Improving Classification of Remotely Sensed Images with the Swin Transformer

Fatema-E- Jannat
dept. of Electrical and Computer Engineering
University of North Carolina at Charlotte
Charlotte, North Carolina, USA
fjannat@uncc.edu

Andrew R. Willis
dept. of Electrical and Computer Engineering
University of North Carolina at Charlotte
Charlotte, North Carolina, USA
arwillis@uncc.edu

SoutheastCon 2022 | 978-1-6654-0652-9/22/$31.00 ©2022 IEEE | DOI: 10.1109/SoutheastCon48659.2022.9764016

***Abstract*—With the recent developments of transformer-based architecture in the image classification domain, the initial Vision Transformer (ViT) model has shown promising results compared to traditional CNN models. Inspired by this, this article reports on the efficacy of transformer-based models on remote sensing images for land cover classification. Our approach applies a variation of the vision transformer named the Swin (Shifted Window) Transformer model for analysis. This is a hierarchical transformer model that computes the representation with shifted windows. Results include an extensive study on the performance of this transformer for three different remote sensing datasets: EuroSat, NWPU-RESISC45, and AID. Findings indicate that the Swin architecture outperforms current state-of-the-art approaches for accurately classifying remote sensing images. Comparative analyses provide insights on the specific margin of improvement and an understanding of the prospect these transformer architectures have for improving image classification tasks of this type.**

***Index Terms*—vision-transformer, swin, remote-sensing, EuroSat, NWPU-RESISC45, AID, classification**

I. Introduction

Remote Sensing is the process of acquiring information from a object or any scene with the help of reflected and emitted electromagnetic radiation from a distance without any physical contact with that object or the scene. These are done by satellite or aircraft. These make the process fast, while making it possible to gather images from dangerous inaccessible locations covering larger area. There are several application of remote sensing images including observing air quality, predicting earthquake, land management, urban planning, and so on.

Land use classification is one of the most important application of remote sensing images. From growing socio-economic development to learning earth's biophysical environments, land use classification plays a very significant role. While maintaining a steady growth in urban area, it is also needed to restrain the unsystematic growth in the city areas. To ensure so, it is indispensable to make the good use of every inch of the lands while making the land management planning models. Remote sensing image comes very efficacious in this regard since it provides information of land areas for a time being and also gives a information about the land changes over the time. Monitoring these changes we can procure the knowledge on global climate changes.

Methods to classify remotely sensed images has been an active research topic in computer vision for more than 40 years. Work up to the past 10 years focused on feature-based and texture analysis classification methods. More recent work has focused on classification using traditional CNN approaches with significant performance improvements. Seminal work in this area includes AlexNet [21] and ResNet [16], GoogleNet [31], SqueezeNet [20], DenseNet [19]. The evolution of CNNs has seen widespread application to problems in computer vision.

With the recent developments of transformer-based architecture in the image classification domain, the initial Vision Transformer (ViT) model has shown promising results compared to traditional CNN models. This transformer model was first introduced by Vaswani et. al. [34] with impressive inference results for Natural Language Processing (NLP) problems. With the breakthrough of this transformer network [35] [33] [11] [12] [3], since then, researchers have applied this method for a number of other computer vision tasks [13] [7] [4] [41] [6] [29] [38] [27] [42] [15] [39] [5] [37].

The traditional CNNs only search for the image features, have no any positional information among the features thus do not possess the global understanding of the whole image. To solve this problem, Alexey Dosovitskiy et al. [13] challenges traditional CNNs approaches by proposing this ViT, based on self-attention layers which allows it to have a global understanding of the images while reducing the image-specific inductive bias . To eliminate the dependencies among input images, Normalization Layer is applied to every block which generalizes the model better than CNNs. [13] introduces Vision Transformer(ViT) architecture for vision applications, which is pretrained on JFT-300M [30], ImageNet dataset [10] and then transferred on mid size dataset for the classification where the authors demonstrate that their approach requires less computational cost in comparison to traditional CNN networks. Touvron et al [33] build on this work to develop a Data-efficient image Transformers (DeiT) that requires less data and less computational cost for inference.

Despite the improvements afforded by new transformer

978-1-6654-0652-9/22/$31.00 ©2022 IEEE

models such as ViT , and DeiT [33], these approaches still have some drawbacks. One specific shortcoming is how these models handle different image domains at different scales. To overcome these problems, a new variation of transformer has been Introduced by Liu et al. [22] named the Swin Transformer. This model is able to infer objects at different scales, i.e., a person in the foreground and in the background. This transformer model has received attention from researchers due to its success in natural image classifications. However, the application of the Swin Transformer model to remotely sensed image classification and analysis has received little attention.

This article fills the important gap in the research field by exploring the performance of the Swin Transformer model to the problem of classifying land use regions from aerial and satellite imagery. Our analysis in this work considers three important datasets used to develop solutions to this problem: (1) EuroSAT [18], (2) AID [40] and (3) NWPU-RESISC45 [8]. Contributions of this work consider the Swin Transformer model and these datasets as an advancement over state-of-the-art in this area. Specific contributions of this article include:

- a comparative performance study that shows the Swin Transformer model outperforms prior work for this problem domain for the (3) databases analyzed.
- an analysis and characterization of the rate of convergence of the trained Swin Transformer model for the (3) databases analyzed.
- we show that training parameters for the network are not sensitive to the sensor type and seem invariant with respect to training results across the (3) databases analyzed.

