Detecting Deepfake Videos using Attribution-Based Confidence Metric

Steven Fernandes, Sunny Raj, Rickard Ewetz, Jodh Singh Pannu, Sumit Kumar Jha

Department of Computer Science, Electrical and Computer Engineering
University of Central Florida, Orlando, FL
{steven, sraj, jodh, jha}@cs.ucf.edu; {rickard.ewetz}@ucf.edu;

Eddy Ortiz, Iustina Vintila, Margaret Salter

Solution Acceleration and Innovation Department
Royal Bank of Canada
{eddy.ortiz, iustina.vintila, margaret.salter}@rbc.com

Abstract

Recent advances in generative adversarial networks have made detecting fake videos a challenging task. In this paper, we propose the application of the state-of-the-art attribution based confidence (ABC) metric for detecting deepfake videos. The ABC metric does not require access to the training data or training the calibration model on the validation data. The ABC metric can be used to draw inferences even when only the trained model is available. Here, we utilize the ABC metric to characterize whether a video is original or fake. The deep learning model is trained only on original videos. The ABC metric uses the trained model to generate confidence values. For, original videos, the confidence values are greater than 0.94.

1. Introduction

Deepfake automatically manipulates the face in a video using a pre-trained generative adversarial network (GAN) to generate fake videos. Deepfake videos are mainly used to manipulate political opinions and create pornographic videos. The most famous among all deepfake video applications is Snapchat application [3]. This application uses an active three-dimensional (3D) model to swap faces in real-time and generate deepfake videos and images. Furthermore, Zao, FaceApp were built on trained deep learning models. The fundamental principle of all these applications is to manipulate the human face, characteristics, such as facial attributes [52]. Face synthesis creates entire non-existent faces using powerful GANs [12].

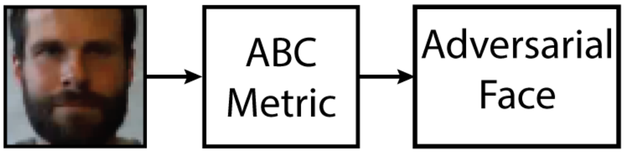


Figure 1: Detecting deepfake videos using ABC metric.

In this study, we propose a novel approach to detect deepfake videos using the state-of-the-art attribution based confidence (ABC) metric [19] as shown in Fig. 1. The ABC metric does not require access to the training data or training the calibration model on the validation data. We train VGGFace2 [5] on ResNet50 [15] model on original face videos. The test face image is provided to the ABC metric, which uses the trained model to generate the attribution score. A threshold value of 0.94 is set to differentiate original and fake videos.

The key contributions of this study are listed below.

- A new database of deepfake videos was created using a commercial website [1] by considering 10 original videos and 10 donor videos. The deepfake database will be made publicly available for research.
- The ABC metric can be used to draw inferences even when only the trained model is available. Hence, it can be used to detect any fake face image without needing to access the training data.

To the best of our knowledge, we are the first to use the state-of-the-art ABC metric [19] to detect deepfake videos without accessing training data

2. Related Work

Fake videos can be mainly generated by manipulating faces in the videos, digital forensics, and face swap.

2.1. Manipulating Faces in Videos

Rossler et al. [43, 44] proposed manipulating facial expressions using a computer graphics approach to transfer the expression from a source video to a target video. The extended version of this approach was presented in FaceForensics++ [44], which is based on natural textures [50]. This technique utilizes original video data to learn the neural texture of the target person using network rendering. Researchers have tried to detect facial expression manipulations using color features [35]. However, this approach was inefficient on most deepfake datasets [26, 43, 29, 61, 8].

Yu et al. [58], proposed a formulation using a learning model based on an attribution network architecture, which maps the image to its respective fingerprint and uses GAN to detect fake images. Furthermore, the learning model learns the fingerprint from each source image to establish a correlation index between each model fingerprint. Wang et al. [53], captured salient features of every layer during activation. These features are important for facial manipulation detection systems [37]. Stehouwer et al. [46], proposed attention mechanisms to enhance obtained convolutional neural network (CNN) feature maps to analyze different types of facial manipulations.

