsat data in an agricultural watershed. *Remote Sensing, 15*, 4020.

Dastour, H., & Hassan, Q.K. 2023. A comparison of deep transfer learning methods for land use and land cover classification. *Sustainability, 15*, 7854.

Ekim, B., & Sertel, E. 2021. Deep neural network ensembles for remote sensing land cover and land use classification. *International Journal of Digital Earth, 14*, 1868-1881.

Fayaz, M., Dang, L.M., & Moon, H. 2024a. Enhancing Land Cover Classification via Deep Ensemble Network. *Knowledge-Based Systems*, 112611.

Fayaz, M., Dang, L.M., & Moon, H. 2024b. Enhancing land cover classification via deep ensemble network. *Knowledge-Based Systems, 305*, 112611.

Fayaz, M., Nam, J., Dang, L.M., Song, H.-K., & Moon, H. 2024. Land-Cover Classification Using Deep Learning with High-Resolution Remote-Sensing Imagery. *Applied Sciences, 14*, 1844.

Frimpong, B.F., Koranteng, A., Atta-Darkwa, T., Junior, O.F., & Zawiła-Niedźwiecki, T. 2023. Land Cover Changes Utilising Landsat Satellite Imageries for the Kumasi Metropolis and Its Adjoining Municipalities in Ghana (1986–2022). *Sensors, 23*, 2644.

He, D., Shi, Q., Xue, J., Atkinson, P.M., & Liu, X. 2023. Very fine spatial resolution urban land cover mapping using an explicable sub-pixel mapping network based on learnable spatial correlation. *Remote Sensing of Environment, 299*, 113884.

He, K., Zhang, X., Ren, S., & Sun, J., 2016. Deep residual learning for image recognition, Proceedings of the IEEE conference on computer vision and pattern recognition. Publishing, pp. 770-778.

Helber, P., Bischke, B., Dengel, A., & Borth, D. 2019. Eurosat: A novel dataset and deep learning benchmark for land use and land cover classification. *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing, 12*, 2217-2226.

Hu, W., Lan, C., Chen, T., Liu, S., Yin, L., & Wang, L. 2024. Scene Classification of Remote Sensing Image Based on Multi-Path Reconfigurable Neural Network. *Land, 13*, 1718.

Hughes, L.H., Schmitt, M., Mou, L., Wang, Y., & Zhu, X.X. 2018. Identifying corresponding patches in SAR and optical images with a pseudo-siamese CNN. *IEEE Geoscience and Remote Sensing Letters, 15*, 784-788.

Hussain, A., Ul Amin, S., Fayaz, M., & Seo, S. 2023. An Efficient and Robust Hand Gesture Recognition System of Sign Language Employing Finetuned Inception-V3 and Efficientnet-B0 Network. *Computer Systems Science & Engineering, 46*.

Iervolino, P., Guida, R., Riccio, D., & Rea, R. 2019. A novel multispectral, panchromatic and SAR data fusion for land classification. *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing, 12*, 3966-3979.

Jamali, A., Mahdianpari, M., Brisco, B., Granger, J., Mohammadimanesh, F., & Salehi, B. 2021. Comparing solo versus ensemble convolutional neural networks for wetland classification using multi-spectral satellite imagery. *Remote Sensing, 13*, 2046.

Khan, Z.A., Ullah, F.U.M., Yar, H., Ullah, W., Khan, N., Kim, M.J., & Baik, S.W. 2025. Optimized cross-module attention network and medium-scale dataset for effective fire detection. *Pattern Recognition, 161*, 111273.

Kulkarni, S.C., & Rege, P.P. 2020. Pixel level fusion techniques for SAR and optical images: A review. *Information Fusion, 59*, 13-29.

Li, L., Zhang, H., Xie, G., & Zhang, Z. 2024. Robust Remote Sensing Scene Interpretation Based on Unsupervised Domain Adaptation. *Electronics, 13*, 3709.

