DOI: 10.15849/IJASCA.241130.18

Int. J. Advance Soft Compu. Appl, Vol. 16, No. 3, November 2024
Print ISSN: 2710-1274, Online ISSN: 2074-8523

A Swin Transformer-based method for Classification of Land Use and Land Cover Images

Enas Ali Mohammed*[1], Amir Lakizadeh[2]

[1]Department of Computer Engineering and Information Technology, University of Qom, University of Kerbala, Iraq
e-mail: enas.ali@uokerbala.edu.iq

[2] Department of Computer Engineering and Information Technology, University of Qom, Iran
e-mail: lakizadeh@qom.ac.ir

Abstract

Satellite image classification plays a crucial role in land use analysis, environmental monitoring, and urban planning. Recent developments in computer vision have led to the development of algorithms for image classification that are becoming increasingly successful. These techniques are known as vision transformer. On the other hand, it is often important to overcome problems related with limited receptive fields and the need for complete training data if one wants optimum performance. This work aims to provide a fresh approach for enhancing the design of the Swin transformer thus improving the classification of land use and land cover on the Eurosat dataset. Depth-wise Separable Convolutional Multi head Self-attention (DWSC-MSA) methods are suggested to be included into Swin transformer blocks. This entails changing the Shifted Window Multi-Head Self-Attention (SW-MSA) in the decoder and encoder blocks respectively. The DWSC-MSA method enables the extraction and prioritizing of specific features, resulting in enhanced classification performance. We performed experiments on the Eurosat dataset using many additional commonly used transformers, including swin-tiny, swin-small, swin-base, crossvit, and convit. The experimental results showcase the efficacy of our suggested framework in capturing spatial relationships and improving feature representation, thus pushing the boundaries of land use and land cover classification.

Keywords: *Depth-wise Separable Convolution, Eurosat, Image Classification, SWIN transformer, Vision Transformers.*

1 Introduction

Remote Sensing is the method of gathering data from an object or scene by using reflected and emitted electromagnetic radiation, without any direct contact with the object or scene. These tasks are performed via satellite or airplane. These devices expedite the procedure, enabling the collection of photos from perilous and hard-to-reach areas, thereby covering a bigger expanse. Remote sensing imagery serves multiple purposes, including air quality

Received 2 September 2024; Accepted 20 November 2024

assessment, earthquake prediction, land management, and urban planning. Remote sensing imagery is extensively utilized for land use classification, rendering it an essential application of this technology. Land use classification is essential for understanding the biophysical ecosystems of the Earth and the effects of socio-economic development. While it is important to continue the steady expansion of urban regions, it is also necessary to control the haphazard growth in metropolitan areas. So, it is crucial to optimize the utilization of every square inch of land while developing land management planning models. Remote sensing imagery is highly effective in this context. It provides current information on land areas and also reveals changes in the land over time. By monitoring these changes, we can acquire understanding about global climate changes. Classification of remotely sensed pictures has been a prominent subject of study in computer vision during four decades. Previous research during last ten years concentrated on feature-based and texture analysis classification techniques. Recent research has mainly concentrated on

classification by utilizing Convolutional Neural Networks (CNNs), resulting in a notable enhancement in performance. Significant contributions in this field are made by AlexNet [1], ResNet [2], GoogleNet [3], SqueezeNet [4], and DenseNet [5]. CNNs have been widely employed in the domain of computer vision. Improving image classification with transformer-based design has made big steps forward. The original Vision Transformer model has exhibited favorable results relative to traditional CNN models.

The transformer model was initially presented by [6] and shown remarkable inference outcomes for Natural Language Processing (NLP) tasks. Since the introduction of the transformer network [6], [7], [8], [9], researchers have subsequently utilized this technique to other computer vision problems [10], [11], [12], [13]. Conventional neural networks (CNNs) just identify image features and lack any positional information between these components, therefore limiting their capacity to understand the whole image. [14] provide an improved approach dubbed Vision Transformer (ViT), which deviates from traditional CNN methods by include self-attention layers, to solve this problem. This helps the model to completely grasp the images and reduces the special assumptions related to every image. Every block has a Normalizing Layer used to eliminate the interdependencies among input images. This method improves the generalization of the model overall than Convolutional Neural Networks. [14] introduces Vision Transformer architecture for vision tasks. First pretrained on the JFT-300M [15] and ImageNet datasets [1], the ViT model then is fine-tuned on a medium-sized dataset for classification. The authors demonstrate that their method is computationally less expensive than conventional convolutional neural networks. By creating a Data-efficient image Transformers (DeiT), which lowers the required data and processing resources for inference, [7] thus expand this study. Recent transformer models such as ViT and DeiT [7] have significant restrictions even if they provide advancements. One clear restriction is the capacity of these models to efficiently handle many picture domains of different sizes. [16] have proposed a new kind of transformer models known as the Swin Transformer to handle these problems. This model could infer or suggest the existence of objects of various diameters. It can therefore identify

a person in both front and the background. Researchers have been interested to the transformer model because of its excellent natural image classification accuracy.
As shown by the references [17], [18], [19], [20], [21], [22], [23], transformers have lately become a main player in the area of remote sensing. Transfer learning starts the parameters of the network model by using large datasets like ImageNet. Remote sensing images then help to fine-tune the network model thereby optimizing processing costs and improving classification results. This paper introduces a method for leveraging the Swin Transformer model [16] and transfer learning to extract features from remote sensing images. This approach derives the weight parameters by pre-training the Swin-T model using ImageNet dataset. Further training makes use of the pre-trained model using a remote sensing image dataset, therefore allowing model fine-tuning to improve remote sensing scene classification accuracy.

The stated contributions in this article are:

- The proposed approach offers a new architecture combining Swin transformer with Depth-wise Separable Convolution. This method enables the simultaneous use of CNNs and transformer features, therefore producing a model with excellent performance in applications requiring feature extraction from spatial and long-range data.
- Five variety of transformers were used in the land use and land cover (LULC) classification tasks using the EuroSAT dataset: Swin-Tiny, Crossvit, Convit, and the suggested DWSC-SWIN Small.
- Practical assessment, both with and without augmented data, evaluates transformers' performance using geometric data augmentations to improve variety of employed datasets.

