DOI: 10.15849/IJASCA.250330.06

Int. J. Advance Soft Compu. Appl, Vol. 16, No. 3, March 2025
Print ISSN: 2710-1274, Online ISSN: 2074-8523

Benchmarking Vision Transformers for Satellite Image Classification based on Data Augmentation Techniques

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

[1]Department of Computer Engineering and Information Technology, University of Qom, Qom, Iran. University of Kerbala, Karbala, 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

This article evaluates the implications on transformer model performance in satellite image classification by means of numerous data augmentation techniques using the Eurosat dataset. We examine the Swin-tiny, Swin-small, Convit-small, and Crossvit-small models under several augmentation methods including MixUp, CutMix, Geometric, WGP-GAN, and DCGAN. Our findings demonstrate that Mixup and WGP-GAN augmentations significantly enhance model performance; Swin small achieves 99.26% test accuracy the best. CutMix was more helpful for Swin-small than for Swin-tiny; geometric augmentation improved Swin-tiny and Crossvit-small. DCGAN behaved differently on many models. These results highlight the significance of selecting appropriate augmentation techniques suited for certain model architectures to increase performance in assignments needing satellite image classification.

Keywords: *Data Augmentation, Vision Transformers, classification, DGAN, CutMix, MixUp, Eurosat.*

1 Introduction

Remote sensing scene classification has shown remarkable improvements with the application of Deep Learning (DL) techniques, as demonstrated by the works of [1] and [2]. However, in cases where there is an insufficient quantity of categorized data presented for training DL models, the performance of these models decreases significantly. When there isn't enough labeled data, DL models suffer from overfitting that is, they learn the training data extremely fine but fail to generalize to testing data that has never been seen before [3]. Nevertheless, the process of labeling huge amount of data needs large material and human resources. Therefore, alternative solutions are required. Several proposed strategies are designed to reduce the issue of overfitting resulting from a lack of labeled data. The first generation of solutions focuses on the network construction, including the implementation of dropout layer [4], Drop Connect [5], L2-regularization [6] , and auxiliary classification branches[7]. Another solution involves training protocols, including well planned

Received 7 November 2024; Accepted 30 January 2025

initialization [8], an appropriate learning rate decay strategy [9], and a well-designed early stopping mechanism [10]. Data augmentation is another solution that can be used to address overfitting and enhance performance [11]. Data augmentation creates additional training samples by transforming the existing ones using geometric and color transformations. By providing the Deep Learning framework more examples to learn from, data augmentation expands the training set and thus minimizes the overfitting issue. Data augmentation never reduces the network's capacity, nor does it expand the computational complication and parameter fine-tuning. Instead, it serves as not explicit regularization technique, which has a greater importance in actual applications [12]. Data imbalance and data augmentation are closely related problems. One of the main approaches to solve unbalanced datasets is data augmentation [13] . Based on the most recent research[14],data augmentation techniques have been classified into two primary types: data-based data augmentation methods and network-based data augmentation methods. Data augmentation methods based on data are generally categorized into single-sample transformation, multi-sample synthesis, deep generative modeling, and virtual sample generating. Conversely, data augmentation techniques grounded on networks are mostly classified into controller network capacity approaches and learning strategies. Deep Generative Models, Multi-Sample Synthesis, and One-Sample Transform are the most often used data augmentation methods depending on data. The One-sample Transform is the most often utilized augmentation method. The single-sample transformation approach enhances the original data by means of geometry, color, sharpness, noise disruption, and random erasure applied on an input single sample dataset. New data produced by this procedure deviate from the original set. The processing of remote sensing picture data depends much on color. Commonly used in data augmentation, the single-sample transformation approach is simple and minimal time required [15]. Geometric transformations change visual shape when pixel values are mapped to new locations. New data come from image rotation, scaling, flipping, moving, cropping. Though its location and direction vary, often the class's fundamental form stays the same [16] [17]. The choice of data augmentation methods should be established on the specific attributes and domains of various image types. These methods should seek to maximize diversity while preserving the semantic information of the images. An example of this is that changing the orientation and reflecting of natural images can alter their semantic meaning. They are rarely employed for tasks using natural images. However, remote sensing images are particularly well-suited for this purpose. It is important to understand that variations in light conditions can significantly affect images captured in natural environments. Multi-Sample Synthesis is the second type of data augmentation. Multi-sample synthesis artificially mixes data from many images, unlike single-sample transformation. Multi-data synthesis is classified into two categories: feature space information synthesis and image space information synthesis. Image space information synthesis algorithms include between-class, Mixup, and sample pairing, whereas feature space algorithms include SMOTE. There are two main types of image spatial information synthesis methods: the multi-image nonlinear blending method and the linear stacking method of multiple images. Many algorithms are used for the linear stacking method of multiple images. These comprise the Mixup [18] and CutMix [19] algorithms as well as the sample pairing and between-class techniques. The third strategy is based on Deep Generative Models. Single-sample transformation and multi-sample synthesis generally use a individual image or many images as input data for producing new images. With little prior knowledge, the freshly created picture only comprises the information from the original image. By use of information about the probability density of the original data, the depth generating model employs a data augmentation method wherein fresh samples are randomly generated. Theoretically ideal, the deep generative model technique combines the whole dataset as previous knowledge. Dealing with the maximum

likelihood problem—that is, the difference between the model distribution and the data distribution—the deep generative model achieves its aim. Deep generative models may be classified into three forms when one considers the usage of maximum likelihood function approaches: the estimation technique, the implicit method, and the deformation method[20]. The implicit method is a technique that ignores the need for maximizing likelihood and can be demonstrated by a generative adversarial network (GAN)[21]. It effectively models the variation between two probability distributions using the learning capacity of a neural network. It effectively avoids the challenge of solving the probability function and has become the most efficient and important generative model. The use of GANs is prevalent among these technologies [21]. Deep learning-based data augmentation is currently gaining a lot of attention among academics. A data augmentation method based on an image style transfer technique was proposed by Mikolajczyk et al. after they examined and compared several strategies for augmentation of data in image classification tasks and gave illustrative examples [17].The first type of data augmentation is widely utilized in remote sensing because it is already part of popular machine learning frameworks like TensorFlow or Pytorch [22]. In a previous study,[23] utilized data augmentation techniques on remote sensing datasets, specifically employing horizontal flipping, vertical flipping, and 90° rotations. These transformations were employed because to their topology-preserving nature and minimal processing complexity. By using these simplest augmentation strategies, they successfully enhanced the Kappa index on the UC Merced dataset from 0.48 to 0.71, when assigning 50% for data training, 30% for data validation, and 20% for data testing.In the latest years, transformer models have proved significant capability in various computer vision operations, including classification of satellite image. Despite their success, the performance of these models can be further enhanced through effective data augmentation techniques. Data augmentation not only helps in mitigating overfitting but also improves the generalization capabilities of models.

