

Detecting GAN-Generated Human Face Images using Explainable AI

Abhil Thadathikunnel Balakrishnan

Dept. of Electrical and Microsystems Engineering OTH Regensburg, Germany

Abstract: Artificial intelligence has made it possible to generate highly realistic human face images using generative models. These synthetic images can be difficult to distinguish from real photographs, which creates risks in areas such as misinformation, identity fraud, and digital security. Detecting whether an image is real or AI-generated has therefore become an important research problem.

Recent advancements in Generative Adversarial Networks (GANs) have enabled the generation of highly realistic synthetic faces that are difficult to detect using traditional methods [4], [5]. This project proposes a deep learning framework for detecting AI-generated face images using a Vision Transformer (ViT-Large) architecture. Vision Transformers process images using a self-attention mechanism that captures global relationships between image patches [1].

The model was trained using real face images and synthetic images generated by StyleGAN1 from the Diverse Deepfake Dataset. To improve dataset consistency, face images were aligned and cropped before training. Balanced sampling was used to address class imbalance during training.

To understand how the model makes decisions, two Explainable Artificial Intelligence (XAI) techniques were used: Attention Rollout and Integrated Gradients. Integrated Gradients is a widely used attribution method that identifies which parts of an input image contribute most to the model's prediction [6].

Experimental results show strong detection performance when tested on StyleGAN1 images. However, performance decreased when testing on StyleGAN2 images due to differences between generative models. This highlights the importance of training deepfake detection systems using diverse datasets.

Index Terms—Deepfake detection, Vision Transformer, explainable artificial intelligence, Integrated Gradients, attention rollout.

1. INTRODUCTION



Figure 1.1: Example of Real & AI-Generated Images

Artificial intelligence has advanced rapidly in recent years, especially in the field of image generation. Generative models are now capable of producing synthetic images that appear extremely realistic. Among these models, **Generative Adversarial Networks (GANs)** are widely used to generate artificial human faces [4], [5].

While these technologies have useful applications such as media generation and visual effects, they also create serious risks. AI-generated images can be used to spread misinformation, impersonate individuals, or manipulate media content. As a result, detecting whether an image is real or AI-generated has become an important research topic.

Early deepfake detection systems relied on **Convolutional Neural Networks (CNNs)** to detect visual artifacts introduced during image manipulation. For example, MesoNet is a CNN architecture designed to detect manipulated face images by analyzing subtle visual inconsistencies [2].

However, modern generative models produce images with fewer visible artifacts, making detection more challenging. Recently, transformer-based models have been introduced for image classification tasks. The **Vision Transformer (ViT)** processes images as sequences of patches and learns global relationships between image regions using self-attention mechanisms [1].

In addition to accurate predictions, it is important to understand how a model makes decisions. Explainable Artificial Intelligence techniques help researchers visualize the regions of an image that influence model predictions. Integrated Gradients is one such technique that provides attribution scores for input features in deep neural networks [6].

This project develops a Vision Transformer-based system for detecting AI-generated face images and uses explainability techniques to visualize the model's decision process. This project focuses on StyleGAN-generated faces and evaluates cross-generator generalization from StyleGAN1 to StyleGAN2.

2. OBJECTIVES

The main objectives of this project are:

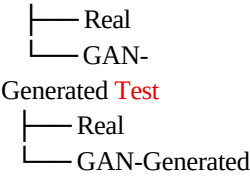
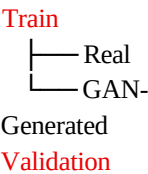
- 1. Develop a deep learning model capable of detecting AI-generated face images.
- 2. Train the model using real images and StyleGAN1-generated images.
- 3. Improve model performance using preprocessing techniques such as face alignment and cropping.
- 4. Apply explainability techniques to understand how the model makes predictions.
- 5. Evaluate the model on images generated by different generative models to study cross-domain performance.

