Representation learning based deepfake detection with facial region features

Ng Jen Neng
School of Computing
Asia Pacific University of Technology and Innovation (APU)
Kuala Lumpur, Malaysia
TP027312@mail.apu. edu.my

Chandra Reka Ramachandiran
School of Computing
Asia Pacific University of Technology and Innovation (APU)
Kuala Lumpur, Malaysia
chandra.reka@apu.edu.my

Mandava Rajeswari
School of Computing
Asia Pacific University of Technology and Innovation (APU)
Kuala Lumpur, Malaysia
prof.dr.mandava@staffemail.apu.edu.my

***Abstract*—As the Deepfake phenomena have become famous today, a more stabilize and versatile Generative Adversarial Network (GAN) architect was purposed every single year. Recently, there is some state-of-art GAN architect that able to face-swap, expression-swap, voice- swap, or crafting a new face on the video by providing just a source image and a target video. It has become a real threat for celebrities and world leaders to fear that the high-realistic Deepfake video may use as a way to defame them in one day. The goal of this study is to propose an improved representation learning framework based Deepfake detector to allow the model to discover the representations need for feature detection automatically. In this work, it reviews the latest GAN architect, face manipulation type, and related work of face manipulation detector. Then addressed the fact that overfitting and less generalization happened in most of the face manipulation detectors. Overall, the study will benefit the Deepfake detector in generalizability and detecting the unseen GAN generated images.**

Keywords—DeepFake, GAN, Representation learning

I. INTRODUCTION

Over the years, image processing, convolution neural network, and Generative Adversarial Networks (GAN) have impacted and influenced multiple industries including film making, game development, and photography due to the ability to generate a high-quality photorealistic image of faces. These generated images can easily fool human eyes and casual observers. As the public is getting mature on creating Face-swap tools based on GANs architect. These kinds of Face-swap tools are easily used to produce "DeepFake" media that synthetic a face picture with a target media provided by the user. Due to the enrichment of cloud services, many modern apps or tools can forge the "Deepfake" media with minimum hardware requirements such as smartphones. People are more easily to create fake media and spread over social media. Anyone can use Deepfake to engage in media manipulation and misinformation. It provides a huge challenge to the social media and news industry to identify the distorted reality before the misinformation spread across millions of users. Thus, the misinformation could be a threat to influence election, politics, or plotting to cyberbully.

II. GENERATIVE ADVERSARIAL NETWORK (GAN)

GAN is originally proposed by Goodfellow et al. [1]. It stated that the GAN framework is purposed to create two models: a generative model to recognize the data distribution and a discriminative model to predict the probability that a sample is from the model G or training data. In a nutshell, it is like a minimax two-player game: one generates a fabrication sample base on training data; another one discriminates the sample whether it is a fabrication or genuine until the discriminator model not able to distinguish whether the sample is a fabrication or genuine. In recent research, a lot of different styles of GAN have been purposed and reviewed such as DCGAN in [2], Conditional GAN in [3], StyleGAN in [4], and StackGAN in [5]. The evolution of GAN has grown into multi-branch since the framework was proposed by Goodfellow et al. [1] in 2014. It has learned to produce a photorealistic synthesis image to optimize the image generation model. In the most recent work from Wang et al. [6], they mentioned that StyleGAN [4] and ProGAN

[7] are getting more appearance in generating high resolution, unprecedented quality, and non-existent faces image. It has overcome the bottleneck of low image resolution suffers at deep learning-based image synthesis technique [8].

TABLE I. LITERATURE REVIEW MATRIX

GAN type	Details	Manipulation type	Dataset Sample
GAN [1] (2014)	1) The first purpose of adversarial nets 2) Concept of two players: generative model and discriminant model	-	MNIST TFD CIFAR-10
DCGAN [2] (2015)	1) Use stride layer instead of pooling 2) Use Batch Normalization 3) Use Stochastic Gradient Descent 4) Applied Deduplication	-	LSUN
StackGAN [5] (2017)	1) Text-to-image based framework 2) Similar to Conditional GANs 3) Using Conditioning Augmentation 4) Generate higher resolution (256x256)	-	COCO CUB Oxford-102
CycleGAN [9] (2018)	1) Image-to-image translation 2) Use cycle-consistency loss formation to stabilize GAN training 3) Can perform super-resolution and style transfer	Attribute Editing	CycleGAN
StarGAN [10] (2018)	1) Image-to-image translation 2) Adopt DIAT, CycleGAN, IcGAN as baseline Models 3) Use auxiliary classifier	Attribute Editing	CelebA RaFD

STGAN [11] (2019)	1) Use selective transfer perspective 2) does not take the full target attribute vectors like StarGAN 3) Implement encoder-decoder to map vector from target to source	Attribute Editing, Expression Manipulation	CelebA
StyleGAN [4] (2019)	1) Introduce Mixing Regularization 2) Copying source style 3) Inherit smaller scale facial features 4) Use Stochastic Variation	Expression Manipulation Entire Face Synthesis	Faces-HQ