Li, R., Gao, X., Shi, F., & Zhang, H. 2023. Scale Effect of Land Cover Classification from Multi-Resolution Satellite Remote Sensing Data. *Sensors, 23*, 6136.

Li, S., Yan, Q., & Liu, P. 2020. An efficient fire detection method based on multiscale feature extraction, implicit deep supervision and channel attention mechanism. *IEEE Transactions on Image Processing, 29*, 8467-8475.

Lin, T.-H., Li, H.-T., & Tsai, K.-C. 2004. Implementing the Fisher's Discriminant Ratio in ak-Means Clustering Algorithm for Feature Selection and Data Set Trimming. *Journal of chemical information and computer sciences, 44*, 76-87.

Liu, Q., Kampffmeyer, M., Jenssen, R., & Salberg, A.-B. 2020. Dense dilated convolutions' merging network for land cover classification. *IEEE Transactions on Geoscience and Remote Sensing, 58*, 6309-6320.

Liu, Z., Lin, Y., Cao, Y., Hu, H., Wei, Y., Zhang, Z., Lin, S., & Guo, B., 2021. Swin transformer: Hierarchical vision transformer using shifted windows, Proceedings of the IEEE/CVF international conference on computer vision. Publishing, pp. 10012-10022.

Lv, Z., Zhang, P., Sun, W., Benediktsson, J.A., Li, J., & Wang, W. 2023. Novel adaptive region spectral-spatial features for land cover classification with high spatial resolution remotely sensed imagery. *IEEE Transactions on Geoscience and Remote Sensing*.

M Sandler, A.H., M Zhu, A Zhmoginov, LC Chen. 2018. MobileNetV2: Inverted Residuals and Linear Bottlenecks. *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2018*, 4510-4520.

M Shafiq, Z.G.-. 2022. Deep residual learning for image recognition: A survey. *applied Sciences, Volume 12*.

McDonnell, M.D. 2018. Training wide residual networks for deployment using a single bit for each weight. *arXiv preprint arXiv:1802.08530*.

Melgani, F., & Bruzzone, L. 2004. Classification of hyperspectral remote sensing images with support vector machines. *IEEE Transactions on geoscience and remote sensing, 42*, 1778-1790.

Meng, X., Xie, S., Sun, L., Liu, L., & Han, Y. 2023. Evaluation of temporal compositing algorithms for annual land cover classification using Landsat time series data. *International Journal of Digital Earth, 16*, 2574-2598.

Moharram, M.A., & Sundaram, D.M. 2023. Land Use and Land Cover Classification with Hyperspectral Data: A comprehensive review of methods, challenges and future directions. *Neurocomputing*.

Obianuju, N.L., Agwu, N., & Ikechukwu, O., 2021. Medium resolution satellite image classification system for land cover mapping in Nigeria: a multi-phase deep learning approach, Intelligent Computing: Proceedings of the 2021 Computing Conference, Volume 2. Publishing, pp. 1056-1072.

Qin, Y., Xiao, X., Dong, J., Zhang, G., Shimada, M., Liu, J., Li, C., Kou, W., & Moore III, B. 2015. Forest cover maps of China in 2010 from multiple approaches and data sources: PALSAR, Landsat, MODIS, FRA, and NFI. *ISPRS Journal of Photogrammetry and Remote Sensing, 109*, 1-16.

Raju, M.N., Natarajan, K., & Vasamsetty, C.S. 2022. Remote Sensing Image Classification Using CNN-LSTM Model. *Rev. d'Intell. Artif, 36*, 147-153.

Rizayeva, A., Nita, M.D., & Radeloff, V.C. 2023. Large-area, 1964 land cover classifications of Corona spy satellite imagery for the Caucasus Mountains. *Remote Sensing of Environment, 284*, 113343.

Shi, Q., He, D., Liu, Z., Liu, X., & Xue, J. 2023. Globe230k: A benchmark dense-pixel annotation dataset for global land cover mapping. *Journal of Remote Sensing, 3*, 0078.

