IADIS International Journal on Computer Science and Information Systems
Vol. 17, No. 2, pp. 1-15
ISSN: 1646-3692

A HYBRID DILATION APPROACH FOR REMOTE SENSING SCENE IMAGE CLASSIFICATION

Anas Tukur Balarabe and Ivan Jordanov
University of Portsmouth, UK

ABSTRACT

While fine-tuning a transfer learning model alleviates the need for a vast amount of training data, it still comes with a few challenges. One of them is the range of image dimensions that the input layer of a model accepts. This issue is of interest, especially in tasks that require the use of a transfer learning model. In scene classification, for instance, images could come in varying sizes that could be too large/small to be fed into the first layer of the architecture. While resizing could be used to trim images to a required shape, that is usually not possible for images with tiny dimensions, for example, in the case of the *EuroSAT* dataset. This paper proposes an *Xception* model-based framework that accepts images of arbitrary size and then resizes or interpolates them before extracting and enhancing the discriminative features using an adaptive dilation module. After applying the approach for scene classification problems and carrying out a number of experiments and simulations, we achieved 98.55% accuracy on the *EuroSAT* dataset, 99.22% on *UCM*, 96.15% on *AID* and 96.04% on the *SIRI-WHU* dataset, respectively. We also monitored the micro-average and macro-average ROC curve scores for all the datasets to further evaluate the proposed model's effectiveness.

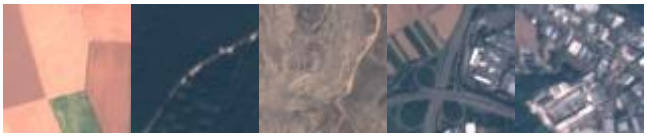
KEYWORDS

Adaptive Dilation, Deep Learning, Interpolation, Scene Classification, Transfer Learning

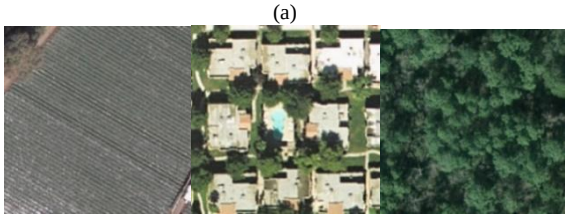
1. INTRODUCTION

Public access to high-resolution remote sensing images has become a reality thanks to the technological advancements recorded in recent years (Helber *et al.*, 2019). Access to satellite images for commercial and research purposes has fuelled interest and innovation in remote sensing and associated fields (Balarabe and Jordanov, 2021). This availability, in turn, triggered an avalanche of applications in domains such as agriculture, environment monitoring, disaster risk analysis, climate change, urban development, surveillance, land mapping, and land use and land cover classification (*LULC*) (Bi *et al.*, 2020; Z. Li *et al.*, 2020; Balarabe and Jordanov, 2021; Broni-Bediako *et al.*, 2021). Interestingly, most of these datasets used for the research and training scene classification systems have images with varying resolutions and dimensions.

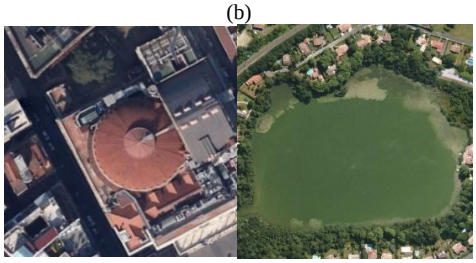
For example, the *EuroSAT* dataset (Helber *et al.*, 2019) has images of 64x64 pixels, while the *AID* dataset (Xia *et al.*, 2017) has images with dimensions up to 600x600 pixels. Generally, the accuracy of image classification models mostly depends on image representations such as image size, colour, shape, texture, and other properties (Bi *et al.*, 2020). More recently, researchers have proven that deep learning classifiers pre-trained on natural images can be repurposed for the task of scene classification (Hu *et al.*, 2015; Bi *et al.*, 2020). However, no one-size-fits-all transfer learning model can work with all the scene classification datasets (Liu and Huang, 2018). The quality of the discriminative features extracted by a transfer learning model could depend on the input data size because the feature maps are generated from the raw image data and fed into the classification layer for inference (Bi *et al.*, 2020). Depending on the dataset image dimensions, resizing or scaling up is often needed to reduce or increase the size of the images to an appropriate level that can produce feature maps containing enough discriminative information. For datasets with relatively larger samples, such as the *AID* dataset (Xia *et al.*, 2017), the images are usually resized to a moderately smaller dimension without throwing away any vital information. In other datasets, the images are either fed into the CNNs models in their original form or resized to reduce the model training time (Broni-Bediako *et al.*, 2021). Other datasets, notably the *EuroSAT* (Helber *et al.*, 2019), have images with a tiny dimension of 64x64 pixels, which cannot fit into the first layer of some transfer learning models.



(a)



(b)



(c)

Figure 1. Images of three different datasets: a) *EuroSAT*; b) *UCM*; and c) *AID*

To experiment on this dataset with transfer learning models, such as the *Xception* (Chollet, 2017), the images must reach at least the minimum size that the input layer of the architectures accepts. We believe that this model, the most advanced evolution of the *Inception* variants, has not received its fair share of attention from computer vision researchers. This paper investigates its effectiveness and efficiency in scene classification problems. In this work, we repurposed a pre-trained *Xception* model (Chollet, 2017) to perform scene classification tasks on the *EuroSAT*, *UCM*, *AID* and *SIRI-WHU* datasets. For the *EuroSAT* dataset, we employed the

LANCZOS image interpolation algorithm to scale up the images of the *EuroSAT* dataset from 64x64 to 71x71 pixels, which is the minimum that the image input layer of the *Xception* accepts. The *UCM* and *SIRI-WHU* images were left in their original form, while the samples of the *AID* dataset were resized to 256x256 pixels. In each case, the proposed framework chooses an appropriate context magnifying module to improve the quality of the extracted features before feeding them into the final classification layer for inference.

