# School of Informatics


## Informatics Research Review
## DeepFake Identity Swap Detection

[redacted]

**January 2021**

**Abstract**

Identity swapped videos or commonly known as 'deepfake' can be seen constantly in social media. Growing public concern over the integrity of videos that exist on the internet leads to the growing development of deepfake detection. This review paper focuses on the current development of the shallow classifier system and the analysis of existing artifacts. Literature review reveals deepfake evolves to synthesize more realistically as different deep learning architectures evolve, thus resulting variety of detection systems to combat the synthesized videos. The conclusion includes a discussion of a different perspective on verifying videos and summarises the overall aspect of this review.

Date: Friday 22nd January, 2021

**Supervisor:** [redacted]

# Contents



# 1 Introduction

In the age of social media and the fast-paced news circulating around human society everyday, fake news has become a threat to public discourse, the society and democracy[1]. Fake news can be in many form, such as: written-form, visual imagery or even recording. Fabricated information produced to deceived the public spreads quickly through social media as the age of technology creates better access toward the most recent news[2]. Manipulated visual imagery is often produced, as the upsurge of new technology in manipulating images and video has given the public the better access to various of tools, such as the popular DeepFake[1] (commonly used to describe any identity swapped videos or images). Credibility of online news in social media become more vulnerable, thus identity swapped detection is required to distinguish whether the videos or the images are real or fake.

The development of identity swapped detection is constantly racing against the development of implementing identity swapped videos and images. This paper is intended to conclude current development of detection system and analyse the potential weakness and advantages of each system. To further perform the analysis, the information regarding different benchmark dataset and types of detection is included to clarify each methodology. Therefore, the intended audience includes anyone who is interested in current development of identity swapped videos and images and have a background in deep learning and machine learning. This review is structured to provide necessary background overview of identity swapped technology, its development and deeper analysis on shallow identity swapped classifier and social impact of fake videos. The goal of this paper is to increase awareness of this topic and motivate more people to research on forgery detection. The follow questions will be investigated in this paper:

- Why is creating identity swap detection can be useful socially?
- What are the methodologies that are currently implemented to detect Deep Fake?
- Are there any limit in each method? How efficient are these methods?

[1] `https://github.com/deepfakes/faceswap`

We have chosen to primarily focus on shallow identity swapped classifier to render how synthesized videos and images still produced visible artifacts however, constant development of generating those synthesized videos and images are in a fast paced manner. This paper wants to additionally highlight that manner and explore the artifacts as well. Therefore, we select papers based on the analysis of the identity swapped videos and their proposed methodology must at least been tested in mentioned benchmark datasets. Hence, in this paper, we will be exploring a variety of methods of combating with identifying identity swapped videos and images.

# 2 Background

## 2.1 DeepFake

Recent advances in video and audio editing tools, DeepFake [1] is a technology that utilized deep learning to produce extremely convincing face-swapped videos. DeepFake utilizes autoencoder architecture to perform face wrapping, which involves an encoder reducing the face from the image to a lower-dimensional latent space, and a decoder reconstructing the face from the encoder. There are developments by combining the autoencoder architecture with a generative adversarial network (GAN) [3] to improves the initial DeepFake algorithm even further. GAN architecture involves training two models simultaneously: a generative model and a discriminative model. The extended version of DeepFake that utilized GAN[2] suggest further improvement of the original autoencoder architecture. They have utilized the autoencoder as the generator, which learns to produce samples that could not be distinguished from the training data distribution. While the discriminator will access the generated images whether they are real images or manipulated images by providing the loss function to the generator.

DeepFake is often referred to as any identity swapped videos or images. Despite the method to produce might not be using any deep learning, the generated images or videos might utilize simple image processing techniques. In this paper, we will refer to identity swapped videos and images as deepfake, since most of the videos and images we mentioned are deep learning generated.

## 2.2 Benchmark Dataset

To compare and understand the difference in detection systems, benchmark dataset is necessary to evaluate each detection respectively. Due to constant development in DeepFakes, the benchmark datasets are split into two generations. We have categorized according to [9], which they mainly focus on the visual deformation aspect of the generated videos and the quantity of videos for training and testing.