Jain et al. [18], proposed the CNN model inspired by ResNet architecture. They employed support vector machine (SVM) to differentiate original and adversarial images. Wang et al. [54], used recurrent neural networks (RNNs) for the detection of face manipulations. They performed face syntheses using the Face-Aware Liquify tool of Adobe Photoshop to manipulate 50 photographs. Zhan et al [59], performed an analysis on the spectrum domain instead of considering raw pixel information. They applied two-dimensional discrete Fourier transform to each RGB image channel to attain one frequency image per channel. This approach performed significantly better compared with GauGAN [17].

Karras et al. [23] extended ProGAN [22] to StyleGAN to generate high-resolution face images. They proposed an intuitive, scale-specific control of synthesis based on the separation of high-level attributes. Facial attributes, including the color of the skin, hair, age, anger, smile, and gender, can also be modified using [11]. StarGAN [6]. The proponents of StarGAN performed an image-to-image translation using a transfer learning network. They trained conditional attributes using cyclic consistency and attribute classification loss. Lample et al. [27] proposed the IcGAN approach using an autoencoder architecture to generate fake images by disentangling salient image information during the training process.

2.2. Digital Forensics

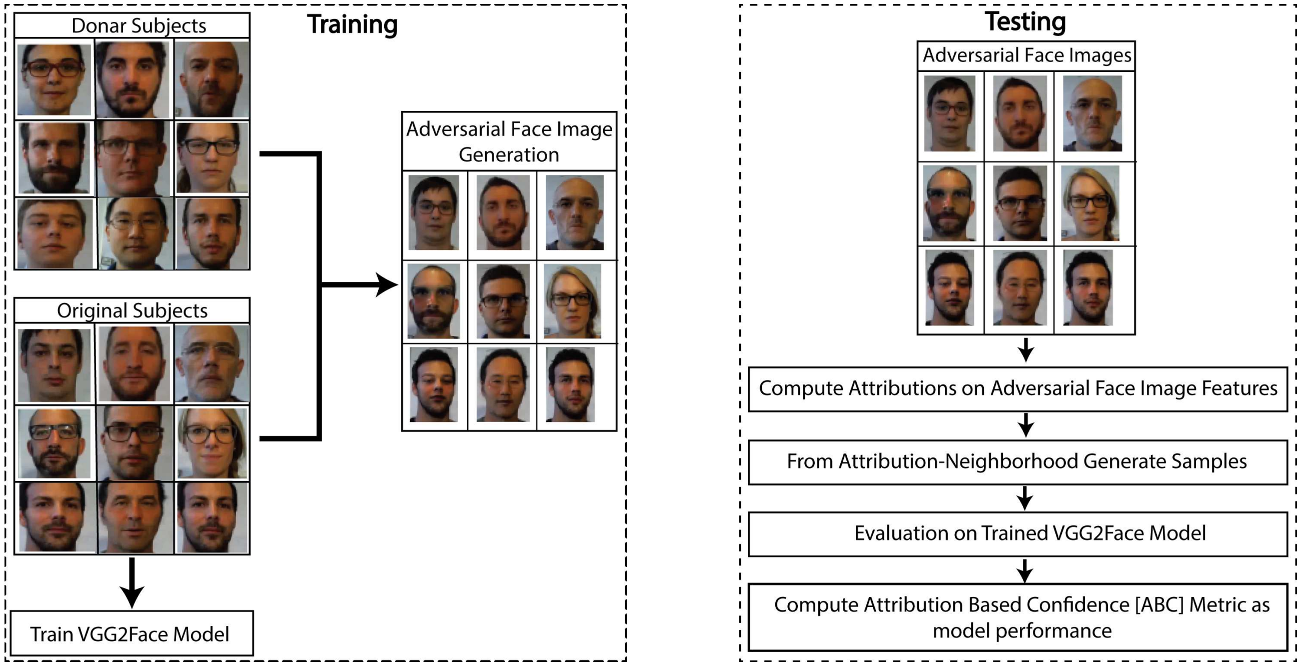
Face recognition systems can also be attacked by cosmetics, makeup, occlusions, and plastic surgery [33, 41]. Bharati et al. [4] employed a restricted Boltzmann machine (RBM) to detect face manipulations using cosmetics. They considered face patches to learn salient features and classify original and adversarial face images. Tariq et al. [49], evaluated the use of their CNN models to detect manipulation using images from the CelebA dataset. Other digital forensics techniques include iCaRL [42], ProGAN [22], StarGAN [6], CycleGAN [61], StyleGAN [23], Glow [25], pixel co-occurrence analysis [36], and Face2Face [51].

