SN Computer Science

GAN-GUARD: Unveiling Synthetic Illusions Using Encoder-Decoder Texture Analysis

--Manuscript Draft--

Manuscript Number:	
Full Title:	GAN-GUARD: Unveiling Synthetic Illusions Using Encoder-Decoder Texture Analysis
Article Type:	Original Research
Section/Category:	Computer Vision
Funding Information:	
Abstract:	Deepfake technology is a rapidly developing field that uses artificial intelligence to create realistic but synthetic videos or images. It has the potential to be used for a variety of beneficial purposes, such as creating special effects in movies and TV shows, or developing new educational tools. However, there are also a number of potential harmful ways that deepfake technology could be used. One of the biggest concerns is that deepfakes could be used to spread misinformation. For example, a deepfake video could be created of a politician saying or doing something that they never actually said or did. Another concern is that deepfakes could be used for criminal purposes. For example, a deepfake video could be used to impersonate someone else in order to commit fraud or blackmail. Most of these generative uses autoencoders and GAN (Generative Adversarial Networks). These models have the ability to generate realistic fakes. However, in order to eliminate such social problems, we propose a model which leverages the imperfection of such models to localize the faked region generated by them.
Corresponding Author:	Shenbagarajan Anantharajan Mepco Schlenk Engineering College INDIA
Corresponding Author Secondary Information:	
Corresponding Author's Institution:	Mepco Schlenk Engineering College
Corresponding Author's Secondary Institution:	
First Author:	Shenbagarajan Anantharajan
First Author Secondary Information:	
Order of Authors:	Shenbagarajan Anantharajan
	Jegan VG
	Davis Joshi A
Order of Authors Secondary Information:	
Author Comments:	Dear Editor, This article is written based on original research
Suggested Reviewers:	Ragavan V, Ph.D Assistant Professor Grade I, Vellore Institute of Technology ragavan.k@vit.ac.in He is an Expert in AI
	Elamparithi p, Ph A, AAA College of Engineering and Technology elamparithi@aaacet.ac.in He is an expert in ML and AI

GAN-GUARD: Unveiling Synthetic Illusions Using Encoder-Decoder Texture Analysis

Dr. A. Shenbagarajan
Department of Artificial Intelligence and Data Science
Mepco Schlenk Engineering College, Sivakasi, Tamil Nadu, India.
E-mail: shenbagarajan@mepcoeng.ac.in

Jegan VG
Department of Artificial Intelligence and Data Science,
Mepco Schlenk Engineering College, Sivakasi, Tamil Nadu, India.
E-mail: jegan7201_ai@mepcoeng.ac.in

Davis Joshi A
Department of Artificial Intelligence and Data Science,
Mepco Schlenk Engineering College, Sivakasi, Tamil Nadu, India.
E-mail: davisabragama_ai@mepcoeng.ac.in

Abstract

Deepfake technology is a rapidly developing field that uses artificial intelligence to create realistic but synthetic videos or images. It has the potential to be used for a variety of beneficial purposes, such as creating special effects in movies and TV shows, or developing new educational tools. However, there are also a number of potential harmful ways that deepfake technology could be used. One of the biggest concerns is that deepfakes could be used to spread misinformation. For example, a deepfake video could be created of a politician saying or doing something that they never actually said or did. Another concern is that deepfakes could be used for criminal purposes. For example, a deepfake video could be used to impersonate someone else in order to commit fraud or blackmail. Most of these generative uses autoencoders and GAN (Generative Adversarial Networks). These models have the ability to generate realistic fakes. However, in order to eliminate such social problems, we propose a model which leverages the imperfection of such models to localize the faked region generated by them.