2. RELATED WORKS

Several research articles have been published on scene classification tasks using the traditional convolutional neural networks as the core models, while others were implemented using repurposed transfer learning architectures (Yuan *et al.*, 2020; Balarabe and Jordanov, 2021). As a rule of thumb, CNNs, traditional or pretrained, downsample images as they traverse a model from the input to the classification layer. This downsampling could cause the suppression of local discriminative information, impacting the overall performance (Z. Li *et al.*, 2020). Some researchers proposed a feature fusion approach to address the challenge by combining two models into a single unit to improve classification efficiency (Wang and Yu, 2020). In contrast, others used inference aggregation to obtain the average classification results of different model streams (F. Li *et al.*, 2020). Li *et al.* embedded context enhancement modules within the repurposed pre-trained models to extract more robust features for better discrimination and improved classification accuracy (Z. Li *et al.*, 2020). In most standard CNNs, an image's low-level and mid-level discriminative features tend to be lost as the depth of a model increases. Bi *et al*. developed a model that learns the feature representations in an image using a multiple instance learning framework that categorises labels by highlighting the semantics relevance of each category and generating a probability value for its prediction (Bi *et al.*, 2020). Unlike in natural images, where the critical discriminative feature is the most dominant aspect, a scene classification image contains the discriminatory features of other images, which could lead to one scene being confused with another by a classifier (Alhichri *et al.*, 2021). Alchiri *et al.* proposed a technique to address misclassifications by paying attention to sections of an image that uniquely identify it (Alhichri *et al.*, 2021).

Another challenge in remote sensing image classification is the discriminative information occlusion due to fused boundaries among objects in an image. Often, the discriminative features are mixed-up with the non-discriminative features. As a result, some authors developed a feature fusion framework that uses an entropy-based technique to fuse selected layers of some pre-trained models into a unified hierarchical scene classification framework (A. *et al.*, 2018). Despite the successes recorded by transfer learning models in scene classification, within-class variability and between-class similarity still need further attention (Balarabe and Jordanov, 2021). Xie *et al.* in (Xie *et al.*, 2019) highlighted that indiscriminate resizing of images could degrade the part of an image containing the information vital for its classification. As a result, the authors proposed a framework that accepts arbitrary image size and effectively extracts quality feature representation for improved classification. Wang *et al.* also developed a technique that extracts and combines local and global discriminative information in an image to increase the classification performance. The method proposed by the authors has two branches for feature extraction, connected by a structured key area localisation mechanism (SKAL) (Wang *et al.*, 2020).

3. PROBLEM STATEMENT AND METHODOLOGY

3.1 Problem Statement

There are many publicly available scene classification datasets (F. Li *et al.*, 2020), such as *UCM*, *AID*, *SIRI-WHU*, and *OPTIMAL-31*, each with its inherent uniqueness: meter(s) per pixel; image dimension; the number of samples per category; etc., to name just a few of the underlined differences among them. At the same time, there is a strong link between a CNNs classifier's accuracy and the size and quality of the input images (Bi *et al.*, 2020). The traditional CNNs and the pre-trained models use either local or global discriminative features for the classification unless engineered to combine the two groups in order to improve the output. Depending on image size, resizing or scaling up to an appropriate dimension is often needed to produce feature maps containing sufficient discriminative information. For datasets with large enough samples, such as the *AID* dataset (Xia *et al.*, 2017), the images are usually scaled down to a relatively smaller dimension without losing the essential information. In other datasets, such as the *UCM* and *SIRI-WHU*, there is the choice of resizing the images or passing them into the CNNs models in their original form (Xie *et al.*, 2019). Other datasets, for example, particularly the *EuroSAT* (Helber *et al.*, 2019), have images with a small dimension of 64x64 pixels. To use this dataset with some transfer learning models, such as the *Xception* model (Chollet, 2017), the images must be interpolated to reach at least the minimum acceptable size for the model input layer. This paper investigates the impact of image interpolation and feature enhancement on the accuracy of the *Xception* model. Section 3.2 briefly describes the overall model architecture, section 3.3 highlights the datasets used, 3.4 explains the training strategy, and 3.5 summarises the metrics employed to evaluate the model.

3.2 Model Architecture

The overall model architecture (Figure 2) consists of four components: the image interpolation module, the feature extraction module, the feature magnification module, and other additional layers for performance enhancement. The backbone of the feature extraction component comprises the *Xception* model's bottom layers, from the image input layer to the last convolutional layer. The dimension of the input image is first checked to ascertain whether interpolation is needed or not. If the image size is below 71x71, then the *LANCZOS* algorithm resizes it from its original dimension to 71x71 pixels; otherwise, it is forwarded to the feature extraction part of the framework. The reason for choosing *LANCZOS* is its ability to scale up an image without compromising its quality. Considering the impact that fused boundaries could have on a classifier's performance, we added a dilation layer and a few fully connected layers to preserve the extracted features as they go further down the pipeline. We included dropout layers with values from 0.25 to 0.5 to mitigate the effect of overfitting before the *softmax* layer. In addition, each fully connected layer has an *l2* regulariser with a 0.001 regularisation value. The output of the feature extraction component goes through one of the context magnifiers, depending on the size of the input from which these features have been extracted. Extractions from images with 71x71 dimensions are fed into a feature magnifier with a 2x2 dilation rate, and for a larger image, a 4x4 dilation rate is used. Then they are fed into the *GlobalAveragePooling 2D* layer, which processes the spatial information therein. Although LeakyReLU has been proposed recently in several deep learning-based applications (Goceri,

2021) to overcome the biasing issue caused by ReLU, we used the ReLU due to its efficiency in the structure of the proposed network with our datasets. The primary motivation behind choosing this model for this research work is that it is lightweight compared to some frequent transfer learning models, such as *VGG-Net, ResNet* and *GoogleNet.* Also, from the articles we have reviewed, it is evident that researchers have neglected this pre-trained architecture despite its efficiency and ease of use.

