# Evaluating the Effectiveness of a Gan Fingerprint Removal Approach in Fooling Deepfake Face Detection

Wasin Alkishri<sup>1\*</sup>, Dr. Setyawan Widyarto<sup>2</sup> and Dr. Jabar H. Yousif<sup>3</sup>

<sup>1\*</sup>Faculty of Computer Studies, Arab Open University, Muscat, Oman. Wasan.k@aou.edu.om, https://orcid.org/0000-0003-0833-5633

<sup>2</sup>Faculty of Communication, Visual Art, and Computing, Universiti Selangor (UNISEL), Selangor, Malaysia. swidyarto@unisel.edu.my, https://orcid.org/0000-0002-6317-9875

<sup>3</sup>Faculty of Computing and IT, Sohar University, Sohar, Oman. Jyousif@su.edu.om, https://orcid.org/0000-0003-2110-8347

Received: October 09, 2023; Accepted: December 15, 2023; Published: February 29, 2024

#### Abstract

Deep neural networks are able to generate stunningly realistic images, making it easy to fool both technology and humans into distinguishing real images from fake ones. Generative Adversarial Networks (GANs) play a significant role in these successes (GANs). Various studies have shown that combining features from different domains can produce effective results. However, the challenges lie in detecting these fake images, especially when modifications or removal of GAN components are involved. In this research, we analyse the high-frequency Fourier modes of real and deep network-generated images and show that Images generated by deep networks share an observable, systematic shortcoming when it comes to reproducing their high-frequency features. We illustrate how eliminating the GAN fingerprint in modified pictures' frequency and spatial spectrum might affect deep-fake detection approaches. In-depth review of the latest research on the GAN-Based Artifacts Detection Method. We empirically assess our approach to the CNN detection model using style GAN structures 140k datasets of Real and Fake Faces. Our method has dramatically reduced the detection rate of fake images by 50%. In our study, we found that adversaries are able to remove the fingerprints of GANs, making it difficult to detect the generated images. This result confirms the lack of robustness of current algorithms and the need for further research in this area.

**Keywords:** GAN-Based Artifacts Detection, GAN, GAN Fingerprint, Discrete Fourier Spectrum, Frequency Spectrum, and Spatial Spectrum.

### **1** Introduction

Generative adversarial networks (GANs) are powerful learning models for modifying digital media to enable generating images and videos that look very realistic. Recently, there has been a growing amount of public concern over images and videos that contain fake face information that was obtained via the use of digital modification. At the same time, we have witnessed massive developments in artificial intelligence and the emergence of software that easily contributes to manipulating multimedia content to spread misinformation online through social media platforms. It has become difficult to trust the

*Journal of Internet Services and Information Security (JISIS)*, volume: 14, number: 1 (February), pp. 85-103. DOI: *10.58346/JISIS.2024.11.006* 

<sup>\*</sup>Corresponding author: Faculty of Computer Studies, Arab Open University, Muscat, Oman.

information published on the Internet, which can lead to dire consequences. The issue of fake facial information obtained by digital manipulation has recently become a major concern among the public (Alkishri et al., 2023). Regardless of the benefits of using GAN algorithms to modify and turn images into real images, many malicious applications allow attackers to exploit GANs for malicious purposes. In the war between Russia and Ukraine, deepFake technology is used to spread fake information on social media. BBC News reports that (Wakefield, 2022), on May 2022, A DeepFake video posted on Twitter showed Russian President Vladimir Putin declaring peace. However, in fact, the Russian president did not announce any of what was mentioned in the fake video. Artificial intelligence-based deepfake software was used to manipulate the movements and voice of the Russian president and merge it into an original video of Putin giving a patriotic speech. And not only that but other fake videos were circulated against the Ukrainian president on social media, asking his soldiers to surrender to Russia. It was also published on the news websites of Radio Ukraine and television by hackers. However, after a short period of time, the video was deleted, and this fake video was reported and removed from all social media platforms.

In this research, we utilise the 140k Real and Fake Faces dataset to evaluate the efficacy of generative adversarial networks (GANs) like StyleGAN in identifying fake images. We use GAN generative models to demonstrate how the high-frequency spectrum's relative size and decay rate may be used to differentiate between genuine and deep network-manipulated images. Then, we show that removing GAN fingerprint artifacts of manipulated images in the frequency spectrum can affect deep-fake detection methods.

In This Context, The Main Contributions of Our Paper Are:

- 1. Analyse the high-frequency Fourier modes of real and deep network-generated images and show that Images generated by deep networks share an observable, systematic shortcoming when it comes to reproducing their high-frequency features.
- 2. Show that removing GAN fingerprint artifacts of manipulated images in the frequency domain can affect deep-fake detection methods.
- 3. In-depth analysis of the recent literature on the GAN-Based Artifacts Detection Approach.
- 4. Empirically evaluate our attacks on CNN model detection methods using StyleGAN architectures, 140k Real and Fake Faces datasets.

#### **2** Background and Related Work

Numerous studies (Ahmed et al. (2022), Mirsky & Lee (2021), Tolosana et al. (2020), Rana et al. (2022), and Zhang (2022)) have reviewed several techniques for detecting fake images. These studies have showcased exceptional results and illustrated promising outcomes in effectively identifying and addressing fake content. However, most of these approaches rely heavily on detecting Generative Adversarial Network (GAN) architectures to identify fake images. The challenge arises when fake images are modified, or GAN components are removed, making it difficult to detect them. Images generated by GANs result from several fixed filters and non-linear processes, which produce similar patterns inside a single instance of a GAN but differ between instances. This suggests that images have unique fingerprints and can be attributed to their GAN sources (Yang & Järvinen 2019).