**First Generation:** UADFV [4] is one of the benchmark dataset presented in Fig. 1, which consists of 49 videos extracted from Youtube and generated 49 fake videos swapping all faces with Nicolas Cage's face by using DeepFake[1] with post-processing to extract only the face before performing the face-swapping. In Fig. 1, FaceForensics++ [5] is displayed and the dataset contains 1000 real videos extracted from Youtube. Similar to UADFV, the fake videos are generated using DeepFake[1] and publicly available FaceSwap algorithm[3], which generates face-swapped videos by performing face alignment, Gauss Newton optimization and image blending

[2] `https://github.com/shaoanlu/faceswap-GAN`
[3] https://github.com/MarekKowalski/FaceSwap

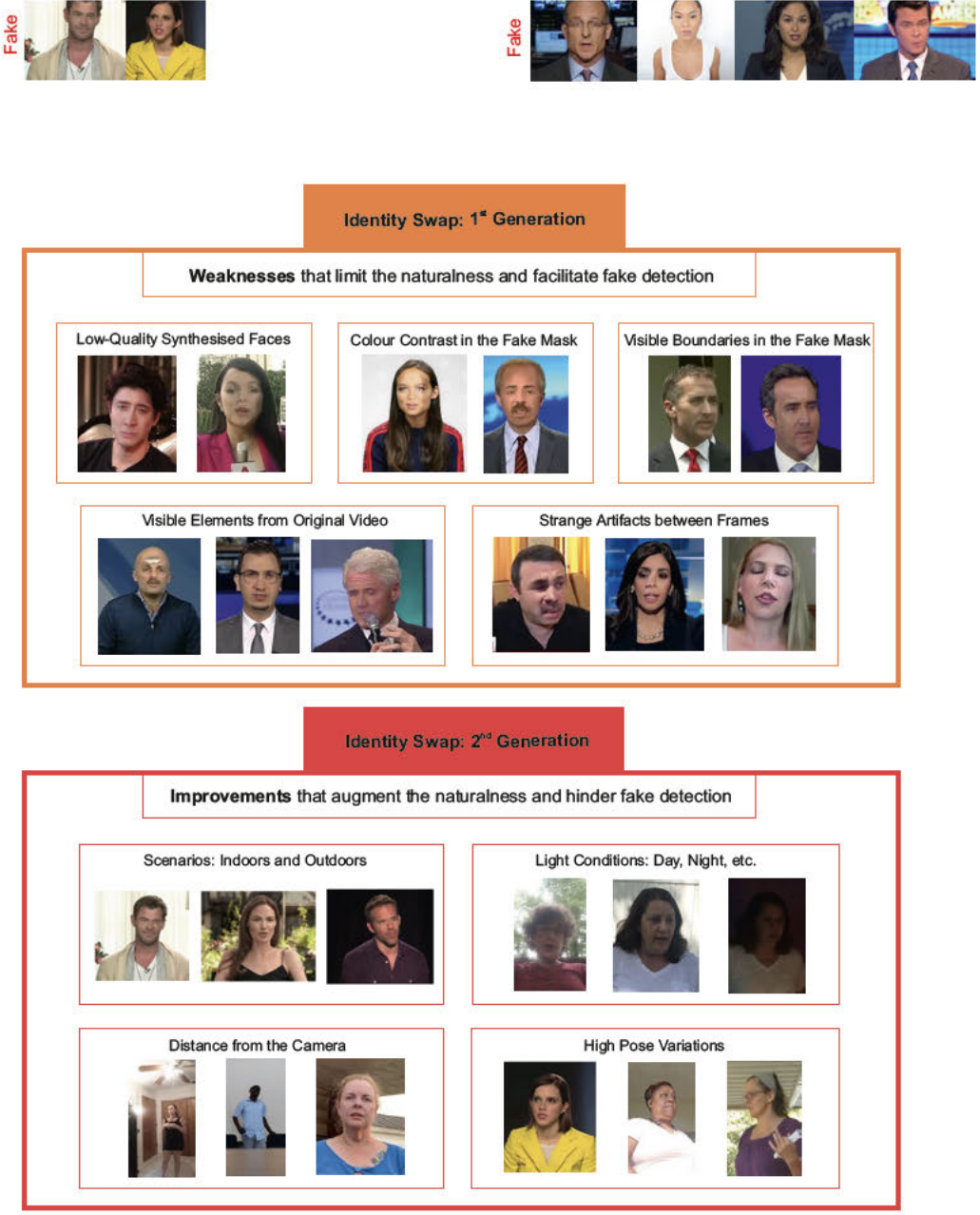
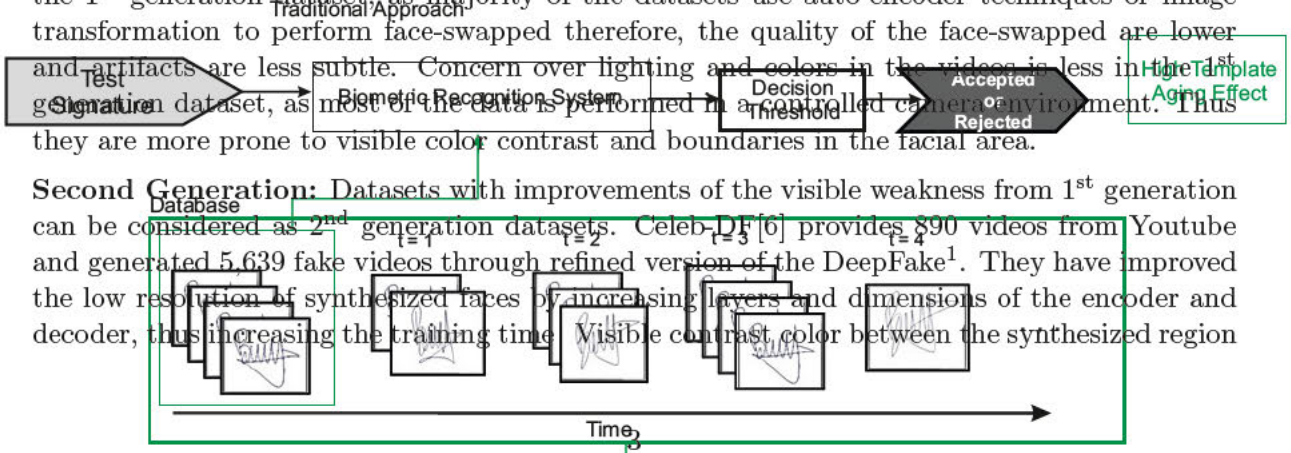