3. DATASET DESCRIPTION

This project uses images from the **Diverse Deepfake Dataset**. The dataset used in this project was obtained from the **Ostbayerische Technische Hochschule (OTH) Regensburg**. The dataset contains two categories:

- Real face images
- AI-generated face images

In this project, synthetic images were generated using **StyleGAN1**, which is a style-based GAN architecture capable of producing highly realistic human faces [5]. The dataset is organized as:



To improve dataset consistency, all images were cropped to **224 × 224 pixels** and centered before training.

3.1 Dataset Statistics

Total Images: 147042		
	REAL	GAN-GENERATED
TRAIN	32414	74785
TEST	5823	12921
VALIDATION	6438	14652

Table 3.1: Dataset statistics for the Diverse Deepfake Dataset (StyleGAN1)

4. UNDERSTANDING VISION TRANSFORMER (ViT)

4.1 Traditional CNN Approach

Most image classification models use **Convolutional Neural Networks (CNNs)**. CNNs scan images using convolution filters to detect patterns such as edges, textures, and shapes.

However, CNNs mainly focus on **local features**, which means they analyze small areas of the image at a time.

4.2 Vision Transformer Concept

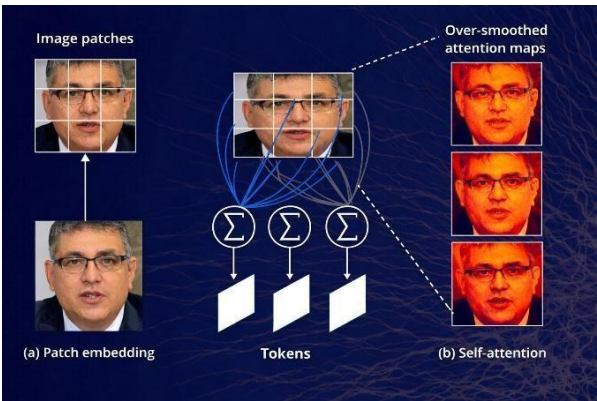


Figure 4.2.1: Vision Transformer Workflow

Vision Transformer (ViT) processes an image by dividing it into fixed-size patches, which are treated as tokens similar to words in natural language processing.

These patch tokens are embedded and passed through transformer encoder layers that use **self-attention** to learn relationships between different regions of the image. This mechanism allows the model to capture global contextual information and extract meaningful features, which are then used by a classifier to predict the final image category.

Vision Transformers apply the transformer architecture, originally developed for natural language processing, to image data [1].

Instead of analyzing images using convolution filters, Vision Transformers divide the image into small patches.

In this project:

An image of size 224×224 is divided into 16×16 **patches**, producing 196 patches.

Each patch is treated as a token, similar to words in a sentence.

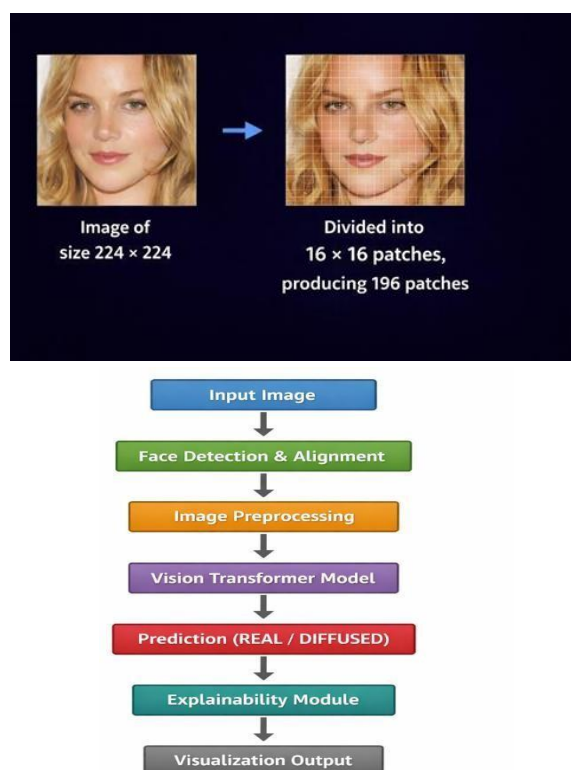


Figure 4.2.2: Illustration of Patch Extraction

4.3 Patch Embedding

Each patch is converted into a numerical vector that represents its visual information. These vectors are then processed by the transformer model.