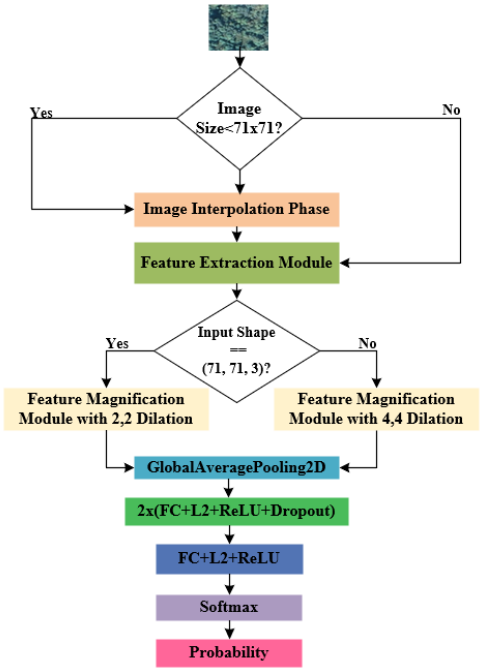


Figure 2. The overall architecture of the proposed model (*FC* – fully connected, *L2* – regularisation function, *ReLU* – transfer function; *Softmax* – output transfer function)

3.3 Datasets

EuroSAT dataset was released in 2019 by (Helber *et al.*, 2019). It consists of different classes of land use and land cover (*LULC*) images extracted across 34 European countries using Sentinel-2 satellite images containing 13 categories of 2000 to 3000 samples. *Annual crops, forests, herbaceous vegetation, residential, and sea lakes* have 3000 scene representations. In contrast, *highway, industrial, permanent crop, and river* have 2500, and *pasture* has the smallest number (200 instances), making 27000 64x64 pixels images for the entire dataset.

The *UCM* dataset (Yang and Newsam, 2010) has 21 categories containing 2100 images of different land use and cover types, with a uniform size of 256x256 pixels. The images in this dataset have a resolution of approximately 0.3m.

AID Dataset (Xia *et al.*, 2017) is one of the most extensive scene classification datasets, with 30 categories containing images with 600x600 pixels. This dataset includes the following categories: *airport, baseball field, bare land, bridge, beach, centre, commercial, church, dense*

residential, farmland, desert, industrial, forest, meadow, mountain, medium residential, parking, playground, park, pond, railway station, port, river, resort, sparse residential school, square storage tanks, viaduct, stadium, and football field. Each category comprises 240 to 420 images.

SIRI-WHU Dataset (Zhao *et al.*, 2016) is a dataset of 12 categories of scene classification images of 200x200 pixels, with a 2m spatial resolution. It has 2,400 images, evenly distributed across its 12 classes. This dataset includes the following classes of scene images extracted from Google Earth across China: agriculture, commercial, harbor, idle land, industrial, meadow, overpass, park, pond, residential, river, and water.

3.4 Training Strategy

We used an adaptive training strategy to train the proposed model on the *EuroSAT*, *UCM*, *SIRI-WHU* and *AID* datasets. For the *EuroSAT* dataset, the main difference is in the dilation rate used to enhance the quality of the extracted features. After scaling up the images of the *EuroSAT* dataset from 64x64 pixels to 71x71 pixels, some data augmentation techniques were used to bolster the training dataset size from 21600 to 151200 samples, 20% of which was reserved for validation and the remaining 20% from the initial split for testing. For the second part of the experiment, which involves the *UCM, SIRI-WHU* and the *AID* datasets, no image interpolation was used since the dimensions of these datasets' images satisfy the requirement of the image input layer of the backbone model. The images in the *UCM and SIRI-WHU* datasets were left in their original dimension of 256x256 and 200x200 pixels, while the *AID* dataset images were resized to 256x256 pixels. The model was trained for 350 epochs using a *batch size* of 32 *Adam* optimiser with a *learning rate* of 0.0001 and decay value of 10-e5. Also, an *early stopping* with a patience value fixed at 300 epochs was added to monitor the model's training progress and avoid overfitting. For the *UCM, SIRI-WHU* and *AID* images, the framework chooses the dilation rate of 4x4, which works better on images with resolutions bigger than the EuroSAT dataset's images. All experiments were carried out using Keras and TensorFlow in the Google Colab Pro environment.

3.5 Evaluation Strategy

The proposed model's performance has been evaluated using some of the most popular deep learning performance evaluation metrics by following the performance evaluation strategy used by the baseline models for ease of comparison. In addition to the *confusion matrix*, *precision*, *recall*, and *F1 score*, we also incorporated the *balance accuracy* metric to evaluate our model further, having split the datasets randomly into the train, validation and test subsets as in (Wang *et al.*, 2019; Zheng, Yuan and Lu, 2019; Z. Li *et al.*, 2020). We also monitored the micro-average and macro-average ROC curve scores for all the datasets to further evaluate the proposed model's effectiveness.

4. EXPERIMENTAL RESULTS

Table 1 compares the results from the first experiment on the *EuroSAT* dataset with the other two methods. The result produced by our model is competitive with what has been published by (Helber *et al.*, 2019).

Table 1. Performance comparison between our model and other approaches on the *EuroSAT* dataset

Model	OA%	Epochs	Training Time	Backbone	Source
EuroSAT	98.57	-	-	*ResNet* -50	(Helber *et al.*, 2019)
SLGE-CNN	99.76	600	9.6GPU days	*EfficientNet*	(Broni-Bediako *et al.*, 2021)
Ours	**98.55**	**400**	**1h 15min**	***Xception***	