2 Related Works on Satellite Image Classification

Considering the strength of deep learning, here is a review of various modern methods for satellite image classification derived from deep learning. Using vision transformers, the authors provided land cover image classification in [24] for the purpose to increase accuracy and efficiency in land cover image study. The work revealed that transformer-based methods—including ViT and Swin Transformer—outperform CNNs, producing cutting-edge results. Training and assessment based on the EuroSAT dataset [25], which consists of 10 land cover classifications derived from Sentinel-2 satellite images. Pre-trained weight models shown better accuracy than those learned from scratch according to validation accuracy curves. MaxViT's validation accuracy is 99.0%; SwinB's is 98.7%; ResNeXt's is 98.1%; DenseNet's is 92.5% when trained from scratch and using pre-trained weights respectively. These results show how well transformer models classify land covers, thereby improving environmental analysis and urban planning.

Combining the Swin Transformer model in [26] with transfer learning provides a fresh approach for remote sensing image scene classification. The model achieves remarkable accuracy on six different remote sensing datasets by means of pre-training on ImageNet datasets and migration learning. On UCM Dataset, validation results indicate incredible classification accuracy rates: 99.99%; on AID, they are 96.80%; on NWPU, they are 95.20%. This method shows how successfully transformer models and transfer learning improve image acquired for remote sensing categorization accuracy. The research presents interesting fresh perspectives for those considering transformer-based technology in remote sensing applications.

Another study [27] using the Swin Transformer model assessed it on three datasets: AID [29], EuroSAT, and NWPU-RESISC45 [28]. The Swin Transformer achieved 99.02% accuracy on the EuroSAT dataset, 95.38% accuracy on the NWPU-RESISC45 dataset, and 95.90% accuracy on the AID dataset with very impressive validation accuracy scores. The findings indicate that in remote sensing image classification the Swin architecture is more effective than present state-of- the-art approaches. The dependability and efficiency of the Swin Transformer for land cover classification tasks are supported by continuously excellent validation accuracies across many datasets.

The authors of [30] presented SCANeXt, a 3D medical image segmentation model combining two forms of attention mechanisms and depth-wise convolution within a transformer framework. On datasets including Synapse, BraTS, and ACDC, SCANeXt proved to be not just effective but also superior than other techniques. This model used channel-wise attention and Swin transformer-based spatial attention for comprehensive feature extraction. Training on datasets such as ACDC produced rather excellent results; SCANeXt obtained a Dice Similarity Coefficient (DSC) of over 95.18%, greater than other approaches under comparison. Better segmentation findings overall from SCANeXt, which indicates that medical image segmentation techniques may be enhanced using it.

The SparTa Block [31] presented to maximize transformer architecture for image classification. On datasets including ImageNet100 (86.96%), CIFAR10 (97.43%), and CIFAR100 (85.35%), this model significantly lowers parameters while nevertheless obtaining great accuracy. SparseSwin improves image classification problems by using the SparTa Block within the Swin Transformer architecture. ImageNet100, CIFAR10, and CIFAR100 among other datasets provide varied images for validation and training. The success of SparseSwin emphasizes how well it may improve performance and efficiency in computer vision applications.

The method SepViT[32], as new type of Vision Transformer, incorporates a depth wise separable self-attention module to improve the efficiency of the model. This module captures both local and global dependencies within and among the windows in a sequential manner. Depth-wise separable self-attention differs from normal self-attention by concentrating on interactions inside narrower windows instead of the complete input sequence, by utilizing the principles of depth wise separable convolution. SepViT achieves efficient computation of attention while preserving its expressive capacity. It improves the interaction between local and global information, therefore addressing the costly computational needs. The work uses the ADE20K dataset for semantic segmentation tasks and the ImageNet-1K dataset for image classification. On ImageNet-1K SepViT achieves an 84.2% validation accuracy and also lowers latency by 40% when compared to similar models like Cross-Shaped Window Transformer (CSWin). Strong performance on typical datasets shows that the model is efficient in balancing performance with computational savings.

In [1] a CNN-based model for trash image classification was created depending on the DSCAM for attention weighting and merging depth-wise separable convolutions. Its Resnet-50 backbone helped to boost feature extraction. Testing the method with five datasets—including Real Scene—had a high validation accuracy of 98.9%. The findings underlined how improved the proposed method is compared to traditional CNN architectures. Attention modules might be included into visual transformers for better classification and probable use in automated waste sorting systems in future directions. The work emphasizes the necessity of advanced techniques in improving garbage image classification accuracy and efficiency.For LULC classification, some relevant research has mainly been focused on the EuroSAT dataset.[25] classified using GoogleNet and ResNet-50 architectures. Reaching a confirmed accuracy of 98.57%, the ResNet-50 model outperformed the GoogleNet model with 98.18%, said the researchers.

The authors in Das et al. (2021), employed and fine-tunned different versions of the Res2Net50 architecture during their evaluation the method on the UC Merced[34], Brazilian Coffee Scenes[35], and EuroSAT[25] datasets. Their classification accuracies were reported as 98.76%, 93.25%, and 97.50% correspondingly. For regions with little data, the authors in [36] suggested a deep learning-based method to map land cover. They employed transfer learning, pre-training a ResNet-50 on EuroSAT (96% accuracy) and fine-tuning on a proposed Nigerian dataset (Nig_Images).

Similar to the image classification tasks, data augmentation and ensemble learning can improve model's generalization. The model attains an accuracy of 80% on Nig_Images, showcasing its potential for the advancement of developing nations. The application of transfer learning to classify LULC from high-resolution remote sensing images was studied by [37]. By means of a comparison between pre-trained VGG16 and Wide Residual Networks (WRNs), the researchers obtained a state-of- the-art validation accuracy of 99.17% for Wide ResNet-50 with geometric augmentation and 99.04% without geometric augmentation on the EuroSAT dataset (RGB bands). This paper shows how well transfer learning performs for chores involving land cover and usage.[38] used EuroSAT dataset to examine how well several deep learning models performed for the classification of remote sensing images. The testing accuracy results on the dataset revealed that ResNET50 achieved a testing accuracy of 92%, EfficientNET B0 achieved a testing accuracy of 91%, and VGG16 demonstrated a testing accuracy of 76%.

3 Methods

The proposed method is based on enhancing of the Swin Transformer by incorporating depth wise separable convolutions. Therefore, in this section, first, the architecture of Swin Transformer and Depth-Wise Separable Convolution is reviewed, then the proposed method is introduced.