Keywords- GAN, Localization, Deep Learning, Deepfake, Face Generation

1. Introduction

In this modern world, humans believe things which are visually represented. Indeed, the power of visual information is undeniable, often conveying complex ideas more efficiently than words alone. However, in today's age, images themselves have become susceptible to manipulation through technologies like deepfakes. Manipulated images and videos can deceive, manipulate, and misinform. The evolution of digital manipulation technologies has given rise to a concerning trend where videos, once seen as credible sources of information, can now be synthesized with relative ease and affordability. Over the past few years, a surge in the creation of synthetic videos, often featuring celebrities, has been observed. These videos are generated using a suite of techniques capable of producing counterfeit images, audio clips, and videos, collectively known as Deepfakes. Deepfake technology, employing synthetic generative models such as GANs and autoencoders, has found applications across various domains, including politics and pornography, where its deceptive capabilities have been harnessed for diverse purposes. In reference AttGAN, Star GAN, StyleGAN and DCGAN, it is stated that GAN models can be modified and trained to adjust according to specific attributes or latent features. This capability allows them to generate synthetic images with characteristics similar to real images but modified based on the desired attributes. In light of the proliferation of Deepfake technology and its potential for misuse, there is an immediate and pressing need to dedicate resources to the research and development of effective and robust methods for detecting and combating face forgery and other forms of multimedia manipulation.

In our investigation, Deepfake methods leave footprints in the generation process. These footprints are due to the imperfection in the process of identifying latent features by the generator of GAN, which then upscales the modified latent feature to match the resolution of the real image. These footprints can be used to detect and localize the faked regions in the generated image. The localization of these regions is more significant and valuable in the field of multimedia forensics. We track the footprints left by the generation method which is universal enough to localize the faked regions produced by unknown GAN techniques.

In GAN-based face generation techniques, the generators typically adopt an encoder-decoder architecture that includes an upsampling component within its decoder. This upsampling step is crucial for enlarging the feature maps generated by the encoder into a full-color image. However, it's important to note that the upsampling process can inadvertently introduce distinctive features into the synthesized images. These features are found in every synthesized images. So, if we track this feature, we can localize the faked regions of the images.

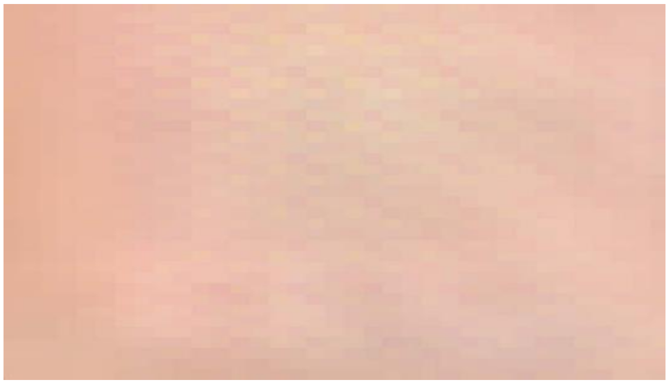


Figure 1. Texture produced by STAR GAN-based face generation methods (checkerboard pattern).

Based on our research findings, we've identified only three primary methods for upsampling. Interestingly, all these upsampling methods introduce certain distinctive textures into the generated images, which we've termed 'fake texture.' In order to attain good localization results, we build our own dataset with existing GANs using the CelebA dataset.

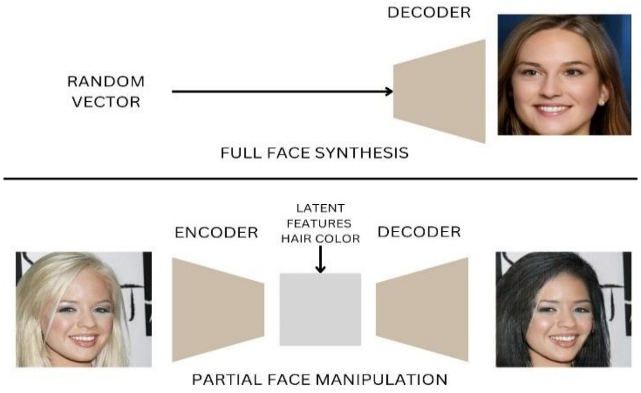


Figure 2. The top subplot shows how full-face synthesis is carried out using style GAN. The bottom subplot shows how partial face manipulation is carried out using AttGAN.