This article aims to evaluate the impact of various data augmentation methods on the performance of transformers (Swin-tiny, Swin-small, Convit-small, and Crossvit-small) using the Eurosat dataset. The remainder of this article is arranged as follows: Section 2 discussion of the related works. In Section 3, “Materials and Methods,” the basic concepts and specific methods of augmentation techniques, transformer models and the model design are presented. Section 4 shows the investigational outcomes. Section 5 reviews the current research outcomes. Lastly, conclusions are illustrated in Section 6.

2 Related Work

[24] The research on Land Cover Image Classification explored the use of deep learning models (ResNet50, ResNeXt, AlexNet, MobileNetV3, and DenseNet), including transformers, to improve accuracy and efficiency in analyzing land cover images. By comparing CNNs and transformer-based methods, the study highlighted the superior performance of transformers, such as ViT and Swin Transformer, in reaching to good outcomes. The Eurosat dataset, comprising ten land cover classes from Sentinel-2 satellite images, was utilized for training and evaluation. Pre-trained weight models shown better accuracy than those learned from scratch according to validation accuracy curves. These results provide important progress in environmental analysis and urban planning as well as highlight the possibilities of transformer models in land cover categorization activities. [25]Combining transfer learning with the Swin Transformer model, the work presents a fresh approach for remote sensing image scene categorization. The model achieves great accuracy on six different remote sensing datasets by means of pre-training on ImageNet datasets and

migration learning. Results of validation demonstrate remarkable classification accuracy rates: 99.99% on UCM, 96.80% on AID, and 95.20% on NWPU. This method shows how well transformer models and transfer learning might be used to increase classification accuracy for remote sensing photos. For academics investigating transformer-based approaches in remote sensing applications, the work offers insightful analysis. [26]present a deep learning deep learning approach to raise satellite picture categorization accuracy. Large volumes of labeled data are a major obstacle in this field as obtaining such quantities may be costly and time-consuming. The scientists tackle issue by creating synthetic satellite pictures using Generative Adversarial Networks (GANs), therefore augmenting the current collection. Moreover, they extract images from Vision Transformers (ViTs), therefore improving the capacity of the classification algorithm to learn from the input. Comparatively to conventional techniques, their methodology shows improvement in categorization accuracy. Using traditional data augmentation techniques, the accuracy of the proposed approach is 76.7 percent; it is 98.7 percent now. The work reported in [27] mostly addresses the use of GANs to generate synthetic satellite pictures to increase the generalizing ability of deep classification models. DCGAN and WGAN-GP produced the synthetic satellite photos. The study showed that both architectures had similar impacts on model performance; WGAN-GP did not show any appreciable advantage over DCGAN in the present situation. Applying geometric augmentations like random horizontal flip, vertical flip, and rotation boosted the model's performance yet further. Two deep classification systems used were Wide Resnet50 and VGG16.The paper showed that using geometric augmentation increased model accuracy on all kinds of testing. Moreover, the combination of geometric augmentation and GAN-generated pictures turned out to be rather helpful, particularly in cases where data was limited as it sufficiently reduced overfitting tendencies. [28]Three datasets—EuroSAT [29], NWPU-RESISC45, and AID [30]—were tested using the Swin Transformer model. The validation accuracy findings were quite exceptional with the Swin Transformer attaining 99.02% efficiency on the EuroSAT dataset, 95.38% efficiency on the NWPU-RESISC45 dataset, and 95.90% efficiency on the AID dataset. These results prove the greater execution of the Swin architecture compared to existing methods in classification of remote sensing image. The consistent high justification accuracies across different datasets further validate the robustness and effectiveness of the Swin Transformer for land cover classification tasks. [31] suggested a vision transformer based remote sensing scene categorization approach while using many data augmentation methods (Geometric augmentation) to improve the performance. They maintained a competitive accuracy despite compressing half of the vision transformer model's layers, hence pruning the model. utilizing many remote sensing datasets—such as Merced, AID, Optimal31, and NWPU datasets—they demonstrated their effectiveness with the classification accuracies of their suggested technique utilizing the RGB pictures. In [32] the authors of the article suggest the use of a GAN named deeply supervised GAN to generate training samples for remote sensing photos covering Anhui Province in China. Applying this strategy to detect soil movement has demonstrated a 5% enhancement compared to the findings obtained without utilizing any data augmentation methodology.

3 Materials and Methods

3.1 The Eurosat Dataset

It includes 13 different spectral bands and 27,000 annotated and geo-referenced sentinel-2 images. In this study, the dataset will be used to train and evaluate classification models

since it contains a large and fairly distributed set of classes. Industrial buildings, residential buildings, rivers, lakes, pastures, forests, highways, and annual and permanent crops are among the ten categories that make up the dataset. The European Urban Atlas uses 64x64 pixel images that cover cities at a spatial resolution of 10 meters per pixel. The dataset's authenticity and integrity are ensured by its direct acquisition from the primary source [29]. An example of the dataset is presented in Fig.1.

Fig.1: Sample images of Eurosat dataset in RGB form.

3.2 Data Augmentation Techniques

One easy and more efficient way to upgrade the variety and scope of a training set is data augmentation. Working with small or unstructured datasets requires this method particularly as it is a necessary start in the preparation process [33]. Using many processing techniques, data augmentation generates new training examples from the initial dataset while preserving the quality of the initial class labels. By helping to avoid overfitting, these enhanced samples act to increase model robustness and generalizability [34]. Data augmentation techniques may be usually categorized into three categories depending on their underlying principles and procedures:

3.2.1 One-Sample Transform

Applying fundamental geometric augmentation techniques like scaling, rotating, translating, and flipping to current samples in order to create new ones is the most often used approach for improving datasets. These changes improve the capacity of the model to learn invariances and thus to generalize to fresh data. The authors in [23] used flip, translate, and rotation augmentation methods to supplement dataset of remote sensing scenes to be used in change detection job. [35] also used picture zooming or scale augmentation method for satellite image collection to be employed in land cover and use changes categorization. For the purpose of categorization, horizontal and vertical flipping methods were used in [22] to enrich satellite photos; they thus contribute to provide great accuracy. This work will simulate real-world differences in picture orientation and perspective using Flipping and Rotation augmentation methods to generate various training examples. The rotation transformation randomly rotates the picture to produce a change in the orientation of the visual elements. Rotation ensures that the model may recognize objects in any direction. The flip transformation turns the picture along either the horizontal or vertical axis.

