**Jilin Daxue Xuebao (Gongxueban)/Journal of Jilin University (Engineering and Technology Edition)**
ISSN: 1671-5497
E-Publication: Online Open Access
Vol: 42 Issue: 02-2023
DOI 10.17605/OSF.IO/BNPC4

# DEEPFAKE IMAGE DETECTION METHODS USING DISCRETE FOURIER TRANSFORM ANALYSIS AND CONVOLUTIONAL NEURAL NETWORK

**WASIN ALKISHRI**

Faculty of Communication, Visual Art, and Computing, University Selangor (UNISEL), Selangor, Malaysia. Email: wasanalkashri63@gmail.com

**Dr. SETYAWAN WIDYARTO**

Faculty of Communication, Visual Art, and Computing, University Selangor (UNISEL), Selangor, Malaysia. Email: swidyarto@unisel.edu.my

**Dr. JABAR H. YOUSIF**

Faculty of computing and information technology, Sohar University, P.O. Box 44, Sohar, Oman. Email: JYousif@su.edu.om

**MAHMOOD AL-BAHRI**

Faculty of computing and information technology, Sohar University, P.O. Box 44, Sohar, Oman. Email: mbahri@su.edu.om

### Abstract

Deepfakes are a type of "artificial intelligence" that involves the use of genuine images or videos that are then transformed into false forms of media for a particular goal. Deep learning algorithms, implemented in software, are used to accomplish this goal. As deep generative models like generative adversarial networks can be visually indistinguishable from real photos, their potential harmful application raises concerns, such as annoyance, embarrassment, provocation, terrorism, extortion, falsification of information, and intimidation. Because of this, industry and governments have become increasingly concerned about distinguishing between them and limiting their use. In this paper, we present an analysis of the high-frequency Fourier transform model of real and deep network-generated images and show that deep network-generated images include some unreal properties, even if these properties are not obvious to the human eye. In order to determine the most effective model to distinguish between original and fabricated images, frequency domain analysis will be applied to two classifiers, custom VGG16 and Dense Net-121. The goal of this study is to evaluate the effectiveness of our technique with the use of the 140 k Real and Fake Faces datasets of deep fake image. The findings of our experiments indicate that the difference between spectra in the frequency domain is a practical artifact that can be used to efficiently recognize different kinds of GAN-based generated images.

**Keywords:** discrete Fourier transform, VGG16 and Dense Net-121, Deepfakes.

## 1. INTRODUCTION

Deepfakes are images and videos that have been digitally changed to give the impression that they were taken in real life. A recent rise in Generative Adversarial Networks (GAN) has allowed deepfakes (extremely realistic fake images) to be easily generated **[1, 2].** These are the forms of media in which it is possible to replace the image of a person in a picture with the image of another person. The prevalence of deepfakes is at an all-time high and has the potential to cause a great deal more trouble in the future if it is not managed in an appropriate manner. Unfortunately, the possibility of malicious usage of deepfakes also grows with such an improvement **[3]**. As a result, it has become vital to be able to identify deepfakes. The scope of deepfakes has expanded to include luring victims into sending money for scams, as well as swapping the face of a celebrity with that of a pornographic model, spreading false information as fake news on social media platforms **[3, 4]**. This is due to the growing popularity of social media and the increased number of people who are connected to each other by the simple act of clicking on a link. The use of bogus digital material in areas such as fake news, financial fraud, political hardship, blackmail, and fake terrorism has brought the topic to the forefront of public consciousness. Deepfake Detection Challenge is a collaborative effort between prominent tech companies and academic institutions **[5]** that aims to raise awareness of the problem at hand and motivate other academics to work on a solution. Because of this, the focus of both industry and the government has been drawn to discovering and limiting its usage. One example of a deepfake is seen in **Fig 1**. As recently as a short while back, such techniques were out of reach for the majority of customers because they were monotonous and time-consuming, and they demanded a high level of spatial aptitude in PC vision. In any event, due to the recent developments in artificial intelligence (AI) and the availability of vast volumes of information that has been prepared, these obstacles have gradually evaporated. As a direct result of this, the perfect possibility for the development and management of computer-generated content has completely decreased. This has made it possible for even novice users to modify the content at their own discretion.