These contributions make Swin transformer models good candidates for solving difficult classification problems where inference of the class label depends in robust multi-scale representations for the image contents. Scale-space analysis proves to be an important component for reliable land-use classification and we predict transformer models will continue to improve the state-of-the-art for this problem domain as they evolve.

II. Related Works

Our review of recent related work on classification of remotely sensed images can is divided into two groups: (1) classifiers that use traditional CNN methods and (2) classifiers that use Vision Transformer methods. In this section we discuss recent related CNNs and transformer models that report the state-of-the-art in terms of their classification accuracy.

A. Traditional Method

While proposing a novel dataset based on Sentinel-2 for land use and classification, Patrick Helber et al. [18] also provided a benchmark using Convolutional Neural Networks(CNN). With their proposed dataset they evaluated their performance with respect to the classification task and achieved 98.57% and 98.18% accuracy while using the Resnet50 and GoogleNet classifier pre-trained with he image classification dataset ILSVRC-2012.

G. Xia et al. [40] introduced a large-scale dataset named Aerial Image Dataset (AID) and provide a benchmark of baseline result conducting on deep learning approaches including CaffeNet, VGG-VD-16, GoogleNet.

Introducing a large-scale dataset named NWPU-RESISC45 covering 45 scene classes and containing total 31,500 images, G. Cheng [8] provided a benchmark for the researchers using current state-of-the-art CNN methods.

M. Wang et al. [36] proposed a method for scene level classification where they used ResNet model to extract the image depth features and then a SVM classifier was trained with those feature combinations to classify the scenes. They implemented their method on UC Merced Land Use (UCM) dataset of 21 classes with 100 images per class.

A. Das et al. [9] used different variants of Res2Net50 architecture and fine tuned it while evaluating their approach on UC Merced, Brazilian Coffee Scenes, and EuroSAT dataset. They reported their classification accuracies as 98.76%, 93.25%, and 97.50% respectively.

M. Schmitt et al. [26] used the multi-spectral data from Sen12MS dataset [25] instead of using the traditional rgb data and used two different CNN model such as ResNet and DenseNet to evaluate their classification performance. This Sen12MS consists of 180,662 multi-sensor remote sensing imagery. While doing several experiments their paper showed that they achieved higher performance on multi-spectral images over the rgb images.

Using a new method of augmentation based on Class Activation Map (CAM) and image manipulation W. Zhang et al. [43] evaluated their method onto NWPU dataset with three different CNN models ResNet-18, SqueezeNet, DenseNet-121. Their experimental result shows improved accuracy by more than 0.4% with compare to original method (Non-augmentation method).

While using an auxiliary classification loss function with fine tuned and pre-trained CNNs such as GoogLeNet, Inception-V3 and EfficientNet. Y. Bazi et al. [1] evaluated their proposed method on (a) Merced, (b) AID, (c) NWPU, (d) Optimal-31, and (e) KSA datasets and showed that their method achieves state-of-the-art results in terms of OA and convergence times . Firstly they used root-mean-square propagation method to fine tune the network then introduced gradient at the earlier layer using an auxiliary classification loss function to overcome the vanishing gradient problem.

B. Transformer Method

As one example of more recent work, Yakoub Bazi et al. [2] proposed a vision transformer based remote sensing scene classification method while implementing different data augmentation techniques to boost the performance. They pruned the model by compressing half of the layers of vision transformer model while maintaining a competitive accuracy. Conducting experiments on several remote sensing datasets for example Merced, AID, Optimal31 and NWPU datasets they showed their efficacy with the classification accuracies of their proposed method using the rgb images.

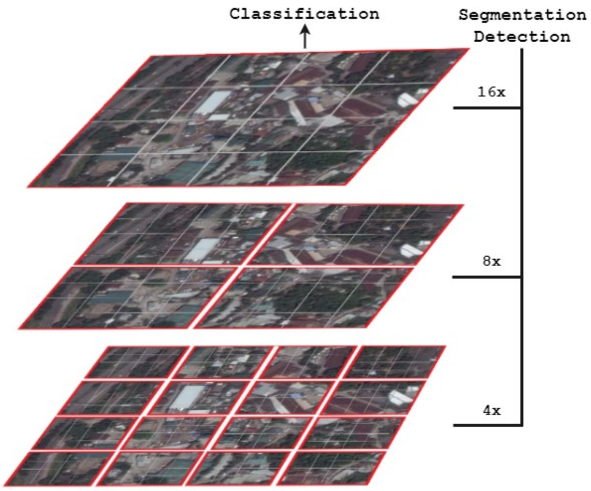

Fig. 1. Hierarchical Feature Maps of Swin Transformer.

While all these works based on traditional CNNs and the ViT make significant strides in recognition accuracy for this important problem of remotely sensed imagery, our work demonstrates that significant additional improvement is possible using a Swin Transformer model for classification because of its ability of handling different image domains at different scales.

III. Methods

The self-attention-based vision transformer architecture has been recently adopted for solving image domain problems and is gaining popularity as an alternative to traditional CNN models. The ViT network extracts small patches from images that form atomic groups of data as network inputs. Image patches are projected to a "flattened" linear array of values includes an embedding of a patch position to allow coherent patch-to-patch relationships and an added class token that is learned to facilitate the transformer encoder network. The transformer encoder consists of multiple layers including normalization layers, a multi-headed attention layer, fully connected layer and a residual block.