3.2.2 Multi-Sample Synthesis

More advanced image-mixing algorithms, like Mixup [18] and CutMix [19], have been created in recent times to produce more instances for the model. Mixup is a method that

produces alternative examples of training by combining pairs of images and their associated labels using linear combinations. Particularly, the process includes combining two images and their corresponding labels by calculating a weighted sum. This results in the creation of a new image that is a combination of the two original images. This methodology enables the model to learn linear interactions among different classes, which enhances its robustness and capability for generalization. Studies have demonstrated that Mixup enhances performance on multiple image tasks related to classification. CutMix is an enhanced version of the Mixup technique that involves cutting out a section from one image and placing it onto another image. The labels from the original photos are combined and weighted based on the area of the patches to create the final image. CutMix improves the model's learning of data from various sections of the image, hence reducing the chance of overfitting to specific features. This method has been demonstrated to be highly successful in dealing with impeded objects in tasks involving object detection and scene classification.

3.2.3 Deep Generative Models

Generative modeling is an interesting approach for augmenting data as models such as GANs [36] learn the data dispersing to create fake samples that very nearly match the pictures collected from the original dataset. Comprising a pair of parallel-trained parallel neural networks (generator and discriminator). Generative Adversarial Networks (GANs) are built from While the discriminator tries to distinguish between real and synthetic pictures, the generator creates fake ones. By implementing an adversarial process, GANs have the capability to produce highly realistic images that enhance the training dataset. DCGAN[37] [38] is an example of a GAN that has been used for satellite images, WGAN-GP[27] and, Cycle GAN[39]. In addition, satellite images have been processed using conditional GAN [40]. Two GAN models are considered: the Wasserstein GAN with Gradient Penalty (WGAN-GP) and the Deep Convolutional GAN (DCGAN). According to our research, DCGAN is not very flexible when it comes to generating images at different resolutions. Nevertheless, our research confirms the results of [38] that it performs excellently when 64x64 images are being generated. Also, DCGAN effectively captures the latent space. After considering the original dataset's 64x64 picture size and the requirement to evaluate performance by adding generated images to the training dataset, DCGAN was selected because of its popularity for effectively handling images of this resolution. We choose WGAN-GP largely to reduce potential instability during the training procedure and achieve earlier convergence. This was noticed in the reduced time of the training period.

3.3 Selected Vision Transformers

One goal of this research is to discover how different data augmentation methods influence the classification performance of certain vision transformers. The transformers selected for this research comprise **Swin** Transformer (**tiny** and **small**), **Convit small**, and **Crossvit small**. The choice of these transformers is based on their distinctive architectural characteristics and proven efficacy. We want to study the responses of every transformer variant to data augmentation and the manner in which these approaches influence the classification accuracy of the data by using many augmentation techniques. The performance of these models on the Eurosat dataset will help to evaluate the effectiveness of augmentation methods in enhancing model robustness and generalization in satellite image classification tasks. Known by many as the Shifted Window Transformer, the Swin

transformer [41] offers a hierarchical framework that transfers certain local windows across many levels and conducts self-attention inside particular local windows. The unique architecture of the Swin Transformer guarantees computational efficiency and helps it to effectively manage many picture resolutions. The changing process captures both local and global input, therefore ensuring that the model understands complex visual patterns. Designed to fit different computational budgets and performance criteria, two popular variants of this architecture are Swin-tiny and Swin-small. Over several benchmarks, the Swin Transformer has shown extraordinary effectiveness in tasks like image classification, semantic segmentation, and object identification, outperforming convolutional neural networks and other transformer models. Combining convolutional layers with vision transformers allows the Convit [42] (Convolutional Vision Transformer) to use both advantages. It combines the worldwide context modeling capability of transformers with the locality and translation equivariance properties of convolutions. Combining the benefits of both techniques, this hybrid approach increases the robustness and efficacy of the model. Convit is very effective for image categorization chores as its architecture maintains the spatial hierarchy of pictures. Including convolutional inductive biases into the model improves its capacity to learn stable and accurate representations, hence improving performance on many different types of datasets. One branch of the Crossvit (Cross-Attention Vision Transformer) manages the input picture at a higher resolution using a smaller patch size, while the other branch manages it at a lower resolution using a larger patch size [43]. Crossvit's dual-path architecture lets it use its cross-attention techniques to gather coarse and fine characteristics. The cross-attention technique helps the model to merge data from both branches, hence enhancing its ability to detect objects and patterns at different sizes. For jobs requiring the extraction of features at many levels, including image categorization and object identification, Crossvit is very effective. The designs of certain transformer variants are reviewed in the Table 1.

Table 1: Architecture of selected transformers.

Variant	Layers' number	Hidden Size	Size of MLP	Heads' number	Parameter Count
Swin-tiny	12	96	384	3	28M
Swin-small	24	96	384	3	50M
ConViT-small	12	384	1536	12	85M
CrossviT-small	12	384 (small),192 (large)	1536 (each)	6 (small), 3 (large)	44M

3.4 Model Description

The model design for this study, as shown in Fig.2, provides a thorough method for assessing how different data augmentation strategies affect the performance of particular visual transformers in a task of classifying land use\cover using the Eurosat dataset. The dataset is the first input data source used in the workflow. Various trials were conducted, both with and without data augmentation. Images created by geometric, Mixup, CutMix, WGAN-GP and DCGAN augmentation are utilized in this study. These augmentation methods are used to enlarge the training set thus as to avoid overfitting of the model.

