2024 International Conference on Intelligent and Innovative Technologies in Computing, Electrical and Electronics (IITCEE) | 979-8-3503-0641-5/24/$31.00 ©2024 IEEE | DOI: 10.1109/IITCEE59897.2024.10467257

GAN Generated Fake Human Face Image Detection

Swati Shilaskar
Department of Electronics and Telecommunication Engineering
Vishwakarma Institute of Technology
Pune, India
swati.shilaskar@vit.edu

Mayur Talewar
Department of Electronics and Telecommunication Engineering
Vishwakarma Institute of Technology
Pune, India
mayur.talewar21@vit.edu

Soham Tak
Department of Electronics and Telecommunication Engineering
Vishwakarma Institute of Technology
Pune, India
soham.tak21@vit.edu

Sidhesh Goud
Department of Electronics and Telecommunication Engineering
Vishwakarma Institute of Technology
Pune, India
sidhesh.goud21@vit.edu

***Abstract*— In recent years, Generative Adversarial Networks (GANs) have revolutionized the generation of synthetic data that closely mimics real-world distributions. This research paper focuses on detecting fake human face images generated through GANs. The paper provides a thorough analysis of the current state of GAN-generated fake human face detection and proposes a novel method for robust detection. Existing detection methods often struggle with newly emerging GAN architectures, lack generalization capabilities, and are prone to adversarial attacks. In this paper, authors propose an efficient Convolutional Neural Network (CNN) architecture that detects StyleGAN3-generated fake human faces. To enhance the robustness of the model the algorithm employs a series of filters to extract image data, performs grayscale normalization and convolutional operations to find out whether the images are fake or real with more accuracy. The outcomes of the experiments demonstrate that the approach outperforms the current systems in terms of robustly identifying fake images. Authors achieved an accuracy of 99.42%. This system can be integrated into social media platforms to identify fake profile pictures or deepfake images that are often used for impersonation or spreading misinformation.**

Keywords: GANs, Convolutional Neural Network, StyleGAN3, Deep Learning, Fake Human Face Detection

I. INTRODUCTION

Along with many beneficial applications of this technological innovation, the potential for abuse of the content generated through Generative Adversarial Networks (GANs) is a severe risk that necessitates the creation of forensic methods that can tell the difference between authentic and false (GAN-generated) images [1]. One of these methods is GAN which was introduced by Ian Goodfellow and his colleagues [2]. The probability game between Discriminator and Generator networks is how GAN operates. The discriminator determines if the data is bogus or real and the generator generates updated information [3].

Many types of GANs are available nowadays which can perform multiple tasks such as producing simulated images sequentially using StyleGAN and StyleGAN2 [4], for increasing image resolution SRGAN is used. Human face detection is one of the important tasks that can be performed using GANs. These categories of fake images created are called DeepFake. The two GANs used to create deep-fake images are (DCGAN) Deep Convolutional GAN and the (PGGAN) Progressive Growing GAN [5]. These models are trained on real time images from multiple datasets such as CelebA which contains over 0.2 million images of more than 10k celebrities, LSUN dataset. Each model has different accuracies and hardware requirements. These models can be used for social media content moderation, deepfake detection, digital forensics and other similar platforms.

II. LITERATURE REVIEW

A novel approach to attribute and detect false pictures produced by known GAN models was proposed. The authors used frequency domain analysis and a similarity metric to discriminate between actual and GAN-generated photos. Experiments on different GAN architectures (ProGAN, StarGAN, and StyleGAN) and real image datasets (FFHQ and CelebA) show promising results in attributing images to their corresponding GAN models [6].

The study investigated how to create remote sensing images using Graphical Generative Adversarial Networks (Graphical-GAN). To create artificial images that closely resemble actual ones, the approach blends probabilistic graphical models and deep generative models. The proposed Graphical-GAN achieves high Inception Scores on different categories of remote sensing images, demonstrating its effectiveness [7]. DCGAN were used in the study to create new images that weren't part of the MNIST dataset. DCGAN is shown to effectively generate new data when there is a limited dataset, which can be beneficial in data augmentation

for machine learning tasks [8]. Another study suggested a technique for distinguishing between actual and GAN-generated facial photos. The approach involves pupil segmentation and boundary detection, using BIoU as a distance metric to estimate irregularities in pupil shapes. GAN-generated images are likely to have lower BIoU values compared to real images. Flickr-Faces- HQ dataset is used [9]. The study uses the CBAM i.e., Convolutional Block Attention Module as well as MFAN i.e., Multilayer Feature Aggregation Module for boosting the power of the Xception model in representing features. Both RGB as well as YCbCr components are utilized to improve the robustness of the model to different post-processing operations [10].