While the ViT vision transformer network has shown promising accuracy for many important applications, there are several shortcomings. For example, ViT transformer performance can degrade when considering image data that crosses between two domains having different image scales. Further, the self-attention of this transformer has a computational complexity that grows as a quadratic function of the input image size. This can impose high computational costs when classifying high resolution images. The Swin transformer, the focus of this work, addresses these problems using a hierarchical vision transformer.

A. Swin Transformer

The two main ideas of Swin Transformer are:

- using a hierarchical feature map and
- using shifted windows.

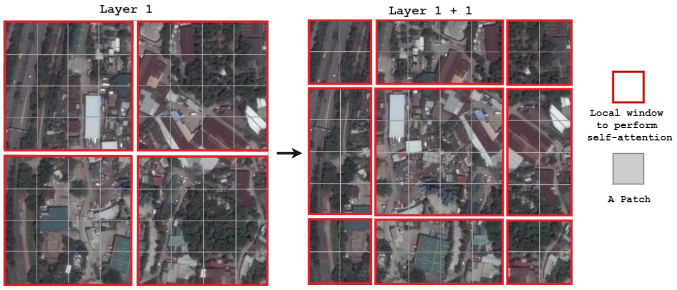

Fig. 2. The shifting window of Swin Transformer.

The hierarchical architecture of this Swin Transformer enables it to model image structures at various scales. Further, the linear computation complexity for inference makes this approach desirable for classification tasks. The key innovation of the Swin transformer is the use of shifted windows for image paches which allows an improved translation-equivariant latent representation of the image data which improves recognition performance of image patches having arbitrary translations.

1) Hierarchical Feature Map: In that Swin Transformer the hierarchical representation is developed through several patch merging processes. Merging features from the 2×2 neighboring patches, it reduces the tokens, as well as a linear transformation is applied to set the dimension two times. With the network going deeper, the resolution of the feature map gets higher as depicted in the Fig.1. Since the computation of self-attention is made locally within the non-overlapping windows, the computational complexity is linear.

2) Shifted Windows: ViT models compute self-attention globally which results in quadratic computational complexity. In contrast, the Swin Transformer shifts the window partition between two successive layers in the hierarchical map and computes the self-attention within local windows.

As illustrated in Fig.2, the left and right both of the layers have identical window size. The left layer is configured with standard window partitioning starting from left-top corner where 8×8 feature maps are uniformly divided into 2×2 windows. The size of each window is 4×4 $(M \times M)$. The right layer is configured with shifted window technique where the window is shifted by the half of the patch size, $\frac{M}{2} \times \frac{M}{2}$ from the left window.

These cross-window connections between two neighboring overlapping windows resemble the behavior of the CNN architecture.

Equation (1) describes how alternating Swin Transformer blocks compute their outputs,

$$\begin{aligned}\hat{z}^l &= W-MSA(LN(z^{l-1}))z^{l-1}\\ z^l &= MLP(LN(\hat{z}^l))\hat{z}^l\\ \hat{z}^{l+1} &= SW-MSA(LN(z^l))z^l\\ z^{l+1} &= MLP(LN(\hat{z}^{l+1}))\hat{z}^{l+1}\end{aligned} \quad (1)$$

where the output features from SW-MSA and MLP at layer l is denoted as $\hat{z}^l$ and z^l as in [22].

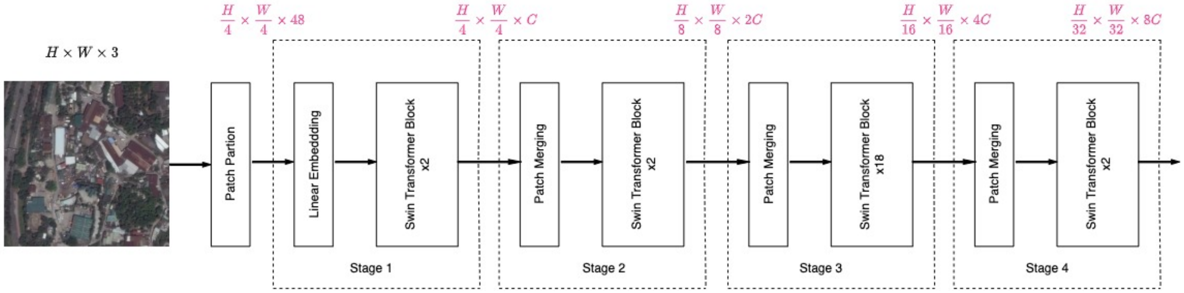
$H \times W \times 3$
$\frac{H}{4} \times \frac{W}{4} \times 48$
$\frac{H}{4} \times \frac{W}{4} \times C$
$\frac{H}{8} \times \frac{W}{8} \times 2C$
$\frac{H}{16} \times \frac{W}{16} \times 4C$
$\frac{H}{32} \times \frac{W}{32} \times 8C$
Patch Partion
Linear Embeddding
Swin Transformer Block x2
Patch Merging
Swin Transformer Block x2
Patch Merging
Swin Transformer Block x18
Patch Merging
Swin Transformer Block x2
Stage 1
Stage 2
Stage 3
Stage 4

Fig. 3. The architecture of Swin-S model.