Yu et al., 2019 studied several GAN fingerprints for picture attribution and employed them in the detector to distinguish between actual and GAN-generated images. They demonstrate that GANs leave

stable fingerprints in the pictures they produce and that these fingerprints are independent of GAN artifacts and include different image frequencies and patches.

Deep learning-based video editing and manipulation techniques have become readily available to the public. People can simply learn. Alternatively, use deepfake software to create fake videos or photos with little or no effort, which poses serious legal, societal, and economic problems. Researchers have made great efforts to confront and expose deepfake techniques in this field. However, these techniques are not generalisable to all deepfake techniques, and some of them cannot reveal the hidden artifact information contained in the manipulated images.

This section will review the most recent studies of the two main types of artifact-based detection approaches in GANs deepfake manipulation: Spatial and Frequency domains. See Table 1.

**A.** The Spatial Domain: Detecting DeepFakes in the spatial domain has recently gained much attention. It is the normal image space in which changes in position in A directly translate into changes in B. that is, mean in A; pixels correspond to actual distances in B. Existing studies mostly use this technique to observe several visible and invisible artifacts and distinguish between real and fake faces.

**Wang et al., 2020** extract spatial features to improve the generalisation capabilities of the Gansynthesised models by using ResNet-50 and ImageNet. They conduct binary classifiers on the PGGAN database, which allows generalisation in other GANs architecture. Based on the obtained results, the proposed model shows the ability to counter perturbation attacks.

A Convolutional LSTM-based Residual Network (CLRNet) was suggested by (**Tariqto et al., 2020**) to extract temporal information from consecutive frames and detect unnatural-appearing artifacts in deepfake manipulated videos. They used the FaceForensics++ dataset for their investigation. Additionally, they suggest a transfer learning-based strategy to generalise several deepfake techniques.

Wu et al., 2020 Investigated and developed a powerful DeepFakes detector, namely SSTNet, that works based on the Spatio-temporal domain and Steganalysis. They applied XceptionNe and the recurrent neural network (RNN) model to extract visible and invisible artifact features of image pixels and inconsistency between consecutive frames. The results show that the SSTNet module achieved accuracy and robustness in the Face-Forensics++ dataset. It has demonstrated its ability to generalise to other GANs architectures.

By utilising global texture features, Gram-Net (Liu et al., 2020) enhances the robustness and generalisation capabilities of current CNNs in discriminating synthesised false faces. According to experimental findings, Gram-Net is resilient to perturbation attacks such as downsampling, JPEG compression, blur, and noise. Gram-Net's generalisation capability has also been demonstrated in real-world applications utilising various GANs.

**B.** Frequency Domain Feature: The frequency domain could also reveal differences between real and synthesised fake faces. Researchers primarily introduce studies that use frequency domain features to distinguish between real and unreal. However, sometimes, these methods fail when dealing with unknown Gan-synthesised DeepFakes.

**Neves et al., 2020** suggested an autoencoder module known as GANprintR to remove the GAN fingerprints from the synthetic images and fool facial manipulation detection systems while maintaining the aesthetic quality of the images that were produced. Their experiment involved multiple datasets, both real (CASIA-WebFace (**Yi et al., 2014**) and VGGFace2 (**Cao et al., 2014**)) and fake (TPDNE, 100K-Faces and PGGAN (**Karras et al., 2017**). The study proved that removing fingerprints may increase the

probability of error and detector deception. XceptionNet detector had an average EER of 9.65%, Steganalysis had an average EER of 14.68%, and Local Artifacts had an average EER of 4.91%. The researchers explained that the error rate in the Local Artifacts detector is low due to the increase in the average error rate in the original fake images compared to the GANprintR images.

**Frank et al., 2020** investigate the frequency domain artifacts of various GAN architectures and datasets. They discover that severe artifacts are introduced into GANs due to up-sampling techniques. Experiments show that a classifier based on a simple linear and a CNN-based model can produce promising results across the entire frequency spectrum. Furthermore, the frequency domain-trained classifier is resistant to common perturbation attacks such as blurring and cropping, and it can deal with future unseen GANs.

FGPD-FA (**Bai et al., 2020**), extracts three types of characteristics (statistical, orientated gradient, and blob) in the frequency domain to distinguish between actual and fake faces. Frequency-aware patterns from frequency-aware image decomposition and local frequency statistics are two complementing frequency-aware cues that are taken into account or detected. For frequency-domain transformation, discrete cosine transform (DCT) is used. Finally, a two-stream collaborative learning architecture successfully sees low-quality DeepFake videos by jointly learning the two frequency hints.

In F3-Net (**Qian et al., 2020**), two complementary frequency-aware clues are used for detection: frequency-aware patterns derived from frequency-aware image decomposition and local frequency statistics. For frequency-domain transformation, discrete cosine transform (DCT) is applied. The final step involves using a two-stream collaborative learning framework to learn the two frequency clues jointly. The system achieves significant performance when it comes to detecting low-quality DeepFake videos.

Wesselkamp et al., 2020 presented a new class of deepfake manipulation attacks. The suggested solution eliminates the need to modify GANs or train complex learning-based systems by changing the frequency spectrum of deep fakes and focusing on a GAN fingerprint. Three different types of detectors are used to evaluate the performance of their proposed attack method. Their experiment uses CelebA (Liu et al., 2018) and LSUN bedrooms (Yu et al., 2015) datasets and four GAN architectures (ProGAN, SNGAN, MMDGAN, CramerGAN). Their findings indicate that GAN models differ in the importance of particular frequency bands. For example, ProGAN and SNGAN appear to evaluate the entire spectrum. In contrast, CramerGAN and MMDGAN focus mostly on higher frequencies. In a similar manner, different datasets for the specific GAN seem to focus differently on other bands of frequencies.

Liu et al., 2021 combine spatial image and phase spectrum to capture the up-sampling artifacts in existing GANs for helping detection and improving the transferability of the face forgery detection method across unknown synthetic methodologies.

