GAN generated Deepfake face detection implementing Transfer learning

Simanta Sarkar

(Student ID: x18201148)
Master of Science in Data Analytics
National College of Ireland
Dublin,Ireland

Shreya Bhattacharya

(Student ID: x18185207)
Master of Science in Data Analytics
National College of Ireland
Dublin,Ireland

***Abstract*— Inappropriate use of the GANs generated realistic images or videos brings in growing concern in social networks and social media. Doctored images or videos of people or events can have detrimental effects on an individual, lead to political or social unrest, cyber crime and many more. Thus, it is important to effectively and efficiently detect the deepfake images or videos. In this research a novel model is proposed to detect deepfake images implementing Inception-ResNet-v2 through transfer learning. The model is trained and tested on Google colab. The environment provides access to free high-performance GPUs. DCGAN is used to generate the sample of fake images. The architecture consists of ADAM model optimizer, sigmoid cross entropy as the loss function, image subsets in mini batches of size 64. To classify the fake face images from the real pretrained Inception-ResNet-v2 is implemented through transfer learning. The model has yielded high accuracy.**

Keywords*— *Computer Vision, Deep Learning, CNN, DCGAN, Deepfake.

I. INTRODUCTION

This One of the growing public concerns is the emergence of high resolution realistic images and videos holding facial and face information. With advancement in computer vision the digital manipulation can be performed with much ease through implementing the booming DeepFake techniques. The term DeepFake[1] coined by a Reddit user came into existence in 2017 after he swapped the celebrity faces to create fake pornography implementing a machine learning algorithm. It can be explained as a method rooted in deep learning, able to generate realistic fake images or videos of face swapping of an individual by the face of another person in images or videos. Several modified methods have taken this to the next level through its contribution in creating head translation, facial movements, and videos from pictures. In addition to this deep learning methods can be implemented to mock an individual's voice through synthesis of speech and voice conversion further synchronizing with visual factors to build audio-visual parody This can have adverse effects on social, political and economic aspects including fake news, fraudulent transactions, and political unrest victimizing a world political leader.

This has resulted in a widened research domain with pressing interest in detecting the deepfake images and videos. This is evident from the growing numbers of workshops and competitions in the domain held by renowned companies like Facebook, NIST, Kaggle and may more. Traditionally, generating realistic fake images was constrained by the lack of availability of advanced tools for editing. For example, the high cost time consuming CGIs used in movies involved expertise in the domain. However, over the recent years availability of huge public data and evolution of deep learning algorithms through extensive research works has made the task of doctoring faces in real images or composing a non-existing face easy. [2]Generative Adversarial Networks (GAN) based on deep learning is a milestone in deep learning to add a new dimension computer vision domain. GANs can be used to simulate images, videos, voice, art and many more. The architecture of the model involves two neural networks namely "generative network" and "discriminative network" respectively. Generative network also termed as the artist is trained to generate credible samples convincingly representing the real samples from the random noise input. As the term suggests the discriminative network learns to classify between generated samples from the real samples. Most recent works have widely used GANs to generate photo-realistic images or content of videos. The modified GANs like PGGAN, the progressive growth of GANs have successfully generated high resolution forged images making it impossible to detect without a compatible detection technique. The GAN generated images are not sourced from the original images posing a challenge to the generic detectors. However, a deep learning based approach has proved successful in forgery detection. The detection can be explained as a binary classification machine learning algorithm. In this regard a convolution neural network (CNN) can be utilised to build the detector.

A. objective and research question

The project aims at contributing in detecting the GANs generated DeepFake face images to control the adverse effect of the forged images on social networks, news, politics and many more.

"Can transfer learning be implemented on Inception-ResNet-v2 to classify between real and GAN generated fake images?"

In recent years GANs have proved success in generating high resolution realistic images posing a challenge in fake images detection. However, various identification problems have been solved implementing deep neural networks.