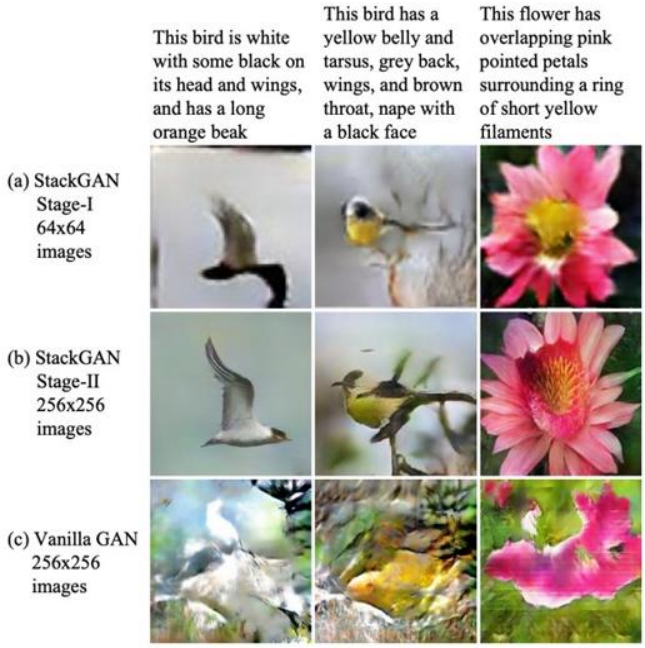

Fig. 1. Text-to-image example by StackGAN [5]

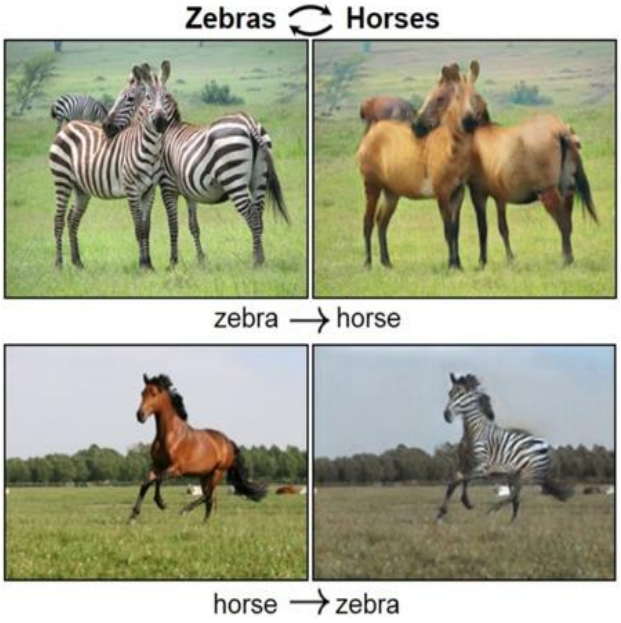

Fig. 2. Image-to-image attribute swap by CycleGAN [9]

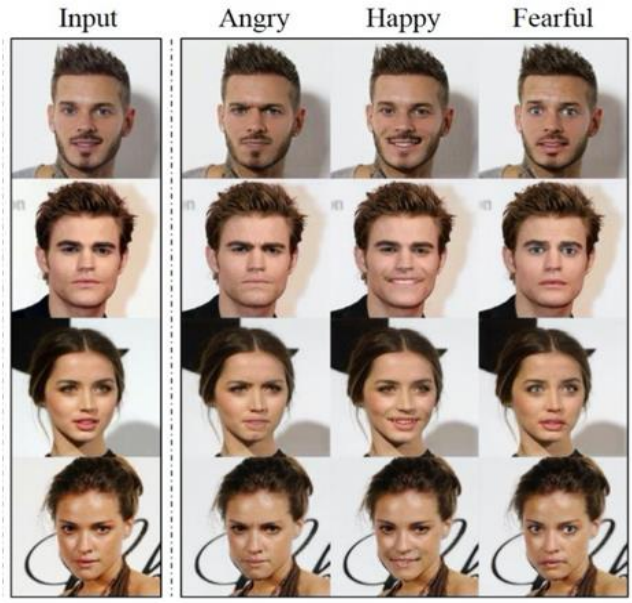

Fig. 3. Image-to-image expression swap by StarGAN [10]

Fig. 4. Entire face synthesis by StyleGAN [4]

III. TYPE OF FACE MANIPULATION

Face manipulations can classify as four main types of group as depicted in Fig 5, there are:

A) Expression Swap

It swaps the facial expression from one person face in a target media with the facial expression from the source media. According to the research from Tang et al. [12], This type of manipulation can apply to the different human race, illumination, background condition, pose, and generate different facial expressions accurately and robustly. The most common example of this technique is using Face2Face and Neural-Textures. This kind of technique is powerful enough to put the words someone never said into their mouth.

B) Attribute Manipulation

It refers to modifying a particular part of facial attributes such as hair color, skin, gender, or adding an accessory. It usually adds some cosmetic, makeup, and glasses. The manipulation is mostly processed by StarGAN framework such as FaceApp mobile application [13].

C) Identity Swap (DeepFake)

This manipulation refers to replace the source face to the target media face while the body and background part remains the same. There are two types of identity swap techniques: 1) Old school technique with computer vision 2) Novel deep learning techniques. These techniques are often found on social media. It could use in the film or game development industry.

D) Entire Face Synthesis

It refers to create non-existent face images by using GAN. The synthesis effect could generate a very high-quality facial image. Ruben et al. has mentioned that the new StyleGAN framework could perform astonishing result with a high level of realism in this synthesis type. It also provides benefits on a video game or 3D-modeling industries to quickly depict character or landscape view.