The proposed work is summarized as follows.

- GAN-based face manipulations and generations leave traces that can be utilized to detect and pinpoint the manipulated areas in a synthetic image.
- We create our dataset using AttGAN, StarGAN to manipulate facial attributes. We define the ground truth map by calculating the absolute difference between the manipulated (fake) image and the real image.
- We fine-tune a pretrained encoder-decoder model (Deeplabv3, commonly used in image segmentation) using our generated fake images and the ground truth map.

Thus, the encoder-decoder model gains the ability to identify the footprints left by GANs, as many GANs utilize upscaling methods such as interpolation to resize the resolution to match that of the real image. This process enables the model to effectively detect and localize manipulated regions within synthetic images generated through GAN-based face manipulations. In our final result, we incorporated a heatmap representation of the faked region, which enhances the localization problem. This heatmap formulation allows for more effective identification of the area of interest, i.e., the faked region present in the modified image.

The main objectives of our model are:

- The localization result map must be of the same size as the input image.
- Cross-attribute universality (unknown facial properties) and Cross-method universality (various GAN methods) must be ensured.
- The localization result map should be in a grayscale format because accurate results can be achieved from grayscale visualization.

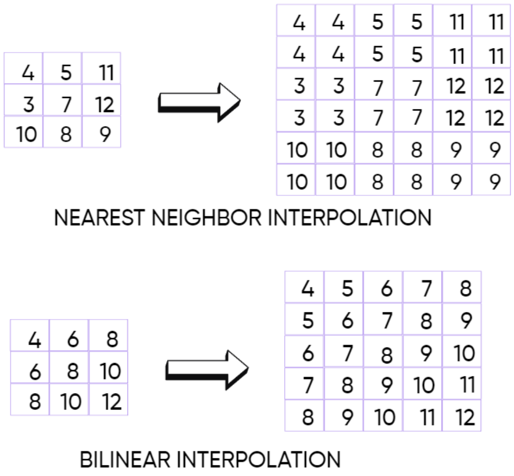
2. Related Work

2.1 GAN-Based Face Generation

GAN-based face generation consist of two techniques: full face synthesis and partial face manipulation methods. AttGAN [2], StarGAN [3], STGAN [4] are used in partial face manipulation techniques. StyleGAN [5] is used in full face synthesis techniques. AttGAN employs a Generative Adversarial Network (GAN) framework where a generator modifies images based on attribute vectors, and a discriminator evaluates their realism. Attribute control is achieved by adjusting attribute vectors, making it ideal for fine-grained image manipulation tasks like partial face modification. StarGAN is a Unified Generative Adversarial Networks for Multi-Domain Image-to-Image Translation which proposes a versatile GAN model capable of translating images across multiple domains with a single generator, enabling seamless transformation between diverse facial attributes and expressions. STGAN is a Unified Generative Adversarial Network for Spatial-Temporal Image Enhancement which presents a GAN framework designed for spatial-temporal image synthesis, offering a unified solution for generating high-quality images with improved spatiotemporal consistency and realism. StyleGAN is a state-of-the-art Generative Adversarial Network (GAN) architecture renowned for its ability to generate high-resolution, realistic images with unparalleled control over various visual attributes.

2.2 Artifacts Left by GAN-Based Face Generation

Attributing Fake Images to GANs [6] shows that GANs carry distinct model fingerprints and leave stable fingerprints in the generated images. Do GANs leave artificial fingerprints? [7] analyzes fingerprints left by pro-GAN and cycle-GAN by the assumption that real image is the combination of fingerprints and noise component.