3) Swin Transformer Block: Computational blocks of the Swin Transformer consist of two alternating window based multi-head self-attention (W-MSA), and a shifted window based multi-head self-attention (SW-MSA). Along with that MSA, It has two layers of MLP. A layer normalization (LN) is applied before each of MSA and MLP. Among these two modules a residual connection is made. The Swin Transformer block is illustrated on Fig.4.

4) Swin Transformer Architecture: Fig.3 shows the overall architecture of the Swin Transformer (Swin-S version). There are four stages in the Swin Transformer model. Each stage consists of several Swin Transformer blocks connected by a patch merging layer.

As illustrated in the Fig.3, the input image goes to the patch partition layer and each image is split into small non-overlapping patches referred to as tokens having a size of $4 \times 4 \times 3$. These tokens are converted into vectors whose length is $4 \times 4 \times 3 = 48$. Then these vectors are then passed through a linear embedding layer which projects it to C dimension, where C is an arbitrary number of dimension.

Then these patch tokens passes through a Swin Transformer block which again consists of alternating multi-head self-attention of W-MSA and SW-MSA as illustrated in Fig.4. This Swin Transformer block along with the linear embedding layer is termed as "Stage 1". In Stage-1, the number of tokens is maintained as $\frac{H}{4} \times \frac{W}{4}$. After this layer feature maps are generated.

These feature maps are processed in 3 steps:

(1) patch merging (where 2×2 adjacent patches are concatenated),
(2) linear embedding layer of the merged patches,
(3) a Swin Transformer block.

The patch merging layer merges 2×2 neighboring patches which results a reduced feature map. A linear layer is applied on top of that to set the output dimension as $2C$. Then Swin Transformer blocks are applied which maintain the token numbers as $\frac{H}{8} \times \frac{W}{8}$. This is termed as "Stage 2". The same steps are repeated two more times referring as "Stage 3" and "Stage 4" where the number of tokens remained as $\frac{H}{16} \times \frac{W}{16}$ and $\frac{H}{32} \times \frac{W}{32}$.

In this way, the hierarchical feature maps are generated and a Swin Transformer block is applied repeatedly which results to get 4 different variations of feature maps at different scales.

5) Variants of Swin Transformer: The Swin Transformer has 4 different versions of models of different sizes and complexity named Swin-B, Swin-T, Swin-S and Swin-L. The default window size is 7 for these models. Keeping the query dimension of each head as 32, the $\alpha = 4$ for the expansion layer for those of each MLP. A detailed summary of the variants of these models are provided in Table.I.

TABLE I
THE VARIANTS OF SWIN ARCHITECTURE.

Name	#params	C(#Channel)	Layer Numbers
Swin-T	28M	96	{2, 2, 6, 2}
Swin-S	50M	96	{2, 2, 18, 2}
Swin-B	88M	128	{2, 2, 18, 2}
Swin-L	197M	192	{2, 2, 18, 2}

6) Our Approach: We conduct recognition experiments using the Swin-S model as depicted in Fig.3 which is pre-trained with ImageNet-1k. This network has been initialized with the pre-trained weights and the last layer of the classifier is modified as per our dataset. We froze the first 3 layers of the network to avoid the weights from being modified while keeping the rest of the layers open for further training on our target datasets. In this method, the number of trainable weights is reduced significantly.

IV. DATASETS

We conduct our experiments on three distinct datasets for a fair evaluation of the performance of the Swin Transformer model: (1) EuroSAT [18], (2)NWPU-RESISC45 [8] and (3)AID [40] . Lately these datasets have been used by many of the researchers in this remote sensing domain which allow us to make an ample and fair comparison of our results to those reported in other work.

A. Eurosat

The EuroSat dataset [18] contains total 27000 labeled and geo-referenced sentinel-2 images covering 13 spectral bands. It has total 10 classes, Industrial Buildings, Residential buildings,

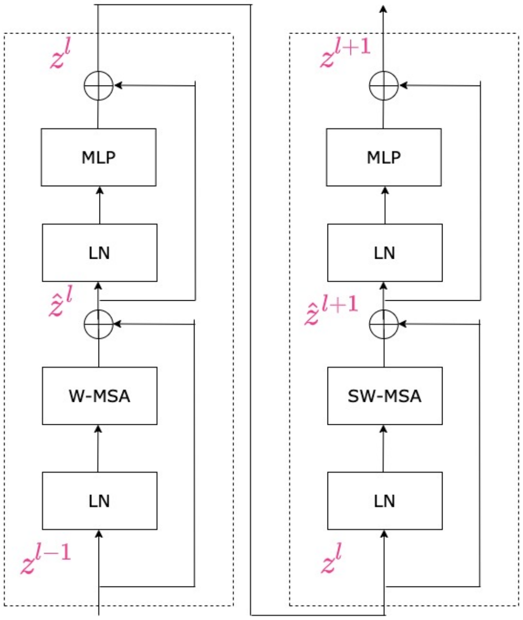

Fig. 4. Two successive blocks of Swin Transformer.

Annual Crop, Permanent Crop, River, Sea and Lake, Herbaceous Vegetation, Highway, Pasture, and Forest. Each images are in size 64X64 with the spatial resolution of 10 meters per pixel.

This dataset is publicly available at https://github.com/phelber/eurosat both in RGB and multi-spectral (MS) version. A few sample of this dataset is shown in Fig.5

B. NWPU-RESISC45

Proposed by Northwestern Polytechnical University (NWPU), this RESISC45 [8] is a learge scale dataset containing 31500 images which covers 46 scene classes and each class has 700 images. While posing high diversity in within-class and high similarity between-class, it holds a large variation in terms of spatial resolution, background, viewpoint and so on. Each of the images are in 256x256 size, spatial resolution range is 30m to 0.2m. From Fig.6 a few samples of this dataset can be observed.