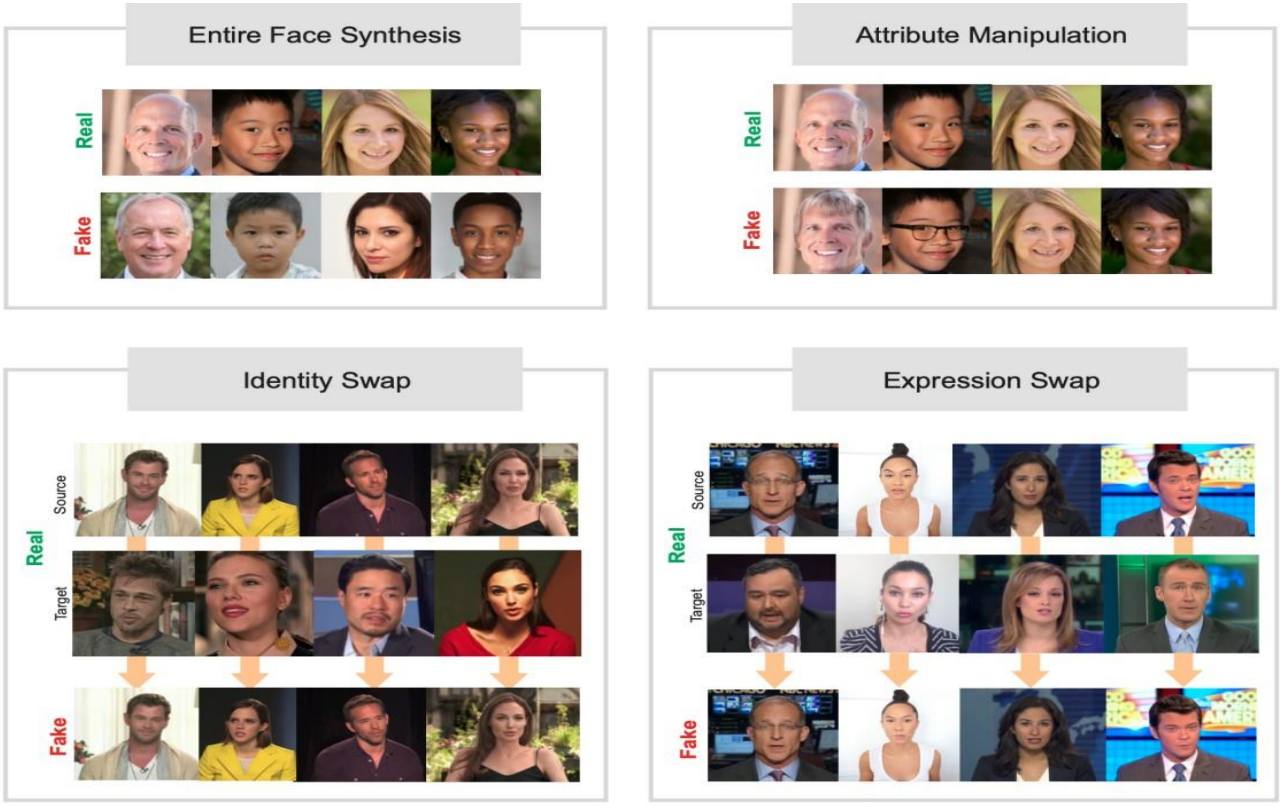

Fig. 5. Real and Fake example for each type of face manipulation [13]

IV. FACE MANIPULATION DETECTION

A. Expression Swap

In a survey carried out by Ruben et al. [13], this manipulation is mostly generated by Face2Face and NeuralTextures. There is a lot of available example dataset in FaceForensics++, where the Face2Face manipulation is transferring the expression from source video to a target video while retaining the same person look. It used to track the first frame obtained identity face with the face in the remaining frame. Then, transferring the facial expression in each frame for the particular faces. Apart from that, Neural Textures is the enhanced version of Face2Face. It applied the GAN-loss technique so that only the mouth is modified based on the facial expression. Ruben et al. [13] also highlighted that these two techniques usually only applied to video. Another approach from Egor et al. [14] only requires a target image and able to produce fantastic results.

According to the survey of Ruben et al. [13], CNN is the most popular classifier for detecting this type of manipulation. By using Deep learning and Mesoscopic based method it can achieve 95+% accuracy. However, because not many researchers are focusing on detect expression swaps, each researcher has proposed a different kind of framework. An approach purpose by Matern et al. [15] is an earlier study looking for the visual feature of a video such as teeth and illumination, but only able to achieve AUC 86%. Another approach purposed by Afchar et al. [16]. A study looking at the mesoscopic and steganalysis analysis called MesoNet. It can take advantage of the raw quality video when analyzing the pixel-level details. A more recent study from Wang et al. [17] has proposed a 3DCNN technique, it analyzes the motion instead of spatial details, but it could only deliver promising results when the determination of its motion consistency. They also mention that the current state-of-the-art 3DCNN method is I3D, 3D ResNet, and 3D ResNeXt to achieve Spatio-temporal modeling. Also, they highlight that expression manipulation is difficult to detect by a human compare to face swap and face synthesis. Apart from that, both Guera and Delp [8], and Sabir et al. [18] have purposed an RNN or temporal-related model to exploit temporal discrepancies across each frame. However, this approach only able to achieve 96+% in FaceForensic++ [19] database DeepFake and FaceSwap methods. The most recent research from Tolosana et al. [20] has used an XceptionNet model to perform discriminant in each facial region, It is well performed in FaceForensics++ and DFDC [21] preview dataset.

B. Attribute Manipulation

This type of manipulation is mostly generated by a GAN-based framework such as Conditional GAN [3] (cGAN), Invertible Conditional GAN [22] (IcGAN). and, StarGAN [10]. cGAN [3] and IcGAN [22] both can edit a complex image. While StarGAN approach provides a promising image-to-image feature with a conditional attribute transfer network to achieve the best synthesis result.