In particular, deep generative models have recently seen widespread application as a tool for the creation of artificial photos that have a plausible look. These models come into being as a result of the merging of a deep generative model with a deep neural network. A neural network is utilized here as a generative model since it has a number of parameters with a lower total than the amount of data that is used to train these models. This allows the network to locate and skillfully use the essence of the information to make these bogus digital media.

This paper uses different machine learning methods to detect deepfake images. Our approach uses classical frequency analysis of real and fake images. We are using the discrete Fourier transform at high frequencies. The organization of the paper is as follows. A review of previous studies and related work is provided in the following section. The proposed solutions are described in **section 3**. **Section 4** describes the experiments we performed and the results we obtained. We conclude our research in the last section.

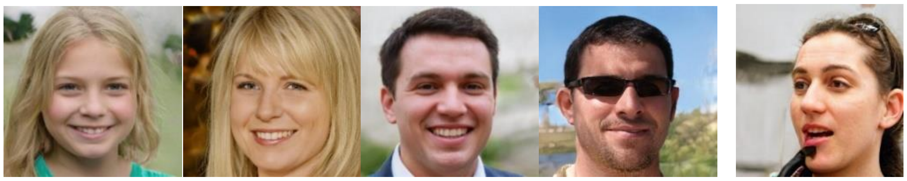


**Fig. 1. Examples of fake faces that do not exist in real life [6].**

## 2. RELATED WORK

Detecting fake images has received wide attention among researchers because images are ubiquitous and appear in various contexts, including social networks, adverts, and other types of online content. In addition, the ability to recognize fraudulent pictures serves as the foundation for many security systems that are capable of functioning in more complex contexts, such as video.

This section reviews recent research on DeepFake detection based on extracted features on spatial and frequency domain. See **table1**

The traditional method of feature extraction involves searching for distortions at the pixel level. Research on deepfakes detection has examined these inconsistencies. They give interpretable hints in the discovery process and demonstrate the distinctions between authentic and fake images. These works suffer from robustness issues when simple transformations are applied to the images or videos.

According to [7] findings, the chrominance components are where the distinctions between synthetic and actual faces become most apparent, particularly in the residual domain. They suggest training a one-class classifier on actual faces by using the chrominance components' variations to challenge hidden GANs. On the other hand, their effectiveness against perturbation assaults such as picture alterations.

Researchers also use DNN-based models to extract spatial features in order to improve the effectiveness of detection and generalization. However, all of these DNN-based detection techniques are vulnerable to adversarial assaults with additive change, and none of the research evaluated their performance in dealing with adversarial noise attacks [8].

Aside from identifying deepfakes, some researchers are aiming to discover modified regions that offer signals as to how legitimate the photo or video is and motivate further work to build DeepFake detectors that are stronger and more robust by focusing on manipulated areas. The study [9] presents an image-specific estimation map to determine the area of forgery in the fake faces. The attention map cannot be fully appreciated unsupervised. This paper proposes not including inverse crosstalk (IINC) as a measure of facial warping localization performance. According to them, forgery detection can be applied to both visible and hidden synthetic techniques. The robustness of the system against disruption attacks still needs to be assessed.

This study [10], used the EM algorithm to extract the local features of the generated facial images to represent the convolutional traces. Any simple classifier can then distinguish real faces from false faces, such as latent Dirichlet allocation (LDA), support vector machine (SVM), and K-nearest neighbours (KNN). Dimension reduction algorithms such as T-SNE might non-linearly distinguish between actual and artificial faces. However, the resilience against perturbation assaults and the generalization capability of many GANs are not well detected.