On the basis of the WIDER FACE and PIPA datasets, the effects of various anonymization techniques on face detectors are assessed. The effectiveness of cutting-edge GAN-based techniques and conventional face anonymisers is compared using RTX 2080 Ti (Single). The findings showed that they could achieve 2.7% mean Average Precision (mAP) at 60000 iterations and 1.6% mAP at 75000 iterations. [11]. To identify false face photos created by humans and GANs, a fresh dataset (HFM)-Handcrafted Facial Manipulation and an advanced neural network-based classifier called (SFFN) Shallow-FakeFaceNet are suggested. The approach shows a promising performance of 72.52%. For training real fake images CelebA dataset was used which contains more than 200K celebrity images [12]. A method for image completion using DCGAN and contextual and perceptual loss is discussed. The goal is to complete images and build a database of covered faces. This work contributes by presenting a novel approach that exploits DCGANs to address the issue of covered faces. The dataset CalebA used in this paper contains 202599 face pictures and matrix operations are used in the image completion task [13].

A generative model G trained to estimate high-resolution counterparts for low-resolution input images. A discriminator network is optimized to differentiate between genuine images and super-resolved images, and the generator network is trained using a loss function. introduced SRGAN and highlighted certain drawbacks of PSNR-focused image super-resolution. The ImageNet database used contains 350 thousand image samples. All networks were trained on a NVIDIA Tesla M40 GPU [14]. The research paper presented novel counterattacks to evade deep-fake detection, focusing on eliminating the GAN fingerprints from the spectrum of frequencies of generated images. Four attack variants were proposed, achieving significant reductions in detection rates compared to traditional image perturbations. However, success varied based on the dataset, GAN architecture, and detection method, highlighting the need for more robust detection techniques to combat evolving deep-fake threats [15].

Overall, this literature covers various approaches to address the problem of detecting and attributing fake images generated by GANs, including frequency domain analysis, graphical GANs, deep convolutional GANs, and various loss functions. The papers also explore the application of these techniques in different domains, such as remote sensing and face detection. The findings suggest promising results and advancements in this field, with some limitations and potential areas for further research highlighted by the authors.

Determining falsified face images generated by GANs is a challenging and evolving field with significant research gaps. Current detection methods are often limited to specific GAN architectures, lacking generalization across other models such as the faces generated by StyleGAN2 are difficult to detect. As GANs continuously evolve, detecting unseen or novel GAN-generated faces poses a challenge since it is difficult to train GANs on large datasets. Adversarial attacks against existing detection models should be studied to develop countermeasures to defend it. Moreover, spatial detection methods, focusing on local patterns, may be less susceptible to frequency-based attacks, which manipulate global frequency characteristics. Combining both frequency-based and spatial analysis can enhance detection accuracy. Also, the model should be efficient and lightweight and should handle the manipulations done with the output images(post-processing) of the generator in order to enhance the realism of generated faces. Extending detection to videos (cross-modal detection) is necessary for real-time uses.

Dataset biases and generalizations across diverse demographics need to be addressed to ensure the robustness of detection models. Models should perform well for different groups of people, ages, ethnicities etc. Addressing these research gaps will lead to more effective and trustworthy fake human face detection methods based on GANs which are built with robust methods and strong generalization abilities, requiring less hardware and offering higher accuracy in detecting fake images. Overcoming these gaps will help us in building more advanced and susceptible systems.

The comparative study (Table I) assesses the performance of different image recognition models on various datasets. In 2020, the CelebA dataset was used, achieving 99.9% accuracy with Xception and MLFA/CBAM modules. In 2021, models were tested on WIDER FACE and PIPA datasets, achieving 2.7% mean Average Precision (mAP).