The recent attribute manipulation detection is more focusing on a neuron in each layer as the modification is mostly in a very small region. A neuron behavior monitoring called FakeSpotter purposed by Wang et al. [6] has used the deep face recognition system to determine fake face synthesis. It is a combination model of VGG-Face, FaceNet, and OpenFace to extract the features. Then using SVM to classify real or fake faces. The researchers have used Face Forensic database to train and validate its accuracy.

Another framework proposed by researchers Nataraj et al. [23] has using part of the steganalysis approach to validate the combination of pixel co-occurrence matrices. This framework also uses the same way as StarGAN to train the model with the CelebA dataset. Apart from that, most of the studies also surround on deep learning. Both face patches or complete face training methods can learn and discriminate by deep learning methods. In the complete face training method, Tariq et al. [24] review different CNN pre- trained models such as XceptionNet, VGG16, VGG19, and ResNet. They perform an experiment with ProGAN [7] machine-generated fake images and human edited images by using Adobe Photoshop CS6. The experiment shows that machine-generated fake images are easy to identify (AUC=99%) while human- edited images are slightly harder to identify (AUC = 74%). Another analysis by Dang et al. [25] performed in a different type of facial manipulation with attention mechanisms. This attention mechanism is an approach to improves the feature map. The attention maps can highlight useful regions to improve the accuracy of the classification model.

C. Identity Swap (Deepfake)

Identity swap manipulation is getting more popular recently due to public concern of creating funny synthetic images such as female face swap with male face in a video. In this type of manipulation, most of the fake face from this category are GAN-based face- swapping algorithm. To create identify swap video, researchers Korshunov and Marcel [26] have adopted CycleGAN and the weight of FaceNet on their database. It consists of a Cascaded convolution network to stabilize the face alignment when generating the video. On the other hand, as the FaceSwap algorithm is publicly available, it consists of image blending functionality to swap the face smoothly. Moreover, the FaceForensics++ [19] database also consists of an identity swap image that is generated based on a face detector to crop and align the source image. Then, two autoencoders to reconstruction the source image face to the target face. Apart from that, Facebook has collaborated with other companies and launch a Deepfake Detection Challenge [21]. The competition also provided a Deepfake dataset that consists of 4,119 fake videos generated by two different unknown Deepfake (identity swap) approaches. For current Identity Swap detection, Mika Westerlund [27] noted that the earlier Deepfake techniques are having issues in mimic the person's eye blinks rate. Li et al. [28] proved that AI-generated face is lacking eye blinking since online face portraits on the internet does not consist of close eyes example. Jung et al. [29] agree with that, they have also proposed a technique to detect eye blinking pattern with Fast-HyperFace and Eye-Aspect-Ration(EAR). On the other hand, Yang et al. [30] purposed a method to reveal the 3D head poses, they use a SVM classifier to determine possible mismatched facial landmarks on the eye and mouth region. Another type of detection method from Matern et al. [15] is using signal processing to determine missing reflection or details in the eye and teeth region. However, all the approaches above require to perform landmark or eye region extraction from each frame in the video time series. Thus, Agarwal and Farid [31] suggest using the OpenFace2 toolkit to extract different facial action units such as cheek raiser and mouth stretch.

In more recent research, Li et al. [32] suggest detecting the surrounded face artifact by using a face wrap approach to identify whether a low resolution is present due to Deepfake transformation. Because the Deepfake generated face in the video is mostly in fixed sizes. Moreover, the resolution inconsistency of the facial region with the surrounding region may also be exploited due to the compression of the auto-encoder performed during the image synthesis process. The third method is using deep learning to perform feature extraction by the model itself. This data-driven approach could be somehow effective when capturing specific artifacts. Another approach from Nguyen et al. has proposed a capsule network to route the image into two output capsule (real or fake) from three primary capsules. This approach can combine different classification and identify pose relationship between different region parts through a contract established between capsules. Apart from that, Rossler et al. [19] have evaluated different types of CNN- based with global pooling layer detection, CNN- MesoInception detection, and CNN-XceptionNet detection. They conclude that CNN-XceptionNet is getting the best accuracy in identity swap and face swap manipulation.

D. Entire Face Synthesis

This manipulation refers to create a high-realistic and entire non-existence face. The approach is mostly based on ProGAN [34] and StyleGAN [4]. ProGAN

is a new training framework to improve the training speed and stability while StyleGAN is a generator framework that providing automatic learn and unsupervised separation of face attributes. Moreover, Neves et al. [35] has created a StyleGAN and ProGAN synthesis image in iFakeFaceDB. In this database, it also removes the GAN-fingerprint details by using GANprintR[35]. GANprintR[35] is a GAN- fingerprint Removal autoencoder to remove GAN "fingerprints" present in the synthetic images that may able to spoof the current facial manipulation detection system.

As for the Entire face detection method, the FakeSporter [6] introduced earlier can detect this type of face manipulation type too. A new detection proposed by Guarnera et al. [36] can determine the forensics trace hidden in images. They have leveraged an algorithm called Expectation-Maximization (EM) to find the maximum posterior of latent variable parameters to extract the features and perform discriminant analysis and SVM in the final step. Apart from that, other studies also focused on detecting fingerprints generated in GAN architectures by using deep learning methods. Research from Yu et al. [37] has proposed a similar approach to map the input image to the GAN-fingerprint image. Then, use a

classifier to compare the correlation index of the test image with the synthesis face image. However, this framework doesn't seem too robust enough against noise, blur, and compressed image. To counter this situation, Marra et al. [38] has purposed an incremental learning detection approach to classifying the new type of GAN generated images, noisy images, and unseen images. This approach is based on the XceptionNet model, which can deliver a promising result.