II. LITERATURE REVIEW

A revolutionary work in the field of Face Swapping which overcame the challenges of face segmentation was the one that showed that instead us using a conventional face segmentation, standard fully convolutional network (FCN) can perform remarkable fast and accurate segmentations, if the model can be trained with a rich dataset. The work was robust under unprecedented conditions and also allowed quantitative tests. The quantitative test results further showed that the intra-subject face swapping had hardly any effect on the face verification accuracy. [3] A two-stage framework named FaceShifter, where in the first stage the swapped face is generated in high fidelity by exploiting and integrating the target image accurately and adaptively. An attribute encoder was developed for extracting multi-level target face attributes, and a new generator was developed with carefully designed Adaptive Attention Denormalization (ADD) layers to adaptively integrate the identity and attributes for the face synthesis. Also, a second stage had been included, consisting of Heuristic Error Acknowledging Refinement Network (HEAR-Net) that was trained to recover anomaly regions in a self-supervised way without any manual annotations.One of the primary advantages of this method is that it is subject agnostic, that is once trained it can be applied to any face pairs without training based on subject. This method has the ability to produce super realistic images.[4] The first work to address the security issues of the face swaps and had projected the techniques that can attack the identification or authentication systems. This research work highlighted the usage of machine learning techniques to determine the swapped images and also to determine the automated solutions for it. From the face images, firstly the key points are detected and each of them are represented by a descriptor capturing the local information, and these key points are independent of each other. Clustering was being applied on each and every descriptor and the centroids of each of these clusters were put to compose a codebook. So, when a new feature has arrived, it is fed into either a linear or a non-linear based machine learning to predict it. [5] The paper to illustrate for the first time a Convolutional Neural Network based system that could detect the fake faces those were generated at the latest best possible ways. The detection method had an accuracy of 99.4% and has a great impact in the field of forensics. In this model they carefully designed a CNN architecture, with focus on high pass filter for the input image, the number of layers and the activation functions that had the ability to detect the fake images efficiently. The input images are transformed into residuals using a high pass filter which are then fed into three layered groups. They are then passed through multiple layers and this method was named as the state-of-the-art method. [6] Another research work highlighted the performance of various image forgery detectors against image to image translation. The study had showed a detection accuracy of about 95 percent on GAN generated fake images. This paper mostly focused on the image to image translation, a process that modifies the attribute of the target image thus making it very realistic. With every possible aspect, to which extent and to what context these attacks can be unveiled and how can be dealt with was the highlighted and illustrated in this paperwork.[7] To accept the challenge of detecting GAN generated fake images, a deep Forgery Discriminator (DeepFD) was developed. The conventional way of detection of fake images has the problem that by learning a classifier directly, sometimes it becomes tricky as it becomes difficult to find a proper discriminator for determining the fake images generated by various types of GANs. To come out of this shortcoming, a contrastive loss was adopted to find the typical features of the synthesized images generated by different GANs and then a classifier was added to detect the computer-generated images. To aid this image forgery detection there were two approaches taken apart from the conventional ways those were the 1. Intrinsic and the 2. Extrinsic approach. The first approach require the original externally signal to check whether it is a forgery or not, whereas the second approach seeks the intrinsic image of the received image, thus finding the anomalous statistical property of it to detect whether it's a forgery or not. The main contribution of this research work is the contrastive loss that could be used to well capture the joint discriminative features of the fake images generated by different GANs with huge precision, accuracy and recall rate. [8]To develop a statistical framework for the detection of deepfakes and error guarantees. A research work had build on the information-theoretic study of authentication to develop deepfake detection as a hypothesis testing problem specifically for outputs of GANs, themselves viewed through a generalized robust statistics framework. It thus helped in prevention of incorrect classification for various types of conditional GANs and for detecting any kind of swaps. [9] Another research work showed first, several state-of-the-art GANs that were collected to generate the fake-real image pairs. After that the contrastive will be used on the proposed common fake feature network (CFFN)to learn the discriminative feature between the fake image and real image (i.e., paired information). At the end, a smaller network was concatenated to the CFFN to determine whether the feature of the input image is fake or real. With the advancement in this field, the PGGANs and BigGAN was able to generate highly hyper realistic photos, which was very difficult for the human eyes to differentiate between the fake and the original. [10] With the advancement in the field of fake images, so the detection of such fake and manipulated images and videos is an important part in the field of digital forensics. To illustrate the use of a multi-tasking learning approach to simultaneously detect fake images and videos and locate the manipulated regions in a subject. Information developed by performing one task was shared with the other task and hence enhance the performance of both tasks. The network had an encoder and a Y-shaped decoder, which were used to detect the manipulated regions of the fake images. One of the primary advantages of the system was to deal with unseen attacks where only few samples were used for fine tuning[11].