4	5	11
3	7	12
10	8	9

4	4	5	5	11	11
4	4	5	5	11	11
3	3	7	7	12	12
3	3	7	7	12	12
10	10	8	8	9	9
10	10	8	8	9	9

4	6	8
6	8	10
8	10	12

4	5	6	7	8
5	6	7	8	9
6	7	8	9	10
7	8	9	10	11
8	9	10	11	12

Figure 3. Interpolation Techniques used in upsampling

By our observations, most of the GAN based face generations uses nearest neighbor interpolation, bilinear interpolation and bicubic interpolation to up sample the feature vectors to match the pixel size of the real image (as shown in Fig. 3)

2.3 Localization of Manipulated Region in Face

There are lot of DeepFake detection methods but only some of them deals with the localization problem. K. Songsri-in and S. Zafeiriou [8] works on enhancing face forensic detection and localization by incorporating facial landmarks, improving the accuracy of identifying manipulated or altered faces. Li, L [9] introduce a novel approach called "Face X-ray" to improve the detection of face forgery by using advanced computer vision techniques, enhancing the identification of manipulated or fake facial images. The datasets used for the detection process mainly focus on face swap and are easier to differentiate from real ones.

In our final result, we incorporated a heatmap representation of the faked region, which enhances the localization problem. This heatmap formulation allows for more effective identification of the area of interest, i.e., the faked region (as shown in Fig. 4).

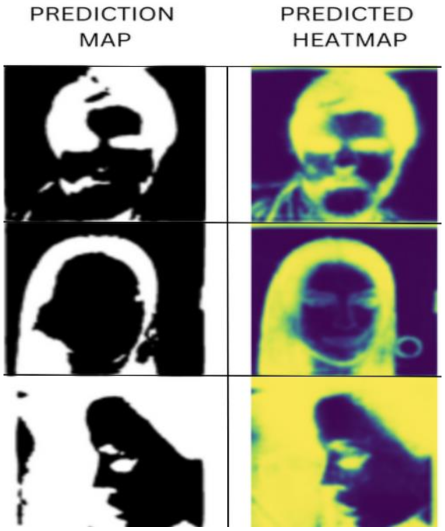


Figure 4. The left column represents prediction map produced by our model. The right column represents the heatmap which accurately shows the faked region.

3. Proposed System

3.1 Generative Adversarial Network (GAN)

GAN architecture consists of a generator and discriminator based on the purpose of generation. The generator part of a GAN learns to create fake data by incorporating feedback from the discriminator. The generators in GAN-based face generation methods are typical encoder-decoder architecture with an upsampling design in its discriminator as the feature vectors generated by the generator are too small to represent an image:

$$\mathbf{I}' = F_{\mathrm{dec}}(F_{\mathrm{enc}}(\mathbf{I}), P) \qquad (1)$$

Here, $\mathbf{I}'$ represents the generated image by GAN, P represents the property added to the image in partial face manipulation, and F_{dec} and F_{enc} is the decoder and encoder used, $\mathbf{I}$ is the real image of height H, width W.

3.2 Gray Scale Fakeness Ground Truth

GAN architecture consists of a generator and discriminator. The generator part of a GAN learns to create fake data by incorporating feedback from the discriminator. The generators in GAN-based face generation methods are typical encoder-decoder architecture with an upsampling design in its discriminator as the feature vectors generated by the generator are too small to represent an image:

$$X = |\mathbf{I}_{i,j} - \mathbf{I}'_{i,j}| \qquad (2)$$

$$\mathbf{X}' = Threshold(X) \qquad (3)$$

Where $1 \leqslant i \leqslant H$, $1 \leqslant j \leqslant W$ and iterated through every pixel of the generated and real image to produce an intermediate image which is then threshold to obtain a grayscale image. Threshold(.) function will convert pixel values less than 127 as 0 and higher as 255. This grayscale image is used in the training of the encoder-decoder architecture for image localization.

3.3 Encoder-Decoder Architecture

In this section, encoder tries to identify the fake textures produced by GAN by providing the manipulated image and ground truth map generated from previous section. The decoder tries to improve the performance of the encoder by minimizing loss.