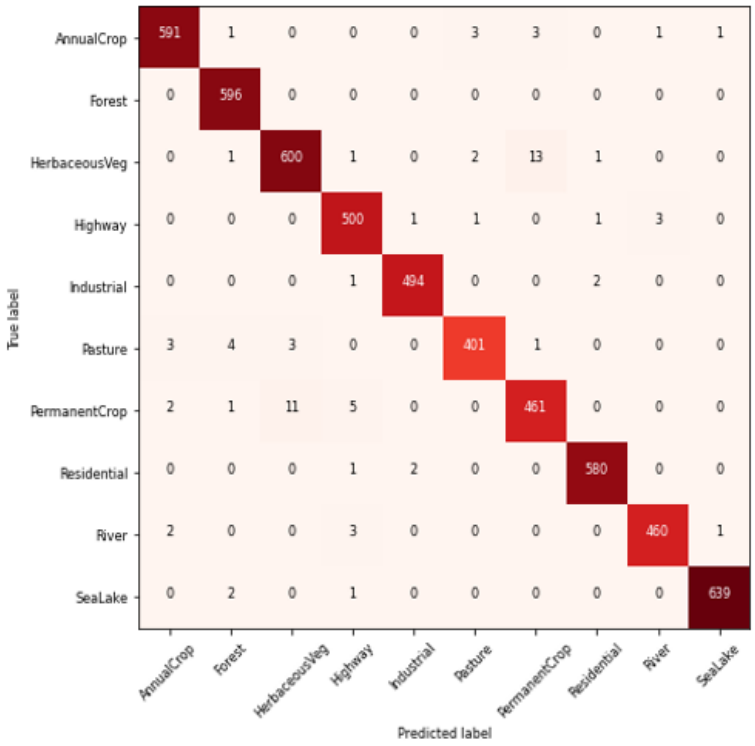
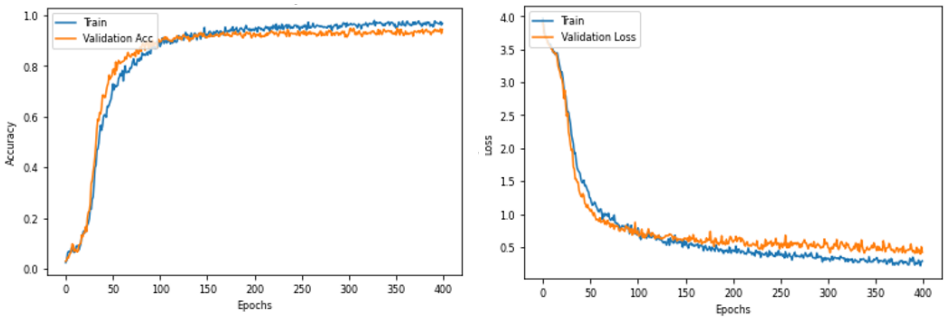


Figure 3. Confusion matrix for the *EuroSAT* dataset



(a) (b)

Figure 4. Model train/validation accuracy and loss on the *EuroSAT* dataset

ResNet-50, which has nearly 3 million more parameters than the *Xception*, produced the best result in the baseline paper. However, the marginal difference of 0.02% between the performance of our model and the state-of-the-arts means the proposed architecture would demand much fewer computing resources than *ResNet*-50. Therefore, the tradeoff is between the margin of 0.02% in classification accuracy and the extra computing power needed to manage an additional 3 million parameters. The *SLGE-CNN* architecture (Broni-Bediako *et al.*, 2021) is 1.21% more accurate than our method. However, it took 9.6 GPU days for *SLGE-CNN* to achieve 600 training epochs compared to the 1h 15mins and 400 training cycles we used to train the framework proposed in this paper. Two classes with very high outer-class similarity affected the performance of our method, as can be observed in Figure 3. Out of the test set, 13 images belonging to the *herbaceous vegetation* class were misclassified as instances of the *permanent crop* class, and 11 images of the *permanent crop* class were misclassified as *herbaceous vegetation*. These misclassifications stemmed from the high spatial similarity between the two categories. Figures 4(a) and (b) give the model training accuracy and validation loss for the dataset, which reflect the percentage accuracy and loss and the number of iterations.

Table 2. Performance comparison between our model and other approaches on the *UCM* dataset

Model	OA%	Epochs	Training Time	Backbone	Source
MSCP-Net	96.56±0.18	-	-	*VGG-Net*	(He, Fang and Li, 2018)
EuroSAT	94.38		-	*GoogleNet*-50	(Helber *et al.*, 2019)
ARCNet	93.10±0.55	50	-	*VGG-Net*	(Wang *et al.*, 2019)
CNN+FV	93.9	-	-	*VGG-Net*	(Zheng, Yuan and Lu, 2019)
MIDC-Net_CS	92.95±0.17	-	-	Dense-Net	(Bi *et al.*, 2020)
CNN+MIL	99.26	90		*VGG-Net*	(Li, Z. *et al.*, 2020)
Ours	**99.22**	**400**	**80 min 47 sec**	***Xception***	

Table 2 shows that our model performed remarkably well on the UCM dataset compared to other techniques. The *VGG-Net*, the backbone used in most baseline models in Table 2, has nearly 120 million more parameters than the *Xception*, making it more computationally costly to run than the proposed architecture's backbone. Our approach outperformed all the models in the table despite having only one dilation layer. For example, the *CNN+MIL* (Z. Li et al., 2020) incorporated a context enhancement module that uses many layers with varying dilation rates

and a summing function to concatenate all the features into a single block and feeds it into a multiple instant learning module. However, the proposed architecture used only one dilation layer and outperformed the *MI-CNN* single model presented in (Z. Li *et al.*, 2020) regarding the accuracy and parameter utilisation. This performance proves that the *Xception* is as effective as other transfer learning models and even better in some areas. Our model classified the *UCM* dataset with very high accuracy, recording only 3 misclassifications, as shown in Figure 5.