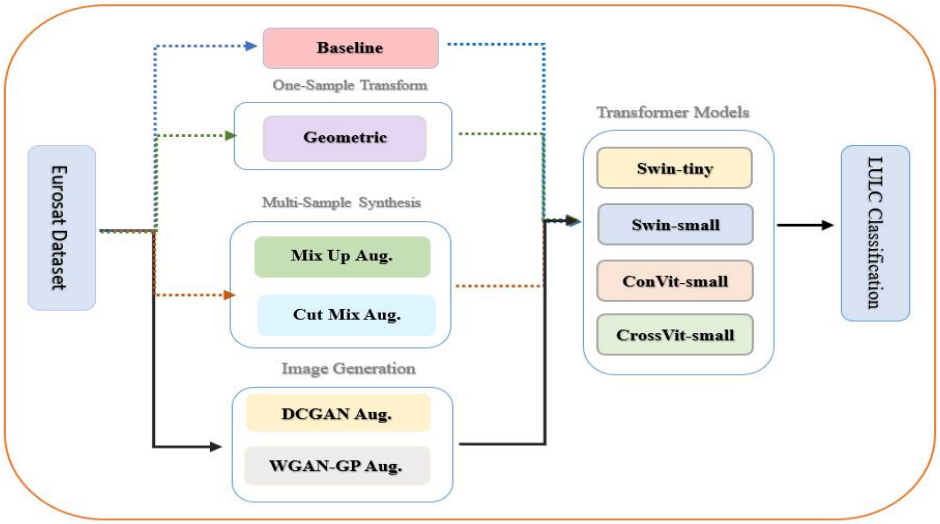


Fig.2: The project's work plan.

3.5 DCGAN and WGAN-GP Training

Comprising 27,000 images from satellites arranged into 10 groups, each shown as a 64x64 RGB image, the collection is All photos have a consistent size, and normalization is used to limit pixel values within the range (-1, 1), a technique recognized to improve GAN training. Inspired by the DCGAN paper [37] and the paper [27], we started convolutional layer weights with mean of 0 and standard deviation of 0.02. Using a mean of 1.0 and a standard deviation of 0.02, batchnorm layers were started from a normal distribution. With five blocks, the generator a deep convolutional network uses transposed convolution, batch norm, and activation layers in each block. In the final layer, a tanh activation is applied to accommodate the normalized image range. The generator creates a 64x64 RGB picture using a 100-dimensional noise vector as input. Designed as a deep convolutional network with five blocks, the discriminator binary classifies images—fake or genuine. Using Adam optimizer and a learning rate of 0.0002, both networks use binary cross-entropy as their loss function. We apply beta coefficients of 0.5 and 0. 999. Training consists of creating photos for every class independently, with 2000 to 3000 images for every class. The WGAN-GP was taught to produce 256 pictures for every class. [44] states that while this strategy may not always provide better pictures than the DCGAN method, it does have the benefit of improved training constancy. A generator construction that is similar to that employed in DCGAN was implemented. The learning rate and optimizer are kept constant to facilitate a comparison with DCGAN. Their results were compared using the code provided by reference [27]. The outcomes are shown in Fig.3 and Fig.4. GAN models are trained for 300 epochs, generating 256 images per class (2560 in total), constituting approximately 10% of the original dataset. The renamed GAN-generated images were stored in a folder to be add to the dataset through training, facilitating a comprehensive set of experimentations. Notably, the imageries retrieved from the Kaggle link were formatted as jpg, whereas the generated images adopted the png format. During image generation, a classification-based folder arrangement was employed, categorizing images by class. For model training, the image label was derived from the filename, which inherently conveyed the respective class affiliation of the image.

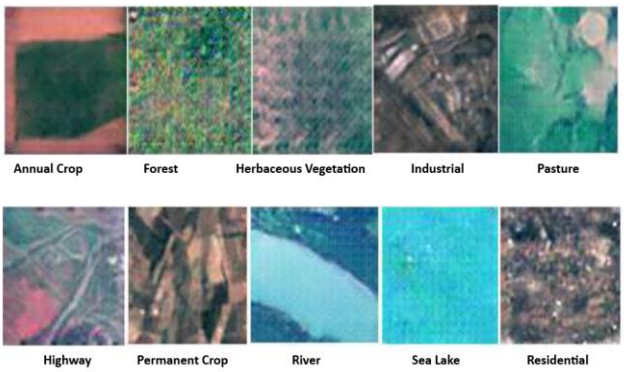


Fig.3: Example Images Produced by DCGAN.

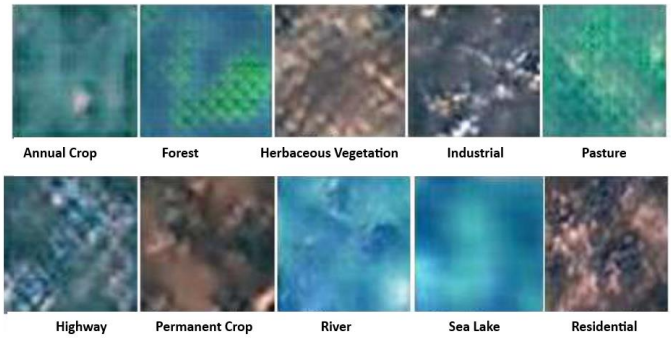


Fig.4: Example Images Produced by WGAN-GP.

3.6 Models Training

We created unique code specifically for effective preparation and data loading. Image resizing and normalizing were part of the preparation tasks. The dataset was randomly divided into three parts: 70% for training, 20% for validation, and 10% for testing to ensure that each subset of the data remained mutually exclusive, meaning no overlap of samples occurred between training, validation, and testing sets. Particularly for deep learning problems, these ratios are generally agreed upon in machine learning research as they provide a strong mix between model training, hyperparameter adjustment, and performance evaluation. Later we trained the same models using transfer learning (Swin tiny, Swin small, Convit, and Crossvit transformer) using pre-trained weights. These weights originated from ImageNet previously trained models. After 25 training cycles, every model was kept with the best validation accuracy. We computed the categorical cross-entropy loss with a 5e-4 learning rate Adam optimizer. The optimizer parameters, such as epsilon (1e-08) and betas set to (0.9, 0.999). The satellite images resized to 224x224 for Swin variants and Crossvit transformers, while images resized to 240 x 240 for Convit transformer. The images were normalized using standard deviations of [0.229, 0.224, 0.225] and mean values of [0.485, 0.456, 0.406]. In the training of transformer models on the Eurosat dataset, geometric augmentation techniques included Random Rotation within a specified degree range set to a maximum of 20 degrees, Random Horizontal Flip with a probability of 0.5, and Random Vertical Flip with a probability of 0.5. These augmentations were used to create diverse training samples by simulating real-world variations in image orientation and perspective. Mixup augmentation involves blending two images and their corresponding labels using a mixing parameter, lambda (λ), which is sampled from a Beta distribution with parameter