The MNIST dataset in 2020 yielded a 99% accuracy using CNN. In 2019, the AID dataset saw a Graphical-GAN achieving a high Inception Score (IS) of 5.96 on CIFAR10. Lastly, the Flickr-Faces-HQ dataset in 2020 resulted in Cross-CoNet and CoNet achieving 99.70% and 98.15% accuracy, respectively. These studies provide insights into model performance across different datasets for diverse image recognition tasks.

TABLE I: Comparison of Prior Art

Reference	Dataset Used	Performance Evaluation
1 (2020)	Flickr-Faces-HQ 52k images	Cross-CoNet 99.70% Acc. CoNet 98.15% Acc.
7 (2019)	AID 10k images, 30 classes	Graphical-GAN: 5.96 IS on CIFAR10
8 (2020)	MNIST 70k Images	CNN 99% Acc.
10 (2020)	CelebA 202k Images	Xception with MLFA and CBAM module 99.9% Acc.
11 (2021)	WIDER FACE 393k images PIPA 60k images	2.7% mAP on 60k iterations

III. METHODOLOGY

In the pursuit of robust and reliable fake human face detection, the methodology employed in this research paper is structured to effectively process and evaluate images with the ultimate goal of distinguishing real from fake human faces. The algorithm is designed to take human face images in RGB format as input and find out whether it is fake or real, consisting of a series of carefully orchestrated steps.

Initially, before applying the CNN model on images we first need to normalize the dataset that is converting the images into grayscale images. Typically, while creating a custom model, the activation function "relu" is utilized in all of the hidden layers for the classification of binary data, and "sigmoid" as an activation function is utilized in the output layer i.e., the dense layer of the model.

In our sequential CNN model design, layer normalization is utilized to normalize the activations at each layer. Following normalization, images undergo filtering with a 3x3 filter size at each layer. To properly train the model, more filters are added at each layer. To avoid data loss from the extracted images, padding is applied to the input image before adding a pooling layer. During model training, overfitting is mitigated using the Dropout function from the Keras layer, which reduces certain percentages of neurons in each epoch. After extracting data from the image using CNN, which produces a 2-dimensional matrix, it is further converted into a 1D matrix to be applied with the Dense layer of the CNN.

Using the 'adam' (adaptive moment optimization) optimizer with 0.001 as standard learning rate and the loss function 'binary_crossentropy,' which provides the difference between true labels and the values predicted, optimization was done during the model compilation phase after the model was designed.

Layer normalization used in the CNN at a given neuron location is given as equations (1), (2).

$$y = \gamma * X_{normalized} + \beta \quad (1)$$

$$X_{normalized} = \frac{x - \bar{x}}{\sqrt{\sigma + \varepsilon}} \quad (2)$$

where, γ and β are learnable parameters and $\bar{x}$ as mean, □ as variance.

Binary cross entropy loss (3), especially used for binary classification.

$$Loss = \begin{cases} -log(1 - \hat{y}) ; & if\ y = 0 \\ -log(\hat{y}) ; & if\ y = 1 \end{cases} \quad (3)$$

where, $\hat{y}$ is a sigmoid function.

The value of (3) is taken as input to binary cross entropy loss function (4).

$$Loss = -y * log(\hat{y} - (1 - y) * log(1 - \hat{y}) \quad (4)$$

In this approach, the dataset used has a total of 140k images out of which we manually crafted 70k fake images using STYLEGAN3 (Fig. 2) and 70k Real images from the Flickr dataset collected by NVIDIA (Fig.1) which is publicly available. Table II specifies the size of the dataset and the resolution of each image.

TABLE II: Dataset Description

Dataset Name	No. of Images	Resolution
FFHQ (Nvidia) - Real	70k	1024*1024
STYLEGAN3 Generated -Fake	70k	1024*1024