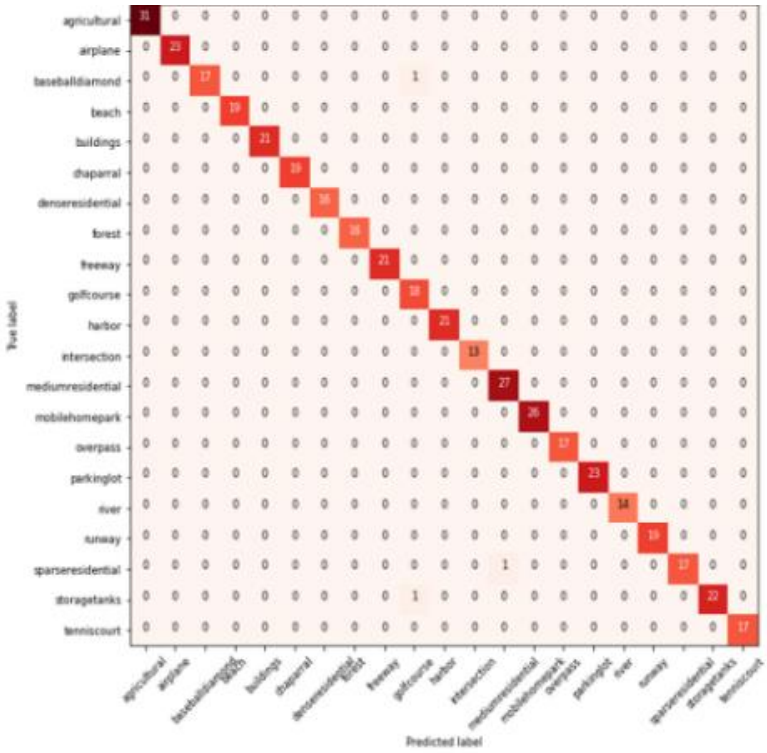
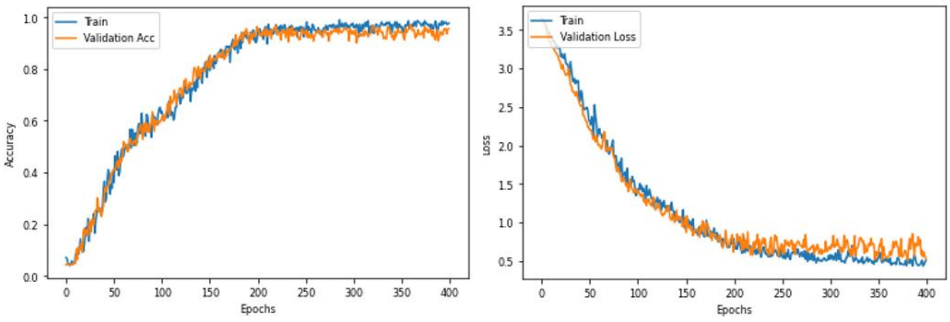


Figure 5. Confusion matrix for the *UCM* dataset



(a) (b)

Figure 6. Model train/validation accuracy and loss on the *UCM* dataset

Table 3. Performance comparison between our model and other approaches on the *AID* dataset

Model	OA%	Epochs	Training Time	Backbone	Sources
MSCP-Net	96.56±0.18	-	-	*VGG-Net*	(He, N, *et al.*, 2018)
EuroSAT Paper	94.38		-	*GoogleNet*-50	(Helber *et al.*, 2019)
ARCNet	93.10±0.55	50	-	*VGG-Net*	(Wang *et al.*, 2019)
CNN+FV	93.9	-	-	*VGG-Net*	(Zheng, Yuan and Lu, 2019)
MIDC-Net_CS	92.95±0.17	-	-	*Dense-Net*	(Bi *et al.*, 2020)
Ours	**96.15**	**400**	**3hrs 57min**	***Xception***	

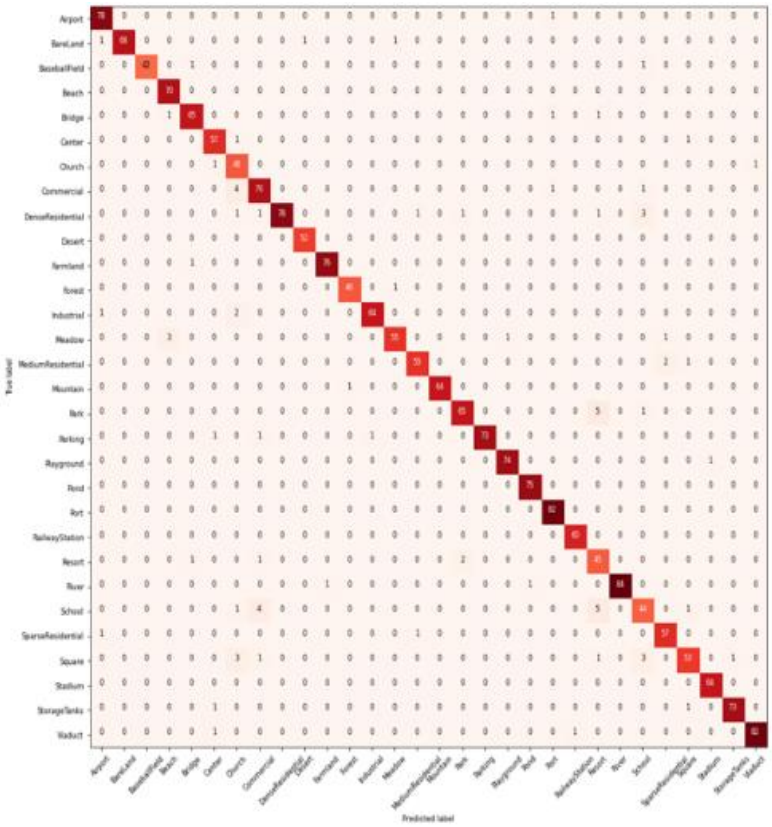
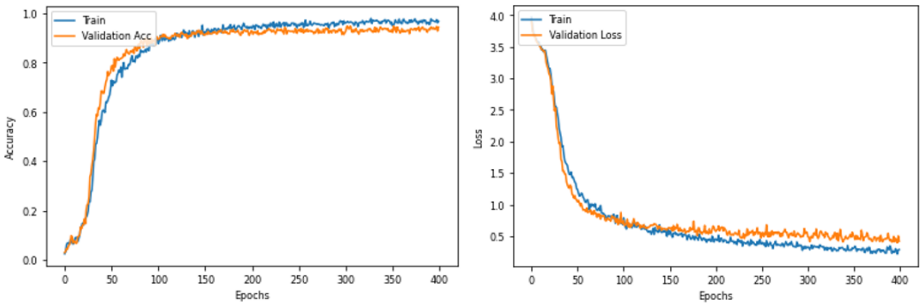