V. DETECTION METHOD GENERALIZABILITY

Hulzebosch et al. [39] have performed different evaluations for state-of-the-art detection such as cross- modal, post-processing, and cross-data. They have highlighted that the current detection method is not robust enough to be used in the real world and images generated by the new GAN framework. Their experiment evaluation on detection method is focusing on:

1) unknown data used to train the generative model (cross-data).
2) unknown type of model used to generate an image (cross-model).
3) unknown type of post-processing technique used to alter an image (post-processing).

They have used ForensicTransfer(FT) and XceptionNet architect to perform the evaluation. The result shows that FT is more robust in cross-modal while XceptionNet is more robust in post-processing. This kind of assessment focuses more on real-world scenarios where one detection approach is not able to perform great in identify multiple types of fake images generated by different GAN frameworks. It proves the importance of generalization is necessary for unseen scenarios. Hulzebosch et al. [39] agreed that most of the fake detection methods will decrease performance substantially when it is exposed to unseen databases. Cozzolino et al. [40] also mention that deep neural network model based-type of fake face detection has proven its effectiveness when it has provided a sufficient number of training samples. However, the underlying neural networks will quickly overfit for manipulation-specific artifacts. They also mention another type of detection is a learning-based approach that having better generalizability. They have proposed a representation learning architect called Forensic Transfer (FT). It is a state-of-the-art fake manipulation detection method. This method is an auto-encoder based architecture that is purposed to act as a form of anomaly detector. Whenever the detector finds out the image cluster is too far away from the real image cluster it will classify it as generated from an unseen method or fake image. In recent research, M. Ye and Y. Guo [41] highlighted that auto-encoder can learn the latent embedding in which the source and target domains are having disjoint label spaces. It is a Zero-Shot Learning scenario where the observer class is not observed during training but needs to predict the category on test scenario. The researcher of Forensic Transfer has proposed to use the Few-Shot Learning scenario to improve generalization with a limited labeled example to classify an image as a pristine or forged image by learning image similarity and applying different augment technique.

TABLE II. LITERATURE REVIEW MATRIX

Reference	Research Topic	Framework	Type	Outcomes	Limitations	Accuracy
Wang et al (2020) [6]	A robust baseline model that could easily spot fake faces	GAN-Pipeline & Neuron monitoring	All	The first neuron coverage for fake detection. Able to achieve high accuracy and low false alarm	Does not consider voice-swap manipulatio n	Acc:84.7% (Mixed)
Nataraj et al (2019) [23]	Use co-occurrence matrices to detect GAN generated images	Deep learning & Steganalysis	Attribute, Entire	Purposed a model by using co-occurrence matrices and able to achieve 99+% accuracy in both StarGAN and cycleGAN database, which is generalizable for both GAN framework	Accuracy drop when using raw image on training and using JPEG images on testing	EER=12.3% (Entire) Acc=99.4% (Attribute)
Tariq et al. (2018) [24]	Evaluate detector accuracy with both machine and human created images	Deep learning	Attribute	Proposed ensemble model with a neural network to detect GAN generated and human modified image. The result shows that the detector getting high accuracy in GAN generated image	Detecting human modified image is getting low accuracy	AUC=99.9% (ProGAN) AUC=74.9% (Photoshop)
Dang et al (2020) [25]	Evaluate the effectiveness of attention mechanism in detecting GAN generated image	Deep learning & Attention mechanism	Attribute, Expression	Purposed an add-on attention-based layer into XceptionNet has increase the detection accuracy significantly. It can highlight the informative regions	Require extra work to collect large-scale of database with numerous types of facial forgeries	AUC=99.9% (DFFD) AUC=99.4% (Face2Face)
Wang et al. (2020) [17]	Evaluate 3DCNN approach in detecting Deepfake image and video	Deep learning & 3DCNN technique with motion	Expression	Purposed an approach of 3DCNN. It is outperform or perform similar to image-based detection	Poor generalisation, accuracy drop when using a different source of GAN dataset in the testing phase	TCR=90.2% (FF+)
Matern et al (2019) [15]	Evaluate exploiting visual artifacts effectiveness	Visual Features	Expression, Identity	Proven that characteristic artifacts can determine face manipulation with	This method only applicable for open eye and visible teeth image	AUC=86% (FF+)