4.4 Self-Attention Mechanism

The key component of the transformer architecture is **self-**

attention. Self-attention allows the model to understand relationships between different patches of the image.

For example, a patch containing an eye may be related to patches containing the nose or mouth. This mechanism enables the model to analyze the **global structure of the face**.

4.5 Classification Token

A special token called the **CLS token** collects information from all patches. This token is used by the model to produce the final prediction.

In this project, the prediction is:

REAL
 or
 GAN-Generated

5. SYSTEM ARCHITECTURE

The proposed system processes images through several stages.

The system not only predicts whether an image is real or fake but also produces visual explanations of the model's decision. In addition to the deep learning classification model, statistical image analysis techniques such as RGB histogram distribution and FFT frequency spectrum visualization are used for

6. DETECTION PIPELINE

When an image is tested, the system processes it through multiple steps.

Step 1: Image Loading

The image is loaded using OpenCV.

Step 2: Face Detection

The system uses **MediaPipe FaceLandmarker** to detect facial landmarks such as the eyes, nose, mouth, and face outline.

These landmarks help locate the face in the image.

Step 3: Face Alignment

Faces may appear in different orientations across images. To make the dataset consistent, the face is rotated so that the eyes are horizontally aligned. This improves the model's ability to focus on facial features.

Step 4: Face Cropping

The aligned face is cropped and resized to 224×224

pixels, which is the input size required by the Vision Transformer model.

Step 5: Image Preprocessing

The image is converted to a tensor and normalized using ImageNet mean and standard deviation values.

Step 6: Statistical Image Analysis (Histogram & FFT)

For qualitative analysis and visualization, additional statistical image analysis can be performed on the image:

- **RGB Histogram:** The color distribution of the image is analyzed by computing histograms for the Red, Green, and Blue channels. GAN-generated images may show subtle color distribution inconsistencies compared to real photographs.
- **FFT Spectrum:** The Fast Fourier Transform (FFT) converts the image from the spatial domain to the frequency domain. GAN-generated images often contain frequency-domain artifacts that can be observed through spectral analysis.

These analyses are used only for interpretation and visualization purposes and do not influence the Vision Transformer's classification output.

Step 7: Model Prediction

The Vision Transformer processes the image and outputs a numerical value called a **logit**.

A sigmoid function converts this value into a probability.

Prediction rule:

Probability $\geq 0.5 \rightarrow$ GAN-GENERATED
Probability $< 0.5 \rightarrow$ REAL

Step 8: Explainability Generation

Two explainability methods are used.

- **Attention Rollout:** This method visualizes how the transformer attends to different image patches.
- **Integrated Gradients:** Integrated Gradients calculates the contribution of each pixel to the model's prediction [6].

These methods produce heatmaps that highlight important regions in the image.

7. TRAINING METHODOLOGY

The model used in this project is **Vision Transformer Large (ViT-Large Patch16-224)** [1].

Training Configuration

- Optimizer: AdamW
- Learning Rate: $3e-5$
- Batch Size: 64
- Epochs: 10

8. DATA AUGMENTATION

Several augmentation techniques were applied:

- Horizontal flip
- Color jitter
- Grayscale
- Gaussian blur
- Auto contrast

Balanced sampling was used to prevent bias between real and GAN-generated images.