3.1. Swin Transformer Architecture

Swin Transformer is a Hierarchical Vision Transformer using Shifted Windows that developed for vision tasks. Fig 1 shows the configuration of the Swin-S model. The dimensions of the supplied image are H×W×3. To begin with, the input image is partitioned into blocks of 4×4 pixels, resulting in a total number of blocks equal to H/4×W/4. Next, each block is compressed in the channel axis to yield a matrix with dimensions of H/4×W/4×48. The image features are ultimately acquired during Stage 1 to Stage 4. Stage 1 consists of two components: Linear Embedding and Swin Transformer Block. Stages 2 to 4 encompass the processes of patch merging and the implementation of the Swin Transformer Block. Stage 3 consists of many Swin Transformer Blocks. The purpose of Linear Embedding is to convert a matrix with dimensions of H/4×W/4×48 into a matrix with dimensions of H/4×W/4×C. Patch Merging in block 2 provides multi-scale feature extraction by reducing the size of the image by a factor of 2 twice. This results in a matrix of dimensions H/8×W/8×2C. The patch merging H and W are halved, while the channel dimension is doubled. After the patch is applied, the result is a merged matrix in Stage 3 with dimensions H/16×W/16×4C.

The primary components of the Swin Transformer Block are the window multi-head self-attention (W-MSA) mechanism and the shift window multi head self-attention mechanism (SW-MSA). The window-based multi-head self-attention mechanism divides the input image into separate windows, each containing several blocks. These windows serve as the units for self-attention calculation, effectively lowering computing complexity.

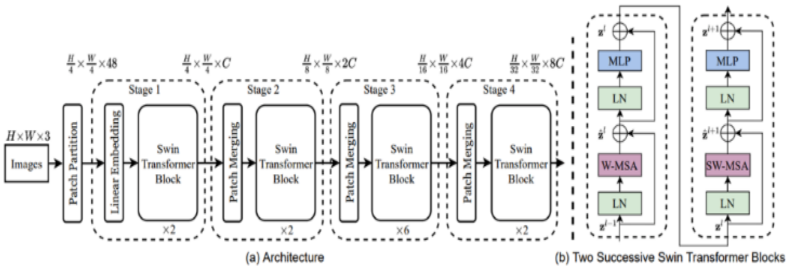


Fig 1. Original architecture of Swin small Transformer [2].

The multi-head self-attention mechanism of the shifted window operates by modifying the arrangement of pixels in an image by window repositioning. This allows for the extraction of feature information from diverse points inside the newly formed windows. The self-attention calculation of the new window facilitates information interaction at various places of the input image. The self-attention mechanism [2] is used to calculate the equation indicated in (1).

$$attention = (Q, K, V) = Softmax(QK^T/\sqrt{d} + B)\mathrm{V} \qquad (1)$$

The query matrix, key matrix, and value matrix are denoted as Q, K and $\mathrm{V} \in R^{M^2*d}$, d correspondingly. B is the relative position offset matrix, and it also belongs to $R^{M^2*M^2}$. M is the number of blocks in the window. The GELU function serves as a non-linear activation function in the Swin Transformer Block. It enhances the training process by accelerating it and has commendable stability and generalization capabilities. The equation (2) described as following:

$$GELU(x) = \frac{1}{2x\left(1+erf\left(\frac{x}{\sqrt{2}}\right)\right)} \qquad (2)$$

$$\approx 1/2(\tanh(\sqrt{\frac{2}{\pi}}(x + 0.044715x^3)))$$

3.2. Depth-Wise Separable Convolution (DWSC)

DWSC is an efficient type of convolution that factorizes a standard convolution operation into two simpler operations: depth-wise convolution and pointwise convolution (1x1 convolution). This factorization significantly reduces the number of parameters and computational cost while preserving the representational power of the convolutional layer. In a standard convolutional layer, the input feature map $X \in R^{H\times W\times C_{in}}$ is convolved with a set of filters $W \in R^{k\times k\times C_{in}\times C_{out}}$, where H and Ware the height and width of the feature map, C_{in} is the number of input channels, C_{out} is the number of output channels, and k is the kernel size. The output feature map $Y \in R^{H\times W\times C_{out}}$ is given by (3):

$$Y(i,j,C_{out}) = \sum_{m=1}^{k}\sum_{n=1}^{k}\sum_{C_{in}=1}^{C_{in}} X(i+m, j+n, c_{in}). W(m,n,c_{in},c_{out}) \qquad (3)$$

The total number of parameters in a standard convolutional layer is $k\times k\times C_{in}\times C_{out}$. DWSC decomposes the standard convolution into two separate layers:

1. Depth-wise Convolution: Each input channel is convolved independently with a set of depth-wise filters $W_{depth} \in R^{k\times k\times C_{in}}$. This operation produces an intermediate feature map $X_{depth} \in R^{H\times W\times C_{in}}$ as shown in (4).

$$X_{depth}(i,j,c_{in}) = \sum_{m=1}^{k}\sum_{n=1}^{k} X(i+m, j+n, c_{in}). W_{depth}(m,n,c_{in}) \qquad (4)$$

The number of parameters in the depth-wise convolution is $k\times k\times C_{in}$.

2. Pointwise Convolution (1x1 Convolution): A 1x1 convolution is then applied to combine the depth-wise-convolved outputs across different channels using pointwise filters
$W_{point} \in R^{1\times 1\times C_{in}\times C_{out}}$ as described in (5)

$$Y(i,j,c_{out}) = \sum_{c_{in}=1}^{c_{in}} X_{depth}(i,j,c_{in}). W_{point}(1,1,c_{in},c_{out}) \qquad (5)$$

The number of parameters in the pointwise convolution is $1\times1\times C_{in}\times C_{out}$. By factorizing the standard convolution into depth-wise and pointwise convolutions, the total number of parameters in a depth-wise separable convolutional layer is significantly reduced. The total number of parameters is: Total Parameters DWSC=$k\times k\times C_{in}+1\times1\times C_{in}\times C_{out}$. Compared to the number of parameters in a standard convolution: Total Parameters Standard=$k\times k\times C_{in}\times C_{out}$. The reduction in the number of parameters is particularly notable when the number of output channels C_{out} is large. To calculate ratio as following:

$$Ratio = \frac{Total\ ParametersStandard}{Total\ ParametersDWSC}$$

$$Ratio = \frac{k^2\times C_{in}\times C_{out}}{C_{in}(k^2+C_{out})} =$$

$$Ratio = \frac{1}{C_{out}} + \frac{1}{k^2} \qquad (6)$$

This ratio described in (6) shows the relative number of parameters in a DWSC compared to a standard convolution.