In addition to differentiating genuine from fake in the spatial domain (extracting features), several research attempts to utilize the frequency domain to distinguish between real and fake. This study [11], examined the design of the generator model and found that the internal value of the generator is normalized, hence limiting the frequency of saturated pixels. After training an SVM-based classifier, false and real faces are distinguished by quantifying how often saturated and under-exposed pixels appear in each picture. According to [12], AutoGAN proposes to detect a unique artifact in GANs that is caused by the upsampling design of popular GAN pipelines. To enhance the generalization capacity of current detectors, a GAN simulator without pre-trained GANs is presented. The artifacts take the form of frequency domain spectra replications. Finally, using the frequency spectrum, a classifier is trained to discriminate Gan-synthesized faces. They also proposed the discovered GAN-based artifacts will likely generalize well in previously unknown synthetic approaches with comparable architectures. Despite this, their robustness against perturbations is not investigated.

This paper utilizes different machine learning methods to detect deepfake images. Our approach uses simple classical frequency analysis at high frequencies to distinguish between real and fake images.

**Table 1. Summary of related work**

| Study | Method | Dataset | Limitation | Accuracy |
|---|---|---|---|---|
| **[7]** | DNN-based models to extract spatial features in RGB color space. | LSUN Bedroom & CelebA-HQ | The existing generative models that used in the study, have not adequately captured many of the inherent color properties of real images. | 94% |
| **[9]** | attention maps to process the feature maps of CNN classifier model | FaceForensics++ & Deepfakes **[13]** | The robustness of the system against disruption attacks still needs to be assessed. | 100% |
| **[10]** | Expectation Maximization (EM) algorithm, using different naive classifiers (KNN, SVM and LDA) | (CELEBA) real faces,<br>&<br>(STARGAN, STYLEGAN, STYLEGAN2, GDWCT, ATTGAN) are fake faces. | the resilience against perturbation assaults and the generalization capability of many GANs are not well detected | 99.81% |
| **[11]** | Extracted features from color Saturation Cues using SVM classifier | NIST MFC2018 | - | 70% |
| **[12]** | Spectrum Domain Features using GAN Discriminator classifier | (CycleGAN) | their robustness against perturbations is not investigated | 100% |

## 3. PROPOSAL APPROACH

In this section, we will detail the methodology we've adopted. Diagram of our approach's design, shown in **Fig 2.**

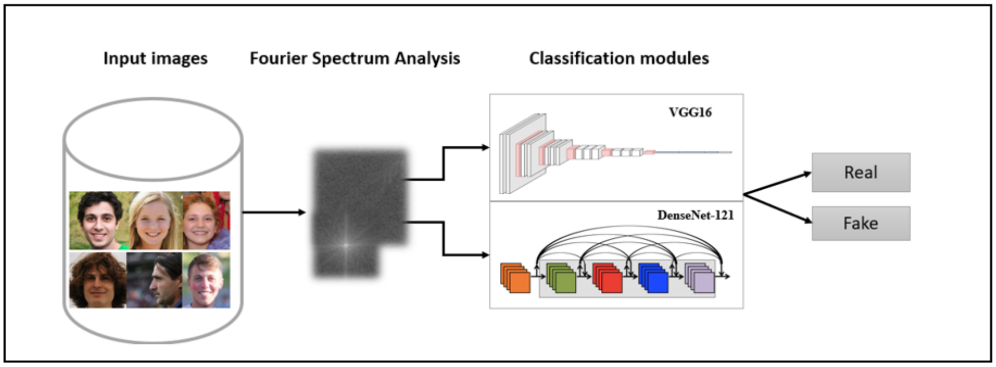


**Fig. 2. Approach's design**

### 3.1 Fourier Spectrum Analysis

The process of transforming an image from its spatial domain into its frequency domain is known as frequency domain analysis. The two-dimensional (2D) Fourier transform is a reliable method for processing images such as enhancing brightness and contrast, blurring, sharpening and noise removal. To decompose a signal into a sum of sinusoids, it uses a variant of the well-known Fourier transform for signals. Since the Fourier transform can reveal the image's frequency content, it is commonly used to analyze images. It's a way to show how the power of a signal can be dispersed across several frequencies. The fundamental concept behind frequency domain analysis is computing the image's discrete two-dimensional Fourier transform. A Fourier transform is needed to assess the features of actual and deep network produced pictures in the frequency domain. Because pixels are not continuous, unlike light waves and sound waves in the actual world, digital pictures are discrete. So, rather of using Fourier Transformation, we should use Discrete Fourier Transformation (DFT).