**Dzanic and Shah, 2020** experimentally demonstrated the bias in high spatial frequencies in order to classify real and deep network-generated images. Le and Woo, 2021 provided an analysis of deeper statistical frequency features. Thus, this large difference in the variable frequencies made the GAN data set easy to detect and thus effectively used to distinguish between real and fake images.

**Durall et al., 2020** presented the critical issue of detecting AI-generated fake images, particularly in the growing concern surrounding the proliferation of such manipulated digital content. With the impressive advancements in deep generative models, there has been a simultaneous increase in the need for automated methods to identify these deceptive visuals. The research focuses on a specific facet of this problem: DeepFakes, which are remarkably realistic but carry subtle artifacts that demand careful

scrutiny. One of the standout features of this work is its simplicity in addressing the complex issue of DeepFake detection. The methodology relies on classical frequency domain analysis followed by a basic classifier. This approach offers a fresh perspective, especially compared to previous systems that require vast amounts of labelled data. The results are impressive, as the research demonstrates the ability to achieve high accuracy with only a limited number of annotated training samples. Even more striking is the successful performance in fully unsupervised scenarios, highlighting the adaptability and robustness of the proposed approach. The creation of the Faces-HQ benchmark dataset, which combines real and fake high-resolution face images, is a notable contribution. This benchmark dataset provides a rigorous testing ground for evaluating DeepFake detection methods. The findings are compelling, with the approach achieving a perfect classification accuracy of 100% in the case of high-resolution face images when trained on a mere 20 annotated samples. The method's adaptability is further underscored by its excellent performance in evaluating medium-resolution images from the CelebA dataset. Additionally, the study extends its evaluation to low-resolution video sequences, where it achieves a commendable 90% accuracy in detecting manipulated videos.

In summary, we drew inspiration from the **Durall et al., 2020** method to enhance our approach, leveraging it as a preprocessing step to identify the GAN fingerprint within manipulated images. Subsequently, we incorporated the strategies outlined in the **Wesselkamp et al., 2022** study to address the removal of the GAN fingerprint, culminating in our comprehensive methodology that is associated with GAN-generated content in our research. In our paper, we will apply Fourier spectrum analysis to identify the high frequency on manipulated images, and then this frequency artifact will be removed, and the effectiveness of the detector will be checked. To the best of our knowledge, no one has used Fourier spectrum analysis to remove the GAN fingerprint and spoof the detector.

Type of GAN-Based Artifacts Detection Approach	Study	Classifier	Dataset	Туре	Accuracy
Spatial Domain	Wang et al., 2020	CNN	FaceForensics++	video	99%
	Tariq et al., 2020	CNN+RNN	FaceForensics++	video	99%
	Wu et al., 2020	CNN+RNN	FaceForensics++	video	98%
	Liu et al., 2020	CNN	Own dataset	image	99%
Frequency Domain	Neves et al.,	CNN	iFakeFaceDB	image	EER =
	2020				4.5%
	Frank et al., 2020	CNN	FFHQ	image	99%
	Bai et al., 2020	CNN	Own dataset	video	97%
	Qian et al., 2020	CNN	FaceForensics++	video	99%
	Wesselkamp et al., 2022	CNN	CelebA and LSUN	image	99%
	Liu et al.,2021	CNN	Celeb-DF,FF++, DFDC	Video & image	95%
	Durall et al., 2020	SVM & Logistic Regression	discrete Fourier transform	Own dataset (FaceForensics++ d)	video

 Table 1: Comparison of state-of-the-art manipulation detection methods based on Spatial and

 Frequency domain

# 3 Experiment and Data Analysis

In this section, we will detail the experiment we have adopted. A diagram of the design of our approach is shown in **Figure 1.** In our work, we will apply Fourier spectrum analysis to identify the high frequency on manipulated images; then, this frequency artifact will be removed, and the detector's effectiveness will be checked. To the best of our knowledge, no one has used Fourier spectrum analysis to remove the GAN fingerprint and spoof the detector.

#### **3.1. Proposal Approach**

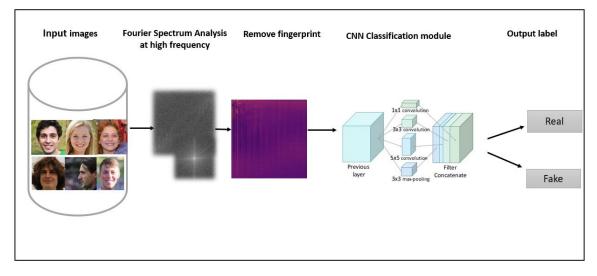


Figure 1: Approach design

#### 3.1.1. Dataset

In our experiment, we used an online publicly available deepfake dataset, named 140k Real and Fake Faces (**Xhlulu, 2020**). In total, 70k REAL faces were collected from the Flickr-Faces-HQ Dataset (FFHQ) by Nvidia with high quality. These images, resized to 256x256 and compressed in JPEG, reflect considerable age, ethnicity, and image background variation. Also, eyeglasses, sunglasses, hats, etc., are well covered. In addition, 70k fake faces were sampled from the 1 million FAKE faces created by StyleGAN that Bojan provided. **See Figure 2.** 

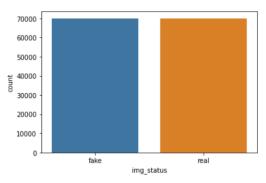


Figure 2: Shows the number of fake and real images that were used in the experiment

#### 3.1.2. GAN Models

StyleGAN: The Style Generative Adversarial Network, also known as StyleGAN, is an addition to the GAN architecture that suggests significant changes to the generator model. As part of these changes, points in latent space are mapped to an intermediate latent space using a mapping network. Each point of the generator model is then controlled for style through the middle latent space, and noise is introduced as a source of variation at each point. Even at 1024\*1024 resolution, it provides realistic and high-quality images. It is employed in approaches for whole-face synthesis.