Figure 7. Confusion matrix for the *AID* dataset



(a) (b)

Figure 8. Model train/validation accuracy and loss on the *AID* dataset

We resized the images of the *AID* dataset from 600x600 pixels to 256x256 to utilise memory by reducing the resources needed to process the images at 600x600 pixels and since keeping the images at that size does not affect the quality of the features extracted by the backbone model. The proposed model performed remarkably well on the *AID* dataset, producing competitive results with the state-of-the-art significantly better than many baseline papers. Table 3 shows the performance of our approach on the *AID* dataset compared to other state-of-the-arts. The *VGG-Net* is the predominantly used pre-trained backbone for feature extraction, as shown in the table. The *Xception* model is more lightweight, despite having more layers than the *VGG-Net*. The latter has more parameters than the former and has more efficient parameter utilisation. Looking at the result published by (He, Fang and Li, 2018), it is evident that *MSCP-Net* is slightly more accurate than the proposed model; however, it is at the expense of 120 additional million parameters and much bigger images. Overall, our model produced a result on this dataset that is competitive and even better than many of the state-of-the-art methods, as shown in Table 3. The general performance of the model, in terms of training accuracy and loss, is given in Figures 8(a) and (b), which show the impact of the anti-overfitting techniques we employed.

Table 4. Performance comparison between our model and other approaches on the *SIRI-WHU* dataset

Model	OA%	Epochs	Training Time	Backbone	Source
MSAA-Net	95.2±0.65	10,000	-	*CNN*	(L. Li *et al.*, 2020)
ResNet-18	92.23±0.9	-	-	*ResNet-18*	(L. Li *et al.*, 2020)
MCNN	93.75±1.3	10,000	-	*CNN*	(Liu, Zhong and Qin, 2018)
Fine Tuned ResNet-50	94.03	-	-	*ResNet-50*	(Shabbir *et al.*, 2021)
Ours	**96.04**	**400**	**54 mins**	***Xception***	

We experimented further on the *SIRI-WHU* dataset to establish the efficacy of the proposed model as an expansion of our earlier work (Balarabe and Jordanov, 2022). On this dataset, as shown in Table 4, our framework classified the images with high accuracy in much fewer epochs than *MSAA-Net* and *MCNN*. Both *MSAA-Net* and *MCNN* models were built using standard CNNs, thus could be more lightweight than what we propose in this paper. However, considering the number of epochs taken to arrive at 95.20% and 97.73% accuracy, our model is arguably more efficient and accurate. The other techniques shown in the table were built on a *ResNet-50* framework, as indicated. These approaches are more computationally costly to run compared to an *Xception*-based framework due to the considerable difference in the number of

parameters used by the two backbones. The performance of our model on this dataset also reflects its robustness against inner-class variability and outer-class similarity. The confusion matrix below shows that only 19 out of the 480 test samples were misclassified.

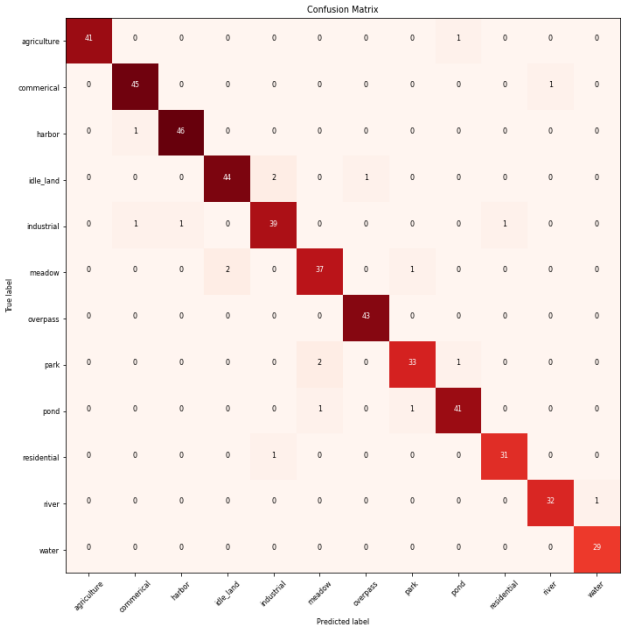
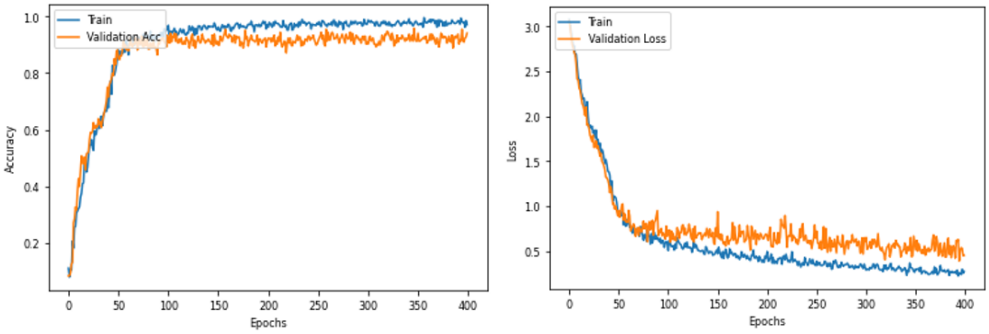