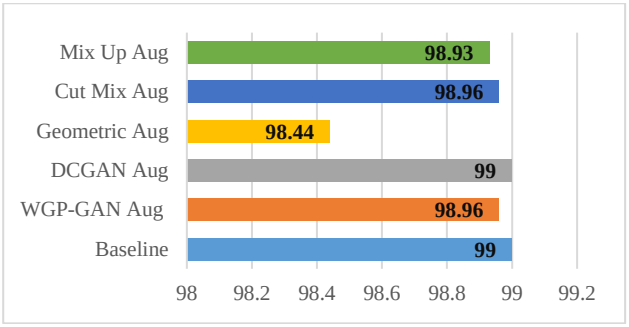
alpha. For this implementation, alpha was set to 0.1, encouraging the creation of blended images that are close to the original samples but still introduce variability. CutMix augmentation involves replacing a randomly selected part of an image with a part from another image in the training batch and adjusting the corresponding labels proportionally. This technique effectively combines aspects of both Cutout and Mixup augmentations by providing localized mixing of image patches and labels, which can enhance the model's robustness and generalization. CutMix augmentation was implemented with a Beta distribution parameter alpha set to 1.0.

4 Evaluation and Experimental Results

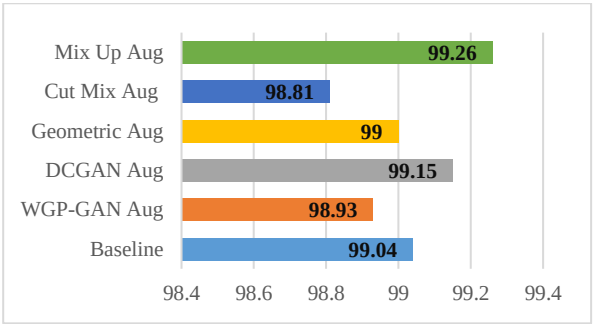
Using a test set—excluded from the training phase—we evaluated every model. We generate the top-1 accuracy as assessment criterion as well as the precision/recall curves. Our experiments made use of Pytorch: A Colab pro+ system using an NVIDIA V100 GPU. In this section, we present and compare the validation accuracy of various vision transformer variants, specifically Swin (tiny and small), ConVit-small, and CrossVit-small. The models were evaluated on a baseline dataset without augmentation and on augmented datasets using various data augmentation techniques, namely Mixup, CutMix, Geometric augmentation, DCGAN, and WGP-GAN. The results are illustrated in Fig.5.

Baseline validation accuracy of the Swin-tiny transformer was 98.96%. Various techniques of data augmentation changed the performance. DCGAN augmentation raised validation accuracy of 99% to show the greatest increase over baseline. With 98.93% for Mixup and CutMix respectively, both Mixup and CutMix augmentations generated validation accuracy very close to the baseline accuracy. On the other hand, geometric augmentation resulted in a somewhat lower accuracy of 98.44%. Equivalent to the baseline accuracy was 98.96% WGP-GAN augmentation. For the Swin-small transformer, the baseline validation accuracy was at 99.04%. Mixup augmentation shown a significant increase in a validation accuracy of 99.26%. DCGAN enhancement also did really well with an accuracy of 99.15%. CutMix augmentation generated a much lower accuracy of 98.81%; geometric augmentation matched the baseline accuracy of 99%. WGP-GAN augmentation was quite below the baseline with a 98.93% accuracy. Baseline validation accuracy for the Convit small arrived at 98.82%. WGP-GAN augmentation among the many augmentation techniques produced the greatest improvement with an accuracy of 99.07%. CutMix augmentation also showed a noteworthy increase with a 99.04% accuracy. Mix-up and DCGAN augmentations generated accuracy of 98.81% and 98.85%, respectively; both are quite below the baseline. Geometric augmentation reached considerably less than the baseline with an accuracy of 98.78%. From baseline validation accuracy of 98.82%, the Crossvit-small WGP-GAN augmentation generated the greatest improvement with an accuracy of 99.07%. DCGAN enhancement amply enhanced over the baseline with a 99% accuracy. Mixup and CutMix augmentations generated accuracy of 98.74% and 98.78%, respectively, both somewhat below the baseline. By use of geometric augmentation, an accuracy of 98.89% was achieved—rather above the baseline. Vision Transformers's performance on classification problems appears to be much improved with data augmentation. Advanced augmentation techniques as DCGAN and WGP-GAN often exhibited the highest gains in validation accuracy across several models because they provide diverse and meaningful training data. Mixup and CutMix augmentations also showed increases even though less consistent in surpassing the baseline performance. Different results came from geometric augmentation; some models performed considerably underperformance while others profited. These findings draw attention to

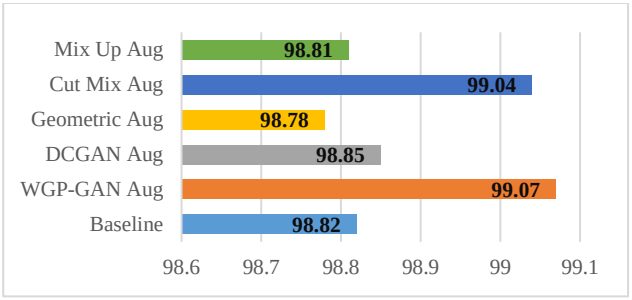
generally the viability of advanced data augmentation techniques in raising the accuracy of Vision Transformers.



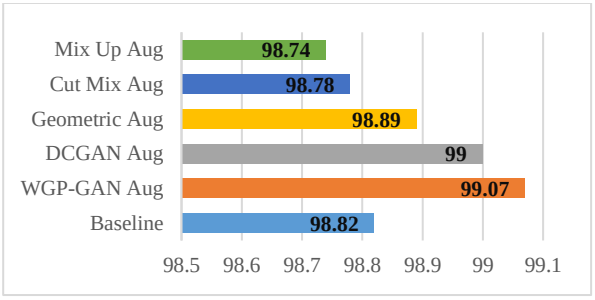
(a) Validation accuracy of Swin-tiny.