III. METHODOLOGY

A. Model

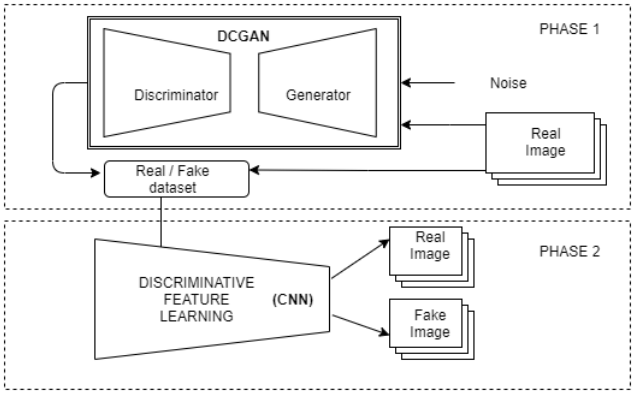

Fig. 1. DCGAN Model

GANs (Generative adversarial networks) have challenged the belief that computers are incapable of being innovative through revolutionizing the field of computer vision. It is a deep learning algorithm which involves two neural networks trained simultaneously to counter each other. The two sub-models are generators and discriminators. The generator is trained to generate credible examples from the domain and the discriminator is trained to classify the input as fake or real. Furthermore, it has addressed the requirement of large data for model training as data can be generated using the networks. DCGAN (Deep Convolution GAN)[12] is an improvement of the GANs through implementation of convolutional-transpose and convolutional layers in the generator and discriminator respectively exclusively. Eliminating the fully connected layers, in the discriminator average global pooled layers is introduced along with the ReLU activation functions Additionally, batch normalisation is performed in generator and discriminator layers, respectively. Several approaches have been proposed in recent years to detect the doctored images generated by GANs. The training images are generated implementing DCGAN, the most recently discussed GAN architecture to ensure the performance of the model. For the fake face classifying network, transfer learning is implemented on Inception-ResNet-v2[13] to identify real and GAN generated fake images. The discriminator network is trained to learn the discriminative features from the training set by introducing the binary cross entropy loss function.

B. Generative Architecture

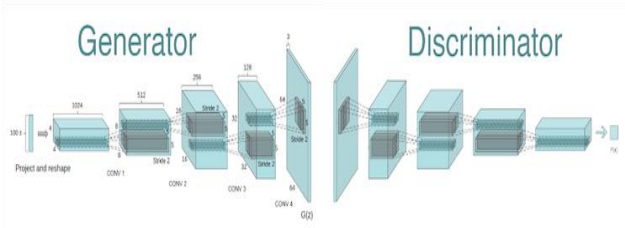

Fig. 2. Generator Architecture

DCGAN is implemented to generate the deepfake face images. The model is trained with the Celeb 100k dataset from Kaggle. It consists of cropped face images of celebrities of 128*128 pixel. Samples were taken in mini batch of size 64 was taken for the purpose. The The architecture consists of ADAM model optimizer, sigmoid cross entropy as the loss function. Details on the usage of each of these datasets are given below. scaling to the range of the tanh activation function [-1, 1] was used to train model. weights were initialized from a form of normal distribution with mean 0 and SD 0.02 . LeakyReLU 0.2 , While previous GAN work has used momentum to accelerate training, we used the Adam optimizer with tuned hyperparameters. Learning rate of 0.002 were used.