Figure 1: General weakness in 1st generation and improvements in 2nd generation benchmark datasets. Fake images are extracted from: UADFV[4], FaceForensics++[5], Celeb-DF[6], and DFDC[7]. Figure in [8].

to seamlessly swap the face. Datasets from this generation usually contain less than 1,000 videos and contains less variety of identity and poses. In Fig. 1, visual characteristics are more distinct in 1st generation datasets. Low quality synthesised faces are visual characteristic of the 1st generation dataset, as majority of the datasets use auto-encoder techniques or image transformation to perform face-swapped therefore, the quality of the face-swapped are lower and artifacts are less subtle. Concern over lighting and colors in the videos is less in the 1st generation dataset, as most of the data is performed in a controlled camera environment. Thus they are more prone to visible color contrast and boundaries in the facial area.




**Second Generation:** Datasets with improvements of the visible weakness from 1st generation can be considered as 2nd generation datasets. Celeb-DF[6] provides 890 videos from Youtube and generated 5,639 fake videos through refined version of the DeepFake[1]. They have improved the low resolution of synthesized faces by increasing layers and dimensions of the encoder and decoder, thus increasing the training time. Visible contrast color between the synthesized region

and other facial regions is reduced as they perform color augmentation to their training data and applying color transfer algorithm, which allows the deep neural network to generalize better and their generated image to appear more realistic. The boundary between the synthesized region and the source facial region is improved as they increase the surface of the facial area to apply a smoothness mask over. DFDC[7] includes eight facial modification methods to generate deepfakes videos. The dataset consists of 124,000 videos that are independently sourced, unlike other datasets that are extracted from the internet. For the training set, they have included videos generated from DFAE, MM/NN face swap, NTH, and FS-GAN methods, which are further described in [7]. Since generalization of the model is an important aspect for training a deepfake detection, they have included additional methods to generate deepfake in their validation dataset and perform various augmentations, such as geometric, color transforms, and overlays objects, on their generated videos. Not only there are visible improvements within the 2$^{nd}$ generation datasets, but various acquisition scenarios are also included with different lighting conditions and distances, unlike the 1$^{st}$ generation.