#### **3.1.3.** Discrete Fourier Spectrum Analysis

The Fourier transform defines the space of the image by computing the sum of complex exponentials of varying magnitudes, frequencies, and phases, known as the frequency domain. One of the Fourier transforms used in image processing includes enhancement, analysis, restoration, and compression. A 2D discrete Fourier transform is used to visualise distributions of signal energy over a range of image frequencies. The Fourier transform is not a real number; it employs a complex exponential that presents in complex matric. Due to the complexity of the DFT of an image, it cannot be shown with a single image. Consequently, as shown in Figure 3, we display the DFT's amplitude (modulus) and phase (argument) separately. In the frequency plane, the amplitude and phase represent the energy distribution. High frequencies are near the boundaries of the image, while low frequencies are at the centre. A Fourier transform is necessary in order to conduct frequency domain analysis of the features of both real and artificially produced pictures generated by deep neural networks. Both the input and output of the discrete Fourier transform are discrete, which facilitates computer manipulation. Consider the case of f(p,q) as a discrete two-dimensional frequencies signal representing distinct colour channels of an image of size m \* n, where p and q are indices for the rows and columns of the image, respectively. F(u, v) is the 2D discrete Fourier transform of f(p, q), where u and v are indices for the frequencies in the x and y directions, respectively. The transform is indexed over  $0 \le k$  x < m and  $0 \le k$  y < n. And the highest frequencies can be calculated by transforming Cartesian coordinates (u, v) into normalised polar coordinates in wavenumber space to create a scale- and rotation-invariant threshold. Let f(x,y) be the image in the spatial domain, F(u,v) be the image in the frequency domain, m and n be the image's dimensions, and kx and ky are the frequency components.

The discrete Fourier transform F(u, v) Can be computed using the equations below:

$$F(\mathbf{u},\mathbf{v}) = \frac{1}{mn} \sum_{p=0}^{m-1} \sum_{q=0}^{n-1} f(p,q) e^{-i2\pi \left(k_x \frac{p}{m} + k_y \frac{q}{n}\right)(1)}$$

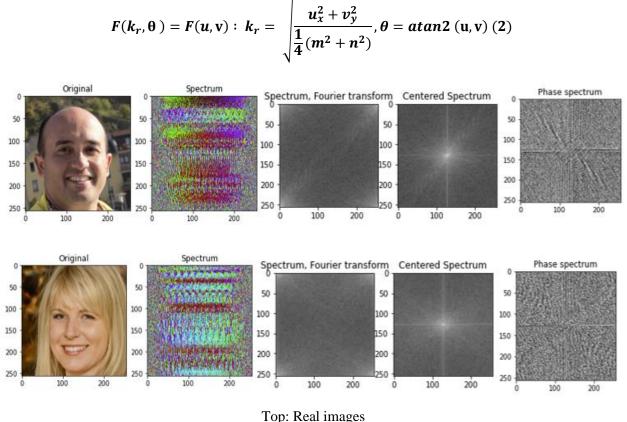
Where,

i: This is the imaginary unit, defined as sqrt(-1).

- e: This is the mathematical constant e, which is approximately equal to 2.71828.
- $\pi$ : This is the mathematical constant pi, which is approximately equal to 3.14159.
- mn: This is the total number of elements in the image, equal to m \* n.

And to normalized polar coordinates, where  $r \in [0; 1], \theta \in [0; 2\pi)$ 

Evaluating the Effectiveness of a Gan Fingerprint Removal Approach in Fooling Deepfake Face Detection



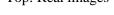




Figure 3: Show The frequency spectra differences of real vs fake image after applying DFT on each channel of images and display the magnitude as a grayscale image, where the brighter pixels correspond to higher frequencies, and the darker pixels correspond to lower frequencies.

#### 3.1.4. Calculating GANs Fingerprint

We average the DFT frequency spectrum of a set of GAN images to compute fingerprints. Using the following equation to Compute the Azimuthal Average of Fourier to reduce the image dimensionality without significant loss in information image. Compute the GAN fingerprint using azimuthal averaging. Azimuthal averaging is a technique used in radial averaging that involves averaging the intensities of the pixels along a set of radial lines originating from the centre of an image. This can be useful for obtaining a radial profile of an image and identifying patterns in the image that have rotational symmetry. As a result of applying the azimuthal average, we can identify patterns or peaks that correspond to the GAN fingerprint. Then, we will use this 1D power spectrum to remove fingerprints because it contains enough information to differentiate between real and fake images. We will be able to obtain a more accurate representation of the sample image. An evaluation of the signal strength in terms of its radial wavenumber. **See Figure 3.** 

The following algorithm is used to compute the GAN fingerprint using azimuthal averaging:

Algorithm: compute the GAN fingerprint using azimuthal averaging in the frequency domain

#### Requirements

Let I(x,y) be the image in the spatial domain, F(u,v) be the image in the frequency domain, R is the radial distance from the origin, and N is the number of radial bins.

1- Calculate the DFT of the image:

F(u, v) = DFT[I(X, Y)] (3)

2- Calculate the magnitude of the frequency components:

 $M(u,v) = |F(u,v)| \, (4)$ 

3- Centre the origin of the coordinate system at the centre of the image, where u\_0 and v\_0 are the coordinates of the centre of the image.

$$R = sqrt((u - u_o))^2 + (v - v_0)^2$$
(5)

4- Bin the frequency components into radial bins:

 $Bin_k = \{(u, v) | R_K \le R_{\{K+1\}\}}$  (6)

5- Average the magnitude of the components in each bin, where Nk is the number of frequency components in Bin\_k, and the sum is taken over all the components in the bin.

$$A_{k} = \frac{1}{N_{k}} * SUM(M(u, v)), for (u, v) in Bin_{k} (7)$$

6- Plot the radial profile of the image. The radial profile of the image can be plotted as a function of the radial distance RK to obtain the GAN fingerprint using azimuthal averaging.