2.3. Face Swap

The first face swap detection was proposed by Zhou et al. [60]. They used GoogLeNet [48] to detect fake face images and SVM for classification. Afchar et al. [2] improved the approach using the Inception model [48]. Guera et al. [14], introduced a combination of CNNs and RNNs to detect face swaps in videos. Yang et al. [56], used a splicing approach to synthesize face regions in 3D head poses. They learned the differences between head poses using facial landmarks and the central face region to distinguish deepfakes and real videos. Matern et al. [34], proposed face swaps by considering missing reflections, eye color, teeth, and eye tears. These missing attributions were fed to logistic regression and multi-layer perceptron models for classification. Nguyen et al. [38], proposed multi-task learning that simultaneously locates the manipulated region and detects face swaps using autoencoders. The other attempts to detect face swap includes [43, 2, 34, 38]. Recently, He et al. [16] proposed attGAN, which uses attribute level classification to guarantee the correct change of attributes of a generated image. attGAN was enhanced by Liu et al. [30].

In this study, we have generated two deepfake datasets using a commercial website [1] by considering 10 original videos and 10 donor videos from the COHFACE database. This dataset was first used to detect the predicted heart rate variations of deepfake videos [10]. We have also generated a new deepfake dataset by giving the original and donor face videos of celebrities from YouTube to the commercial website [1]. For every deepfake video, a minimum of 300 min of GPU time is purchased on [1]. The loss value obtained for the donor and original videos by [1] were analyzed.

To the best of our knowledge, we are the first to generate two deepfake datasets using a commercial website [1]. The datasets generated will be made publicly available for the research community. Furthermore, we are the first to use the state-of-the-art ABC metric [19] to detect deepfake videos without accessing training data.


Training
Donar Subjects
Original Subjects
Adversarial Face Image Generation
Train VGG2Face Model
Testing
Adversarial Face Images
Compute Attributions on Adversarial Face Image Features
From Attribution-Neighborhood Generate Samples
Evaluation on Trained VGG2Face Model
Compute Attribution Based Confidence [ABC] Metric as model performance

(a) Generating deepfake videos using the commercial website [1] and training VGG2Face-ResNet50 model only on original videos.

(b) Detecting deepfake videos using the state-of-the-art ABC metric [19] without accessing training data.

Figure 2: Block diagram of the proposed approach used to detect deepfake videos using the state-of-the-art ABC metric

3. Attribution-based Confidence (ABC) Metric

In this study, we apply the state-of-the-art ABC metric for detecting deepfake videos as shown in Fig. 2. The ABC metric is highly motivated by the Dual Process Theory [9, 13] and Kahneman's Decomposition [21] which classifies cognition into System 1 and System 2. The traditional deep neural network model uses a bottom-up approach, i.e., System 1. The ABC metric calculation uses the top-down approach, i.e., System 2. System 2 uses the attribution of System 1 to produce new samples in the neighborhood of the original input. Kilbertus et al. [24] investigated basic differences between causal and anti-causal systems, with the findings of causal systems being continuous. Machine learning problems, such as prediction, are anti-causal. The Deepfake face videos generate adversarial face images that are less resilient compared with original videos that generate original face images. The face region is cropped from each of the frames. We assume that the lack of resilience is attributed to the non-occurrence of learning in the causal direction; rather, it occurs in the anti-causal direction. The experimental and theoretical results are provided as evidence to support the claims [19]. Merging anti-causal System 1 of deep learning networks with attribution driven System 2 allowed the calculation of the ABC metric. The ABC metric provides a cognition model that is comparatively much more resilient.