### 2.3 Types of Detection

**Temporal Features across Video Frames:** Video manipulation is produced frame-by-frame basis, therefore temporal artifacts across the video produced by face manipulations, should be inconsistent across the generated frames. Observing those temporal artifacts can help to identify whether the video is deepfake or not. The use of eye blinking frequency throughout the video, as the physiological signal, was proposed in [10] based from observing abnormal lower frequency blinking rate in deepfake videos. Blinking rate is calculated by detecting the eye region using face detection and compute the EAR of each eye. The proposed method compared the average EAR to their integrity verification database. Depending on current activity that person participate, their demographic details and psychological related details, blinking rate do vary, therefore the proposed method relies on further detail of the person within the detail, which can be difficult to automatically extract those features from the video. Detection based on observing temporal artifacts in deepfake videos can be less future-proof as constant evolution of deepfake videos solves visible defects or require much more intense infrastructure, like long short term memory (LSTM), to process sequentially. Thus, currently development of deepfake videos focuses more into exploring visual artifacts within frames, rather than finding overall temporal artifacts.

**Visual Artifacts within Video Frames:** Similar to observing temporal features, this approach decomposes videos into frames and explores visual artifacts existing within single frame to extract appropriate features. Those features are then input into either shallow or deep classifier to differentiate between fake and real videos.

For deep classifier, MesoNet[11] proposed a compact facial video forgery detection network. The method analyses frame at mesoscopic level instead of microscopic level, since deepfake videos are commonly compressed and degraded the image quality strongly. The proposed architecture, MesoInception-4, utilized inception modules with multiple convolutional layers with ReLU as the activation function. The model is tested in both 1$^{st}$ and 2$^{nd}$ generation datasets in [12], which the result suggest mesoscopic analysis perform better with 1$^{st}$ generation datasets more than 2$^{nd}$ generation datasets. Other popular deep classifier tested with the benchmark datasets, Xception network[12] utilized CCN architecture inspired by Inception modules, which is replaced with depthwise separable convolutions. The proposed method differ from original Xception[13] by modifying last fully-connected layer, which was designed for ImageNet to classify two classes. The result of Xception shows promising result for both 1$^{st}$ generation and 2$^{nd}$ generation datasets. Despite that, further analysis in [8] suggests there is video quality

dependency, as the proposed method does not perform well in lower quality videos. Lower result in 2nd generation dataset shows the possibility that Xception cannot generalized well when performing in dataset that include more variety of environmental settings. Most deep classifier have encounter generalization problem.

Shallow classifier relies on artifacts or inconsistency of intrinsic features to classifier between real and fake images or videos.[9]. Analysis of artifacts before classifying is required to confirm the proposed hypothesis, thus it requires more manual work than deep classifier. In this paper, we will focus on the shallow classifier in the following section. The proposed method we will be exploring are: head pose inconsistency[14] and GAN fingerprints[15].

# 3 Literature Review

## 3.1 Shallow Classifier

Analysing different manipulation artifacts is required to produce shallow classifier. In this section, we investigate different analysis on visual artifacts that exist in deepfake videos and how each proposed method produce a classifier using the intrinsic features.