 $A(R_K) = A_k(8)$ 

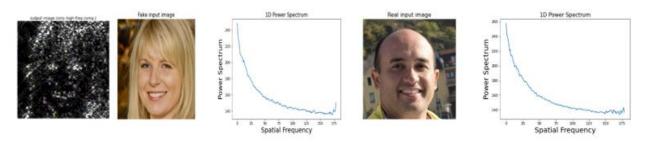
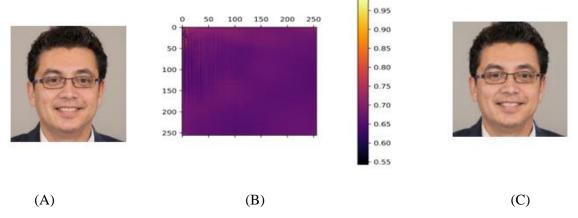
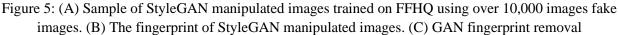


Figure 4: Appling the azimuthal component of the DFT power spectrum to analyse spectral distributions of images. It Shows The frequency spectra differences of real vs fake images after applying DFT on each channel of images. There is a difference between a real and fake image based on any higher frequency distinctive patterns or structures in the frequency spectrum that are consistent across multiple generated images and that are not present in real images.

#### 3.1.5. GAN Fingerprint Removal

As a result of applying azimuthal average on manipulated and real image AK, we will now use this 1D power spectrum to remove the fingerprint. **See Figures 3, and 4** because they contain enough information to differentiate between real and fake photos. Firstly, we need to Identify the frequency components that correspond to the fingerprint in the 1D power spectrum by creating a threshold to separate the low and high-frequency components in the 1D power spectrum. This threshold can be set based on the fingerprint's expected frequency range and the image's general distribution of frequencies. Then, we remove the high-frequency components that exceed the threshold by setting the high-frequency components to zero. Finally, we apply an Inverse Fourier Transform to the filtered frequency domain representation to obtain the image without the fingerprint. **As presented in Figure 5**.





#### **3.2. Deepfake Detectors**

Classification accuracy was measured by how well the classifier could tell if an image was real or fake. This was done to show how the spectrum disagreement could be used to define a characteristic. Xception is a deep learning model developed for image classification and computer vision tasks. Xception has been used in several recent studies (**Dolhansky et al., (2019), Rossler et al., (2019), Carlini et al., (2020) and Tariq et al., (2019));** they obtained the best detection results. Our training approach was similar to that in (**Le rt al., 2021**) The last fully-connected layer of the ImageNet model has been replaced with a new one (two classes, real or fake image). A 200-epoch training process was used to select the best-performing model based on accuracy in table 2.

Table 2: Hyperparameter tuning analysis Xception network' architectures used in this study

Model compile	Value		
Optimizer	Adam (0.0002)		
loss Function	Binary cross entropy		
Metrics	Accuracy		
Epochs	200		

# **4** Evaluation

#### 4.1. GAN-fingerprint-removed

To evaluate the performance of the model on the "GAN-fingerprint-removed" dataset, we computed several evaluation metrics, including accuracy, precision, recall, and F1-score, and compared them to the corresponding metrics on the original dataset. In addition, we computed the ROC curve and AUC on the "gan-fingerprint-removed" dataset.

Our results showed that the fingerprint removal technique successfully fooled the model. Specifically, the accuracy, precision, recall, and F1-score of the model on the "gan-fingerprint-removed" dataset were similar to or slightly worse than the corresponding metrics on the original dataset. This suggests that the model could not distinguish between the "gan-fingerprint-removed" fake images and real images and classified both as real. After we remove the Gan fingerprint from the manipulated image and store in the nofingerprint folder, we run a pre-trained deep learning model (xception model) in a directory containing only the nofingerprint image. The model classified images and predicted whether the image is "fake" or "real". The predictions are based on the model's output, which is the probability of the image being real (output value between 0 and 1). The prediction checks if the output value exceeds 0.5 and returns 0 for fake and 1 for real. The output of the classifier shows that 99.778% of the nofingerprint images are classified as "real," while only 0.222% are classified as "fake."This result suggests that the removal of the GAN fingerprint from the synthetic images has made them very difficult to distinguish from real images using the particular classifier(XceptionNet). We evaluate the performance of a pre-trained deep learning model on a directory of images consisting of 50% fake and 50% no fingerprint images. The goal is to evaluate how well the model can distinguish between fake and nofingerprint images. The model's accuracy on the folder consisting of 50% nofingerprint data and 50% fake data is 0.9932. The model's precision, recall, and F1-score are 0.9897, 0.9968, and 0.9932 respectively. See figuer 6 and 7. The ROC AUC of the model is 0.9932. These evaluation metrics indicate that the model is considering nofingerprint image as real with high precision, recall, and F1score and a high ROC AUC. These metrics indicate how accurately the model can classify fake and nofingerprint images. This is conclusive evidence that the images are free of GAN fingerprints, and the detector considered them real images. See table 3 and figure 6. Then, We ran the model on the folder consisting of 50% real and 50% nofingerprint data.