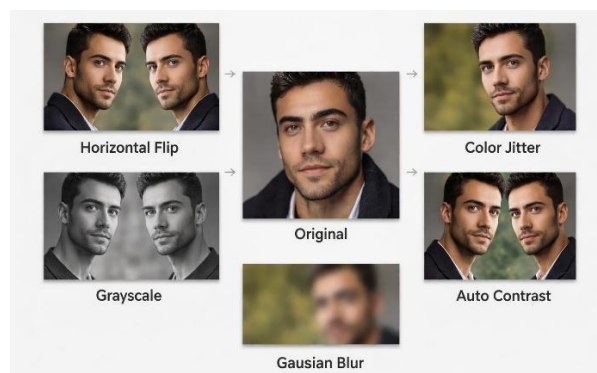


Figure 8.1: Visualization of Image Augmentation Techniques

9. CHALLENGES FACED DURING THE PROJECT

- **Problem 1: Model Focused Outside the Face**

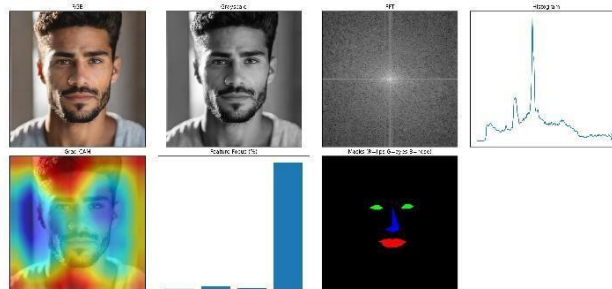


Figure 9.1.1: Output of First training set

Grad-CAM (Gradient-weighted Class Activation Mapping) was initially tested during early CNN-based experiments to visualize model attention. However, the final system uses Vision Transformer-specific explainability methods, namely Attention Rollout and Integrated Gradients.

Instead of focusing on semantic facial landmarks—such as the eyes, nose, or mouth—the model's attention was dispersed across the periphery of the image. The highest

activation (indicated by the red "hot" zones) is concentrated on the background, the hair, and the transition areas between the subject and the frame.

Solution Implemented

Initially, the images were not properly aligned during training. To address this, all images were cropped to 240×240 pixels, and faces were aligned to the center.

This significantly improved model performance.

• Problem 2: Poor Prediction

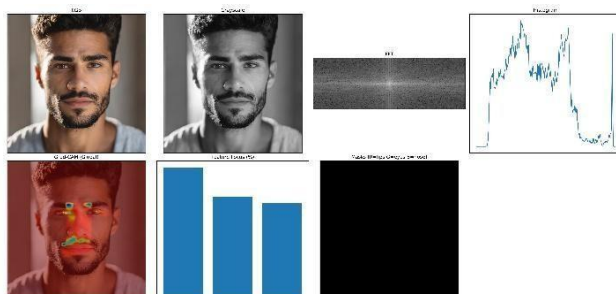


Figure 9.1.2: Output of Second training set

In the second model, the model's interpretability significantly improved; the Grad-CAM visualizations shifted from the background to critical facial landmarks, with decision-making weights now distributed across the eyes (39.9%), nose (31.0%), and lips (29.1%).

Solution Implemented

During training, a dataset organization problem was identified. Some GAN-generated images had been placed in the real-image folder because the files had not been separated correctly according to their naming convention. This caused label noise and led to incorrect predictions by the model.

Images with the suffix "R" were moved to the Real folder, and images with the suffix "F" were moved to the GAN-Generated folder.

• Final Prediction and Visualization Outputs

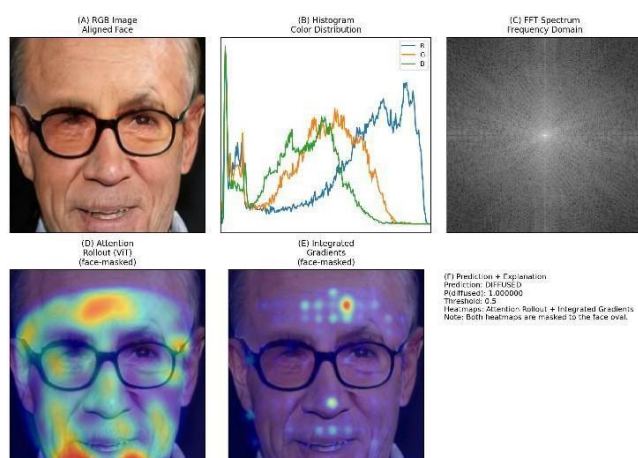


Figure 9.1.3: Final prediction and visualization outputs: (A) RGB image, (B) RGB histogram, (C) FFT spectrum, (D) Attention Rollout (ViT), and (E) Integrated Gradients.