3.3. The Proposed Method

In this study, we introduce a novel approach to enhancing the Swin Transformer architecture for land use and land cover classification using the Eurosat dataset. Our methodology was done on encoder and decoder blocks. The proposed model comprises four stages, with each stage utilizing a Swin Transformer Block. The primary enhancement involves substituting the SW-MSA with DWSC-MSA within these blocks. The detailed architecture is illustrated in Fig 2. The proposed mechanism for DWSC-MSA blocks can be detailed as follows:

1. *Input Processing:* The input image $X \in R^{B\times H\times W\times C}$ is first normalized using layer normalization. Where B is the batch size, H and W are the height and width of the feature map, and C is the number of channels. This step ensures that the input features are standardized across the feature dimension, which helps stabilize and accelerate training.

2. *Depth-wise Separable Convolution (DWSC) Layer:* The normalized input is processed through the DWSC layer, which consists of two main components:
   - Depth-wise Convolution**:** Each input channel is convolved independently using a depth-wise convolutional filter. This operation captures spatial relationships within each channel without combining information across channels, as in (7).

$$X^{(c)}_{depthwise} = X^{(c)} * K^{(c)}_{depthwise} \qquad (7)$$

For each chanel c, where * denotes the convolution operation and $K^{(c)}_{depthwise}$ is the depth-wise convolution kernel for channel c.

Output: $X_{depthwise} \in R^{B\times H\times W\times C}$.

- Pointwise Convolution: Following depth-wise convolution, a 1x1 pointwise convolution is applied to combine the depth-wise-convolved outputs across different channels. This step ensures efficient interaction and combination of features from different channels, as in (8).

$$\mathrm{X_{pointwise}} = \mathrm{X_{depthwise}} * \mathrm{K_{pointwise}} \qquad (8)$$

$K_{pointwise}$ is the pointwise convolution kernel with dimensions 1×1×C×M.
Output: $\mathrm{X_{dwsc}} \in \mathrm{R^{B\times H\times W\times M}}$ where M is the number of output channels.

3.*Multi-Head Self-Attention (MultiHead) Layer:* The output from the DWSC layer serves as the input for the multi-head self-attention mechanism. This layer operates as follows:

- Linear Projections: The input $\mathrm{X_{dwsc}}$ is projected into queries (Q), keys(K), and values (V) matrices using learned linear projection weights.

$$Q = X_{dwsc}W^Q, K = X_{dwsc}W^K, V = X_{dwsc}W^V.$$

Where $W^Q, W^K, W^V \in R^{M\times d_k}$ are learned projection matrices and dk is the dimensionality of the queries and keys.

- Scaled Dot-Product Attention: For each attention head, the attention weights are computed using the scaled dot-product of the queries and keys. These attention weights are then applied to the value matrix, allowing the model to focus on different parts of the input feature map simultaneously, as described in (9).

$$Attention(Q, K, V) = softmax\left(\frac{QK^T}{\sqrt{d_k}}\right)V \qquad (9)$$

- Multi-Head Attention: The outputs from all attention heads are concatenated and linearly projected back to the original input dimension, resulting in the multi-head attention output.

$$\mathrm{MultiHead(Q, K, V) = Concat(head_1, head_2, .., head_h)}$$

Where $head_i = Attention\left(Qw_i^Q, KW_i^K, VW_i^V\right)$. Then, project the concatenated output back to the original dimension, as in equation (10).

$$MultiHeadOutput = Concat(head_1, head_2, \dots, head_h)W^O \quad (10)$$

Where $W^O \in R^{hd_k\times M}$ is the output projection matrix.

4.*Layer Normalization and Multi-Layer Perceptron (MLP):* The output of the multi-head self-attention layer undergoes another layer normalization, followed by a feedforward neural network (MLP) consisting of two fully connected layers with a ReLU activation in between. This step further refines the feature representation as in (11).

$$\mathrm{X_{norm} = LN(MultiHeadOutput)}$$

$$X_{mlp} = MLP(X_{norm})$$

$$MLP(X) = FC_2\left(ReLU\left(FC_1(x)\right)\right) \quad (11)$$

5.*Residual Connection:* An element-wise sum is performed to add the output of the MLP to the original input, creating a residual connection that helps mitigate the vanishing gradient problem and facilitates the flow of information through the network. The final output of DWSC-MSA is shown in (12):

$$X_{out} = X_{mlp} + X \qquad (12)$$

The enhanced model is implemented in four stages, each comprising multiple DWSC-MSA blocks and linear embedding layers to reduce the feature map dimensions progressively. The final output is used for classification and segmentation tasks, leveraging the extracted hierarchical feature representations.

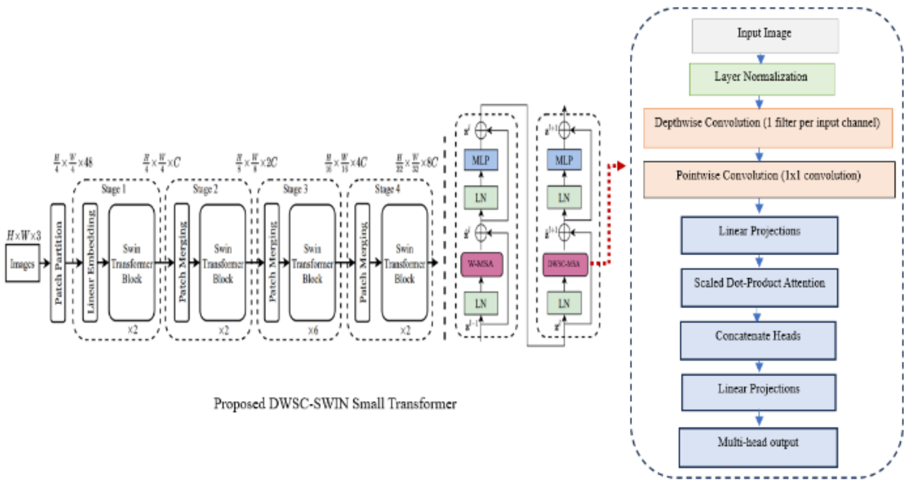


Fig 2. The architecture of the proposed method.

4 Experiments and Results

4.1. Dataset

The EuroSat collection comprises a total of 27,000 annotated and geo-referenced sentinel-2 pictures including 13 spectral bands. This dataset contains a comprehensive set of classes, which are mostly evenly distributed, that are essential for training and assessing our classification algorithms. The dataset consists of a total of 10 classes, including Industrial Buildings, Residential Buildings, Annual Crop, Permanent Crop, River, Sea and Lake, Herbaceous Vegetation, Highway, Pasture, and Forest. Each image in the European Urban Atlas covers cities with a spatial resolution of 10 meters per pixel and has a size of 64X64. The dataset was acquired directly from the primary source [3], guaranteeing its validity and the integrity of the data. The dataset will be divided into three parts: 70% for training, 20% for validation, and 10% for testing. A subset of this dataset is depicted in Fig 3.