	in detecting Deepfake			little data and low hardware requirement		
Yang et al. (2019) [43]	Evaluate inconsistent head poses effectives in detecting Deepfake	Deep learning & 3D Head Pose	Identity	Proposed a method to reveal 3D head poses of the image and classify using SVM. Able to achieve AUROC of 0.89	This method only applicable for identity swap scenario	AUC=89% (UADFV) AUC=47.3% (FF++)
Guera and Delp (2019) [8]	Evaluate RNN effectives in detecting Deepfake video	Image & Temporal (RNN)	Identity	Proposed a Conv + LSTM method to detect face manipulation video by using only 2 seconds of data. Able to achieve 97% accuracy	The experiment only perform on HOHA datasets and using the same training and testing data source	Acc=97.1% (Mixed)
Sabir et al (2019) [18]	Evaluate RCNN effectives in detecting Deepfake video	Image & Temporal (RNN)	All	Proposed CNN + RNN method to detect face manipulation video with cropping and face alignment. Able to achieve 98% AUC scores for all manipulation type while evaluate using FaceForensic ++ database	Does not evaluate performance using multiple datasets and different compression format	AUC=96% (FF++)
Korshunov and Marcel (2018) [26]	Exploiting the audio-visual inconsistence in Deepfake	Audio Visual features	Identity	Proposed audio-visual approach model, but is not able to distinguish Deepfake video by exploiting lip movement and audio speech (high error rate)	Low quality video (64x64) will fail to detect	ERR=3.3% (DeepfakeTIMIT(LQ))
Reference	Research Question	Framework	Type	Outcomes	Limitations	Accuracy
Li et al. (2018) [28]	Exploiting the abnormal eye blinking rate in Deepfake	Eye blinking features	Identity	Proposed Eye blinking detection method with CNN + LSTM, it shows a little accuracy increase on top of baseline CNN model.	Only evaluate in CEW database, which is the same source as training data. Also, easy to bypass the detection by adding a realistic eye blinking effect	-
Jung et al. (2020) [29]	Exploiting the abnormal eye blinking pattern in Deepfake	GAN-Pipeline & Eye blinking features	Identity	Proposed Eye blinking detection method by GAN framework and analysing eye blinking frequency distribution. Able to achieve 87% accuracy on a different type of Deepfake video	Not applicable for people with mental illness and dopamine activity	Acc=87.5% (Mixed)
Li et al. (2019) [32]	Evaluate Face Warping artifacts effectiveness in detecting Deepfake	Deep learning & Face wrap features	Identity	Proposed Face-warping method to distinguish low resolution or blurred region created by GAN framework. Achieve higher accuracy compare to MesoInception model.	Only work on identity swap	AUC=93% (FF++) AUC=64.6% (Celeb-DF)
Nguyen et al. (2019) [33]	Evaluate Capsule Network effectiveness in detecting Deepfake	Deep learning & Dynamic Routing Algorithm Capsule Network	Identity	Proposed Capsule network method with CNN. It shows 92% accuracy when evaluating on FF++ database. The number of parameters is far lower than Xceptionnet (faster training time)	Low accuracy for unseen datasets	AUC=96.9% (FF+) AUC=54.3% (Celeb-DF)
Cozzolino et al. (2018) [40]	Using transfer-learning technique for GAN forgery detection	Deep learning & Auto Encoder representation learning	All	Proposed a new auto-encoder based architect with a transfer learning concept. Able to achieve 85% accuracy for unseen examples and 95% for seen examples	Accuracy drop when tested on video frame with compressed	Acc=90.5% (StyleGAN) (train test different source)
Guarnera et al. (2020) [36]	Detect DeepFake by analysing convolutional traces	GAN-Pipeline & Expectation Maximization	Entire Face	Analysing GAN fingerprint (forensics trace) by extracting Convolutional Traces with unsupervised learning and Expectation-Maximization algorithm. The approach can achieve high accuracy in both StyleGAN and StyleGAN2	Does not evaluate performance in different compression format	Acc=99.81% (Mixed)

Yu et al. (2019) [37]	Analyse GAN generated image fingerprints	GAN-Pipeline	Entire Face	Proposed GAN Fingerprint analysis architect to determine fake images from different frequencies and patch size. Able to achieve high accuracy in all generated image from CelebA and LSUN	Fail to overcome noise and JPEG compression	Acc=99.5% (Mixed)
Marra et al. (2019) [38]	Incremental learning for GAN-generated images detection	Deep learning & Incremental Learning	Entire Face	Inspired by Incremental and Representation Learning, Their purposed method can detect newly purposed StyleGAN generated images by increment the detector capability. Mean it can adapt to a new class without forgetting the old one.	Getting lower accuracy in ProGAN	Acc=99.3% (Mixed)
Neves et al. (2020) [35]	Use GAN generated image fingerprint to improved face manipulation detection	Deep learning	Entire Face	A method to determine GAN fingerprint, it is getting perfect performance when using the same source in the training phase	Poor generalisation, accuracy drop when using a different source of GAN dataset in the testing phase	EER=0.3% (StyleGAN)

VI. DISCUSSION AND CONCLUSION

Nowadays, GAN architecture can generate different "feel and look" images. The most recent GAN framework such as StyleGAN and StarGAN has become the main target for most of the research. Based on the summary of Tables 2 and 3, most of the researchers are focusing on identity swaps (Deep fake), which is because the real-world problem is focusing on protecting the celebrities and world leaders. In terms of technical view, most of the proposed techniques found the sweet spot in using deep learning feature extraction to achieve very high accuracy. By using the recent deep learning model only, the accuracy can be near to 90+% when testing using FaceForensics++ images. Hence, some of the add-on techniques also increase significant accuracies such as Steganalysis (Pixel domain), attention mechanism, LSTM, incremental learning, Expectation-Maximization, GAN-Fingerprint, and representation learning. Steganalysis and GAN- Fingerprint are both similar techniques where it is looking for GAN modification trace. It is easier to determine fake images as the GAN-Fingerprint trace has provided absolute fact unless a new technique can cover it up from time-to-time. As for LSTM, it can on apply on video and consume more time on training as it added a recurrent LSTM layer in the model. On the other hand, Incremental learning and Representation learning are both fascinate ideas. It can solve most of the deep learning algorithm overfit problems when the test data are unseen images. Hulzebosch et al. (2020) has highlighted the best point where the importance of robustness and real- world scenario for face manipulation detector. Every single year it will be a lot of GAN-based synthesis framework proposed by a different researcher. It is difficult for an industrial face manipulation detector to keep increasing its detection capability by using more ensemble models or adding more training data over time. From this point of view, a face manipulation detector with an "automatic learning" attribute is most suitable for this research to find out. Moreover, Forensic Transfer has proven that the auto-encoder based framework can automatically discover the Deepfake feature and increase the generalizability of Deepfake detection. Therefore, the direction is getting clear that the auto-encoder network can be customized and further improved to increase more generalizability.