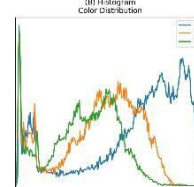
(A) RGB Image (Aligned Face)



The RGB image represents the preprocessed input image after face alignment and cropping. Face alignment ensures that facial landmarks such as the eyes, nose, and mouth are consistently positioned across samples before being passed to the model.

Purpose of the Analysis: Aligned face images allow the neural network to focus on **facial features rather than pose variations**, which is crucial for detecting subtle artifacts introduced by GAN-based image generation.

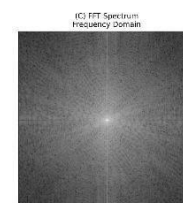
(B) Histogram-Color Distribution



The histogram shows the **distribution of pixel intensities across the Red (R), Green (G), and Blue (B) channels** in the image.

Purpose of the Analysis: Color distribution analysis helps identify **statistical inconsistencies between real and generated images**, which can serve as useful cues for synthetic image detection.

(C) FFT Spectrum

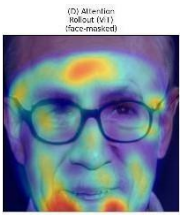


The FFT (Fast Fourier Transform) spectrum visualizes the image in the **frequency domain**, representing how spatial patterns in the image are distributed across different frequencies.

Purpose of the Analysis: GAN-generated images often contain distinct frequency artifacts, making FFT analysis

useful for identifying synthetic content.

(D) Attention Rollout (ViT), Face-Masked



Attention Rollout visualizes which regions of the image the Vision Transformer (ViT) model focuses on when making predictions.

Observed Focus Areas

The model typically attends to:

- Forehead
- Eyes
- Mouth region
- Skin texture patterns

These areas often contain generation artifacts in synthetic images.

(E) Integrated Gradients



Integrated Gradients is an Explainable AI (XAI) technique that highlights the pixels most responsible for the model's prediction.

Purpose of the Analysis: This method confirms whether the model focuses on meaningful facial features rather than irrelevant background noise.

10. CROSS-DOMAIN EVALUATION

The model was trained using StyleGAN1 images. When tested on StyleGAN1 images, the model achieved high accuracy.

File Name	Dataset Name	True Label	Model Prediction	Result Type (TP/TN/FP/FN)	Confidence Score
Data Set 1 StyleGAN1 Total 370 (TP-362 FN-4) 4 IMAGES NOT TESTED					
0.691315.png	StyleGAN1	AI	GAN-GENERATED	TP	1
0.765591.png	StyleGAN1	AI	GAN-GENERATED	TP	1
0.796213.png	StyleGAN1	AI	GAN-GENERATED	TP	1
0.796441.png	StyleGAN1	AI	GAN-GENERATED	TP	0.9998
0.836086.png	StyleGAN1	AI	GAN-GENERATED	TP	1
0.842027.png	StyleGAN1	AI	GAN-GENERATED	TP	1
0.856839.png	StyleGAN1	AI	GAN-GENERATED	TP	1
0.868483.png	StyleGAN1	AI	GAN-GENERATED	TP	0.9959
0.903764.png	StyleGAN1	AI	GAN-GENERATED	TP	0.9999
0.906319.png	StyleGAN1	AI	GAN-GENERATED	TP	1
n 021186.png	StyleGAN1	AI	GAN-GENERATED	TP	1

However, when tested on StyleGAN2 images, performance decreased.

File Name	Dataset Name	True Label	Model Prediction	Result Type (TP/TN/FP/FN)	Confidence Score
-----------	--------------	------------	------------------	---------------------------	------------------

Data Set Stylegan2 Total 303(TP-45 FN-251) 7 IMAGES NOT TESTED

00005.png	StyleGAN2	AI	REAL	FN	0.8887
seed0111.png	StyleGAN2	AI	GAN-GENERATED	TP	0.7709
seed0049.png	StyleGAN2	AI	GAN-GENERATED	TP	0.9836
seed0147.png	StyleGAN2	AI	GAN-GENERATED	TP	0.9976
seed0018.png	StyleGAN2	AI	GAN-GENERATED	TP	0.7465
seed0146.png	StyleGAN2	AI	REAL	FN	0.9999
00124.png	StyleGAN2	AI	REAL	FN	0.9998
00126.png	StyleGAN2	AI	REAL	FN	1.0000
00092.png	StyleGAN2	AI	REAL	FN	0.9945
seed0003.png	StyleGAN2	AI	GAN-GENERATED	TP	0.9942
seed0075.png	StyleGAN2	AI	REAL	FN	0.9990
00127.png	StyleGAN2	AI	REAL	FN	0.9994