Fig 3. Sample images from Eurosat dataset in RGB form.

4.2. Data Preprocessing and Training

We created specialized code to optimize the loading, preparation, and enhancement of data. The preprocessing procedures encompassed the normalizing and scaling of photographs. We conducted training on five transformer-based architectures: Swin tiny, Swin small, Swin base [2], Convit small [4], and Crossvit small [5]. Additionally, we trained the suggested architecture DWSC-SWIN small. All models were trained for 25 epochs and the model with the highest validation accuracy was selected. We employed the categorical cross-entropy loss function, a batch size of 32, and the Adam optimizer with a learning rate of 5e-4. Afterwards, we utilized transfer learning by training the models with pre-trained weights. The weights were acquired using models that were pre-trained on ImageNet. The dataset was randomly divided into three parts: 70% for training, 20% for validation, and 10% for testing.

4.3. Model Evaluation

We evaluated each model on the test set, which was not used during the training phase. As evaluation metrics, we compute the Top-1 accuracy and the precision/recall curves. For our experiments, we used Pytorch: A Colab pro+ machine with NVIDIA V100 GPU.

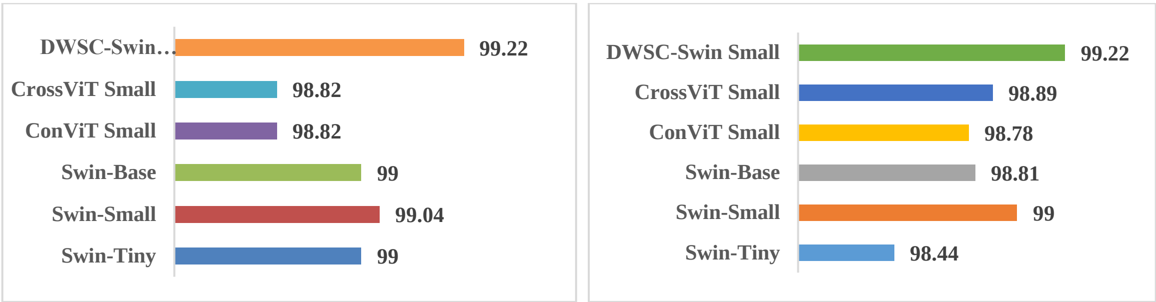
4.4. The Effect of Using Geometric Data Augmentation Techniques

Our enhanced DWSC-Swin Transformer model demonstrates superior classification performance on the EuroSat dataset, surpassing the baseline and other state-of-the-art Transformer architectures. The integration of DWSC-MSA blocks significantly improves the model's ability to capture and prioritize relevant features, resulting in more accurate land use and land cover classification. Geometric augmentation techniques, including rotation, horizontal flipping, and vertical flipping, were employed in this study to enhance the robustness of the model. Fig 4-(a) and Fig 4-(b) show the best validation accuracies for Transformer models with and without implementing geometric augmentation. As illustrated, the proposed DWSC-Swin

Small model achieved the highest accuracy in both scenarios, demonstrating its robustness and effectiveness.

- No Data Augmentation (Baseline Data): The DWSC-Swin Small model achieved a validation accuracy of 99.22%, outperforming CrossViT Small (98.82%), ConViT Small (98.82%), Swin-Base (99%), Swin-Small (99.04%), and Swin-Tiny (99%).
- Geometric Augmentation: With geometric augmentation, the DWSC-Swin Small model maintained its leading performance with a validation accuracy of 99.22%, compared to CrossViT Small (98.89%), ConViT Small (98.78%), Swin-Base (98.81%), Swin-Small (99%), and Swin-Tiny (98.44%).

These results confirm that our proposed model not only excels in accuracy but also demonstrates robustness across various data augmentation strategies, making it highly effective for land use and land cover classification on the EuroSat dataset.



(a) **(b)**

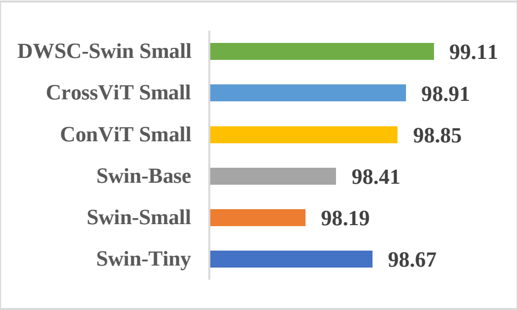
Fig 4. The comparison results in case of using Validating data without(a)\(b) with using Geometric data augmentation techniques.

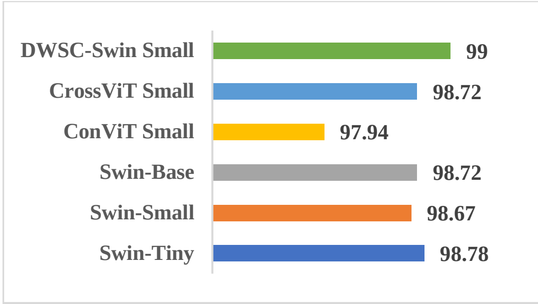
5 Discussion

Our results reveal that the proposed model much outperforms competing transformer-based designs in land use and land cover classification tasks. We have overcome the constraints of current models and thereby raised classification accuracy and efficiency by including DWSC-MSA methods into the Swin Transformer design.

5.1. Comparative Analysis

Our enhanced Swin Transformer, incorporating the Depth-wise Separable Convolutional Multi-Head Self-Attention (DWSC-MSA) mechanism, was evaluated against several state-of-the-art transformer architectures, including the original Swin Transformer, ConViT, and CrossViT. The evaluation metrics included Area Under the Curve (AUC), F1-score, and overall classification accuracy on the Eurosat dataset as shown in Fig (5), Table (1), and Table (2).





(a) **(b)**

Fig 5. Using testing data both with (a) and without (b) geometric data augmentation methods produce comparative results.

A- Baseline Eurosat Dataset

- Test Accuracy: The proposed DWSC-Swin Small model achieved a classification accuracy of 99.11%, outperforming Swin-Tiny (98.67%), Swin-Small (98.19%), Swin-Base (98.41%), ConViT Small (98.85%), and CrossViT Small (98.91%) as shown in Fig (5-a). This substantial improvement demonstrates the effectiveness of the DWSC-MSA mechanism in capturing spatial relationships and channel-wise dependencies more accurately than traditional self-attention mechanisms.