C. Detection Architecture

```
Model: "sequential_1"
_________________________________________________________________
Layer (type)                 Output Shape              Param #
=================================================================
inception_resnet_v2 (Model)  (None, 2, 2, 1536)        54336736
_________________________________________________________________
global_average_pooling2d_1 ( (None, 1536)              0
_________________________________________________________________
dense_1 (Dense)              (None, 2)                 3074
=================================================================
Total params: 54,339,810
Trainable params: 54,279,266
Non-trainable params: 60,544
_________________________________________________________________
```

Fig. 3. Detection Architecture

Detection of GANs generated deepfake is challenges as the fake images are generated from the low-dimensional random noise instead of modifications of original images. Combination of residual connections and updated version of the Inception[13] architecture is a milestone in image classification. Residual[14] connection is argued to be important in extreme deep learning architecture training. The filter concatenation phase of very deep inception network architecture is commonly substituted with residual connection. Thus inception network architecture is added with the benefits of residual connection along with inherited computational ability. The Inception-ResNet-v2 is used in the project. This competes with the recently introduced Inception-v4 architecture with respect to the raw high cost but has higher step time in practice probably because of lesser number of layers. The compounded version 4283 of Inception-ResNet-v2 has enhanced performance in image recognition. According to our understanding, the combination of residual connection with the Inception architecture dramatically enhances the speed of model training. Thus, Inception-ResNet-v2 is used for image classification in the project. The primary assumption of most machine learning and data mining algorithms of training and test dataset having the same distribution and belonging to the same space of feature is violated by real world implementation. Suppose for example we are interested in implementing classification algorithms in a particular domain but training data is available from a disjoint domain of interest belonging to a different feature sample space having diverse data distribution. In this regard knowledge transfer can successfully enhance the training performance eliminating high cost involved in data-labeling. This led to the emergence of a new learning architecture called transfer learning. [15]Transfer learning can be explained as follows. Say for a given domain source D_s and learning task T_s with target domain D_t and learning task T_t the goal of transfer learning is to enhance the learning process of target predictive function in D_t utilizing the information in D_s and Ts for $D_s \neq D_t$ or $T_s \neq T_t$.

D. Data Mining Methodology

This project adopts the CRISP-DM[16] model or the Cross Industry Standard Process for Data Mining methodology to implement the machine learning techniques in detecting GAN (Generative Adversarial Networks) generated deepfake faces in images. The main objective for choosing CRISP-DM model to implement this project is that it provides a framework to design, develop, build, test and finally deploy the machine learning solutions for the problems associated with GAN concerning the generation of realistic fake images. A machine learning project is segmented through implementation of the framework into six stages: business understanding, data understanding, data preparation, modelling, evaluation and deployment.

- **Business Understanding:** To effectively outline the problem and the goal of any machine learning projects it is important to understand the business perspective or social, political and economic perspective of it followed by initial designing of the project plans to achieve the goals. The phase not only involves deciding on the working data for the project. There is a whole load of business, social, political and economic benefits that have come with the detection of GANs generated deepfake face images which mainly focuses on favoring the social networks. This can facilitate the spreading of fake news, hoaxes, political unrest and many more. This can eventually suppress the alleged forging of images. To address the issue it is decided to build the model on the popular Kaggle Celeb 100k.
- **Data Understanding:** Understanding the perspective of the project and deciding on the data is important to explore the data to access the data quality and its feasibility for implementing the desired model. The dataset name celeb 100k containing cropped face images of celebrities is sourced from Kaggle. It is collected by utilizing Kaggle API. The dataset consists of 1 lakh celebrity face images. Each image is of size 128 * 128 * 3 in JPG and PNG format. The image quality is verified through multiple subplots.
- **Data Preparation:** This stage plays the crucial role in reconstructing the raw data to feed the model. This basically involves subsetting the data, data cleaning, data merging. Data preparation involved collection of all the image names and image path in the text file ensuring well-constructed data prior to modelling. The images are sampled in mini batches to train the generator network.
- **Model:** This phase involves selection and implementation of modelling algorithms on the processed data. The models are built implementing the relevant modelling tools. The trained model is put to test on an exclusive set of data to determine the model performance. The proposed architecture involves two consecutive model training. The initial phase involves the DCGAN architecture consisting of ADAM model optimiser, sigmoid cross entropy as the loss function, image subsets in mini batches of size 64 to generate the fake images. The following classification model to perform the detection task using the Inception-ResNet-v2 architecture in transfer learning. The binary cross entropy loss function is used in the classification network.
- **Evaluation:** Emphasis is laid in evaluating the implemented model in this phase. It is thoroughly accessed if the model has successfully addressed the proposed problems. To evaluate the performance of CNN train and validation accuracy is plotted against the epochs. Further for assessing the classification performance train and validation loss is plotted against echos.
- **Deployment:** The deployment platform is a common repository where the final model will be set up where all the associated files and process models are organized. This platform gives a road map to various groups or stakeholders to collaborate and access the required data. There are several web services offered with respect to deployment where the final architectures are stored in a specific repository that is made accessible widely among different stakeholders. In that perspective the detecting architecture developed in this process can be integrated with social networking service based companies like Facebook, NIST, Kaggle and many more stakeholders as this can contribute towards the challenges faced by them in controlling the adverse effects caused by forged imaged in social media networks, news and various other platforms.