The Results indicate that the model is not able to distinguish between real and nofingerprint images in the given dataset. The accuracy score of 0.5004 suggests that the model is performing no better than random guessing. The precision score of 0.5002 indicates that when the model predicts an image as real, there is a 50.02% chance that it is actually real. The recall score of 0.9982 suggests that the model is able to correctly identify 99.82% of the real and nofingerprint images. The F1-score of 0.6664 suggests that the model performs moderately in identifying real and nofingerprint images. The ROC AUC score of 0.5004 confirms the model's poor performance in distinguishing between real and nofingerprint in the manipulated image, presented in **Figures 8 and 9.** Our results showed that the fingerprint removal technique successfully fooled the model. **See Table 4. In Table 5**, we compare our method with a similar study. We found that removing GAN artifacts from images directly in the frequency spectrum can highly mislead the detector by 32.7%.

Table 3: The results of the model run on the folder consisting of 50% fake and 50% nofingerprint fake image

Dataset	Detection model	Accuracy	Precision	Recall	F1-score
140k Real and Fake Faces	Xception	99.3%	98%	99.6%	99.3%

Table 4: The model results run on the folder consisting of 50% real and 50% nofingerprint fake images

Dataset	<b>Detection model</b>	Accuracy	Precision	Recall	F1-score
140k Real and Fake Faces	Xception	50.04%	50.02%	99.82%	66.64%

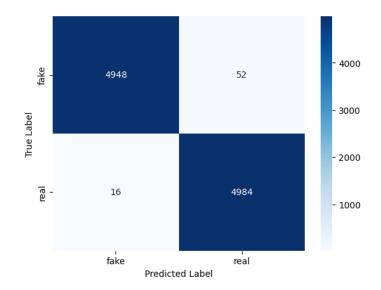


Figure 6: Confusion matrix shows how well the fake face detection model performs in attempting to correctly categorise fake and no fingerprint images, along with the types of misclassifications that are most common

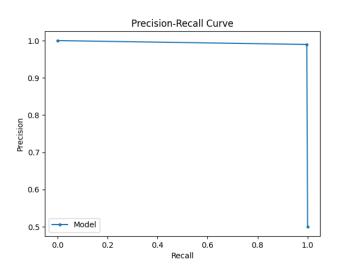


Figure 7: Shows that the model considers nofingerprint images as real with high precision and recall

Wasin Alkishri et al.

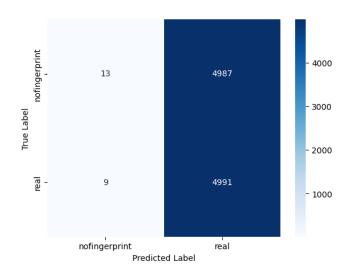


Figure 8: Shows that model was not able to distinguish between the "gan-fingerprint-removed" fake images and real images and classified both as real

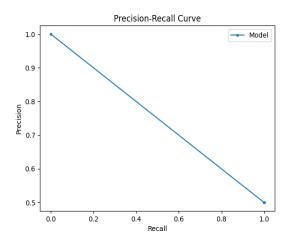


Figure 9: Shows that a precision score of 0.5002 indicates that when the model predicts an image as real, there is a 50.02% chance that it is actually real. The recall score of 0.9982 indicates that the model is able to identify 99.82% of the real and nofingerprint images correctly.

Study	Dataset	GAN type	GAN Removal technique	Detection model	Accuracy before fingerprint removal	Accuracy after fingerprint removal in frequency domain
o ur method	140k Real and Fake Faces	StyleGAN	Discrete Fourier spectrum	XceptionNet	99.7%	50.02%
Refer Neves et al., 2020	PGGAN	PGGAN	AutoEncoder	XceptionNet	99.7%	82.7%

Table 5: Summary of results compareing with similier study

#### 4.2. Image Perturbations

We also examined whether the artifacts are still present in the modified images during the processing processes that occur to images automatically when uploading images to websites like Facebook or Instagram. So, we've added some common perturbations like blurring, cropping, and adding noise. To ensure that the detector can detect the forgery of the image. We prepared and created a data set for the three types of s perturbations (blurring, cropping, and adding noise). This data was again divided into training and validation sets. After that, each type was examined separately from the other. Then test and report the detector's ability to detect images after adding these perturbations. We train the dataset on the best model we obtained after removing the fingerprints in the frequency domain. Below are descriptions of the corruption applied to the images. **See Figure 10**.

- **Cropping:** we used the formula height//2 and width//2 to crop the image to one-half of its original size.
- **Blurring:** we applied Gaussian filtering with kernel size 5 X 5.
- Noise: we applied Gaussian distribution with a mean of 0 and a standard deviation of 50



Figuer 10: Perturbations image on Fake Faces

#### 4.2.1. Results of Perturbations Images

The images in nofingerprint are data that are fake. When cropping was added to the nofingerprint images, the detector deteriorated significantly compared to other perturbations, with only a 49% detection rate. **See Figure 11.** Prior to the classification between real and nofingerprint, the detector was hugely misclassified. However, after blurring was applied, the outcome improved by about 24% as fake. It's possible that the noise has removed some of the fine details that were in no fingerprint images. **See Figure 13.** 

Furthermore, the accuracy decreases significantly when noise is added, as opposed to blurring, when the model is employed to distinguish between real and fake images. Initially, when the model was fed nofingerprint data and fake images, the majority of which were classified as real, the accuracy rate was high. However, this is no longer the case after introducing noise, and some of the nofingerprint images are now identified as fake. Interestingly, when we tested the model on 50% real and 50% nofingerprint noised data, the accuracy increased after adding noise, and more nofingerprint images are currently being classified as fake than before. We can see that in the confusion matrix in **Figure 12**, more nofingerprint images are predicted to be fake. In conclusion, adding perturbations impacts fooling the detector; after adding noise, the detector seems to perform well and starts predicting more nofingerprint as a fake; a similar thing happened with blur as well. **See Table 6**.