- AUC and F1-Score: Our model achieved an AUC of 0.9999 and an F1-score of 0.9911. In comparison, Swin-Tiny had an AUC of 0.9997 and an F1-score of 0.9867, Swin-Small had an AUC of 0.9996 and an F1-score of 0.9819, Swin-Base had an AUC of 0.9996 and an F1-score of 0.9840, ConViT Small had an AUC of 0.9997 and an F1-score of 0.9885, and CrossViT Small had an AUC of 0.9997 and an F1-score of 0.9891. The DWSC-Swin Small model's superior F1-score indicates a more balanced performance across different classes, essential for robust land use and land cover classification.

B- Eurosat Dataset with Geometric Augmentation

- Test Accuracy: On the augmented dataset, the DWSC-Swin Small model maintained its lead with a classification accuracy of 99.00%, compared to Swin-Tiny (98.78%), Swin-Small (98.67%), Swin-Base (98.72%), ConViT Small (97.94%), and CrossViT Small (98.72%), as shown in Fig(5-b). This result underscores the model's robustness to geometric transformations.

- AUC and F1-Score: The DWSC-Swin Small model achieved an AUC of 0.9998 and an F1-score of 0.99. Swin-Tiny had an AUC of 0.9997 and an F1-score of 0.9878, Swin-Small had an AUC of 0.9997 and an F1-score of 0.9867, Swin-Base had an AUC of 0.9997 and an F1-score of 0.9872, ConViT Small had an AUC of 0.9996 and an F1-score of 0.9795, and CrossViT Small had an AUC of 0.9996 and an F1-score of 0.9872. The superior performance of our model highlights its ability to maintain high accuracy and balanced precision and recall even under augmented conditions.

Overall, the proposed DWSC-Swin Small model consistently outperforms other transformers across various metrics, demonstrating its effectiveness for land use and land cover classification on the Eurosat dataset. The improvements can be attributed to the DWSC-MSA mechanism, which enhances feature extraction by capturing fine-grained spatial relationships and channel-wise dependencies more effectively than conventional self-attention mechanisms.

Table 1: The comparison results of the proposed method with the others based on Eurosat dataset.

Transformer	AUC	F1
Swin-Tiny	0.9997	0.9867
Swin-Small	0.9996	0.9819
Swin-Base	0.9996	0.9840
ConViT Small	0.9997	0.9885
CrossViT Small	0.9997	0.9891
DWSC-Swin Small	0.9999	0.9911

Table 2: Performance measures of Transformers on Eurosat dataset with Geometric Augmentation.

Transformer	AUC	F1
Swin-Tiny	0.9997	0.9878
Swin-Small	0.9997	0.9867
Swin-Base	0.9997	0.9872
ConViT Small	0.9996	0.9795
CrossViT Small	0.9996	0.9872
DWSC-Swin Small	0.9998	0.99

5.2. Computational Efficiency

Apart from improved accuracy, our proposed model exhibits less number of parameters than the original Swin Small model.

Faster training times and less memory use resulting from this reduction assist the model to be more scalable and efficient. Good depth-wise separable convolution parameter application aids in the decrease's accomplishment.

- The 49,606,258 parameters count of the first Swin small model.
- The new DWSCMSA-Swin compact model has 48,844,948 parameters.

This parameter reduction along with the higher validation accuracy highlights the efficacy of our approach in enhancing the Swin Transformer architecture for the specific goal of land use and land cover categorization.

5.3. Qualitative Analysis

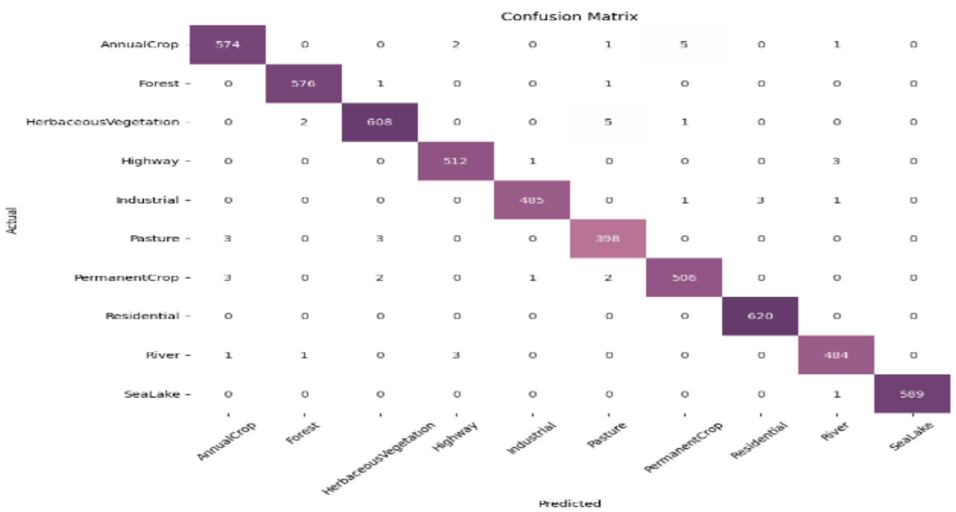
A visual examination of the classification results confirmed the efficacy of our strategy. The improved DWSC-Swin Transformer generated more precise and reliable segmentation maps, particularly in complex scenarios with varied land cover types. The model showed higher performance in differentiating between comparable categories, such as various agricultural field kinds, outperforming previous transformers and underscoring its outstanding feature extraction and classification abilities.

We provide confusion matrices in Fig. 6 that provide a complete picture of the model's classification performance to help to understand all of this. These matrices exhibit the perfect accuracy of the model in separating many land cover types, therefore demonstrating its dependability and strength. Including the confusion matrices gives a more complete and graphic view of the performance of the model, therefore strengthening your qualitative evaluation.