C. AID

The Aerial image dataset (AID) [40] dataset is a large-scale dataset containing high intra-class diversity and low inter-class dissimilarity of 10,000 aerial scene images, It was published in 2017 by Wuhan University. It has a total of 30 different classes of 220 to 420 images per class. Images are in size of 600 × 600 pixels and the resolution range is 8m to 0.5m. For the reference samples from this dataset is shown in Fig.7

A general overview of the characteristic of these three datasets are provided in Table.II.

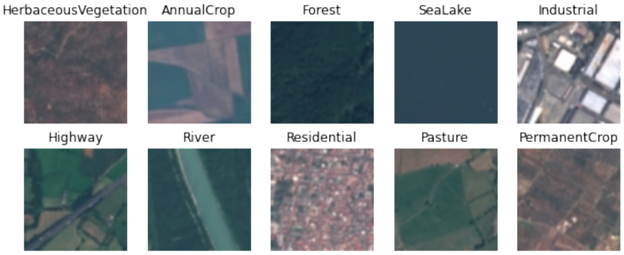

Fig. 5. Sample images from Eurosat dataset in RGB form.

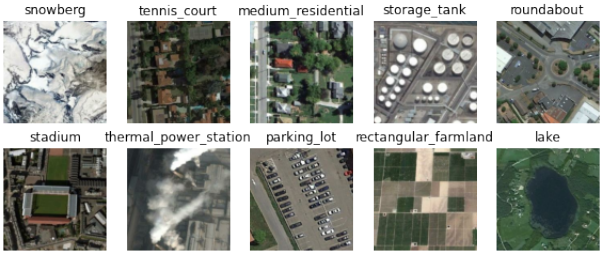

Fig. 6. Sample images from NWPU dataset in RGB form.

V. Experiments

We present results of fine-tuned Swin Transformer model on three datasets in the task of image classification in this section.

A. Implementation Details

We conducted all the experiments on a graphical processing unit (GPU) named NVIDIA GeForce GTX 1660 SUPER using Ubuntu 18.04.6 LTS . All the codes were implemented using Pytorch 1.7.1, python 3.7.11, with CUDA 11.5.

For the training purpose a pre trained Swin transformer trained on ImageNet-1k has been used. Fine tuning the network, a batch of 4 images has been used as the input, initializing the learning rate as 1e-3, setting the Stochastic gradient descent algorithm for optimizing the parameters, Cross-Entropy loss as the loss function. A scheduling technique named Cosine AnnealingLR has been used to converge the model, which was proposed by I. Loshchilov et al [23], in this method it starts with a very high learning rate then decays rapidly and then raises it's value again which makes it faster finding the global minima. Images are resized into 224x224 and the model was run for 25 epochs.

B. Evaluation Metrics

For the assessment of the performance of our experiments classification accuracy has been used as the evaluation metric. Accuracy is the proportion of currently classified samples among the total number of samples that have been used for the testing. Since all three datasets that we conducted our experiments are well balanced, we choose this metric for our evaluation.

$$\frac{TP + TN}{TP + TN + FP + FN} \tag{2}$$

TABLE II
OVERVIEW OF THE DATASETS.

Dataset Name	Number of Images	Number of Classes	Image Size	Year of Publication
Eurosat	27,000	10	64X64	2019
NWPU-RESISC45	31,500	46	256X256	2017
AID	10,000	30	600X600	2017

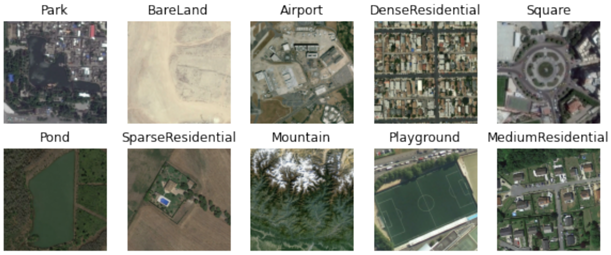

Fig. 7. Sample images from AID dataset in RGB form.

The accuracy formula is given in equation (2) where TP, TN, FP, FN are True Positive, True Negative, False Positive, False Negative respectively.

C. Performance Comparison

We have performed several experiments to evaluate the performance of our proposed method using three different datasets EuroSat, NWPU-RESISC45, and AID, the details of these datasets are already described in the Datasets section.

To make a valid comparison on our results we compare it with following state-of-the-art results,

- ResNet [17] : To solve the vanishing gradient problem in deep neural network K. He et al proposed this ResNet which takes input from earlier layer and connect to the later layer using skip connection.
- Res2Net [14]: Res2Net is a variation of residual blocks that builds a hierarchical residual connection within the single residual block allowing to represent the multi-scale features at a granular level. This feature thereby increases the span of receptive fields for each network layer.
- VGG [28]: It's a convolutional neural network that was proposed as a improvement over alexNet by replacing large sized filter to multiple smaller sized filter together.
- GoogleNet [31]:GoogleNet architecture is built based on Inception network which utilize different size of filters in the same layer to infer more information and the number of parameters is reduced in this network which make this less prone to overfitting.
- EfficientNet [32]: A convolutional neural network that uses a compound coefficient to scale network dimension.
- ViT [13]: A transformer based model that splits the input image into small patches and then those patches are linearly embedded and along with that a positional embedding is added then those are fed to the network.