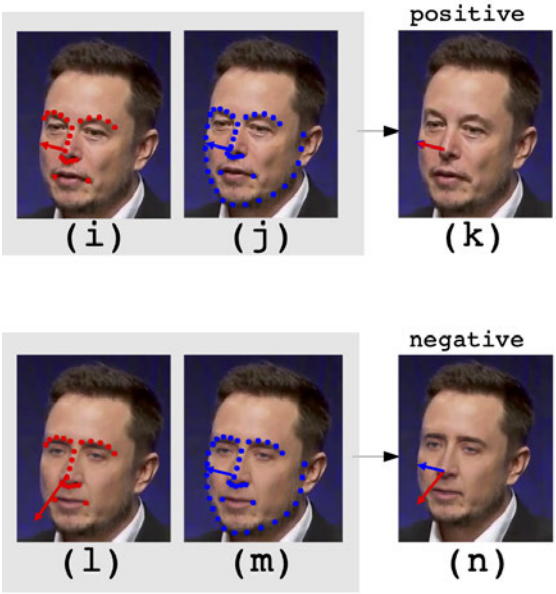



Figure 2: Head pose of authentic image in 1st row and deepfake in 2nd row. Central facial region **(i), (l)** and whole facial region **(j), (m)** are separated by red and blue. [14]

**Head pose inconsistency:** In [14], the proposed method hypothesizes that deepfake often generate an incorrect head pose, visualized in Fig.2, due to the limitation of the affine transformation and the difference in facial structure between the target and the source face, therefore the method can introduce head pose estimation. In preparation, tracking facial landmarks is required to find the whole facial region and the center region. The proposed method utilizes world coordination from the facial landmark to calculate the rotation of the the head.

$$\min_{R,\vec{t},s} \sum_{i=1}^{n} \left\| s \begin{bmatrix} x_i \\ y_i \\ 1 \end{bmatrix} - \begin{bmatrix} f_x & 0 & c_x \\ 0 & f_y & c_y \\ 0 & 0 & 1 \end{bmatrix} \left( R \begin{bmatrix} U_i \\ V_i \\ W_i \end{bmatrix} + \vec{t} \right) \right\|^2 \tag{1}$$

In Eq. 1, the head pose estimation computes by estimating the image coordination of the 2D image, which involves the camera pose $R$ - the rotation of the camera with regards to world coordinate. The camera pose can be reversed $R^T$ to estimate the head pose. The process involves transforming world coordination of the image to camera coordination, which is later transform to the image coordination that the proposed method requires.

To transform the world coordination to camera coordination, the estimation requires certain variables which are: the world coordination $U, V, W$, the camera pose $R$, and translation vector $\vec{t}$. To transform camera coordination to image coordination system, camera's focal length $f_x, f_y$, optical center $c_x, c_y$ and scaling factor $s$ are required. World coordination can be obtained from facial landmark generated from the preprocessing process. Focal length is estimated using the image's width, similar to optical center, which uses the image center. The unknown variables are camera pose $R$, translation vector $\vec{t}$ and scaling factor $s$. Eq. 1 define the optimization problem that can be solved using Levenberg-Marquardt algorithm. Therefore, head pose for both whole facial region and center region can be estimated using the optimization equation and find their difference in the same manner shown in Fig.2.

The estimate head pose for central facial region and whole facial region are the intrinsic features for the SVM classifier. The experiment trains using UADFV[4] dataset and test with UADFV and DARPA GAN dataset. The performance is evaluated frames individually with Area Under ROC (AUROC) as the evaluated metric. The experiment conclude that the classifier can perform highest AUROC at 0.890 with the difference in head pose rotation matrix, represent as flatten Rodrigues' rotation vector, and difference of translation vectors, as the features.

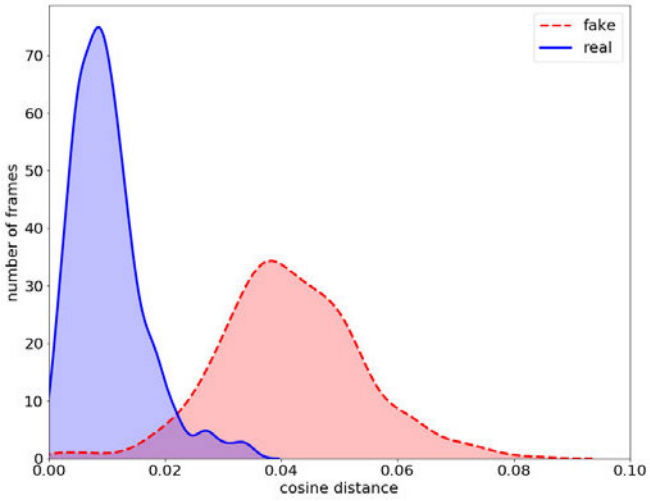



Figure 3: Caption

In [12], the proposed method is evaluated with other both generation benchmark dataset. The review paper have tested with FaceForensic[5], UADFV[4], DFDC[7], and Celeb-DF[6] and evaluated based on AUROC score. The performance for UADFV is exceptionally higher than