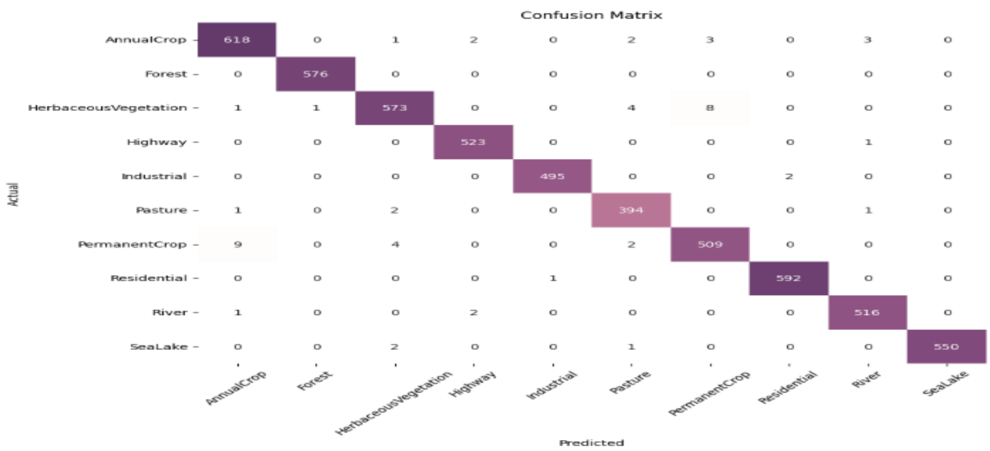
6 Conclusion and Future Work

Depth-wise Separable Convolutional Multi-Head Self-Attention combined into the proposed Swin Transformer architecture has shown quite remarkable efficiency in land use and land cover categorization. The potential of our technique for numerous remote sensing applications are highlighted by the increased accuracy, precision, recall, and processing economy.

Future research will examine more enhancements to the DWSC-MSA method, interactions with other advanced transformer designs, and other applications for extra remote sensing data. The positive results facilitate the continued use of transformer-based models in several fields of computer vision including geospatial analysis.



(a) No-Aug.



(b) Geo-Aug.

Fig 6. Confusing matrices result from test results on transformers on Baseline/Geometric Augmentation.

Data Availability Statement: Online copies of the relevant research datasets are available. For free one may get the EuroSAT dataset [25]. Upon request, the first author might provide datasets generated for this study.

References

[1] A. Krizhevsky, I. Sutskever, and G. E. Hinton, "ImageNet classification with deep convolutional neural networks," Commun ACM, vol. 60, pp. 84–90, 2012, [Online]. Available: https://api.semanticscholar.org/CorpusID:195908774

[2] K. He, X. Zhang, S. Ren, and J. Sun, "Delving Deep into Rectifiers: Surpassing Human-Level Performance on ImageNet Classification," 2015 IEEE International

Conference on Computer Vision (ICCV), pp. 1026–1034, 2015, [Online]. Available: https://api.semanticscholar.org/CorpusID:13740328

[3] C. Szegedy et al., "Going deeper with convolutions," 2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pp. 1–9, 2014, [Online]. Available: https://api.semanticscholar.org/CorpusID:206592484

[4] F. N. Iandola, M. W. Moskewicz, K. Ashraf, S. Han, W. J. Dally, and K. Keutzer, "SqueezeNet: AlexNet-level accuracy with 50x fewer parameters and <1MB model size," ArXiv, vol. abs/1602.07360, 2016, [Online]. Available: https://api.semanticscholar.org/CorpusID:14136028

[5] G. Huang, Z. Liu, and K. Q. Weinberger, "Densely Connected Convolutional Networks," 2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pp. 2261–2269, 2016, [Online]. Available: https://api.semanticscholar.org/CorpusID:9433631

[6] A. Vaswani et al., "Attention is All you Need," in Neural Information Processing Systems, 2017. [Online]. Available: https://api.semanticscholar.org/CorpusID:13756489

[7] H. Touvron, M. Cord, M. Douze, F. Massa, A. Sablayrolles, and H. J'egou, "Training data-efficient image transformers & distillation through attention," in International Conference on Machine Learning, 2020. [Online]. Available: https://api.semanticscholar.org/CorpusID:229363322

[8] J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova, "BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding," in North American Chapter of the Association for Computational Linguistics, 2019. [Online]. Available: https://api.semanticscholar.org/CorpusID:52967399

[9] T. Brown et al., "Language Models are Few-Shot Learners," in Advances in Neural Information Processing Systems, H. Larochelle, M. Ranzato, R. Hadsell, M. F. Balcan, and H. Lin, Eds., Curran Associates, Inc., 2020, pp. 1877–1901. [Online]. Available: https://proceedings.neurips.cc/paper_files/paper/2020/file/1457c0d6bfcb4967418bfb8ac142f64a-Paper.pdf

[10] X. W. Gao, Y. Qian, and A. Gao, "COVID-VIT: Classification of COVID-19 from CT chest images based on vision transformer models," ArXiv, vol. abs/2107.01682, 2021, [Online]. Available: https://api.semanticscholar.org/CorpusID:235731661

[11] H. Wu et al., "CvT: Introducing Convolutions to Vision Transformers," 2021 IEEE/CVF International Conference on Computer Vision (ICCV), pp. 22–31, 2021, [Online]. Available: https://api.semanticscholar.org/CorpusID:232417787

[12] C.-F. Chen, Q. Fan, and R. Panda, "CrossViT: Cross-Attention Multi-Scale Vision Transformer for Image Classification," 2021 IEEE/CVF International Conference on Computer Vision (ICCV), pp. 347–356, 2021, [Online]. Available: https://api.semanticscholar.org/CorpusID:232404237

[13] W. Wang et al., "Pyramid Vision Transformer: A Versatile Backbone for Dense Prediction without Convolutions," 2021 IEEE/CVF International Conference on

Computer Vision (ICCV), pp. 548–558, 2021, [Online]. Available: https://api.semanticscholar.org/CorpusID:232035922

[14] A. Dosovitskiy et al., "An Image is Worth 16x16 Words: Transformers for Image Recognition at Scale," ArXiv, vol. abs/2010.11929, 2020, [Online]. Available: https://api.semanticscholar.org/CorpusID:225039882

[15] C. Sun, A. Shrivastava, S. Singh, and A. K. Gupta, "Revisiting Unreasonable Effectiveness of Data in Deep Learning Era," 2017 IEEE International Conference on Computer Vision (ICCV), pp. 843–852, 2017, [Online]. Available: https://api.semanticscholar.org/CorpusID:6842201

[16] Z. Liu et al., "Swin Transformer: Hierarchical Vision Transformer using Shifted Windows," 2021 IEEE/CVF International Conference on Computer Vision (ICCV), pp. 9992–10002, 2021, [Online]. Available: https://api.semanticscholar.org/CorpusID:232352874

[17] D. Hong et al., "SpectralFormer: Rethinking Hyperspectral Image Classification With Transformers," IEEE Transactions on Geoscience and Remote Sensing, vol. 60, pp. 1–15, 2021, [Online]. Available: https://api.semanticscholar.org/CorpusID:235755242