Figure 10. Confusion matrix for the *SIRI-WHU* dataset



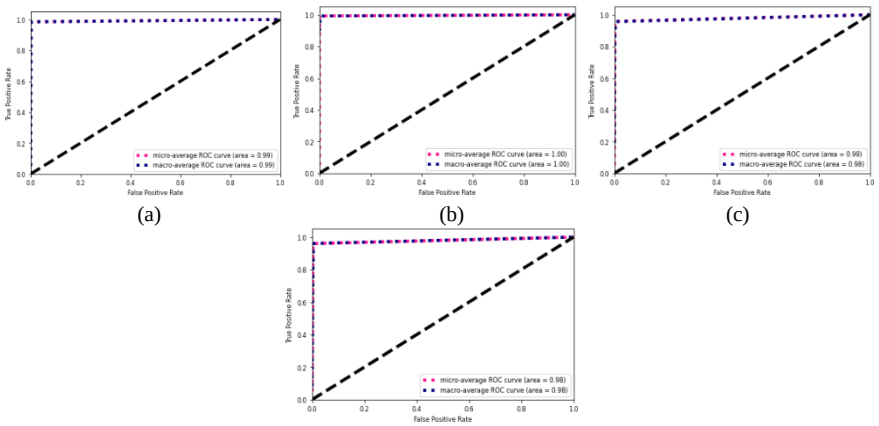
(a) (b)

Figure 9. Model train/validation accuracy and loss on the *SIRI-WHU* dataset

Table 5. Other Performance evaluation metrics on *EuroSAT*, *UCM*, *AID and SIRI-WHU* Datasets

Dataset	Precision	Recall	F1 Score	Kappa	Balance Accuracy
EuroSAT	0.9852	0.9849	0.9851	0.9839	0.9849
UCM	0.9935	0.9926	0.9931	0.9925	0.9926
AID	0.9558	0.95689	0.9563	0.9601	0.9607
SIRI-WHU-12	0.9641	0.9606	0.9607	0.9563	0.9609

Table 5 shows that the other performance metrics we monitored to evaluate this framework corroborate the overall test accuracy result. Figure 9(a) gives the model train and validation accuracy plot, while 9(b) shows the training and validation loss. In each case, the impact of the measures carefully applied to checkmate overfitting is evident in slight differences between the train and validation accuracy and train and validation loss, which are within the acceptable limit.



(a) (b) (c)

(d)

Figure 11. Micro-average and macro-average ROC curve for (a): the EuroSAT, (b): the UCM, (c) the AID and (d) the SIRI-WHU datasets

As an extension of our earlier work (Balarabe and Jordanov, 2022), we added ROC curve plots to gain more insight (at all classification thresholds) into the performance of our framework. Figures 11(a), (b), (c), and (d) show the ROC curve plots for the *EuroSAT*, *UCM*, *AID* and *SIRI-WHU* datasets, respectively. In each plot, the micro and macro-ROC curve average values are given. Since we consider a multiclass problem with random data split (see the confusion matrices), the micro-average, which computes the average ROC curve results by aggregating the contribution of each class and taking the average, gives a better idea of the proposed model's stability. Each experiment's area under the ROC curve value stands between 98% and 100%. These results show that the slight imbalance in data distribution does not affect our model.

5. CONCLUSION

This paper proposes a framework for classifying remote sensing images using a hitherto unused pre-trained model. We experimented on the *EuroSAT* dataset, one of the most extensive publicly available datasets for scene classification. It has images with the smallest dimension among the remote sensing (RS) datasets. We scaled up the dataset images from 64x64 pixels to 71x71 pixels using the *LANCZOS* image interpolation algorithm to meet the minimum input size requirement of the backbone model. In the second experiment, the performance of the proposed approach on some of the most popular satellite image classification datasets (*UCM*, *SIRI-WHU* and *AID*) was further evaluated. Our framework produced competitive results with state-of-the-art, better accuracy and computational efficiency for some of our comparisons. The

performance of the proposed model shows that *Xception* can also be efficiently utilised for tasks of satellite image classification. It also shows that a careful selection of dilation rate and hyperparameters can significantly reduce the impact of misclassification by CNNs classifiers. Despite the computational efficiency of the proposed framework, we believe there is still room for improvements concerning reducing the training time and increasing the accuracy. As part of future work, we intend to enhance the model by incorporating an ensemble approach using *Xception* and *EfficientNet*, which will help to achieve these objectives.

ACKNOWLEDGEMENT

This work is part of a PhD research sponsored by the Petroleum Technology Development Fund (PTDF), Nigeria.

REFERENCES

A., F. K. G. *et al.* (2018) 'A deep heterogeneous feature fusion approach for automatic land-use classification', *Information Sciences*. Elsevier Inc., 467, pp. 199–218. doi: 10.1016/j.ins.2018.07.074.

Alhichri, H. *et al.* (2021) 'Classification of Remote Sensing Images Using EfficientNet-B3 CNN Model with Attention', *IEEE Access*, 9, pp. 14078–14094. doi: 10.1109/ACCESS.2021.3051085.

Anwer, R. M. *et al.* (2018) 'Binary patterns encoded convolutional neural networks for texture recognition and remote sensing scene classification', *ISPRS Journal of Photogrammetry and Remote Sensing*. International Society for Photogrammetry and Remote Sensing, Inc. (ISPRS), 138, pp. 74–85. doi: 10.1016/j.isprsjprs.2018.01.023.

Balarabe, A. T. and Jordanov, I. (2021) 'LULC IMAGE CLASSIFICATION WITH CONVOLUTIONAL NEURAL NETWORK Anas Tukur Balarabe and Ivan Jordanov School of Computing, University of Portsmouth, UK', *International Geoscience and Remote Sensing Symposium (IGARSS)*. IEEE, pp. 5985–5988.

Balarabe, A. T. and Jordanov, I. (2022) 'Interpolation and Context Magnification Framework for Classification of Scene Images', in *International Conferences Computer Graphics, Visualization, Computer Vision and Image Processing (CGVCVIP)*, pp. 93–100.

Bi, Q. *et al.* (2020) 'A Multiple-Instance Densely-Connected ConvNet for Aerial Scene Classification', *IEEE Transactions on Image Processing*, 29, pp. 4911–4926. doi: 10.1109/TIP.2020.2975718.