another dataset by half. Therefore, we can conclude that head pose inconsistency does not generalized well with different methods of generating deepfake. Constant deepfake evolution can clarify why the proposed method cannot distinguish deepfake videos from other datasets since each dataset is generated differently and may develop better facial region transformation. The training set, the UADFV dataset, is relatively smaller and contains less variety of environmental settings when compared to FaceForenic++ and others. This results from the analysis of the hypothesis to be more biased toward deepfake that are generated in the same manner as the UADFV dataset. In Fig.3, the proposed method[14] confirms their hypothesis on the cosine distance between head pose rotation in real and synthesized videos. However, their analysis can suggest the possibility of classifying between the two classes incorrectly, as there are regions where the cosine distance of the head poses are similar. This method is limited to only the frontal view of the person, which hinders the ability of the method to be implemented in real scenarios.

**GAN fingerprint:** Inspired by the photo-response non-uniformity (PRNU) pattern[16], the proposed method investigates the possibility of deepfake, which are generated using GAN[3], produced any significant pattern or not[15]. In [17], the paper exposes the fingerprint of GAN-produced images in a similar pipeline as how the PRNU pattern is generated. Given the generated image $X_i$ by a given GAN, they have applied denoising filter $f(\cdot)$ to the given image. The fingerprint of the image represents the disturbance in the image, therefore they have subtracted the generated image and the denoised image to generate the fingerprint of the given image or the residual $R_i$ of the image.

$$R_i = X_i - f(X_i) \tag{2}$$

The overall fingerprint of the given GAN is the average of the computed residuals. A number of residuals can affect the quality of the generated fingerprint. They have investigated that 512 residuals will give optimal visible fingerprint for given GAN.

However, in [15], the method to find the fingerprint utilizes deep learning, instead of the PRNU pattern approach. The proposed method trains an autoencoder to map the source image to the generated image of the given GAN. The fingerprint is the reconstruction residual, which computes from finding difference from reconstruction mapping and the source image. Reconstruction mapping is optimized based on pixel-wise loss with adversarial loss from the trained discriminator. The proposed method is trained to find the given image-pair fingerprint and the overall fingerprint of the given GAN.

Both proposed methods distinguished GAN-produced images from real images by computing correlation. In [17], they compute the correlation between the corresponding residual of the image, and overall fingerprint of multiple GAN architecture and different cameras. The most correlated result will be the classification result. Similar to [15], in Fig.4, they perform pixel-wise multiplication the image fingerprint to each model fingerprints and classify the image based on the correlation index. For [17], they have evaluted their method in Forensics GAN Challenge[4] and successfully cluster certain group of GAN-produced images. They did not identify which dataset they have tested on [17], since their objective was to expose GAN's fingerprint existence. Despite that, in [15] have perform evaluation on generated CelebA dataset[18] with various of state-of-art GAN architecture, such as: ProGAN[19]. They have compared their method to [17] using accuracy rate as the evaluating metric. Their method can distinguished real and synthesized images better than the PRNU method by 99.43% accuracy.

Due to rapid development of GAN architecture, the proposed methods will require verification database of each model fingerprints to compute the correlation. Proven in [15], different ini-