### Discrete Fourier Transform:

For a discrete two-dimensional signal of image $f(x, y)$ representing individual color channels size M x N will be represented in the frequency domain$F(u, v)$.

The equation for the discrete Fourier transform (DFT) in two dimensions is:

$$F(u,v)=\frac{1}{mn}\sum_{x=0}^{m-1}\sum_{y=0}^{n-1}f(x,y)e^{-2\pi(\frac{ux}{m}+\frac{vy}{n})} \qquad (1)$$

By expanding the exponential in the above formula, sines and cosines are calculated, with the variables u and v the image frequencies will be determined. The frequency domain is responsible for representing the information regarding the amplitude and phase of the signal at each frequency. **Fig 3** displays the results of applying DFT on the data. We will use azimuthal average to get an average of the 2D spectrum from the center to the radii without losing any important data or features. So, if we use this, we can get a more accurate picture of the sample image. Starting by convert the original image to grayscale and calculating the 2-dimensional Spectrum Fourier Transform. The white symmetric patterns in the spectrum picture represents the high frequency power.

The low frequencies are represented by the corners of the spectrum picture. As a result of integrating the two points mentioned above, the white pattern can be centered on spectrum to shows that there is a lot of energy in low frequencies. **Fig. 4.** Shows applying the Gaussian filter. It is smoother cutoff high frequency and identify the changes in an image. The cutoff between passed and filtered frequencies is blurry, producing smoother processed images.

## 3.2 Binary Classifiers

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. We classified real and deepfake images using two neural network's models in the analyzed problem, VGG16 and Desnet 121. Each classification module accuracy was calculated separately from a subset of training and testing data.

**VGG16:** It is a powerful detection object and classification model. It is developed by the Visual Geometry Group (VGG) is an examples of CNN architecture. The presence of a significant number of hyper-parameters is the aspect of VGG16 that stands out the most. The number 16 in VGG16 refers to the fact that the structure is comprised of 16 layers of varying densities. VGG16 has 33 filter convolution layers, each with a stride of 1, and has consistently utilized the same padding and max pool layer, which both have 22 filter strides of 2. The convolution and max pool layers are laid up in the exact same way over the entirety of the design. It begins with two FC, which stand for completely connected layers, and then it moves on to a Soft Max as the output. This network is rather extensive, with an estimated total of around 138 million parameters **[14].**