Deep neural networks, specifically Generative Adversarial Networks (GANs), have proven to be highly effective in producing realistic images across various domains with remarkable flexibility. They possess the ability to learn complex patterns from data efficiently, requiring less input data to generate diverse and convincing visuals. However, leveraging these networks comes with its own set of challenges. The training process demands significant computational resources and time, making maintaining output quality and diversity challenging. Mode collapse is also a concern, as networks may produce limited variations, and ethical considerations, such as the creation of deceptive content, demand vigilance. Additionally, these networks may not generalize well beyond the training data, limiting their adaptability to diverse scenarios. Despite their impressive capabilities, careful management and control are crucial when using deep neural networks to generate realistic images.

Datasets	Results on 140k Real and Fake no fingerprint Faces dataset					
	Accuracy Precision Recall F1-score					
CCropping	49.7%	49.8%	99.4%	66.4%		
Noise	55.7%	53.5%	87.14%	66.3%		
Blurring	70%	68%	76%	72%		

Table 6: Summary results of perturbations images on 140k Real and Fake (**no fingerprint**) Faces dataset

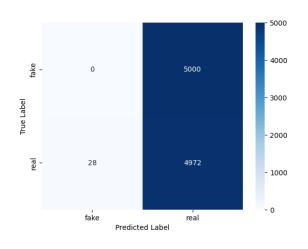
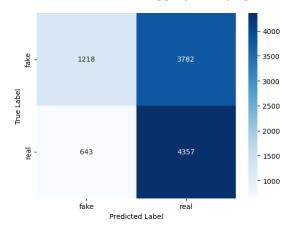


Figure 11: Confusion matrix on cropping nofingerprint fake images



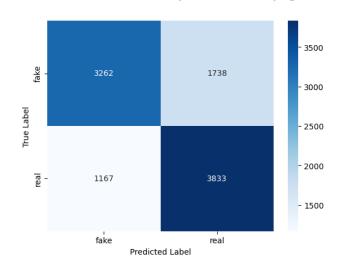


Figure 12: Confusion matrix on adding noise on nofingerprint fake images

Figure 13: Confusion matrix on blurring nofingerprint fake images

## 5 Conclusion

Deepfake images offer a substantial danger to a variety of applications, including security, journalism, and entertainment. As a result, various studies have been conducted to create approaches for recognising deepfake manipulation. Therefore, many studies have focused on developing techniques to detect deep-fake images; however, these approaches can be easily avoided by adversaries who can utilise a variety of tactics to deceive detection. In this work, we analyse the effectiveness of a GAN fingerprint removal approach in fooling a real vs fake image detection algorithm. Our results demonstrated that the proposed approach successfully evades the model, indicating that adversaries may exploit it to make deepfake pictures that can dodge detection by existing deepfake detection approaches. As a result, we confirm that these approaches are far from robust and that further research is needed to be developed. It is important to note that this study has some limitations that can be addressed in future research. This study focused on StyleGAN-generated images and a narrow evaluation scope, which might restrict its generalizability to diverse scenarios. The work we have done highlights areas for improvement. Future research should explore various GAN architectures beyond StyleGAN, evaluate them across diverse datasets and delve deeper into unexplored perturbation types, enhancing the adaptability of Deepfake detection. Moreover, continuous innovation is crucial for developing more effective and adaptive detection methods to keep pace with evolving Deepfake technology. The adoption of these paths will strengthen Deepfake detection against the constantly evolving digital deception landscape.

# References

- [1] Ahmed, S.R., Sonuç, E., Ahmed, M.R., & Duru, A.D. (2022). Analysis survey on deepfake detection and recognition with convolutional neural networks. *In International Congress on Human-Computer Interaction, Optimization and Robotic Applications (HORA)*, 1-7.
- [2] Alkishri, W., Widyarto, S., Yousif, J.H., & Al-Bahri, M. (2023). Fake Face Detection Based on Colour Textual Analysis Using Deep Convolutional Neural Network. *Journal of Internet Services and Information Security* 13(3), 143-155.

- [3] Bai, Y., Guo, Y., Wei, J., Lu, L., Wang, R., & Wang, Y. (2020). Fake generated painting detection via frequency analysis. *In IEEE international conference on image processing (ICIP)*, 1256-1260.
- [4] Cao, Q., Shen, L., Xie, W., Parkhi, O.M., & Zisserman, A. (2018). Vggface2: A dataset for recognising faces across pose and age. In 13th IEEE international conference on automatic face & gesture recognition (FG 2018), 67-74.
- [5] Carlini, N., & Farid, H. (2020). Evading deepfake-image detectors with white-and black-box attacks. *In Proceedings of the IEEE/CVF conference on computer vision and pattern recognition workshops*, 658-659.
- [6] Dang, H., Liu, F., Stehouwer, J., Liu, X., & Jain, A.K. (2020). On the detection of digital face manipulation. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern recognition, 5781-5790.
- [7] Dolhansky, B., Howes, R., Pflaum, B., Baram, N., & Ferrer, C.C. (2019). The deepfake detection challenge (dfdc) preview dataset. *arXiv preprint arXiv:1910.08854*.
- [8] Durall, R., Keuper, M., & Keuper, J. (2020). Watch your up-convolution: CNN-based generative deep neural networks are failing to reproduce spectral distributions. *In Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 7890–7899.
- [9] Frank, J., Eisenhofer, T., Schönherr, L., Fischer, A., Kolossa, D., & Holz, T. (2020). Leveraging frequency analysis for deep fake image recognition. *In International conference on machine learning*, 3247-3258.
- [10] Generated Humans. Generated Photos. (n.d.). https://generated.photos/humans
- [11] Karras, T., Aila, T., Laine, S., & Lehtinen, J. (2017). Progressive growing of gans for improved quality, stability, and variation. *arXiv preprint arXiv:1710.10196*.
- [12] Le, B.M., & Woo, S.S. (2021). Exploring the Asynchronous of the Frequency Spectra of GAN-generated Facial Images. arXiv preprint arXiv:2112.08050.
- [13] Liu, H., Li, X., Zhou, W., Chen, Y., He, Y., Xue, H., & Yu, N. (2021). Spatial-phase shallow learning: rethinking face forgery detection in frequency domain. *In Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 772-781.
- [14] Liu, Z., Luo, P., Wang, X., & Tang, X. (2018). Large-scale celebfaces attributes (celeba) dataset. https://mmlab.ie.cuhk.edu.hk/projects/CelebA.html?
- [15] Liu, Z., Qi, X., & Torr, P.H. (2020). Global texture enhancement for fake face detection in the wild. In Proceedings of the IEEE/CVF conference on computer vision and pattern recognition, 8060-8069.
- [16] Mirsky, Y., & Lee, W. (2021). The creation and detection of deepfakes: A survey. ACM Computing Surveys (CSUR), 54(1), 1-41.
- [17] Neves, J.C., Tolosana, R., Vera-Rodriguez, R., Lopes, V., Proença, H., & Fierrez, J. (2020). Ganprintr: Improved fakes and evaluation of the state of the art in face manipulation detection. *IEEE Journal of Selected Topics in Signal Processing*, 14(5), 1038-1048.
- [18] Qian, Y., Yin, G., Sheng, L., Chen, Z., & Shao, J. (2020). Thinking in frequency: Face forgery detection by mining frequency-aware clues. *In European conference on computer vision*. Cham: Springer International Publishing, 86-103.
- [19] Rana, M.S., Nobi, M.N., Murali, B., & Sung, A.H. (2022). Deepfake detection: A systematic literature review. *IEEE Access*, *10*, 25494-25513.
- [20] Rossler, A., Cozzolino, D., Verdoliva, L., Riess, C., Thies, J., & Nießner, M. (2019). Faceforensics++: Learning to detect manipulated facial images. *In Proceedings of the IEEE/CVF international conference on computer vision*, 1-11.