[4] `https://www.nist.gov/itl/iad/mig/open-media-forensics-challenge`

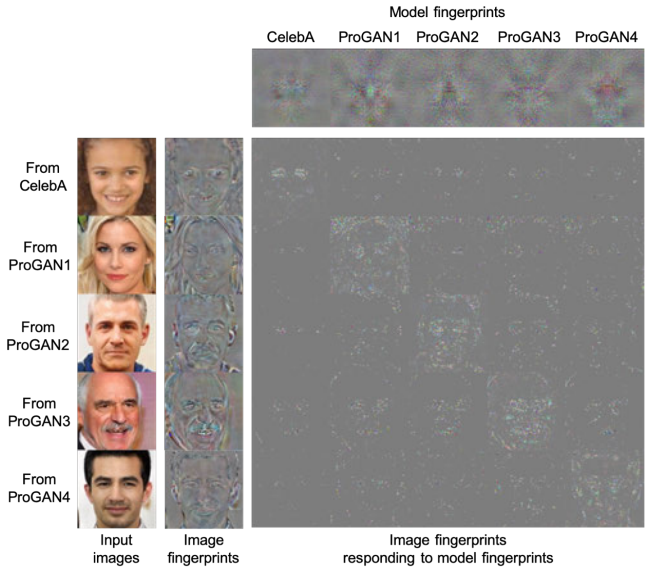



Figure 4: Comparing image fingerprints to each model fingerprints. Each ProGan model's seed is differently initialized. [15]

tialization seeds or other setting in the network will affect the produced model fingerprints. Therefore, it will be difficult to distinguished GAN-generated images apart and update the fingerprint database. Most deepfake videos or images exist in the internet, which often are compressed, losing image detail. Therefore, image's noise and other attribute will exist more and disrupt the existing fingerprint. Removal of GAN fingerprint is introduced by [20] by utilizing autoencoder to reconstruct the fake images or videos with similar fingerprint exist in non-synthesized images. Despite that, the existence of fingerprints suggests microscopic feature of GAN architecture. It is comprehensible that the proposed methods did not tested with any benchmark dataset hence proven to be more difficult to investigate the potential of this intrinsic feature.

Head pose estimation and GAN fingerprint detection are both promising methods to distinguished deepfake from the rest of organic images. They both analyse the intrinsic features that exist in the images or videos. It is difficult to compare both of the proposed methods, however we understand that both methods implemented alone are not efficient enough to distinguished deepfake videos or images. The proposed features can be further evaluated in the future, as they might be useful to combine with other method to produce even more efficient detection system. We want to emphasize the hidden artifacts within deepfake can be noticeable or in image attribute level.

### 3.2 Social Impact on DeepFake

In social media, many have witness the infamous deepfake video of President Obama given a forge speech[21], or various of videos that have the original identity swapped. There are potential harm of deepfake being circulating around the internet that we clearly witness occasionally. In [1], most possible threat of deepfake are usually associated with democracy issue and stability of human society. Misleading information can be harmful for many occasion, such as the election or division within the country. We want to raise awareness of the harmful side of developing these technology. For GAN[3], there are certainly multiple benefit of developing such state-of-

art, but there exists potential misuse of the technology. Identity swap detection will be required to verify what we see everyday in internet.

# 4 Summary & Conclusion

The purpose of this paper is to summarise the current development of identity swap detection system, understanding how each proposed methods are implemented and its potential used. Despite the fact that both proposed methods do not suggest any potential usage, the motion of analysing visual artifacts do emphasized the current development of deepfake generated by autoencoder or combination of autoencoder and GAN architecture[3].

Inconsistent head pose estimation[14] analyses the potential visible visual defect exist in deepfake videos and images. However, when evaluated with other benchmark dataset, the proposed method cannot distinguished the synthesized videos. Therefore, we introduced more microscopic analysis in the image attribute, GAN fingerprint[15]. The method suggests the existence of general GAN model fingerprint in all its generated images and classify by computing the correlation between image fingerprint and model fingerprint. The fingerprint is viewed to represent the disturbance in the image [17], which is similar concept of PRNU pattern[16]. Evaluating the proposed method is difficult due to the difference in testing, as it is strictly GAN-generated images only. Deepfake videos or images can be generated using deep learning without the state-of-art GAN architecture or simple image processing. We conclude that GAN fingerprint analysis and head pose inconsistency cannot be implemented in real scenarios, however the intrinsic features produced from both methods can be further utilized in other identity swapped detection system.

Furthermore, we emphasize on the importance of social impact of deepfake. The spreading of certain deepfake videos have the potential to be harmful [1] to human society and the stability of certain country's democracy. Thus, to conclude, the current state of development of deepfake detection is certainly racing with the deepfake evolution and proves there is a need for reliable detection system in human society to improve our trust in the internet again.