The Encoder-Decoder structure gets the following inputs:

- Manipulated image ($\mathbf{I}'$)
- Gray Scale Fakeness Ground Truth Map ($\mathbf{X}'$)

$$\mathbf{M}_{\mathrm{Pred}} = F_{\mathrm{dec}}(F_{\mathrm{enc}}(\mathbf{X}', \mathbf{I}')) \qquad (4)$$

Where $\mathbf{M}_{\mathrm{Pred}}$ is the predicted Grayscale fakeness map which contains the probable manipulated regions in the image provided by the user.

The loss functions which are used in the decoder were (i) Dice loss and (ii) IoU score as our job is an image segmentation which segments the faked texture from the whole image.

The **Dice loss** is derived from the Dice coefficient (also known as the Sørensen–Dice coefficient), which is a measure of the similarity between two sets. In the context of image segmentation, the Dice coefficient measures the overlap between the predicted segmentation mask and the ground truth mask.

$$\text{Dice Coefficient } = \frac{2\times|X\cap Y|}{|X|+|Y|} \qquad (5)$$

Where X is the predicted mask and Y is the ground truth mask.

Dice Loss = 1 – Dice Coefficient (6)

The **Intersection over Union (IoU) score**, commonly known as the Jaccard index, serves as a pivotal metric in image segmentation and object detection tasks. It quantifies the similarity between the predicted and ground truth regions by measuring the overlap between them. By calculating the ratio of their intersection to their union, the IoU score provides a concise assessment of the accuracy and precision of a model's predictions. A higher IoU score indicates a closer alignment between predicted and ground truth regions, signifying superior performance in identifying objects or delineating regions within an image.

The IoU score is calculated as the ratio of the intersection of the predicted and ground truth regions to the union of these regions.

$$IoU = \frac{|X\cap Y|}{|X\cup Y|} \qquad (5)$$

Where $|X \cap Y|$ is the area of overlap between predicted and ground truth mask (intersection) and $|X \cup Y|$ is the total area covered by both regions (union).

Dice loss functions as a crucial component during model training, guiding the optimization process by penalizing discrepancies between predicted and ground truth segmentation masks, thus fostering more precise segmentation outputs. In contrast, Intersection over Union (IoU) operates as an evaluation metric utilized post-training to quantify the degree of overlap between predicted and ground truth regions, providing a quantitative measure of the model's segmentation accuracy. While Dice loss shapes the learning trajectory, IoU offers insights into the model's performance.

4. Proposed Methodology

4.1 Fake Image Generation

Using CelebA and FaceForensics++ [11] dataset, we generate more than 7 manipulated images for each image in the dataset using AttGAN and StarGAN. Each manipulated image has a certain facial attribute changed. Pretrained AttGAN and StarGAN is used to generate 256x256 standard manipulated images. The tagging and classification are done to store the images in respective folders as fake or real images. We specially selected AttGAN and StarGAN as they can produce partial face manipulation and full-face synthesis out of all available GAN models. We can specify annotations (facial attributes like bald, beard, glasses, bangs, gender) by means of AttGAN.

4.2 Ground Truth Formulation

The real image and manipulated images are then subjected to ground truth formulation by the help of absolute difference and thresholding the difference. This helps in the preparation of grayscale image which is used to train the encoder-decoder model. Out of all localization methods, we choose grayscale in order to accurately provide the faked region (even the pixels modified). This enhances the integrity of images circulated in modern world.