- [21] Tariq, S., Lee, S., & Woo, S.S. (2020). A convolutional lstm based residual network for deepfake video detection. https://capitol.texas.gov/tlodocs/86R/analysis/html/SB00751F.htm
- [22] Tolosana, R., Vera-Rodriguez, R., Fierrez, J., Morales, A., & Ortega-Garcia, J. (2020). Deepfakes and beyond: A survey of face manipulation and fake detection. *Information Fusion*, 64, 131-148.
- [23] Wakefield, J. (2022). *Deepfake presidents used in Russia-ukraine war*. BBC News. https://www.bbc.com/news/technology-60780142
- [24] Wang, S.Y., Wang, O., Zhang, R., Owens, A., & Efros, A.A. (2020). CNN-generated images are surprisingly easy to spot... for now. *In Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 8695-8704.
- [25] Wesselkamp, V., Rieck, K., Arp, D., & Quiring, E. (2022). Misleading Deep-Fake Detection with GAN Fingerprints. *In IEEE Security and Privacy Workshops (SPW)*, 59-65.
- [26] Wu, X., Xie, Z., Gao, Y., & Xiao, Y. (2020). Sstnet: Detecting manipulated faces through spatial, Steganalysis and temporal features. In ICASSP 2020-2020 IEEE international conference on acoustics, speech and signal processing (ICASSP), 2952-2956.
- [27] Xhlulu. (2020). *140k real and fake faces*. Kaggle. https://www.kaggle.com/datasets/xhlulu/140k-real-and-fake-faces
- [28] Yang, Z., & Järvinen, K. (2019). Towards modeling privacy in WiFi fingerprinting indoor localization and its application. *Journal of Wireless Mobile Networks, Ubiquitous Computing, and Dependable Applications (JoWUA)*, 10(1), 4-22.
- [29] Yi, D., Lei, Z., Liao, S., & Li, S.Z. (2014). Learning face representation from scratch. *arXiv* preprint arXiv:1411.7923.
- [30] Yu, F., Seff, A., Zhang, Y., Song, S., Funkhouser, T., & Xiao, J. (2015). Lsun: Construction of a large-scale image dataset using deep learning with humans in the loop. *arXiv preprint arXiv:1506.03365*.
- [31] Yu, N., Davis, L.S., & Fritz, M. (2019). Attributing fake images to gans: Learning and analyzing gan fingerprints. *In Proceedings of the IEEE/CVF international conference on computer vision*, 7556-7566.
- [32] Zanic, T., Shah, K., & Witherden, F. (2020). Fourier spectrum discrepancies in deep network generated images. *Advances in neural information processing systems*, *33*, 3022-3032.
- [33] Zhang, T. (2022). Deepfake generation and detection, a survey. *Multimedia Tools and Applications*, 81(5), 6259-6276.

# **Authors Biography**



**Wasin Alkishri** is a PhD candidate and researcher in computer vision and image processing. A lecture at an Arab Open University - OMAN. She holds a master's degree. And BA in Computer Science. She has experience in various courses in the field of computer science—extensive knowledge of computer vision and image processing. Actively participate as a speaker in international and local conferences and professional research.



**Dr. Setyawan Widyarto** is an Associate Professor of Industrial Computing. He has Sarjana's Certificate (IPB, Bogor), Masters (University of Bradford, England), PhD (UTM, Malaysia). He is the Dean of the Department of Computing, School of Communication, Visual Arts, and Computing. His areas of research expertise are (Modeling and Simulation, Digital, Image Processing, Computer Science, and Systems Engineering and Management).



**Dr. Jabar H. Yousif** is an Associate Professor at the Faculty of Computing and Information Technology at Sohar University. He completed his postdoctoral work on computer graphics and virtual/augmented reality at the National University of Malaysia (UKM) in 2007–2008. He earned a PhD in Artificial Intelligence and Natural Language Processing from the Faculty of Information Science and Technology, National University of Malaysia (UKM) in 2007.