Shu, X., Yang, J., Yan, R., & Song, Y. 2022. Expansion-squeeze-excitation fusion network for elderly activity recognition. *IEEE Transactions on Circuits and Systems for Video Technology, 32*, 5281-5292.

Simonyan, K., & Zisserman, A. 2014. Very deep convolutional networks for large-scale image recognition. *arXiv preprint arXiv:1409.1556*.

Stanimirova, R., Tarrio, K., Turlej, K., McAvoy, K., Stonebrook, S., Hu, K.-T., Arévalo, P., Bullock, E.L., Zhang, Y., & Woodcock, C.E. 2023. A global land cover training dataset from 1984 to 2020. *Scientific Data, 10*, 879.

Sukawattanavijit, C., Chen, J., & Zhang, H. 2017. GA-SVM algorithm for improving land-cover classification using SAR and optical remote sensing data. *IEEE Geoscience and Remote Sensing Letters, 14*, 284-288.

Talukdar, S., Singha, P., Mahato, S., Pal, S., Liou, Y.-A., & Rahman, A. 2020. Land-use land-cover classification by machine learning classifiers for satellite observations—A review. *Remote sensing, 12*, 1135.

Thakur, P., Khanna, P., Sheorey, T., & Ojha, A. Explainable vision transformer enabled convolutional neural network for plant disease identification: PlantXViT. arXiv 2022. *arXiv preprint arXiv:2207.07919*.

Ullah, W., Ullah, A., Hussain, T., Khan, Z.A., & Baik, S.W. 2021. An efficient anomaly recognition framework using an attention residual LSTM in surveillance videos. *Sensors, 21*, 2811.

Vinaykumar, V., Babu, J.A., & Frnda, J. 2023. Optimal guidance whale optimization algorithm and hybrid deep learning networks for land use land cover classification. *EURASIP Journal on Advances in Signal Processing, 2023*, 13.

Wang, Q., Gao, J., Lin, W., & Li, X. 2020. NWPU-crowd: A large-scale benchmark for crowd counting and localization. *IEEE transactions on pattern analysis and machine intelligence, 43*, 2141-2149.

Wang, X., Jiang, W., Deng, Y., Yin, X., Peng, K., Rao, P., & Li, Z. 2023. Contribution of land cover classification results based on Sentinel-1 and 2 to the accreditation of wetland cities. *Remote Sensing, 15*, 1275.

Xu, X., Li, W., Ran, Q., Du, Q., Gao, L., & Zhang, B. 2017. Multisource remote sensing data classification based on convolutional neural network. *IEEE Transactions on Geoscience and Remote Sensing, 56*, 937-949.

Xu, Y., Du, B., & Zhang, L., 2018. Multi-source remote sensing data classification via fully convolutional networks and post-classification processing, IGARSS 2018-2018 IEEE International Geoscience and Remote Sensing Symposium. Publishing, pp. 3852-3855.

Yamashkin, S.A., Yamashkin, A.A., Zanozin, V.V., Radovanovic, M.M., & Barmin, A.N. 2020. Improving the efficiency of deep learning methods in remote sensing data analysis: geosystem approach. *IEEE Access, 8*, 179516-179529.

Yan, X., Li, J., Smith, A.R., Yang, D., Ma, T., & Su, Y. 2023. Rapid Land Cover Classification Using a 36-Year Time Series of Multi-Source Remote Sensing Data. *Land, 12*, 2149.

Yang, Y., & Newsam, S., 2010. Bag-of-visual-words and spatial extensions for land-use classification, Proceedings of the 18th SIGSPATIAL international conference on advances in geographic information systems. Publishing, pp. 270-279.

Yar, H., Hussain, T., Agarwal, M., Khan, Z.A., Gupta, S.K., & Baik, S.W. 2022. Optimized dual fire attention network and medium-scale fire classification benchmark. *IEEE Transactions on Image Processing, 31*, 6331-6343.

Declaration of interests

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