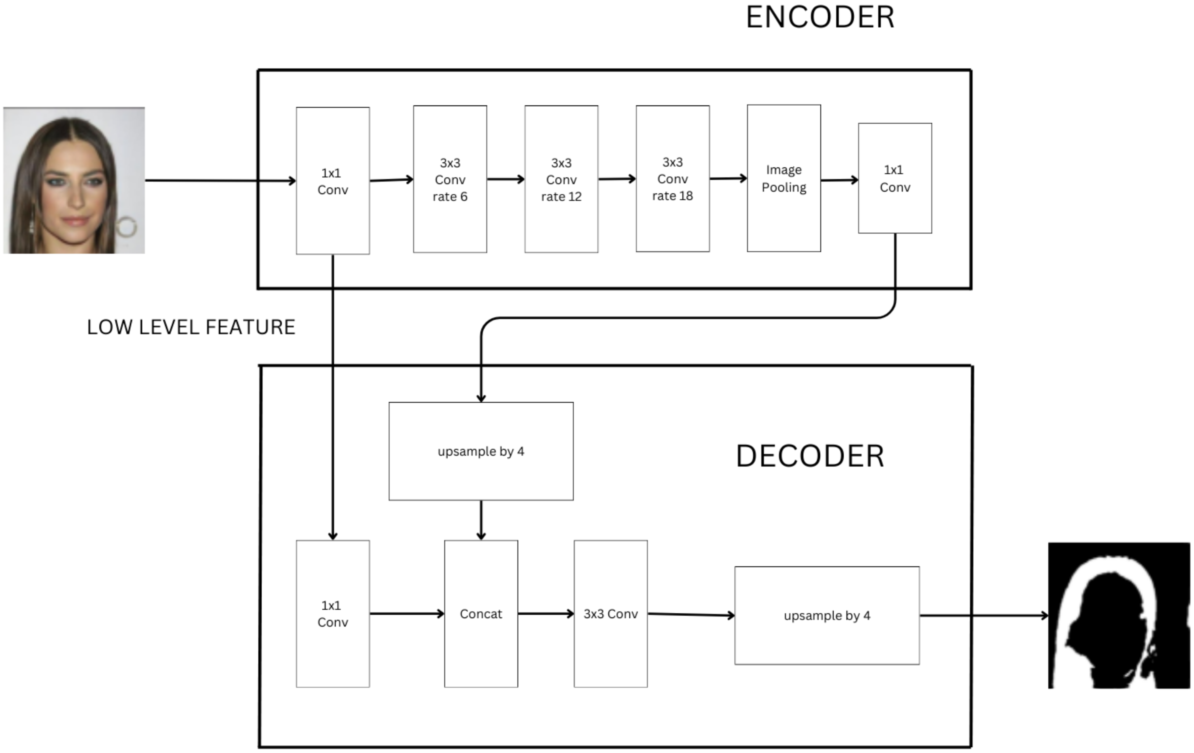


Figure 5. The Architecture of Proposed Encoder-Decoder Model (DeepLabV3 [10]) for Localization of the Manipulated Region in Images which gets input as any image.

4.3 Segmentation of Fake Texture using Encoder-Decoder

As now we have the ground truth and faked image, we can use them to train the Encoder to gain the ability of classifying such texture in any images. Decoder tries to minimize the Dice loss and improve the IoU score of the predicted map as mentioned above.

4.3 Calculating Loss and Performance Factors

Using Dice loss and IoU score, we are assessing the performance of our model across diverse attributes falsified by GAN models. Given the task of segmenting faked regions, our objective aligns with image segmentation. After segmenting these regions, we meticulously scrutinize the epoch at which the loss ceases to diminish. Subsequently, we bypass the subsequent epoch, recognizing that our model has attained its zenith of performance.

5. Results

The training of the encoder-decoder model for image segmentation using Dice loss and IoU score as evaluation metrics demonstrated promising results over 150 epochs. Throughout the training process, the Dice loss exhibited a consistent downward trend, indicating the progressive improvement in the model's ability to minimize discrepancies between predicted and ground truth segmentation masks. Simultaneously, the IoU score showed a steady increase, highlighting the model's enhanced accuracy in delineating manipulated regions within facial images. This positive correlation between Dice loss reduction and IoU score enhancement underscores the efficacy of the training strategy employed. Despite encountering challenges such as convergence issues, the model's final performance showcased significant advancements in detecting and localizing manipulated textures, thereby holding considerable promise for applications in face manipulation detection and image integrity verification. Further research may explore optimization techniques and dataset augmentation strategies to refine the model's performance and extend its applicability in real-world scenarios.