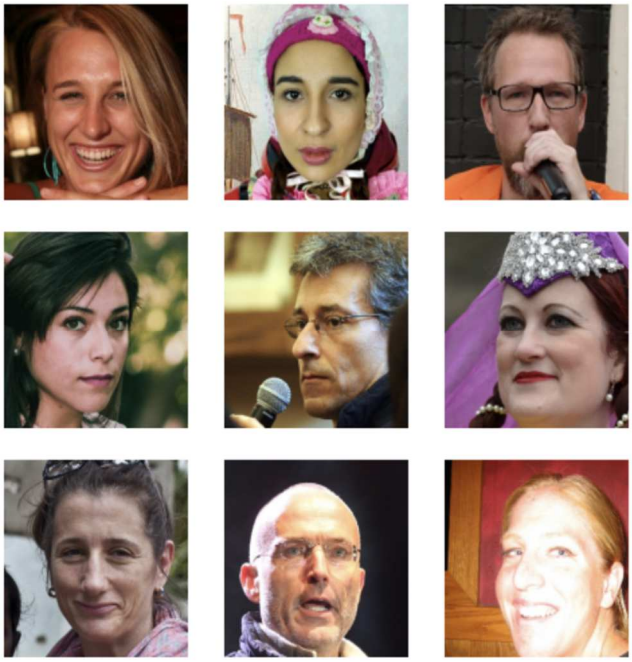

Fig 1: Real Images from NVIDIA Dataset
(Publicly available and adult images used)

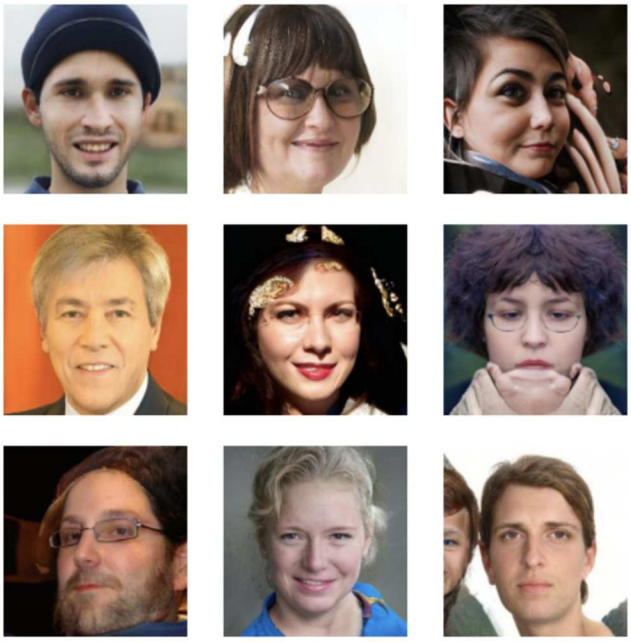

Fig 2: Fake Images StyleGAN3 Generated

Data pre-processing involves reducing the image size to 14 kb and adjusting the resolution to 256x256. Prior to applying CNN on image clusters, we performed grayscale image normalization. This entailed converting the images into grayscale for uniformity and consistency in further analysis.

Algorithm 1: CNN-Based Classification Model

Input: Human face image in RGB format
1: Image normalization into grayscale.
2: **for** image in folder:
3: Apply n filters (f*f) to extract image (p*p) data.
4: Activation using f(x) = max (0, x)
5: Applying 2*f filter size till p.
6: Drop neurons by a factor of 0.1.
7: Flat 2D output into 1D matrix.
8: Pass 1D matrix to a dense layer.
9: **end for**
Output: Classify the image as real or fake.

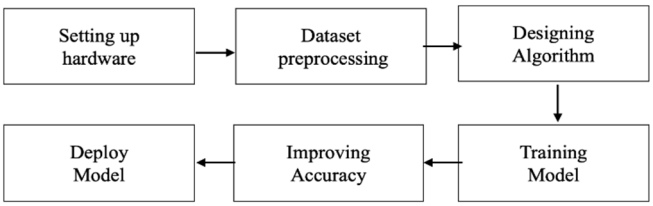

Fig 3: GAN-Generated Human Fake Face Detection

In the proposed system, before deploying the actual model a few steps are followed by the authors like setting up hardware before dataset preprocessing. Designing and applying a model on a preprocessed dataset to evaluate accuracy and improve performance (Fig. 3).

IV. Result And Discussion

In this paper we have introduced significant methodological advancements aimed at enhancing the accuracy of fake human face detection (Fig.4). The accuracy of the model is dependent on several critical factors, each of which contributes to the overall effectiveness of our approach. The Graph below (Fig.5) shows the decrease of loss on the training set as on increase in the number of epochs but this is not the case with validation loss that could be monotonically decreased by the number of epochs that followed.