This happens because StyleGAN2 generates more realistic images with fewer visible artifacts, resulting in a domain shift problem. Since the model was trained using StyleGAN1 images, it did not learn the artifact patterns specific to StyleGAN2.

11. EXPERIMENTAL RESULTS

Main metrics:	
Accuracy:	0.9820 (98.20%)
Precision:	0.9883 (98.83%)
Recall:	0.9856 (98.56%)
F1 Score:	0.9869 (98.69%)
ROC-AUC:	0.9980 (99.80%)
Confusion Matrix:	
[[5672 151]	
[186 12735]]	

11.1 Confusion Matrix

The confusion matrix provides a detailed overview of the classification performance of the model by comparing predicted labels with the actual labels in the test dataset. It summarizes the number of correct and incorrect predictions made by the model for each class.

11.2 Accuracy

Accuracy measures the overall proportion of correctly classified images among all test samples. It represents how often the model makes correct predictions for both classes (Real and GAN-Generated). Accuracy is calculated as the ratio of correct predictions to the total number of predictions.

11.3 Precision

Precision measures how many of the images predicted as AI-generated by the model are actually AI-generated. It evaluates the reliability of positive predictions made by the model. A high precision value indicates that when the model predicts an image as synthetic, it is very likely to be correct.

Precision is calculated as the ratio of True Positives to the total number of predicted positives, which includes both

True Positives and False Positives. In deepfake detection, precision is important because it reduces the number of real images that are incorrectly classified as AI-generated.

11.4 F1 Score

The F1 Score is the harmonic mean of **precision and recall**. It provides a balanced measure when evaluating classification

models. It is particularly useful when the dataset contains **class imbalance**, as it considers both false positives and false negatives.

11.5 Recall

Recall measures the ability of the model to correctly identify **all AI-generated images** in the dataset. It shows how effectively the model detects synthetic images.

11.6 ROC-AUC

The **Receiver Operating Characteristic – Area Under Curve (ROC-AUC)** measures the model's ability to distinguish between real and synthetic images across different classification thresholds. The ROC curve plots the **True Positive Rate (Recall)** against the **False Positive Rate**.

12. DISCUSSION

This project demonstrates that Vision Transformer models can effectively detect AI-generated images when the training and testing data come from the same generative model. However, cross-domain testing revealed that models trained on one generator may not generalize well to images produced by different generators.

13. CONCLUSION

This project developed an explainable deepfake detection system using a Vision Transformer architecture. Face alignment and balanced sampling improved detection performance. Explainable AI techniques helped visualize the model's decision-making process. Cross-domain testing revealed challenges in generalization, highlighting the importance of training models with diverse datasets.

REFERENCES

- [1] A. Dosovitskiy et al., "An Image is Worth 16×16 Words: Transformers for Image Recognition at Scale," ICLR, 2021.
- [2] D. Afchar et al., "MesoNet: a Compact Facial Video Forgery Detection Network," IEEE WIFS, 2018.
- [3] Y. Li and S. Lyu, "Exposing DeepFake Videos by Detecting Face Warping Artifacts," CVPR Workshops, 2019.
- [4] T. Karras et al., "Progressive Growing of GANs for Improved Quality, Stability, and Variation," ICLR, 2018.

- [5] T. Karras et al., "A Style-Based Generator Architecture for Generative Adversarial Networks," CVPR, 2019.
- [6] M. Sundararajan et al., "Axiomatic Attribution for Deep Networks," ICML, 2017.
- [7] S. Jha et al., "Explaining Deep Learning-Based Image Classification," IEEE Access, 2021.
- [8] H. Rossler et al., "FaceForensics++," ICCV, 2019.