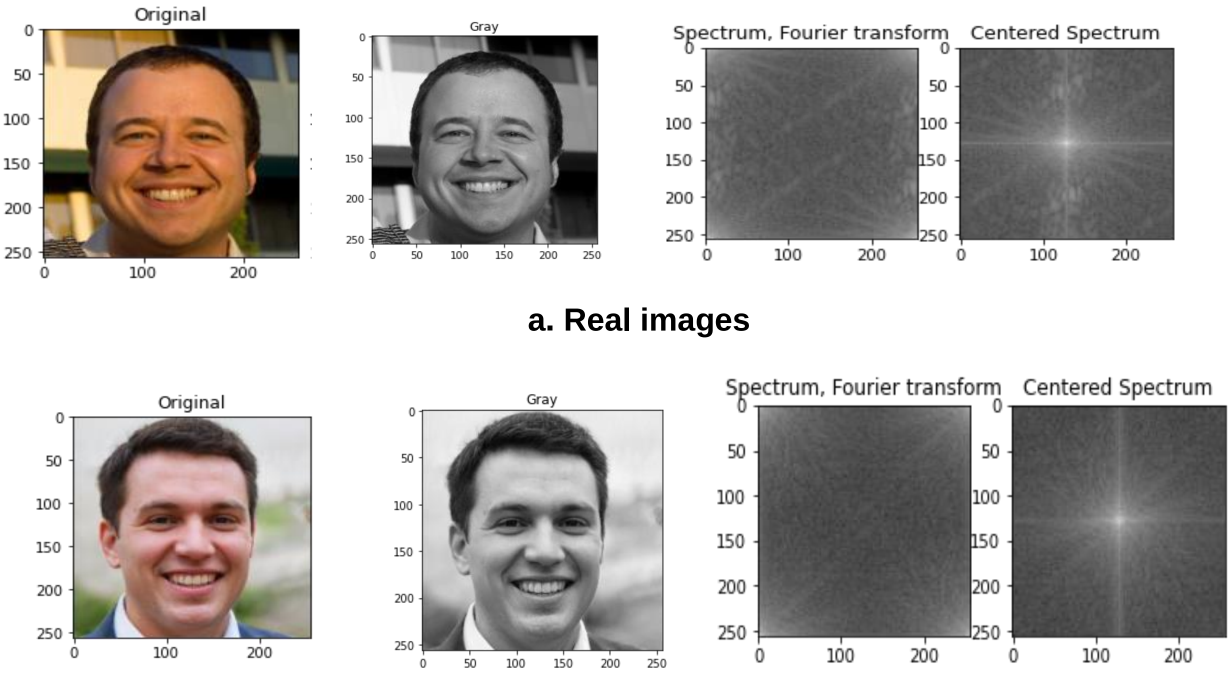
**Dense Net-121:** The term "Dense Net" comes from the Densely Connected Convolutional Network design, in which every layer is directly connected to every other layer. There are 120 convolutions in Dense Net-121, and there are 4 average pools.

The feature maps from all the preceding layers are concatenated and utilised as inputs in each layer rather than being summed. As a result, Dense Nets need less parameters than a comparable classical method. Dense Nets are divided into Dense Blocks, where the size of the feature maps inside a block is kept constant, but the number of filters between them varies. Transition Layers are the layers in between the blocks that cut the number of channels in half compared to the number of channels used and allow deeper layers to use features extracted early on.

Our approach involves the following steps: first, the information obtained from a 140k real and fake faces dataset gathered from Kaggle. The design approach performs the discrete Fourier transform of the image followed by binning the magnitudes of the Fourier coefficients along the radial direction and averaging azimuthally to obtain a reduced spectrum. A frequency-domain signal presents information about its amplitude and phase at each frequency. In order to predict whether an image is real or fake, different binary classifiers are trained and applied to the decay parameters of the image.

### 3.3 Experiments and results

This Experiment illustrates how the entire approach design is carried out, along with the result, and demonstrates how successfully our technique works. In this paper, our approach will be evaluated using photos of the medium resolution, of which there are 140K accessible in both actual and fake datasets.



**a. Real images**

**b. Fake images**

**Fig. 3. a) Real and b) fake visualization image applying DFT on each channel of images.**

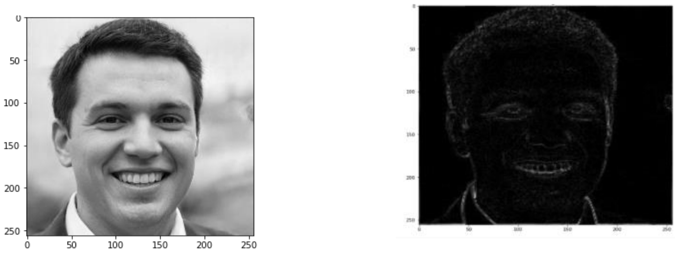


**Fig. 4. Applying the Gaussian filter.**

**Data Set:** In our experimentation, 140 k real and fake face benchmark datasets were used to evaluate the proposed approaches. This dataset can be accessed online on the Kaggle website. This dataset has a total of 1 million FAKE faces created by StyleGAN and 1 million REAL faces acquired from Flickr by Nvidia. Additionally, this dataset contains 70k FAKE faces sampled from the 1 million REAL faces that Bojan provided. The dataset used in this experiment is described in **table 2**.