The key contribution of the ABC metric is that it uses attribution over features for the decision making of machine learning models. ABC builds a constructor that can sample the attribution neighborhood of the specified input and observe the validity of the model in the neighborhood. Although learning is an anti-causal process, ABC appends with causal System 2 that reasons the validity of the model.

Computing the ABC metric

ABC metric computation on machine learning models requires accurate sampling in the neighborhood of high-dimensional input data. This problem can be tackled by performing sampling over low-dimensional or output layers of deep neural networks [39, 7] or depending upon manifold-based and topological analysis of the data [20]. However, it is not always possible to have large training data during testing. Hence the ABC metric utilizes axioms on Shapley values [47]. Recent research has shown that in deep learning systems, few features have very high attributions, and they significantly contribute towards the prediction output [47]. Hence, no change can be obtained in the prediction output by sampling the low attribution features. Low attribution reflects the equivariance of the deep learning model with these features. The ABC metric considers the high-attribution features during sampling instead of considering low-attribution features.

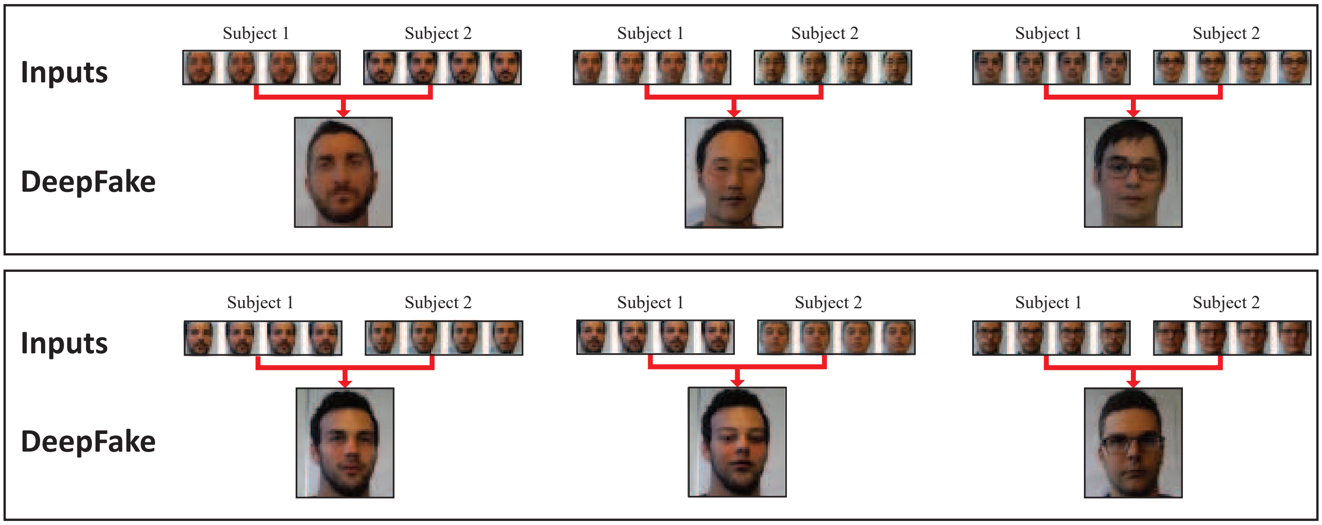


Figure 3: Creating deepfake videos under non-varying head pose movements using the commercial website [1]