IV. IMPLEMENTATION

We have trained and tested our networks in the google colab environment. to access the free high performance GPUs. Both high high RAM and baseline environment were used according to the needs of the task. All the conducted experiments on the environment were supported by the GPUs: NVidia Tesla K80, Tesla P100, Tesla T4.

A. DCGAN: Fake face generation

The DCGAN architecture is inspired from (ref) which consists of ADAM model optimiser, sigmoid cross entropy as the loss function, image subsets in mini batches of size 64. The model was trained for 60 epochs. But as Google colab allows 12 hours timed out sessions. Thus the training task was split into 6 each of 10 epochs at a time. After every 10th epoch a check point was exported to personal Google drive to save the model. . It was observed that the average time taken in training one epoch varied for different GPUs. Tesla K80 running on a Kepler architecture took 98 minutes on average. Tesla P100 running on a Pascal architecture took 72 minutes on average. Tesla T4 running on a Turing architecture took 47 minutes on average. The trained model was used to generate the fake images.

B. Inception-ResNet-v2: Fake face detection

Transfer learning was implemented on pretrained Inception-ResNet-v2 to classify fake face images as it has enhanced classification performance. Binary cross entropy is used as the loss function on Inception-ResNet-v2 to perform the classification task. Subsets of real and fake images each having 29,500 samples are used to train and test the model. The subset is used to eliminate the task of training in mini batches. Samples more than 29,500 were exceeding the RAM. This was trained in Google colab's high RAM environment. The training consisted of 20 epochs , for each epoch average

runtime was 8.30 min,53sec ,2.45 min and minutes on Tesla k80 running on a Kepler architecture , Tesla p100 running on a Pascal architecture and T4 running on a Turing architecture respectively.

C. Fake Face generation:

DCGAN is implemented to generate deepfake face images. The model is trained for 60 epochs on the Celeb 100k dataset. After every epoch a random sample of 5 images are generated followed by plotting a loss function to check the model performance.

Fig. 4. Sample images (1st epoch).

Fig. 5. Sample images (10th epoch).

Loss function is utilized to evaluate the model. However, the loss curve is often non-intuitive because of the fact that the generator and discriminator counters each other in the learning process. Mostly, it is observed that the generator loss and the discriminator loss converge at a point. This is considered to be the optimum point where the learning process of the model cannot be further improved. At the initial training point in the first epoch there is a significant difference in generator and the discriminator loss stating that the discriminator majorly fails to detect the fake generated images. However, the respective loss function converges at approximately batch 300. There might be an instant after several iterations with significant improvement of the discriminator resulting in high generator loss. The significant fall in discriminator loss curve for the final 60th epoch illustrates even at early batch hints at improved model train of the discriminator. From the notable difference between the generator and the discriminator loss hints the model has reached an optimum learning point where it can successfully generate realistic fake images.

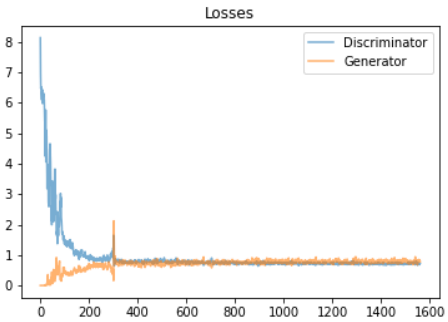

Fig. 6. loss (1st epoch).