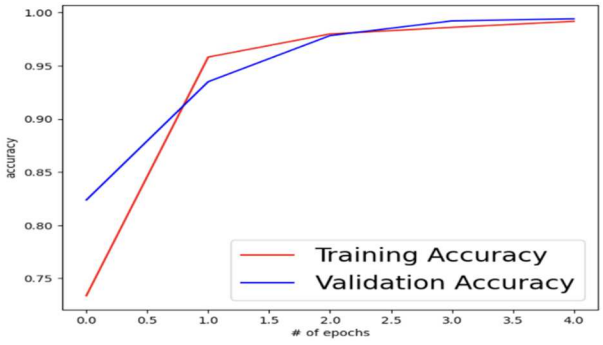

Fig 4: Accuracy Vs No. of Epoch

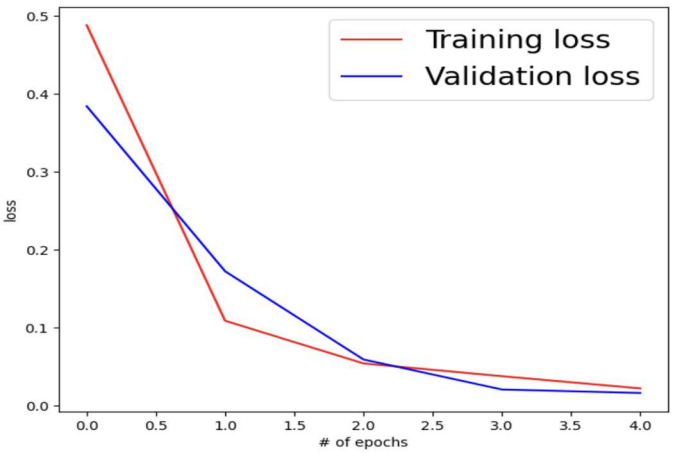

Fig 5: Loss Vs No. of Epoch

Table III illustrates that while training loss and validation loss reduce as the number of epochs grows, training accuracy also increases along with validation accuracy.

TABLE III: Model Accuracy

Epochs	Training Loss (%)	Validation Loss (%)	Training Accuracy (%)	Validation Accuracy (%)
1	48.83	38.43	73.87	82.37
2	10.92	17.27	95.82	93.50
3	5.45	5.95	98.00	97.84
4	3.81	2.10	98.62	99.22
5	2.25	1.66	99.18	99.42

Testing model with test dataset, where fake images predicted as fake Fig. 6 and real images predicted as real Fig. 7.

Fig 6: Fake images detected as Fake

Fig 7: Real images detected as Real

V. Conclusion And Future Scope

A novel method for fake human face detection is presented in this paper. This paper has introduced significant methodological advancements in the realm of fake human face detection, particularly in the context of GAN-generated images. Our approach leverages deep learning techniques to enhance detection accuracy, but challenges and research gaps persist. As we look into the future, addressing these gaps will be crucial for building more advanced and robust systems capable of effectively detecting fake human face images generated through evolving GAN architectures.

The method outlined in this paper represents a remarkable leap forward in the field. It signifies a pivotal moment in the ongoing battle against deceptive content across the internet. By incorporating cutting-edge deep learning techniques, our approach has significantly elevated the accuracy of fake human face detection which is 99.42%, particularly of GAN-generated images.

The future work needs to be conducted in a similar manner with regard to fake videos. In light of the increasing prevalence of fake video allegations in today's world, the work conducted in a similar vein will aid society in safeguarding privacy. Similarly various detection methods can be adapted to keep pace with evolving GAN architectures and exploring multimodal approaches that combine visual, audio, and text analysis for more comprehensive fake face detection.

References

[1] M. Barni, K. Kallas, E. Nowroozi and B. Tondi, "CNN Detection of GAN-Generated Face Images based on Cross-Band Co-occurrences Analysis," 2020 IEEE International Workshop on Information Forensics and Security (WIFS), New York, NY, USA, 2020, pp. 1-6, doi: 10.1109/WIFS49906.2020.9360905.

[2] A. Aggarwal, M. Mittal, G. Battineni, "Generative adversarial network: An overview of theory and applications," 2021 International Journal of

Information Management Data Insights, 2021 vol 1, issue 1, doi: 10.1016/j.jjimei.2020.100004.