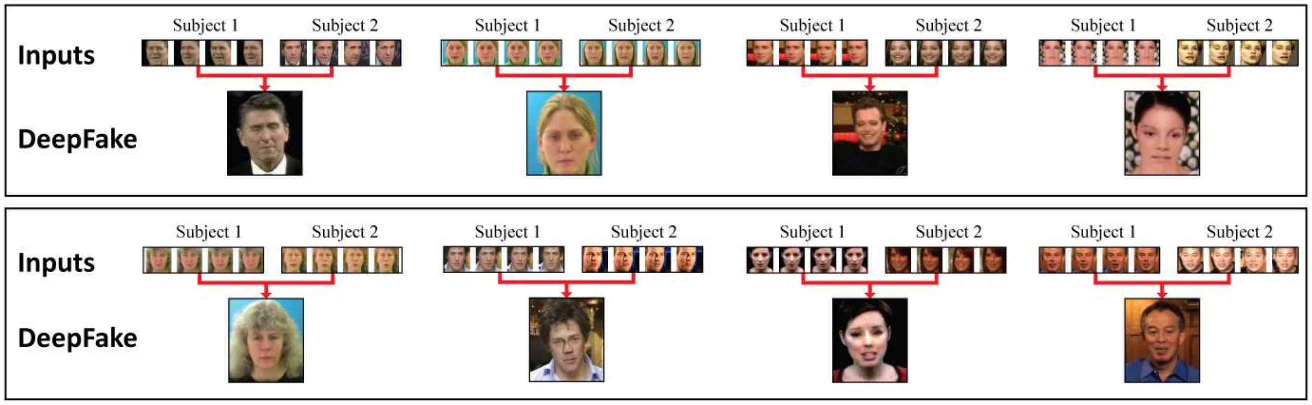


Figure 4: Creating deepfake videos with varying head pose movements using the commercial website [1]

Mathematical formulations of the ABC metric

Consider that the ResNet50 model M trained on VGGFace2 has a input, such that M_i demonstrates the i-th logit output of the ResNet50 model. The attribution of feature a_j of a for label i is calculated as $\mathcal{A}^i_j(\mathrm{a})$. The steps involved in computing the ABC metric are neighborhood sampling and model conformance computation.

- Neighborhood sampling: Select x_j with probability $\frac{|\mathcal{A}^i_j(\mathbf{a})/\mathbf{a}_j|}{\sum_j |\mathcal{A}^i_j(\mathbf{a})/\mathbf{a}_j|}$ and replace it to flip the label from i, i.e., alter the model's decision.

- Conformance computation: Confirm the fraction of the sample in the neighborhood that does not change the ResNet50 model's decision. The original decision is thus conformed as the appropriately obtained confidence value.

The feature attributions are used by the ABC metric to reduce the dimensionality of the input space. On the reduced dimensionality neighborhood, importance sampling is performed in the input to estimate the ResNet50 model's conformance. The traditional dimensionality reduction technique is principal component analysis (PCA). PCA searches the features that are important in the entire image. However, the ABC metric does not search the entire image but identifies the features that are relevant locally for the input. Hence, even in a very high-dimensional input image neighborhood, the ABC metric can appropriately estimate the conformance of the model and effectively compute the confidence. In this study, we have used high-dimensional face images obtained from original and deepfake videos. The deepfake videos are generated for varying/non-varying head poses using the commercial website [1], as shown in Figs. 3 and 4, respectively.

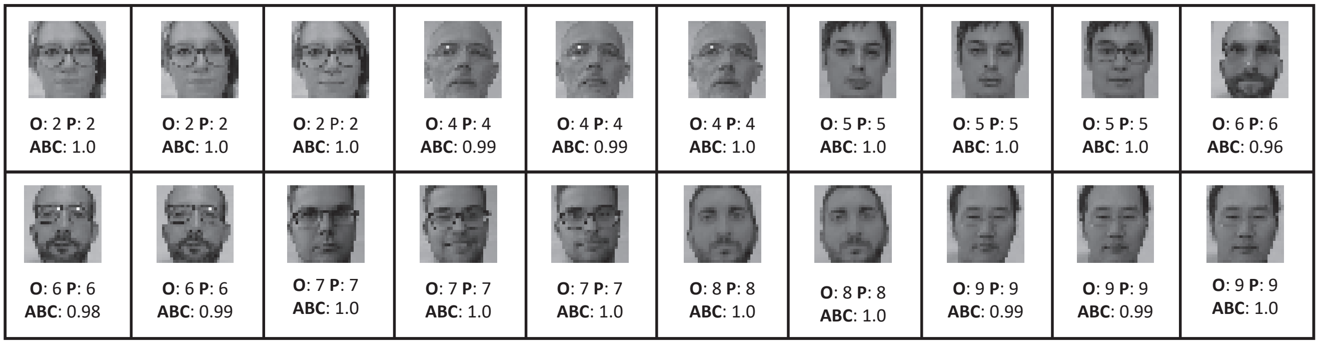


Figure 5: ABC metric values obtained for the original videos from the COHFACE database is greater than 0.94