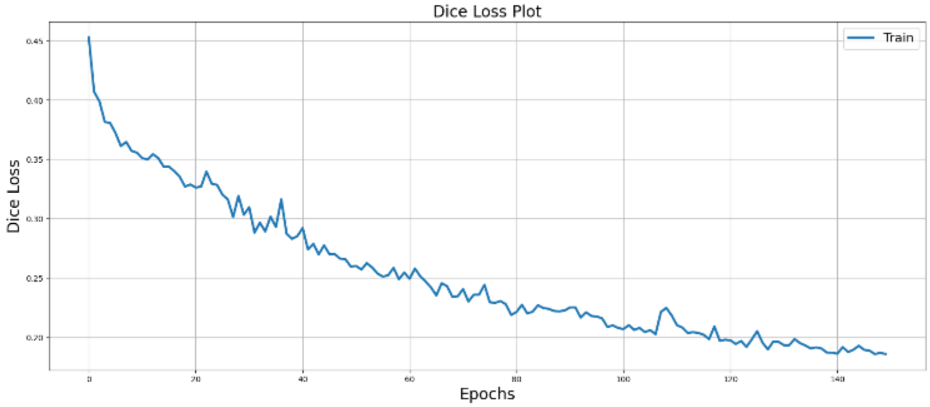


Figure 6. The Plot shows that the Dice Loss decreases from 0.45 to 0.1 after training the model for 150 epochs

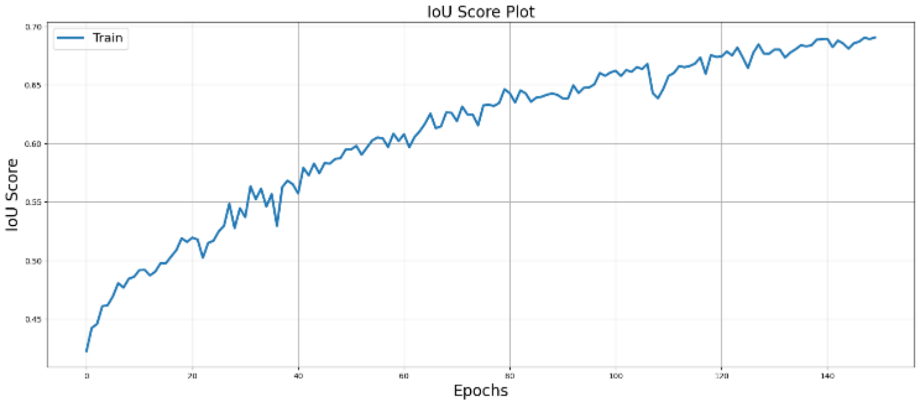


Figure 7. The Plot shows that the IoU Score increases from 0.4 to 0.8 after training the model for 150 epochs