[18] Z. Xue, X. Tan, X. Yu, B. Liu, A. Yu, and P. Zhang, "Deep Hierarchical Vision Transformer for Hyperspectral and LiDAR Data Classification," IEEE Transactions on Image Processing, vol. 31, pp. 3095–3110, 2022, [Online]. Available: https://api.semanticscholar.org/CorpusID:248100455

[19] Y. Qing, W. Liu, L. Feng, and W. Gao, "Improved Transformer Net for Hyperspectral Image Classification," Remote. Sens., vol. 13, p. 2216, 2021, [Online]. Available: https://api.semanticscholar.org/CorpusID:235421060

[20] A. Jamali and M. Mahdianpari, "Swin Transformer and Deep Convolutional Neural Networks for Coastal Wetland Classification Using Sentinel-1, Sentinel-2, and LiDAR Data," Remote. Sens., vol. 14, p. 359, 2022, [Online]. Available: https://api.semanticscholar.org/CorpusID:245980047

[21] H. Dong, L. Zhang, and B. Zou, "Exploring Vision Transformers for Polarimetric SAR Image Classification," IEEE Transactions on Geoscience and Remote Sensing, vol. 60, pp. 1–15, 2022, [Online]. Available: https://api.semanticscholar.org/CorpusID:245430425

[22] J. Yao, B. Zhang, C. Li, D. Hong, and J. Chanussot, "Extended Vision Transformer (ExViT) for Land Use and Land Cover Classification: A Multimodal Deep Learning Framework," IEEE Transactions on Geoscience and Remote Sensing, vol. 61, pp. 1–15, 2023, [Online]. Available: https://api.semanticscholar.org/CorpusID:259215916

[23] M. Iman, K. M. Rasheed, and H. R. Arabnia, "A Review of Deep Transfer Learning and Recent Advancements," ArXiv, vol. abs/2205.10356, 2022, [Online]. Available: https://api.semanticscholar.org/CorpusID:246240811

[24] A. Rangel, J. R. Terven, D. M. C. Esparza, and E. A. Chavez-Urbiola, "Land Cover Image Classification," ArXiv, vol. abs/2401.09607, 2024, [Online]. Available: https://api.semanticscholar.org/CorpusID:267034663

[25] P. Helber, B. Bischke, A. R. Dengel, and D. Borth, "EuroSAT: A Novel Dataset and Deep Learning Benchmark for Land Use and Land Cover Classification," IEEE J Sel Top Appl Earth Obs Remote Sens, vol. 12, pp. 2217–2226, 2017, [Online]. Available: https://api.semanticscholar.org/CorpusID:11810992

[26] Y. Qiao, J. Ge, Y. Zhang, and Y. Ling, "Remote sensing image scene classification based on transfer learning and Swin transformer mode," in International Conference on Remote Sensing, Mapping, and Geographic Systems, 2023. [Online]. Available: https://api.semanticscholar.org/CorpusID:265254442

[27] F. Jannat and A. R. Willis, "Improving Classification of Remotely Sensed Images with the Swin Transformer," SoutheastCon 2022, pp. 611–618, 2022, [Online]. Available: https://api.semanticscholar.org/CorpusID:248518578

[28] G. Cheng, J. Han, and X. Lu, "Remote Sensing Image Scene Classification: Benchmark and State of the Art," Proceedings of the IEEE, vol. 105, pp. 1865–1883, 2017, [Online]. Available: https://api.semanticscholar.org/CorpusID:3046524

[29] G.-S. Xia et al., "AID: A Benchmark Data Set for Performance Evaluation of Aerial Scene Classification," IEEE Transactions on Geoscience and Remote Sensing, vol. 55, pp. 3965–3981, 2016, [Online]. Available: https://api.semanticscholar.org/CorpusID:15298934

[30] Y. Liu, Z. Zhang, J. Yue, and W. Guo, "SCANeXt: Enhancing 3D medical image segmentation with dual attention network and depth-wise convolution," Heliyon, vol. 10, 2024, [Online]. Available: https://api.semanticscholar.org/CorpusID:268192616

[31] K. Pinasthika, B. S. P. Laksono, R. B. P. Irsal, S. H. Shabiyya, and N. Yudistira, "SparseSwin: Swin Transformer with Sparse Transformer Block," Neurocomputing, vol. 580, p. 127433, 2023, [Online]. Available: https://api.semanticscholar.org/CorpusID:261682513

[32] W. Li et al., "SepViT: Separable Vision Transformer," ArXiv, vol. abs/2203.15380, 2022, [Online]. Available: https://api.semanticscholar.org/CorpusID:247778932

[33] F. Liu, H. Xu, M. Qi, D. Liu, J. Wang, and J. Kong, "Depth-Wise Separable Convolution Attention Module for Garbage Image Classification," Sustainability, 2022, [Online]. Available: https://api.semanticscholar.org/CorpusID:247375813

[34] Y. Yang and S. Newsam, "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, in GIS '10. New York, NY, USA: Association for Computing Machinery, 2010, pp. 270–279. doi: 10.1145/1869790.1869829.

[35] O. A. B. Penatti, K. Nogueira, and J. A. dos Santos, "Do deep features generalize from everyday objects to remote sensing and aerial scenes domains?," 2015 IEEE

Conference on Computer Vision and Pattern Recognition Workshops (CVPRW), pp. 44–51, 2015, [Online]. Available: https://api.semanticscholar.org/CorpusID:9560434

[36] N. Obianuju Lynda, N. Agwu Nnanna, and M. Mahamat Boukar, "Remote Sensing Image Classification for Land Cover Mapping in Developing Countries: A Novel Deep Learning Approach," IJCSNS International Journal of Computer Science and Network Security, vol. 22, no. 2, p. 214, 2022, doi: 10.22937/IJCSNS.2020.22.2.28.

[37] R. Naushad, T. Kaur, and E. Ghaderpour, "Deep Transfer Learning for Land Use and Land Cover Classification: A Comparative Study," Sensors (Basel), vol. 21, 2021, [Online]. Available: https://api.semanticscholar.org/CorpusID:238408385

[38] M. Belcaid, "Comparison of transformer-based and convolutional neural network-based (CNN) models for remote sensing image classification," 2023.

[39] S. d'Ascoli, H. Touvron, M. L. Leavitt, A. S. Morcos, G. Biroli, and L. Sagun, "ConViT: improving vision transformers with soft convolutional inductive biases," Journal of Statistical Mechanics: Theory and Experiment, vol. 2022, 2021, [Online]. Available: https://api.semanticscholar.org/CorpusID:232290742