As for the researchers that perform facial and head features such as Li et al. [32], Wang et al. [17], Yang et al. [42], and Matern et al. [15], their method does not perform as good as the deep learning-based approach. It may because the new pre-trained model on CNN is getting more efficient and accurate. Performing manual feature extract is inefficient as different GAN framework can generate a different kind of face manipulation images. Furthermore, the latest GAN framework generated image is so realistic that determining facial and head artifact does not provide enough evidence to discriminant fake images. Moreover, the facial region extracted may be difficult to identify by classifier due to low resolution and caused huge accuracy drop. In that case, SRGAN [43] can help to enhance the image with super-resolution ability. It will be a good approach to experiment whether it can solve the limitation of the recent Deepfake detector by providing a super-resolution face region image as input instead of providing the whole face image. It could also solve the accuracy drop when using a frame image from a huge compression video.

REFERENCES

[1] I. J. Goodfellow et al., "Generative adversarial nets," Adv. Neural Inf. Process. Syst., vol. 3, no. January, pp. 2672– 2680, 2014.

[2] 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.

[3] M. Mirza and S. Osindero, "Conditional Generative Adversarial Nets," pp. 1–7, 2014.

[4] T. Karras, S. Laine, and T. Aila, "A style-based generator architecture for generative adversarial networks," Proc. IEEE Comput. Soc. Conf. Comput. Vis. Pattern Recognit., vol. 2019-June, pp. 4396–4405, 2019, doi: 10.1109/CVPR.2019.00453.

[5] H. Zhang et al., "StackGAN: Text to Photo-Realistic Image Synthesis with Stacked Generative Adversarial Networks," in Proceedings of the IEEE International Conference on Computer Vision, 2017, vol. 2017-Octob, doi: 10.1109/ICCV.2017.629.

[6] R. Wang et al., "FakeSpotter: A Simple yet Robust Baseline for Spotting AI-Synthesized Fake Faces," pp. 3444–3451, 2020, doi: 10.24963/ijcai.2020/476.

[7] T. Karras, T. Aila, S. Laine, and J. Lehtinen, "Progressive growing of GANs for improved quality, stability, and variation," 6th Int. Conf. Learn. Represent. ICLR 2018 - Conf. Track Proc., pp. 1–26, 2018.

[8] D. Guera and E. J. Delp, "Deepfake Video Detection Using Recurrent Neural Networks," Proc. AVSS 2018 - 2018 15th IEEE Int. Conf. Adv. Video Signal-Based Surveill., 2019, doi: 10.1109/AVSS.2018.8639163.

[9] J.-Y. Zhu, T. Park, P. Isola, A. A. Efros, and B. A. Research, "Unpaired Image-to-Image Translation using Cycle-Consistent Adversarial Networks Monet Photos."

[10] Y. Choi, M. Choi, M. Kim, J. W. Ha, S. Kim, and J. Choo, "StarGAN: Unified Generative Adversarial Networks for Multi-domain Image-to-Image Translation," Proc. IEEE Comput. Soc. Conf. Comput. Vis. Pattern Recognit., pp. 8789–8797, 2018, doi: 10.1109/CVPR.2018.00916.

[11] M. Liu et al., "STGAN: A unified selective transfer network for arbitrary image attribute editing," Proc. IEEE Comput. Soc. Conf. Comput. Vis. Pattern Recognit., vol. 2019-June, pp. 3668–3677, 2019, doi: 10.1109/CVPR.2019.00379.

[12] H. Tang et al., "Expression Conditional Gan for Facial Expression-to-Expression Translation," Proc. - Int. Conf. Image Process. ICIP, vol. 2019-Septe, pp. 4449–4453, 2019, doi: 10.1109/ICIP.2019.8803654.

[13] R. Tolosana, R. Vera-Rodriguez, J. Fierrez, A. Morales, and J. Ortega-Garcia, "DeepFakes and Beyond: A Survey of Face Manipulation and Fake Detection."

[14] E. Zakharov, A. Shysheya, E. Burkov, and V. Lempitsky, "Few-shot adversarial learning of realistic neural talking head models," Proc. IEEE Int. Conf. Comput. Vis., vol. 2019-Octob, pp. 9458–9467, 2019, doi:

[15] 10.1109/ICCV.2019.00955.

[16] F. Matern, C. Riess, and M. Stamminger, "Exploiting visual artifacts to expose deepfakes and face manipulations," Proc. - 2019 IEEE Winter Conf. Appl. Comput. Vis. Work. WACVW 2019, pp. 83–92, 2019, doi: 10.1109/WACVW.2019.00020.

[17] D. Afchar, V. Nozick, J. Yamagishi, and I. Echizen, "MesoNet: A compact facial video forgery detection network," 10th IEEE Int. Work. Inf. Forensics Secur. WIFS 2018, 2019, doi: 10.1109/WIFS.2018.8630761.

[18] Y. Wang, A. Dantcheva, Y. Wang, A. Dantcheva, Y. Wang, and A. Dantcheva, "A video is worth more than 1000 lies . Comparing 3DCNN approaches for detecting deepfakes," 2020.

[19] E. Sabir, J. Cheng, A. Jaiswal, W. AbdAlmageed, I. Masi, and P. Natarajan, "Recurrent Convolutional Strategies for Face Manipulation Detection in Videos," 2019.