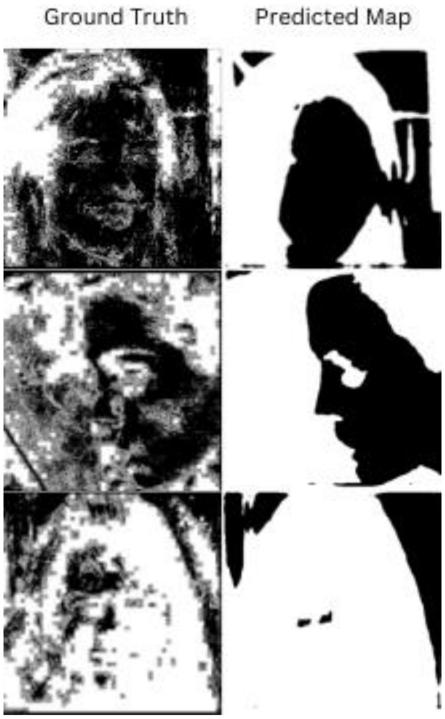


Figure 8. Prediction Results from the Encoder-Decoder Model

6. Conclusion and Future Scope

In conclusion, the development of our image manipulation detection model represents a significant step forward in addressing the pressing need for safeguarding digital image integrity and combating the spread of misinformation. By employing advanced techniques such as Dice loss and IoU score evaluation metrics, our model has demonstrated promising results in accurately detecting and localizing manipulated regions within facial images. This achievement holds profound implications across various sectors, including journalism, digital forensics, security, and AI ethics, where the ability to authenticate digital content is paramount. Moving forward, our model not only serves as a valuable tool for preserving image integrity but also contributes to fostering a more trustworthy and responsible digital ecosystem, thereby enhancing the reliability and credibility of digital media in the modern age.

The future scope for our image manipulation detection model is extensive, spanning various dimensions of innovation and application. Enhancements in model architecture, training procedures, and evaluation metrics offer avenues for achieving higher accuracy and robustness in detecting manipulated images. Adapting the model to different domains beyond facial images broadens its applicability, while real-time integration into online platforms enables immediate detection of fake content. Anticipating and countering emerging threats like deepfakes ensures the model's relevance, while ethical considerations guide responsible deployment. Collaboration across disciplines and industries enriches insights and fosters comprehensive approaches to address digital misinformation. Overall, the model's future trajectory holds promise for advancing digital integrity and societal resilience against image manipulation.

Conflict of Interest

The authors confirm that there is no conflict of interest to declare for this publication.

Acknowledgments

This research did not receive any specific grant from funding agencies in the public, commercial, or not-for-profit sectors. The authors would like to thank the editor and anonymous reviewers for their comments that help improve the quality of this work.

References

Choi, Y., Choi, M., Kim, M., Ha, J. W., Kim, S., & Choo, J. (2018). StarGAN: Unified generative adversarial networks for multi-domain image-to-image translation. In Proceedings of the IEEE conference on computer vision and pattern recognition (pp. 8789-8797).

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 (pp. 5781-5790).

He, Z., Zuo, W., Kan, M., Shan, S., & Chen, X. (2019). AttGAN: Facial attribute editing by only changing what you want. IEEE Transactions on Image Processing, 28(11), 5464-5478.

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).

Li, L., Bao, J., Zhang, T., Yang, H., Chen, D., Wen, F., & Guo, B. (2020). Face x-ray for more general face forgery detection. In Proceedings of the IEEE/CVF conference on computer vision and pattern recognition (pp. 5001-5010).

Li, Y., Yang, X., Sun, P., Qi, H., & Lyu, S. (2020). Celeb-DF: A large-scale challenging dataset for deepfake forensics. In Proceedings of the IEEE/CVF conference on computer vision and pattern recognition (pp. 3207-3216).

Liu, M., Ding, Y., Xia, M., Liu, X., Ding, E., Zuo, W., & Wen, S. (2019). StGAN: A unified selective transfer network for arbitrary image attribute editing. In Proceedings of the IEEE/CVF conference on computer vision and pattern recognition (pp. 3673-3682).

Marra, F., Gragnaniello, D., Verdoliva, L., & Poggi, G. (2019, March). Do GANs leave artificial fingerprints?. In 2019 IEEE Conference on Multimedia Information Processing and Retrieval (MIPR) (pp. 506-511).

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 (pp. 1-11).

Songsri-in, K., & Zafeiriou, S. (2019). Complement face forensic detection and localization with facial landmarks. arXiv preprint arXiv:1910.05455.

Yang, Y., Liang, C., He, H., Cao, X., & Gong, N. Z. (2021). FaceGuard: Proactive deepfake detection. arXiv preprint arXiv:2109.05673.

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 (pp. 7556-7566).

Yurtkulu, S. C., Şahin, Y. H., & Unal, G. (2019, April). Semantic segmentation with extended DeepLabv3 architecture. In 2019 27th Signal Processing and Communications Applications Conference (SIU) (pp. 1-4).