Broni-Bediako, C. *et al.* (2021) 'Searching for CNN Architectures for Remote Sensing Scene Classification', *IEEE Transactions on Geoscience and Remote Sensing*. IEEE, pp. 1–13. doi: 10.1109/tgrs.2021.3097938.

Cheng, G. *et al.* (2018) 'When Deep Learning Meets Metric Learning: Remote Sensing Image Scene Classification via Learning Discriminative CNNs', *IEEE Trans. Geosci. Remote Sens*. IEEE, 56(5), pp. 2811–2821.

Chollet, F. (2017) '*Xception*: Deep learning with depthwise separable convolutions', *Proceedings - 30th IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2017*, 2017-Janua, pp. 1800–1807. doi: 10.1109/CVPR.2017.195.

Goceri, E. (2021) 'Diagnosis of skin diseases in the era of deep learning and mobile technology', *Computers in Biology and Medicine*. Elsevier Ltd, 134(January), p. 104458. doi: 10.1016/j.compbiomed.2021.104458.

He, N. *et al.* (2018) 'Remote sensing scene classification using multilayer stacked covariance pooling', *IEEE Transactions on Geoscience and Remote Sensing*. IEEE, 56(12), pp. 6899–6910. doi: 10.1109/TGRS.2018.2845668.

Helber, P. *et al.* (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(7), pp. 2217–2226. doi: 10.1109/JSTARS.2019.2918242.

Hu, F. *et al.* (2015) 'Transferring deep convolutional neural networks for the scene classification of high-resolution remote sensing imagery', *Remote Sensing*, 7(11), pp. 14680–14707. doi: 10.3390/rs71114680.

Li, F. *et al.* (2020) 'High-Resolution Remote Sensing Image Scene Classification via Key Filter Bank Based on Convolutional Neural Network', *IEEE Transactions on Geoscience and Remote Sensing*, (Cv), pp. 1–16. doi: 10.1109/tgrs.2020.2987060.

Li, L. *et al.* (2020) 'A multiscale self-adaptive attention network for remote sensing scene classification', *Remote Sensing*, 12(14). doi: 10.3390/rs12142209.

Li, Z. *et al.* (2020) 'Deep Multiple Instance Convolutional Neural Networks for Learning Robust Scene Representations', *IEEE Transactions on Geoscience and Remote Sensing*, 58(5), pp. 3685–3702. doi: 10.1109/TGRS.2019.2960889.

Liu, Y. and Huang, C. (2018) 'Scene classification via triplet networks', *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing*. IEEE, 11(1), pp. 220–237. doi: 10.1109/JSTARS.2017.2761800.

Liu, Y., Zhong, Y. and Qin, Q. (2018) 'Scene classification based on multiscale convolutional neural network', *IEEE Transactions on Geoscience and Remote Sensing*. IEEE, 56(12), pp. 7109–7121. doi: 10.1109/TGRS.2018.2848473.

Pathak, A. R., Pandey, M. and Rautaray, S. (2018) 'Application of Deep Learning for Object Detection', *Procedia Computer Science*. Elsevier BV, 132(Iccids), pp. 1706–1717. doi: 10.1016/j.procs.2018.05.144.

Shabbir, A. *et al.* (2021) 'Satellite and Scene Image Classification Based on Transfer Learning and Fine Tuning of ResNet50', *Mathematical Problems in Engineering*, 2021. doi: 10.1155/2021/5843816.

Wang, H. and Yu, Y. (2020) 'Deep Feature Fusion for High-Resolution Aerial Scene Classification', *Neural Processing Letters*. Springer US, 51(1), pp. 853–865. doi: 10.1007/s11063-019-10119-4.

Wang, Q. *et al.* (2019) 'Scene classification with recurrent attention of VHR remote sensing images', *IEEE Transactions on Geoscience and Remote Sensing*. IEEE, 57(2), pp. 1155–1167. doi: 10.1109/TGRS.2018.2864987.

Wang, Q. *et al.* (2020) 'Looking Closer at the Scene: Multiscale Representation Learning for Remote Sensing Image Scene Classification', *IEEE Transactions on Neural Networks and Learning Systems*, pp. 1–15. doi: 10.1109/TNNLS.2020.3042276.

Xia, G. S. *et al.* (2017) 'AID: A benchmark data set for performance evaluation of aerial scene classification', *IEEE Transactions on Geoscience and Remote Sensing*. IEEE, 55(7), pp. 3965–3981. doi: 10.1109/TGRS.2017.2685945.

Xie, J. *et al.* (2019) 'Scale-Free Convolutional Neural Network for', *IEEE Transactions on Geoscience and Remote Sensing*. IEEE, PP(9), pp. 1–13. doi: 10.1109/TGRS.2019.2909695.

Yang, Y. and Newsam, S. (2010) 'Bag-of-visual-words and spatial extensions for land-use classification', *GIS: Proceedings of the ACM International Symposium on Advances in Geographic Information Systems*, pp. 270–279. doi: 10.1145/1869790.1869829.

Yu, Y. and Liu, F. (2018) 'A Two-Stream Deep Fusion Framework for High-Resolution Aerial Scene Classification', *Computational Intelligence and Neuroscience*, 2018. doi: 10.1155/2018/8639367.

Yuan, B. *et al.* (2020) 'Multi-deep features fusion for high-resolution remote sensing image scene classification', *Neural Computing and Applications*. Springer London, 7, pp. 49–51. doi: 10.1007/s00521-020-05071-7.

Zhao, B. *et al.* (2016) 'Dirichlet-derived multiple topic scene classification model for high spatial resolution remote sensing imagery', *IEEE Transactions on Geoscience and Remote Sensing*. IEEE, 54(4), pp. 2108–2123. doi: 10.1109/TGRS.2015.2496185.

Zheng, X., Yuan, Y. and Lu, X. (2019) 'A Deep Scene Representation for Aerial Scene Classification', *IEEE Transactions on Geoscience and Remote Sensing*. IEEE, 57(7), pp. 4799–4809. doi: 10.1109/TGRS.2019.2893115.