(b) Validation accuracy of Swin-small.



(c) Validation accuracy of Convit-small.



(d) Validation accuracy of Crossvit-small

Fig.5: The comparison results of using different data augmentation techniques on baseline Transformers.

5 Discussion

The evaluation of various augmentation methods on different transformer architectures (Swin-tiny, Swin-small, Convit-small, and Crossvit-small) using the Eurosat dataset provides valuable insights into their performance in terms of AUC, F1-score, and test accuracy. Below, we discuss the results as presented in Tables 2, 3, 4, and 5.

For **the baseline (no augmentation),** the Swin-tiny transformer achieved an AUC of 0.9997, with an F1-score of 0.9867 and a test accuracy of 98.67% (Table 2). The Swin-small transformer had an AUC of 0.9996, with an F1-score of 0.9819, and a lower test accuracy of 98.19% (Table 3). The Convit-small transformer showed better performance an AUC of 0.9997, with an F1-score of 0.9885, and a test accuracy of 98.85%, (Table 4). The Crossvit-small transformer had the highest baseline effectiveness with an AUC of 0.9997, an F1-score of 0.9881 and a test accuracy of 98.91%, (Table 5).

With **geometric augmentation**, the Swin-tiny transformer achieved an AUC of 0.9998, with an F1-score of 0.9878 and a test accuracy of 98.78% (Table 2). The Swin-small transformer showed an AUC of 0.9997, an F1-score of 0.9867, and a test accuracy of 98.67% (Table 3). The Convit-small transformer had an AUC of 0.9995, an F1-score of 0.9795, and a test accuracy of 97.94% (Table 4). The Crossvit-small transformer also achieved an AUC of 0.9997, with an F1-score of 0.9872 and a test accuracy of 98.78% (Table 5). Geometric

augmentation improved the effectiveness of Swin tiny and Crossvit-small, but Convit-small showed a decline.

Mixup data augmentation provided the highest boost in test accuracy for the Swin small transformer, reaching 99.26%, with an AUC of 0.9996 and an F1-score of 0.9871 (Table 3). The Convit-small transformer also benefited significantly, achieving a test accuracy of 98.98%, an AUC of 0.9995, and an F1-score of 0.9880 (Table 4). The Crossvit-small transformer showed a test accuracy of 98.74%, an AUC of 0.9998, and an F1-score of 0.9880 (Table 5). The Swin-tiny transformer had a test accuracy of 98.41%, an AUC of 0.9996, and an F1-score of 0.9841 (Table 2). Mixup augmentation significantly enhanced the performance of Swin small and Convit-small, while Swin-tiny showed less improvement.

CutMix data augmentation resulted in the highest performance for the Swin small transformer with a test accuracy of 98.96%, an AUC of 0.9998, and an F1-score of 0.9896 (Table 3). The Convit-small transformer showed a test accuracy of 98.50%, an AUC of 0.9997, and an F1-score of 0.9850 (Table 4). The Crossvit-small transformer had a test accuracy of 98.28%, an AUC of 0.9997, and an F1-score of 0.9828 (Table 5). The Swin tiny transformer experienced a significant drop in effectiveness with a test accuracy of 97.07%, an AUC of 0.9994, and an F1-score of 0.9703 (Table 2). CutMix was highly effective for Swin-small but resulted in a performance decline for Swin-tiny.

WGP-GAN augmentation performed well across all models. The Swin-tiny transformer achieved a test accuracy of 98.70%, with an AUC of 0.9998 and an F1-score of 0.9870 (Table 2). The Swin small transformer had a test accuracy of 99.00%, an AUC of 0.9997, and an F1-score of 0.9900 (Table 3). The Convit-small transformer showed a test accuracy of 98.83%, an AUC of 0.9997, and an F1-score of 0.9893 (Table 4). The Crossvit-small transformer had a test accuracy of 98.84%, with an AUC of 0.9999 and an F1-score of 0.9889 (Table 5). WGP-GAN was particularly effective for Swin-small and Crossvit-small, providing notable improvements in test accuracy and F1-scores.

DCGAN augmentation showed mixed results. The Swin tiny transformer had a test accuracy of 98.56%, an AUC of 0.9998, and an F1-score of 0.9855 (Table 2). The Swin small transformer achieved a test accuracy of 98.70%, with an AUC of 0.9997 and an F1-score of 0.9870 (Table 3). The Convit-small transformer displayed a noticeable drop in test accuracy to 97.73%, with an AUC of 0.9995 and an F1-score of 0.9783 (Table 4). The Crossvit-small transformer had a test accuracy of 98.54%, an AUC of 0.9996, and an F1-score of 0.9854 (Table 5). DCGAN was less effective for ConVit small, showing inconsistent results across the models. Across all augmentation methods, Mixup and WGP-GAN generally provided the best improvements in test accuracy. The Swin small transformer benefited significantly from Mixup augmentation, achieving the highest test accuracy of 99.26% (Table 3). WGP-GAN also performed well across various models, particularly enhancing Swin small and Crossvit-small (Tables 3 and 5). Geometric augmentation was most effective for Swin tiny and Crossvit-small (Tables 2 and 5), while CutMix was notably less effective for Swin tiny but beneficial for Swin small (Tables 2 and 3). DCGAN showed inconsistent results, being less effective for Convit-small but relatively better for Swin models (Table 4). Tailoring augmentation strategies to specific model architectures is crucial for optimizing performance.

Table 2: Results of applying Swin-tiny transformer on Eurosat Dataset.

Augmentation Method	Epoch	Validation Acc.	AUC	F1	Test Acc.
Baseline	8	99.00	0.9997	0.9867	98.67%
WGP-GAN Aug	20	98.96	0.9998	0.9870	98.70%
DCGAN Aug	15	99.0	0.9998	0.9855	98.56%
Geometric Aug	15	98.44	0.9998	0.9878	98.78%
Cut Mix Aug	11	98.96	0.9994	0.9703	97.07
Mix Up Aug	24	98.93	0.9996	0.9841	98.41

Table 3: Results of applying Swin-small transformer on Eurosat Dataset.