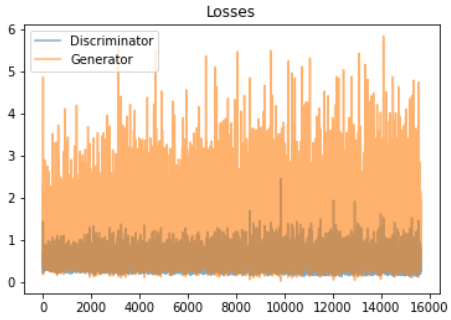

Fig. 7. loss (60th epoch).

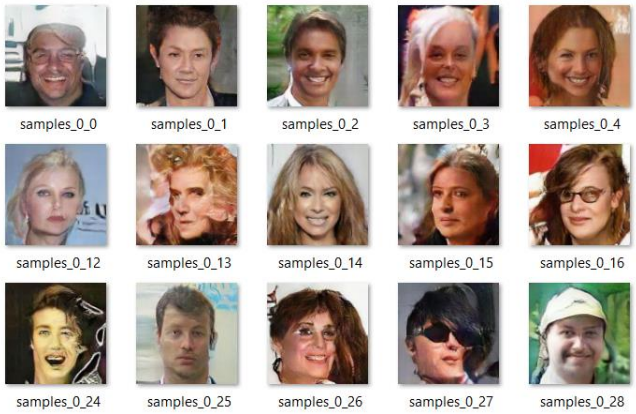

Fig. 8. Generated Fake Faces

D. . Face classification:

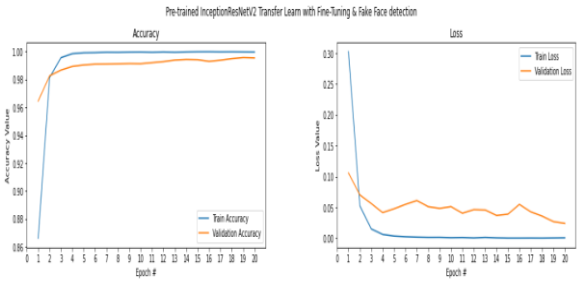

Fig. 9. Accuricy and Loss (Inception-Resnet v2)

The classification task is performed by implementing the *Inception-ResNet-v2* through transfer learning. The model is trained with a subset of real and fake images for 20 epochs. The real images are sampled from the Celeb 100k dataset and the fake images are sampled from the DCGAN generated images. The accuracy graph and the confusion matrix for the model is presented below. The model has resulted in significant high accuracy.

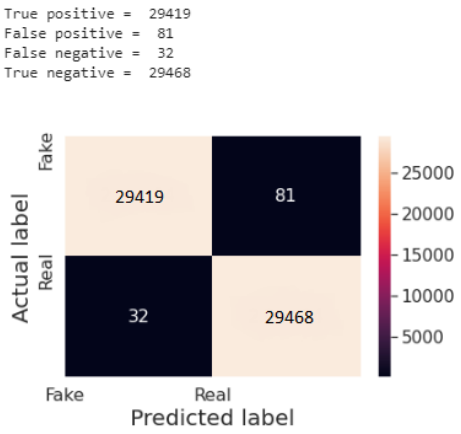

Fig. 10. Confusion Matrix

V. Conclusion

The proliferation of fake images and videos in social networks has been a growing social concern. GAN, the most advanced technique to generate realistic manipulation. This has facilitated anyone to build realistic fake images and videos through open applications and software like FaceApp, ZAO and so on. There is an increasing effort in research work to detect the forgery. Extensive research work has led to several novel detection techniques. However, the evolving modified GANs have been posing challenges to the detection techniques with unseen threats. The proposed model has successfully performed the detection task through implementing transfer learning on Inception-ResNet-v2 utilizing binary cross entropy as the loss function. It resulted in high accuracy in detection tasks. This can be further optimised to detect face manipulation in deepfake videos and segment the doctored regions in the images.

Acknowledgment

I take this opportunity to extend my heartfelt gratitude to the module guides Prof. Dr. Anu Sahni of Machine learning for Data Analytics for guiding and supporting with relevant suggestions and study materials throughout the semester.