1) Evaluation: EuroSat: In the EuroSat dataset, we set the training and validation ratio to be 80% and 20%, respectively. We ran the model for 25 epochs. The training and validation accuracy and loss curve of the proposed model has been shown in 8, where the x-axis is for the number of epochs and the y-axis is for the accuracy or loss. From this curve, it is clear that this model achieves its highest accuracy and converges within a very short time, specifically within 10 epochs, while maintaining a very short gap between the training and validation accuracy.

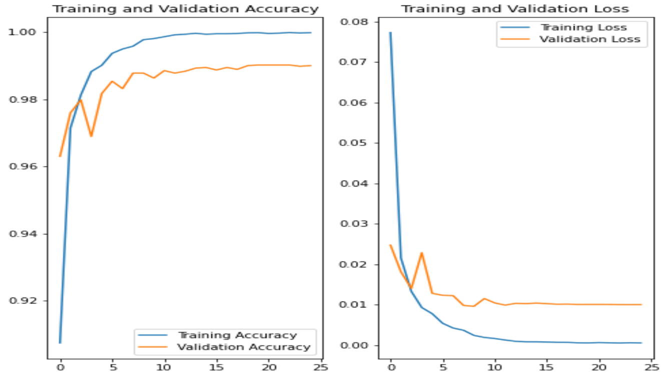

Fig. 8. Accuracy and loss curve with Swin transformer network for Eurosat dataset.

TABLE III
COMPARISON OF OUR WORK WITH THE STATE-OF-THE-ART METHODS ON EUROSAT DATASET IN TERMS OF CLASSIFICATION ACCURACY (%).

Method Name	Accuracy(%)
GoogLeNet [18]	98.18
Resnet-50 (scratch) [18]	96.43
Resnet-50 (pre-trained) [18]	98.57
Res2Net-50 (pre-trained) [9]	97.50
Swin (Fine-Tuned)	99.02

The overall best accuracy that is achieved for this dataset is 99.02% which appears to be the best when compared to the existing results, the performance comparison with the existing state-of-the-art methods is listed in Table. III.

2) Evaluation: NWPU-RESISC45: While performing the experiments on NWPU-RESISC45 we kept the training and validation ratio as 80% and 20% and ran the model for 25 epochs using the same hyper-parameters that have been used in our previous experiments. From the 9, the analysis of the training and validation accuracy and loss depicts the robustness of the network in terms of convergence. The model reaches its peak accuracy within 10 epochs and achieves 95.38% accuracy which is 2% more than the reported result from the works of Bazi et al. [2] who used the initial vision transformer model.

IV shows the comparison among the accuracy of our approach to the state-of-the-art methods which is a clear

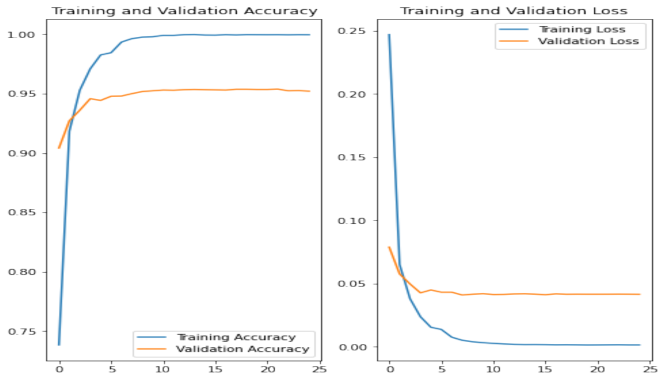

Fig. 9. Accuracy and loss curve with Swin transformer network for NWPU-RESISC45 dataset.

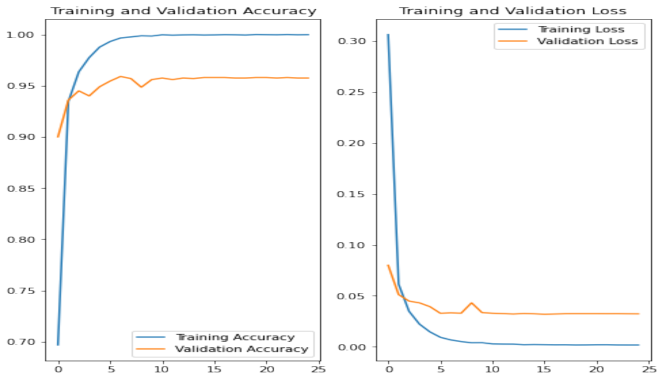

Fig. 10. Accuracy and loss curve with Swin transformer network for AID dataset.

TABLE IV
COMPARISON OF OUR WORK WITH THE STATE-OF-THE-ART METHODS ON NWPU-RESISC45 DATASET IN TERMS OF CLASSIFICATION ACCURACY (%).

Method Name	Accuracy(%)
ResNet-18 with CAM-Augmentation [43]	91.49
Fine-tuned VGGNet-16 [8]	90.36±0.18
Fine-tuned GoogLeNet [8]	86.02±0.18
ResNet-34-A [24]	92.79±0.56
ResNet-34-B [24]	92.04±0.10
ViT pruned (Bazi et al) [2]	93.83
Swin (Fine-Tuned)	95.38

indication of the applicability of the Swin transformer model over any other methods.