**Training and testing:** Our method used both types of images, fake and real, to train the classifier for deepfake detection. This paper uses the 140k real and fake images dataset. As seen in **Fig 2,** our pipeline design will perform the discrete Fourier transform of the image followed by binning the magnitudes of the Fourier coefficients along the radial direction and averaging azimuthally to obtain a reduced spectrum. A frequency-domain signal presents information about its amplitude and phase at each frequency. In order to predict whether an image is real or fake, we will use different binary classifiers to train the decay parameters of the image. The classifier will use a 1D power spectrum to distinguish between real and fake images. See **Fig3, Fig4**

**Evaluation Measures:** To confirm the correctness of the results obtained in this study and to evaluate the success of the program activities in achieving the expected objectives. Our experiment evaluated the proposed approaches using 140 k real and fake face benchmark datasets. We used 1000 real and fake images for each dataset, with 80% and 20% used for training and evaluation, respectively. To evaluate a classification model, we use three main metrics: accuracy, precision, and recall.

**1. Accuracy:** it refers to the percentage of correctly predicted test data. Calculated by dividing the number of correct predictions by the total number of predictions.

$$\textbf{Accuracy} = \frac{\text{Correct predictions}}{\text{Total number of predictions}} \tag{2}$$

**2. Precision:** it measures of how many examples are relevant (true positives) among all the ones predicted to belong to a particular group.

$$\textbf{Precision} = \frac{\text{True positives}}{\text{True positives + False positives}} \tag{3}$$

**3. Recall:** it indicates the percentage of examples correctly predicting that they belong to a class.

$$\textbf{Recall}= \frac{\text{True positives}}{\text{True positives + False negatives}} \tag{4}$$

This experiment will use two different classifiers, which are VGG16 and Dense Net-121 respectively. Both classifiers will split the information that has been processed into two parts: one will be used for training, and the other will be used for testing. The remaining twenty per cent of the data will be used to evaluate the efficacy of our methodology and assess the system's accuracy.

**Jilin Daxue Xuebao (Gongxueban)/Journal of Jilin University (Engineering and Technology Edition)**
ISSN: 1671-5497
E-Publication: Online Open Access
Vol: 42 Issue: 02-2023
DOI 10.17605/OSF.IO/BNPC4

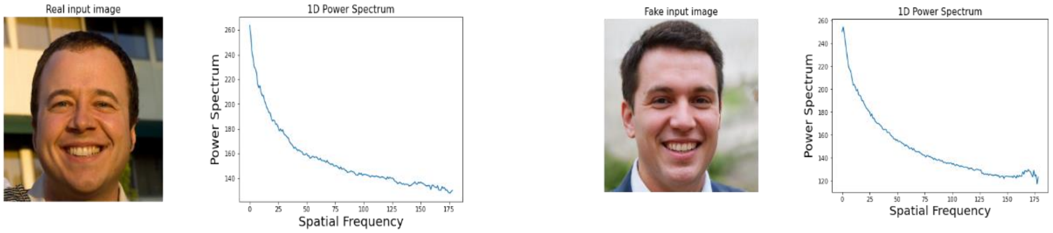


**Fig. 5. Appling the azimuthal component of the DFT power spectrum to analyze spectral distributions of images. It Show the frequency spectra differences of real vs. fake image after applying DFT on each channel of images. There is a difference between a real and fake image based on the high-frequency components of its power spectrum.**

**Table 2. Detailed information about the dataset used in this experiment**.

| **Data Set** | **Resolution** | **Training Size** | **Testing Size** |
|---|---|---|---|
| 140k Real and Fake face | 224 X 224 | 1000 | 1000 |

**Table 3. Experimental results of VGG16 and Dense Net – 121 classifier on 140k Real and Fake face datasets**

| | **VGG16** | | | | **DenseNet-121** | | | |
|---|---|---|---|---|---|---|---|---|
| **140k Real and Fake face** | Accuracy | Recall | Precision | F1 | Accuracy | Recall | Precision | F1 |
| | 99 | 98 | 99 | 99 | 92 | 91 | 92 | 92 |

## Results

According to **Fig 5,** we can clearly distinguish between real and "fake" faces in the high-frequency range of our space feature spectrum. The results of the experiments reported in **Table 3** confirm that distortions in the energy spectrum caused by sampling units are common and can be detected easily. A comprehensive analysis of the results reveals that both architectures had excellent efficiency in detecting and classifying GAN-generated images due to the artifact that GAN generators had on the generated media. Even though VGG-16 may not be the most computationally efficient model, it performed competitively better than the other studied model and yielded encouraging findings when taking into account its performance and behavior. This shows that, in terms of the crucial technological and legal conditions that establish the admissibility of evidence, VGG-16 may be a more acceptable backbone architecture for deepfake detection.