References

[1] P. Korshunov and S. Marcel, "DeepFakes: a New Threat to Face Recognition? Assessment and Detection," pp. 1–5, 2018.

[2] Y. Nirkin, Y. Keller, and T. Hassner, "FSGAN: Subject agnostic face swapping and reenactment," *Proc. IEEE Int. Conf. Comput. Vis.*, vol. 2019-Octob, pp. 7183–7192, 2019, doi: 10.1109/ICCV.2019.00728.

[3] Y. Nirkin, I. Masi, A. T. Tuăn, T. Hassner, and G. Medioni, "On face segmentation, face swapping, and face perception," *Proc. - 13th IEEE Int. Conf. Autom. Face Gesture Recognition, FG 2018*, pp. 98–105, 2018, doi: 10.1109/FG.2018.00024.

[4] L. Li, J. Bao, H. Yang, D. Chen, and F. Wen, "FaceShifter: Towards High Fidelity And Occlusion Aware Face Swapping," 2019.

[5] Y. Zhang, L. Zheng, and V. L. L. Thing, "Automated face swapping and its detection," *2017 IEEE 2nd Int. Conf. Signal Image Process. ICSIP 2017*, vol. 2017-Janua, pp. 15–19, 2017, doi: 10.1109/SIPROCESS.2017.8124497.

[6] H. Mo, B. Chen, and W. Luo, "Fake faces identification via convolutional neural network," *IH MMSec 2018 - Proc. 6th ACM Work. Inf. Hiding Multimed. Secur.*, pp. 43–47, 2018, doi: 10.1145/3206004.3206009.

[7] F. Marra, D. Gragnaniello, D. Cozzolino, and L. Verdoliva, "Detection of GAN-Generated Fake Images over Social Networks," *Proc. - IEEE 1st Conf. Multimed. Inf. Process. Retrieval, MIPR 2018*, pp. 384–389, 2018, doi: 10.1109/MIPR.2018.00084.

[8] C. C. Hsu, C. Y. Lee, and Y. X. Zhuang, "Learning to detect fake face images in the wild," *Proc. - 2018 Int. Symp. Comput. Consum. Control. IS3C 2018*, pp. 388–391, 2019, doi: 10.1109/IS3C.2018.00104.

[9] S. Agarwal and L. R. Varshney, "Limits of Deepfake Detection: A Robust Estimation Viewpoint," pp. 1–7, 2019.

[10] C. C. Hsu, Y. X. Zhuang, and C. Y. Lee, "Deep fake image detection based on pairwise learning," *Appl. Sci.*, vol. 10, no. 1, pp. 1–10, 2020, doi: 10.3390/app10010370.

[11] H. H. Nguyen, F. Fang, J. Yamagishi, and I. Echizen, "Multi-task Learning For Detecting and Segmenting Manipulated Facial Images and Videos," 2019.

[12] A. Radford, L. Metz, and S. Chintala, "Unsupervised representation learning with deep convolutional generative adversarial networks," *4th Int. Conf. Learn. Represent. ICLR 2016 - Conf. Track Proc.*, pp. 1–16, 2016.

[13] C. Szegedy, V. Vanhoucke, S. Ioffe, J. Shlens, and Z. Wojna, "Rethinking the Inception Architecture for Computer Vision," *Proc. IEEE Comput. Soc. Conf. Comput. Vis. Pattern Recognit.*, vol. 2016-Decem, pp. 2818–2826, 2016, doi: 10.1109/CVPR.2016.308.

[14] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," *Proc. IEEE Comput. Soc. Conf. Comput. Vis. Pattern Recognit.*, vol. 2016-Decem, pp. 770–778, 2016, doi: 10.1109/CVPR.2016.90.

[15] S. J. Pan and Q. Yang, "A survey on transfer learning," *IEEE Trans. Knowl. Data Eng.*, vol. 22, no. 10, pp. 1345–1359, 2010, doi: 10.1109/TKDE.2009.191.

[16] R. Wirth, "CRISP-DM : Towards a Standard Process Model for Data Mining," *Proc. Fourth Int. Conf. Pract. Appl. Knowl. Discov. Data Min.*, no. 24959, pp. 29–39, 2000, doi: 10.1.1.198.5133.