# References


[1] Mika Westerlund. The emergence of deepfake technology: A review. *Technology Innovation Management Review*, 9(11), 2019.

[2] Katie Elson Anderson. Getting acquainted with social networks and apps: combating fake news on social media. *Library Hi Tech News*, 2018.

[3] Ian J. Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. Generative adversarial networks, 2014.

[4] Y. Li, M. Chang, and S. Lyu. In ictu oculi: Exposing ai created fake videos by detecting eye blinking. In *2018 IEEE International Workshop on Information Forensics and Security (WIFS)*, pages 1–7, 2018.

[5] Andreas Rössler, Davide Cozzolino, Luisa Verdoliva, Christian Riess, Justus Thies, and Matthias Nießner. Faceforensics++: Learning to detect manipulated facial images. *CoRR*, abs/1901.08971, 2019.

[6] Yuezun Li, Xin Yang, Pu Sun, Honggang Qi, and Siwei Lyu. Celeb-df: A new dataset for deepfake forensics. *CoRR*, abs/1909.12962, 2019.

[7] Brian Dolhansky, Joanna Bitton, Ben Pflaum, Jikuo Lu, Russ Howes, Menglin Wang, and Cristian Canton Ferrer. The deepfake detection challenge (dfdc) dataset, 2020.

[8] Rubén Tolosana, Rubén Vera-Rodríguez, Julian Fiérrez, Aythami Morales, and Javier Ortega-Garcia. Deepfakes and beyond: A survey of face manipulation and fake detection. *CoRR*, abs/2001.00179, 2020.

[9] Thanh Thi Nguyen, Cuong M. Nguyen, Dung Tien Nguyen, Duc Thanh Nguyen, and Saeid Nahavandi. Deep learning for deepfakes creation and detection. *CoRR*, abs/1909.11573, 2019.

[10] T. Jung, S. Kim, and K. Kim. Deepvision: Deepfakes detection using human eye blinking pattern. *IEEE Access*, 8:83144–83154, 2020.

[11] Darius Afchar, Vincent Nozick, Junichi Yamagishi, and Isao Echizen. Mesonet: a compact facial video forgery detection network. *CoRR*, abs/1809.00888, 2018.

[12] Ruben Tolosana, Sergio Romero-Tapiador, Julian Fierrez, and Ruben Vera-Rodriguez. Deepfakes evolution: Analysis of facial regions and fake detection performance, 2020.

[13] François Chollet. Xception: Deep learning with depthwise separable convolutions. *CoRR*, abs/1610.02357, 2016.

[14] Xin Yang, Yuezun Li, and Siwei Lyu. Exposing deep fakes using inconsistent head poses. *CoRR*, abs/1811.00661, 2018.

[15] Ning Yu, Larry S Davis, and Mario Fritz. Attributing fake images to gans: Learning and analyzing gan fingerprints. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 7556–7566, 2019.

[16] J. Lukas, J. Fridrich, and M. Goljan. Digital camera identification from sensor pattern noise. *IEEE Transactions on Information Forensics and Security*, 1(2):205–214, 2006.

[17] F. Marra, D. Gragnaniello, L. Verdoliva, and G. Poggi. Do gans leave artificial fingerprints? In *2019 IEEE Conference on Multimedia Information Processing and Retrieval (MIPR)*, pages 506–511, 2019.

[18] Ziwei Liu, Ping Luo, Xiaogang Wang, and Xiaoou Tang. Deep learning face attributes in the wild. In *Proceedings of International Conference on Computer Vision (ICCV)*, December 2015.

[19] Tero Karras, Timo Aila, Samuli Laine, and Jaakko Lehtinen. Progressive growing of gans for improved quality, stability, and variation, 2018.

[20] João C. Neves, Rubén Tolosana, Rubén Vera-Rodríguez, Vasco Lopes, and Hugo Proença. Real or fake? spoofing state-of-the-art face synthesis detection systems. *CoRR*, abs/1911.05351, 2019.

[21] Shruti Agarwal, Hany Farid, Yuming Gu, Mingming He, Koki Nagano, and Hao Li. Protecting world leaders against deep fakes. In *CVPR Workshops*, pages 38–45, 2019.