[3] R. Patel, "A brief overview on generative adversarial networks. Data and Communication Networks," Proceedings of GUCON 2018, pp. 267-77, doi:10.1007/978-981-13-2254-9_24.

[4] X. Wang, R. Ni, W. Li and Y. Zhao, "Adversarial Attack on Fake-Faces Detectors Under White and Black Box Scenarios," 2021 IEEE International Conference on Image Processing (ICIP), Anchorage, AK, USA, 2021, pp. 3627-3631, doi: 10.1109/ICIP42928.2021.9506273.

[5] R. K., V. R., M. Wilscy, "Detection of Deepfake images created using Generative Adversarial Networks: A review," 2021 Second International Conference on Networks and Advances in Computational Technologies, pp. 25-35, doi: 10.1007/978-3-030-49500-8_3.

[6] M. Joslin and S. Hao, "Attributing and Detecting Fake Images Generated by Known GANs," 2020 IEEE Security and Privacy Workshops (SPW), San Francisco, CA, USA, 2020, pp. 8-14, doi: 10.1109/SPW50608.2020.00019.

[7] G. Wang, G. Dong, H. Li, L. Han, X. Tao and P. Ren, "Remote Sensing Image Synthesis via Graphical Generative Adversarial Networks," IGARSS 2019 - 2019 IEEE International Geoscience and Remote Sensing Symposium, Yokohama, Japan, 2019, pp. 10027-10030, doi: 10.1109/IGARSS.2019.8898915.

[8] Z. Liu, M. Tong, X. Liu, Z. Du and W. Chen, "Research on Extended Image Data Set Based on Deep Convolution Generative Adversarial Network," 2020 IEEE 4th Information Technology, Networking, Electronic and Automation Control Conference (ITNEC), Chongqing, China, 2020, pp. 47-50, doi: 10.1109/ITNEC48623.2020.9085221.

[9] H. Guo, S. Hu, X. Wang, M. -C. Chang and S. Lyu, "Eyes Tell All: Irregular Pupil Shapes Reveal GAN-Generated Faces," ICASSP 2022 - 2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), Singapore 2022, pp. 2904-2908, doi: 10.1109/ICASSP43922.2022.9746597.

[10] B. Chen, X. Liu, Y. Zheng, G. Zhao and Y. -Q. Shi, "A Robust GAN-Generated Face Detection Method Based on Dual-Color Spaces and an Improved Xception," in IEEE Transactions on Circuits and Systems for Video Technology, vol. 32, no. 6, pp. 3527-3538, June 2022, doi: 10.1109/TCSVT.2021.3116679.

[11] S. R. Klomp, M. Van Rijn, R. G. J. Wijnhoven, C. G. M. Snoek and P. H. N. De With, "Safe Fakes: Evaluating Face Anonymizers for Face Detectors," 2021 16th IEEE International Conference on Automatic Face and Gesture Recognition (FG 2021), Jodhpur, India, 2021, pp. 1-8, doi: 10.1109/FG52635.2021.9666936.

[12] Lee, S., Tariq, S., Shin, Y., and Woo, S. S., "Detecting handcrafted facial image manipulations and GAN-generated facial images using Shallow-FakeFaceNet," 2021 Applied soft computing, vol. 105,10725, https://doi.org/10.1016/j.asoc.2021.107256.

[13] Y. Xiao, M. Lu, and Z. Fu, "Covered Face recognition based on deep convolution Generative Adversarial Networks," 2020 International Conference on Artificial Intelligence and Security (ICAIS), Lecture Notes in Computer Science, vol 12239. Springer, Cham. doi:10.1007/978-3-030-57884-8_12.

[14] C. Ledig et al., "Photo-Realistic Single Image Super-Resolution Using a Generative Adversarial Network," 2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Honolulu, HI, USA, 2017, pp. 105-114, doi: 10.1109/CVPR.2017.19.

[15] V. Wesselkamp, K. Rieck, D. Arp and E. Quiring, "Misleading Deep-Fake Detection with GAN Fingerprints," 2022 IEEE Security and Privacy Workshops (SPW), San Francisco, CA, USA, 2022, pp. 59-65, doi: 10.1109/SPW54247.2022.98338.