Augmentation Method	Epoch	Validation Acc.	AUC	F1	Test Acc.
Baseline	9	99.04	0.9996	0.9819	98.19
WGP-GAN Aug	18	98.93	0.9997	0.9900	99.00
DCGAN Aug	25	99.15	0.9997	0.9870	98.70
Geometric Aug	19	99.00	0.9997	0.9867	98.67
Cut Mix Aug	16	98.81	0.9998	0.9896	98.96
Mix Up Aug	20	99.26	0.9996	0.9871	98.70

Table 4: Results of applying Convit-small transformer on Eurosat Dataset.

Augmentation Method	Epoch	Validation Acc.	AUC	F1	Test Acc.
Baseline	20	98.82	0.9997	0.9885	98.85
WGP-GAN Aug	22	99.07	0.9998	0.9883	98.83
DCGAN Aug	13	98.85	0.9995	0.9783	97.83
Geometric Aug	12	98.78	0.9996	0.9795	97.94
Cut Mix Aug	24	99.04	0.9997	0.9850	98.50
Mix Up Aug	17	98.81	0.9995	0.9898	98.98

Table 5: Results of applying Crossvit- small transformer on Eurosat Dataset.

Augmentation Method	Epoch	Validation Acc.	AUC	F1	Test Acc.
Baseline	14	98.82	0.9997	0.9891	98.91
WGP-GAN Aug	24	99.07	0.9999	0.9889	98.89
DCGAN Aug	9	99.00	0.9999	0.9854	98.54
Geometric Aug	25	98.89	0.9996	0.9872	98.72%
Cut Mix Aug	6	98.78	0.9997	0.9828	98.28
Mix Up Aug	5	98.74	0.9998	0.9874	98.74

6 Conclusion

The results demonstrate that data augmentation techniques can enhance the performance of transformers for the classification of satellite image. However, the success of these techniques is highly dependent on the model architecture. Mixup and WGP-GAN augmentations generally provided the best improvements, making them strong candidates for further exploration in future studies. Geometric augmentation and CutMix also showed promise but require careful consideration of the model being used. Tailoring augmentation strategies to specific models is crucial for optimizing performance and achieving the best possible results. In conclusion, this study provides effective visions into the usage of data augmentation techniques for improving transformer models in various tasks of remote sensing. Future research should continue to explore and refine these techniques, considering the unique characteristics of different model architectures and datasets.

7 References

[1] G. Cheng, X. Xie, J. Han, L. Guo, and G. Xia, "Remote Sensing Image Scene Classification Meets Deep Learning: Challenges, Methods, Benchmarks, and Opportunities," *IEEE J Sel Top Appl Earth Obs Remote Sens*, vol. 13, pp. 3735–3756, 2020, [Online]. Available: https://api.semanticscholar.org/CorpusID:218486791

[2] H. S. Alhichri, A. S. Alswayed, Y. Bazi, N. Ammour, and N. A. Alajlan, "Classification of Remote Sensing Images Using EfficientNet-B3 CNN Model With Attention," *IEEE Access*, vol. 9, pp. 14078–14094, 2021, [Online]. Available: https://api.semanticscholar.org/CorpusID:231725395

[3] X. Wu, F. Yang, T. Zhou, and X. Lin, "Rethinking the Impacts of Overfitting and Feature Quality on Small-scale Video Classification," *Proceedings of the 29th ACM International Conference on Multimedia*, 2021, [Online]. Available: https://api.semanticscholar.org/CorpusID:239011486

[4] N. Srivastava, G. E. Hinton, A. Krizhevsky, I. Sutskever, and R. Salakhutdinov, "Dropout: a simple way to prevent neural networks from overfitting," *J. Mach. Learn. Res.*, vol. 15, pp. 1929–1958, 2014, [Online]. Available: https://api.semanticscholar.org/CorpusID:6844431

[5] L. Wan, M. D. Zeiler, S. Zhang, Y. LeCun, and R. Fergus, "Regularization of Neural Networks using DropConnect," in *International Conference on Machine Learning*, 2013. [Online]. Available: https://api.semanticscholar.org/CorpusID:2936324

[6] A. Kendall and Y. Gal, "What Uncertainties Do We Need in Bayesian Deep Learning for Computer Vision?," *ArXiv*, vol. abs/1703.04977, 2017, [Online]. Available: https://api.semanticscholar.org/CorpusID:71134

[7] 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

[8] 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

[9] G. A. Carpenter and W. D. Ross, "ART-EMAP: A neural network architecture for object recognition by evidence accumulation," *IEEE Trans Neural Netw*, vol. 6 4, pp. 805–18, 1995, [Online]. Available: https://api.semanticscholar.org/CorpusID:22677117

[10] Q. Zhang, A. Liu, and X. Tong, "Early stopping criterion for belief propagation polar decoder based on frozen bits," *Electron Lett*, vol. 53, pp. 1576–1578, 2017, [Online]. Available: https://api.semanticscholar.org/CorpusID:116412641

[11] C. Shorten and T. M. Khoshgoftaar, "A survey on Image Data Augmentation for Deep Learning," *J Big Data*, vol. 6, pp. 1–48, 2019, [Online]. Available: https://api.semanticscholar.org/CorpusID:195811894

[12] H. A. Alhaija, S. K. Mustikovela, L. M. Mescheder, A. Geiger, and C. Rother, "Augmented Reality Meets Computer Vision: Efficient Data Generation for Urban Driving Scenes," *Int J Comput Vis*, vol. 126, pp. 961–972, 2017, [Online]. Available: https://api.semanticscholar.org/CorpusID:3731652

[13] A. Fernández, S. García, M. Galar, R. C. Prati, B. Krawczyk, and F. Herrera, "Learning from Imbalanced Data Sets," in *Cambridge International Law Journal*, 2018. [Online]. Available: https://api.semanticscholar.org/CorpusID:53046396

[14] X. Hao, L. Liu, R. Yang, L. Yin, L. Zhang, and X. Li, "A Review of Data Augmentation Methods of Remote Sensing Image Target Recognition," Feb. 01, 2023, *MDPI*. doi: 10.3390/rs15030827.

[15] D. A. Ma, P. Tang, L. J. Zhao, and Z. Zhang, "Review of Data Augmentation for Image in Deep Learning," *J. Image Graph*, vol. 26, pp. 0487–0502, 2021.