[20] A. Rossler, D. Cozzolino, L. Verdoliva, C. Riess, J. Thies, and M. Niessner, "FaceForensics++: Learning to detect manipulated facial images," Proc. IEEE Int. Conf. Comput. Vis., vol. 2019-Octob, pp. 1–11, 2019, doi: 10.1109/ICCV.2019.00009.

[21] R. Tolosana, S. Romero-Tapiador, J. Fierrez, and R. Vera- Rodriguez, "DeepFakes Evolution: Analysis of Facial Regions and Fake Detection Performance," 2020.

[22] B. Dolhansky, R. Howes, B. Pflaum, N. Baram, and C. C. Ferrer, "The Deepfake Detection Challenge (DFDC) Preview Dataset," 2019.

[23] G. Perarnau, J. van de Weijer, B. Raducanu, and J. M. Álvarez, "Invertible Conditional GANs for image editing," no. Figure 1, pp. 1–9, 2016.

[24] L. Nataraj et al., "Detecting GAN generated Fake Images using Co-occurrence Matrices," IS T Int. Symp. Electron. Imaging Sci. Technol., vol. 2019, no. 5, 2019, doi: 10.2352/ISSN.2470-1173.2019.5.MWSF-532.

[25] S. Tariq, S. Lee, H. Kim, Y. Shin, and S. S. Woo, "Detecting both machine and human created fake face images in the wild," Proc. ACM Conf. Comput. Commun. Secur., pp. 81–87, 2018, doi: 10.1145/3267357.3267367.

[26] H. Dang, F. Liu, J. Stehouwer, X. Liu, and A. Jain, "On the Detection of Digital Face Manipulation," 2019.

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

[28] M. Westerlund, "The Emergence of Deepfake Technology: A Review," Technol. Innov. Manag. Rev., vol. 9, no. 11, pp. 39–52, Nov. 2019, doi: 10.22215/timreview/1282.

[29] Y. Li, M. C. Chang, and S. Lyu, "In Ictu Oculi: Exposing AI created fake videos by detecting eye blinking," 10th IEEE Int. Work. Inf. Forensics Secur. WIFS 2018, 2019, doi: 10.1109/WIFS.2018.8630787.

[30] T. Jung, S. Kim, and K. Kim, "DeepVision: Deepfakes Detection Using Human Eye Blinking Pattern," IEEE Access, vol. 8, 2020, doi: 10.1109/ACCESS.2020.2988660.

[31] X. Yang, Y. Li, and S. Lyu, "Exposing Deep Fakes Using Inconsistent Head Poses," in ICASSP 2019 - 2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), 2019, vol. 2019-May, pp. 8261– 8265, doi: 10.1109/ICASSP.2019.8683164.

[32] S. Agarwal, H. Farid, Y. Gu, M. He, K. Nagano, and H. Li, "Protecting world leaders against deep fakes," Cvprw, pp. 38–45, 2019.

[33] Y. Li and S. Lyu, "Exposing DeepFake Videos By Detecting Face Warping Artifacts," 2018.

[34] H. H. Nguyen, J. Yamagishi, and I. Echizen, "Use of a Capsule Network to Detect Fake Images and Videos," 2019.

[35] T. Karras, T. Aila, S. Laine, and J. Lehtinen, "Progressive growing of GANs for improved quality, stability, and variation," in 6th International Conference on Learning Representations, ICLR 2018 - Conference Track Proceedings, 2018.

[36] J. C. Neves, R. Tolosana, R. Vera-Rodriguez, V. Lopes, H. Proença, and J. Fierrez, "GANprintR: Improved Fakes and Evaluation of the State of the Art in Face ManipulationDetection," pp. 1–11, 2019, doi: 10.1109/JSTSP.2020.3007250.

[37] L. Guarnera, O. Giudice, and S. Battiato, "DeepFake Detection by Analyzing Convolutional Traces," 2020.

[38] N. Yu, L. Davis, and M. Fritz, "Attributing fake images to GANs: Learning and analyzing GAN fingerprints," in Proceedings of the IEEE International Conference on Computer Vision, 2019, vol. 2019-October, doi: 10.1109/ICCV.2019.00765.

[39] F. Marra, C. Saltori, G. Boato, and L. Verdoliva, "Incremental learning for the detection and classification of GAN-generated images," 2019 IEEE Int. Work. Inf. Forensics Secur. WIFS 2019, 2019, doi: 10.1109/WIFS47025.2019.9035099.

[40] N. Hulzebosch, S. Ibrahimi, and M. Worring, "Detecting CNN-Generated Facial Images in Real-World Scenarios," 2020.

[41] D. Cozzolino, J. Thies, A. Rössler, C. Riess, M. Nießner, and L. Verdoliva, "ForensicTransfer: Weakly-supervised Domain Adaptation for Forgery Detection," 2018.

[42] M. Ye and Y. Guo, "Zero-shot classification with discriminative semantic representation learning," Proc. - 30th IEEE Conf. Comput. Vis. Pattern Recognition, CVPR 2017, vol. 2017-Janua, pp. 5103–5111, 2017, doi: 10.1109/CVPR.2017.542.

[43] X. Yang, Y. Li, and S. Lyu, "Exposing Deep Fakes Using Inconsistent Head Poses," ICASSP, IEEE Int. Conf. Acoust. Speech Signal Process. - Proc., vol. 2019-May, pp. 8261– 8265, 2019, doi: 10.1109/ICASSP.2019.8683164.

[44] C. Ledig et al., "Photo-realistic single image super- resolution using a generative adversarial network," in Proceedings - 30th IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2017, 2017, vol. 2017- January, doi: 10.1109/CVPR.2017.19.