3) Evaluation: AID: We repeated the experiment for the AID dataset while maintaining the same settings, and achieved 95.90% accuracy which is slightly higher than the accuracy reported by Bazi et al [2]. with their ViT network. The Table. V illustrates the comparative analysis of the accuracy of AID dataset with the well known existing methods. From this table it can be concluded that the Swin Transformer model outperformed the other results.

The training and validation accuracy and loss analysis is shown in Fig. 10 which is consistent with our previous experiments in terms of stability and convergence.

VI. CONCLUSION

In this work, we investigated the accuracy of the vision transformer model in the analysis of remote sensing images. The vision transformer has received increased research interest but few attempts have been mode to analyse its use for remote sensing image classification. This article shows how the newly proposed Swin transformer model performs relative to state-of-the-art in classification of remote sensing images.

TABLE V
COMPARISON OF OUR WORK WITH THE STATE-OF-THE-ART METHODS ON AID DATASET IN TERMS OF CLASSIFICATION ACCURACY (%).

Method Name	Accuracy(%)
GoogLeNet [40]	83.44±0.40
EfficientNetB3-aux [1]	94.19 ± 0.15
ViT pruned (Bazi et al) [2]	95.86
Swin (Fine-Tuned)	95.90

Our experiments based on three different datasets showed promising results of using this Swin Transformer in the improvement of classification accuracy by providing a consistent best accuracy. This work reports an accuracy of 99.02%, 95.38% and 95.90% on EuroSat, NWPU-RESISC45 and AID dataset respectively. Moreover, to make our approach computationally efficient we used the pre-trained model and then fine tuned on our target dataset while freezing the first three of layers.

REFERENCES

[1] Yakoub Bazi, Mohamad M. Al Rahhal, Haikel Alhichri, and Naif Alajlan. Simple yet effective fine-tuning of deep cnns using an auxiliary classification loss for remote sensing scene classification. *Remote Sensing*, 11(24), 2019.

[2] Yakoub Bazi, Laila Bashmal, Mohamad M. Al Rahhal, Reham Al Dayil, and Naif Al Ajlan. Vision transformers for remote sensing image classification. *Remote Sensing*, 13(3), 2021.

[3] Tom B. Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, Sandhini Agarwal, Ariel Herbert-Voss, Gretchen Krueger, Tom Henighan, Rewon Child, Aditya Ramesh, Daniel M. Ziegler, Jeffrey Wu, Clemens Winter, Christopher Hesse, Mark Chen, Eric Sigler, Mateusz Litwin, Scott Gray, Benjamin Chess, Jack Clark, Christopher Berner, Sam McCandlish, Alec Radford, Ilya Sutskever, and Dario Amodei. Language models are few-shot learners. *CoRR*, abs/2005.14165, 2020.

[4] Nicolas Carion, Francisco Massa, Gabriel Synnaeve, Nicolas Usunier, Alexander Kirillov, and Sergey Zagoruyko. End-to-end object detection with transformers. *CoRR*, abs/2005.12872, 2020.

[5] Chun-Fu Chen, Quanfu Fan, and Rameswar Panda. Crossvit: Cross-attention multi-scale vision transformer for image classification. *CoRR*, abs/2103.14899, 2021.

[6] Hanting Chen, Yunhe Wang, Tianyu Guo, Chang Xu, Yiping Deng, Zhenhua Liu, Siwei Ma, Chunjing Xu, Chao Xu, and Wen Gao. Pre-trained image processing transformer. *CoRR*, abs/2012.00364, 2020.

[7] Mark Chen, Alec Radford, Jeff Wu, Heewoo Jun, Prafulla Dhariwal, David Luan, and Ilya Sutskever. Generative pretraining from pixels. 2020.

[8] Gong Cheng, Junwei Han, and Xiaoqiang Lu. Remote sensing image scene classification: Benchmark and state of the art. *CoRR*, abs/1703.00121, 2017.

[9] Arijit Das and Saravanan Chandran. Transfer learning with res2net for remote sensing scene classification. pages 796–801, 2021.

[10] Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. Imagenet: A large-scale hierarchical image database. pages 248–255, 2009.

[11] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. BERT: pre-training of deep bidirectional transformers for language understanding. *CoRR*, abs/1810.04805, 2018.

[12] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. BERT: pre-training of deep bidirectional transformers for language understanding. *CoRR*, abs/1810.04805, 2018.

[13] Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, Jakob Uszkoreit, and Neil Houlsby. An image is worth 16x16 words: Transformers for image recognition at scale. *CoRR*, abs/2010.11929, 2020.

[14] Shanghua Gao, Ming-Ming Cheng, Kai Zhao, Xinyu Zhang, Ming-Hsuan Yang, and Philip H. S. Torr. Res2net: A new multi-scale backbone architecture. *CoRR*, abs/1904.01169, 2019.

[15] Xiaohong Gao, Yu Qian, and Alice Gao. Covid-vit: Classification of covid-19 from ct chest images based on vision transformer models. 2021.

[16] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. *CoRR*, abs/1512.03385, 2015.

[17] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. *CoRR*, abs/1512.03385, 2015.

[18] Patrick Helber, Benjamin Bischke, Andreas Dengel, and Damian Borth. Eurosat: A novel dataset and deep learning benchmark for land use and land cover classification. *CoRR*, abs/1709.00029, 2017.