[16] L. Perez and J. Wang, "The Effectiveness of Data Augmentation in Image Classification using Deep Learning," *ArXiv*, vol. abs/1712.04621, 2017, [Online]. Available: https://api.semanticscholar.org/CorpusID:12219403

[17] A. Mikołajczyk and M. Grochowski, "Data augmentation for improving deep learning in image classification problem," *2018 International Interdisciplinary PhD Workshop (IIPhDW)*, pp. 117–122, 2018, [Online]. Available: https://api.semanticscholar.org/CorpusID:49348179

[18] H. Zhang, M. Cissé, Y. Dauphin, and D. Lopez-Paz, "mixup: Beyond Empirical Risk Minimization," *ArXiv*, vol. abs/1710.09412, 2017, [Online]. Available: https://api.semanticscholar.org/CorpusID:3162051

[19] S. Yun, D. Han, S. J. Oh, S. Chun, J. Choe, and Y. J. Yoo, "CutMix: Regularization Strategy to Train Strong Classifiers With Localizable Features," *2019 IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 6022–6031, 2019, [Online]. Available: https://api.semanticscholar.org/CorpusID:152282661

[20] M. F. Hu, X. Zuo, and J. W. Liu, " Survey on Deep Generative Model," *Acta Autom. Sin*, vol. 48, pp. 40–74, 2022.

[21] I. J. Goodfellow *et al.*, "Generative Adversarial Nets," *In Advances in Neural Information Processing Systems*, pp. 2672–2680, 2014.

[22] M. Abdelhack, "A Comparison of Data Augmentation Techniques in Training Deep Neural Networks for Satellite Image Classification," *ArXiv*, vol. abs/2003.13502, 2020, [Online]. Available: https://api.semanticscholar.org/CorpusID:214714063

[23] X. Yu, X. Wu, C. Luo, and P. Ren, "Deep learning in remote sensing scene classification: a data augmentation enhanced convolutional neural network framework," *GIsci Remote Sens*, vol. 54, pp. 741–758, 2017, [Online]. Available: https://api.semanticscholar.org/CorpusID:64616771

[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] 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

[26] A. Alzahem, W. Boulila, A. Koubaa, Z. Khan, and I. Alturki, "Improving satellite image classification accuracy using GAN-based data augmentation and vision transformers," *Earth Sci Inform*, vol. 16, pp. 4169–4186, 2023, [Online]. Available: https://api.semanticscholar.org/CorpusID:265412833

[27] O. Adedeji, P. Owoade, O. Ajayi, and O. F. Arowolo, "Image Augmentation for Satellite Images," *ArXiv*, vol. abs/2207.14580, 2022, [Online]. Available: https://api.semanticscholar.org/CorpusID:251196989

[28] 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

[29] 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

[30] 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

[31] Y. Bazi, L. Bashmal, M. M. Al Rahhal, R. Al Dayil, and N. Al Ajlan, "Vision Transformers for Remote Sensing Image Classification," *Remote. Sens.*, vol. 13, p. 516, 2021, [Online]. Available: https://api.semanticscholar.org/CorpusID:231992220

[32] N. Lv *et al.*, "Remote Sensing Data Augmentation Through Adversarial Training," *IEEE J Sel Top Appl Earth Obs Remote Sens*, vol. 14, pp. 9318–9333, 2021, doi: 10.1109/JSTARS.2021.3110842.

[33] E. D. Cubuk, B. Zoph, D. Mané, V. Vasudevan, and Q. V Le, "AutoAugment: Learning Augmentation Strategies From Data," *2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 113–123, 2019, [Online]. Available: https://api.semanticscholar.org/CorpusID:196208260

[34] E. A. Mohammed and A. Lakizadeh, "A Swin Transformer-based method for Classification of Land Use and Land Cover Images," *International Journal of Advances in Soft Computing and its Applications*, vol. 16, no. 3, pp. 328–347, 2024, doi: 10.15849/IJASCA.241130.18.

[35] B. Hu, C.-H. Lei, D. Wang, S. Zhang, and Z. Chen, "A Preliminary Study on Data Augmentation of Deep Learning for Image Classification," *Proceedings of the 11th Asia-Pacific Symposium on Internetware*, 2019, [Online]. Available: https://api.semanticscholar.org/CorpusID:195750783

[36] C. Bowles *et al.*, "GAN Augmentation: Augmenting Training Data using Generative Adversarial Networks," *ArXiv*, vol. abs/1810.10863, 2018, [Online]. Available: https://api.semanticscholar.org/CorpusID:53024682

[37] A. Radford, L. Metz, and S. Chintala, "Unsupervised Representation Learning with Deep Convolutional Generative Adversarial Networks," *CoRR*, vol. abs/1511.06434, 2015, [Online]. Available: https://api.semanticscholar.org/CorpusID:11758569

[38] A. Gautam, M. A. Sit, and I. Demir, "Realistic River Image Synthesis Using Deep Generative Adversarial Networks," in *Frontiers in Water*, 2020. [Online]. Available: https://api.semanticscholar.org/CorpusID:211677974

[39] C. X. Ren, A. Ziemann, A. M. S. Durieux, and J. Theiler, "Cycle-Consistent Adversarial Networks for Realistic Pervasive Change Generation in Remote Sensing Imagery," *2020 IEEE Southwest Symposium on Image Analysis and Interpretation (SSIAI)*, pp. 42–45, 2019, [Online]. Available: https://api.semanticscholar.org/CorpusID:208512998

[40] A. Kulkarni, T. Mohandoss, D. Northrup, E. Mwebaze, and H. Alemohammad, "Semantic Segmentation of Medium-Resolution Satellite Imagery using Conditional Generative Adversarial Networks," *ArXiv*, vol. abs/2012.03093, 2020, [Online]. Available: https://api.semanticscholar.org/CorpusID:227335388

[41] 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

[42] 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

[43] 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

[44] H. Yassine, K. Tout, and M. Jaber, "IMPROVING LULC CLASSIFICATION FROM SATELLITE IMAGERY USING DEEP LEARNING – EUROSAT DATASET," *The International Archives of the Photogrammetry, Remote Sensing and Spatial Information Sciences*, vol. XLIII-B3-2021, pp. 369–376, 2021, doi: 10.5194/isprs-archives-XLIII-B3-2021-369-2021.