## 3. CONCLUSION

In this study, we described and evaluated the efficacy of a simple method to expose AI-generated deepfake face images. The foundation of our strategy is a high-frequency component analysis. We conducted in-depth tests to show that our pipeline is reliable regardless of the source image. We demonstrate the capability of our technique to identify medium resolution deepfake images using 140k Kaggle datasets of genuine and fake faces. A comprehensive analysis of the results reveals that both architectures had excellent efficiency in detecting and classifying GAN-generated images due to the artifact that GAN generators had on the generated media. Our experiment shows that VGG-16can detect fake faces with 99% accuracy. However, it is challenging to identify low-resolution content due to its limited frequency spectrum and small size. This opens up further future work for us in improving our approach to fit low, medium, and high-quality images and experimenting with them on a larger and more diverse data set, including the architecture of different GANs.

## References

- ❖ Karras, T.; Laine, S.; Aittala, M.; Hellsten, J.; Lehtinen, J.; and Aila, T. 2020. Analyzing and Improving the Image Quality of StyleGAN. CVPR.
- ❖ Nguyen, T. T.; Nguyen, C. M.; Nguyen, D. T.; Nguyen, D. T.; and Nahavandi, S. 2019. Deep Learning for Deepfakes Creation and Detection. arXiv preprint arXiv:1909.11573.
- ❖ Tolosana, R.; Vera-Rodriguez, R.; Fierrez, J.; Morales, A.; , Karras, T., Laine, S., & Aila, T. (2019). A style-based generator architecture for generative adversarial networks. In Proceedings of the IEEE/CVF conference on computer vision and pattern recognition (pp. 4401-4410).
- ❖ 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 (pp. 7890-7899).
- ❖ D. Afchar, V. Nozick, J. Yamagishi, and I. Echizen. Mesonet: a compact facial video forgery detection network. In 2018 IEEE International Workshop on Information Forensics and Security (WIFS), pages 1–7. IEEE, 2018.
- ❖ Xhlulu. (2020, February 10). 140k real and fake faces. Kaggle. Retrieved November 20, 2022, from https://www.kaggle.com/datasets/xhlulu/140k-real-and-fake-faces
- ❖ Li H, Li B, Tan S, Huang J (2020a) Identification of deep network generated images using disparities in color components. Signal Processing 174:107616
- ❖ Carlini N, Farid H (2020) Evading deepfake-image detectors with whiteand black-box attacks. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops, pp 658–659
- ❖ Dang H, Liu F, Stehouwer J, Liu X, Jain AK (2020) on the detection of digital face manipulation. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp 5781–5790
- ❖ Guarnera L, Giudice O, Battiato S (2020a) Deepfake detection by analyzing convolutional traces. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops, pp 666–667
- ❖ McCloskey S, Albright M (2019) Detecting gan-generated imagery using saturation cues. In: 2019 IEEE International Conference on Image Processing (ICIP), IEEE, pp 4584–4588
- ❖ Zhang X, Karaman S, Chang SF (2019) Detecting and simulating artifacts in gan fake images. In: 2019 IEEE International Workshop on Information Forensics and Security (WIFS), IEEE, pp 1–6
- ❖ Deepfakes github. https://github.com/ deepfakes/faceswap. Accessed: 2022-12-11. 2, 3, 5
- ❖ Tammina, S. (2019). Transfer learning using vgg-16 with deep convolutional neural network for classifying images. International Journal of Scientific and Research Publications (IJSRP), 9(10), 143-150.