[19] Gao Huang, Zhuang Liu, and Kilian Q. Weinberger. Densely connected convolutional networks. *CoRR*, abs/1608.06993, 2016.

[20] Forrest N. Iandola, Matthew W. Moskewicz, Khalid Ashraf, Song Han, William J. Dally, and Kurt Keutzer. Squeezenet: Alexnet-level accuracy with 50x fewer parameters and <1mb model size. *CoRR*, abs/1602.07360, 2016.

[21] Alex Krizhevsky, Ilya Sutskever, and Geoffrey E. Hinton. Imagenet classification with deep convolutional neural networks. page 1097–1105, 2012.

[22] Ze Liu, Yutong Lin, Yue Cao, Han Hu, Yixuan Wei, Zheng Zhang, Stephen Lin, and Baining Guo. Swin transformer: Hierarchical vision transformer using shifted windows. *CoRR*, abs/2103.14030, 2021.

[23] Ilya Loshchilov and Frank Hutter. Sgdr: Stochastic gradient descent with warm restarts. 2017.

[24] Baogui Qi, He Chen, Yin Zhuang, Shaorong Liu, and Liang Chen. A network pruning method for remote sensing image scene classification. pages 1–4, 2019.

[25] Michael Schmitt, Lloyd Haydn Hughes, Chunping Qiu, and Xiao Xiang Zhu. SEN12MS - A curated dataset of georeferenced multispectral sentinel-1/2 imagery for deep learning and data fusion. *CoRR*, abs/1906.07789, 2019.

[26] Michael Schmitt and Yu-Lun Wu. Remote sensing image classification with the SEN12MS dataset. *CoRR*, abs/2104.00704, 2021.

[27] Debaditya Shome, T. Kar, Sachi Nandan Mohanty, Prayag Tiwari, Khan Muhammad, Abdullah AlTameem, Yazhou Zhang, and Abdul Khader Jilani Saudagar. Covid-transformer: Interpretable covid-19 detection using vision transformer for healthcare. *International Journal of Environmental Research and Public Health*, 18(21), 2021.

[28] Karen Simonyan and Andrew Zisserman. Very deep convolutional networks for large-scale image recognition. 2015.

[29] Chen Sun, Austin Myers, Carl Vondrick, Kevin Murphy, and Cordelia Schmid. Videobert: A joint model for video and language representation learning. *CoRR*, abs/1904.01766, 2019.

[30] Chen Sun, Abhinav Shrivastava, Saurabh Singh, and Abhinav Gupta. Revisiting unreasonable effectiveness of data in deep learning era. *CoRR*, abs/1707.02968, 2017.

[31] Christian Szegedy, Wei Liu, Yangqing Jia, Pierre Sermanet, Scott E. Reed, Dragomir Anguelov, Dumitru Erhan, Vincent Vanhoucke, and Andrew Rabinovich. Going deeper with convolutions. *CoRR*, abs/1409.4842, 2014.

[32] Mingxing Tan and Quoc V. Le. Efficientnet: Rethinking model scaling for convolutional neural networks. *CoRR*, abs/1905.11946, 2019.

[33] Hugo Touvron, Matthieu Cord, Matthijs Douze, Francisco Massa, Alexandre Sablayrolles, and Hervé Jégou. Training data-efficient image transformers & distillation through attention. *CoRR*, abs/2012.12877, 2020.

[34] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Lukasz Kaiser, and Illia Polosukhin. Attention is all you need. *CoRR*, abs/1706.03762, 2017.

[35] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Lukasz Kaiser, and Illia Polosukhin. Attention is all you need. *CoRR*, abs/1706.03762, 2017.

[36] Mingchang Wang, Xinyue Zhang, Xuefeng Niu, Fengyan Wang, and Xuqing Zhang. Scene classification of high-resolution remotely sensed image based on resnet. *Journal of Geovisualization and Spatial Analysis*, 3(2), 2019.

[37] Wenhai Wang, Enze Xie, Xiang Li, Deng-Ping Fan, Kaitao Song, Ding Liang, Tong Lu, Ping Luo, and Ling Shao. Pyramid vision transformer: A versatile backbone for dense prediction without convolutions. *CoRR*, abs/2102.12122, 2021.

[38] Xinpeng Wang, Chandan Yeshwanth, and Matthias Nießner. Sceneformer: Indoor scene generation with transformers. *CoRR*, abs/2012.09793, 2020.

[39] Haiping Wu, Bin Xiao, Noel Codella, Mengchen Liu, Xiyang Dai, Lu Yuan, and Lei Zhang. Cvt: Introducing convolutions to vision transformers. *CoRR*, abs/2103.15808, 2021.

[40] Gui-Song Xia, Jingwen Hu, Fan Hu, Baoguang Shi, Xiang Bai, Yanfei Zhong, and Liangpei Zhang. AID: A benchmark dataset for performance evaluation of aerial scene classification. *CoRR*, abs/1608.05167, 2016.

[41] Fuzhi Yang, Huan Yang, Jianlong Fu, Hongtao Lu, and Baining Guo. Learning texture transformer network for image super-resolution. *CoRR*, abs/2006.04139, 2020.

[42] Lei Zhang and Yan Wen. A transformer-based framework for automatic covid19 diagnosis in chest cts. pages 513–518, October 2021.

[43] Wei Zhang and Yungang Cao. A new data augmentation method of remote sensing dataset based on class activation map. 1961(1):012023, jul 2021.