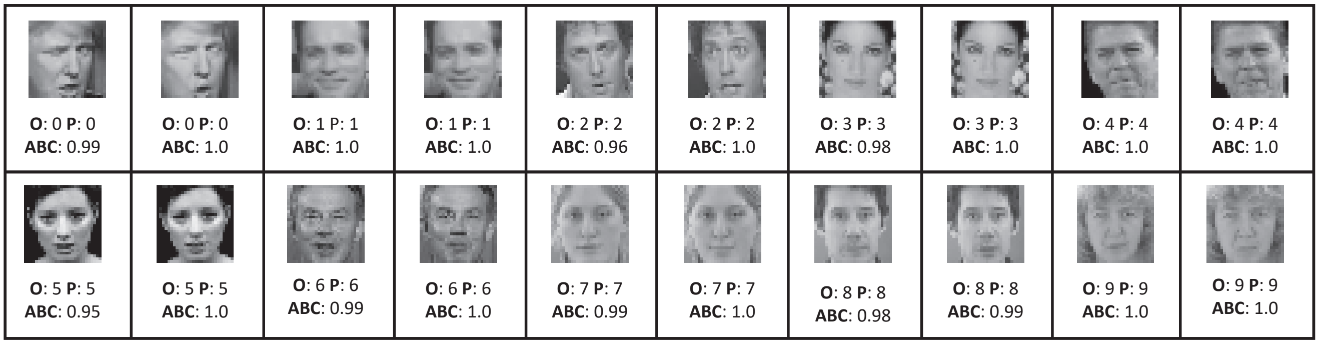


Figure 6: ABC metric values obtained for the original videos from the YouTube database is greater than 0.94

Algorithm 1 ABC metric confidence values $c(M, a)$ of ResNet50 Model M trained on facial input images a

```
Input: ResNet50 Model $M$, Face image $a$ with facial features set $a_1, a_2, a_3, \cdots, a_n$, Sample Size $T$
Output: ABC metric $c(M, a)$
$A_1, ..., A_n \leftarrow$ Attributions of facial features $a_1, a_2, a_3, ..., a_n$ from face input $a$
$i \leftarrow M(a)$ Get ResNet50 model prediction
for $j = 1$ to $n$ do
    $P(a_j) \leftarrow \frac{|A_j/a_j|}{\sum_{k=1}^{n} |A_k/a_k|}$
end for
Generate samples $T$ by mutating facial feature $a_j$ of input $a$ to baseline $a_j^b$ with probability $P(a_j)$
ResNet50 model $(M)$ output on $T$ is obtained.
$c(M, a) \leftarrow T_{conform}/T$ where model $(M)$ ResNet50 output on $T_{conform}$ samples is $i$
return $c(M, a)$ as confidence metric (ABC) of prediction by the ResNet50 model $M$ on face input image $a$
```

ABC metric for detecting deepfakes

The ABC metric presented in [19] is applied to detect deepfake videos. The attribution methods use Shapley values to consider the baseline input of a^b. The baseline can be a completely dark image. It can also contain a set of randomly selected input values, and the computation of expected values is considered an attribution. Suppose the attribution of the j-th feature having i as the output label be as $\mathcal{A}_j^i(\mathbf{a})$. The attribution obtained for the j-th feature does not depend only on a_j but on the entire image a. The deep learning networks considers each logit similarly. Hence, in an ABC metric calculation, the class/logit is dropped. The output of the deep learning network is denoted only using $M(\cdot)$ and the attribution as $A_j(\boldsymbol{a})$. The the proponents of this study [19] have considered the baseline input $\boldsymbol{a}^b = 0$. The steps involved in calculating the ABC metric for detecting deepfake videos is presented in Algorithm 1. We have trained the ResNet50 model on VGGFace2. In Figs. 5, 6, and 7, O indicates the original class label, and P indicates the predicted class label.

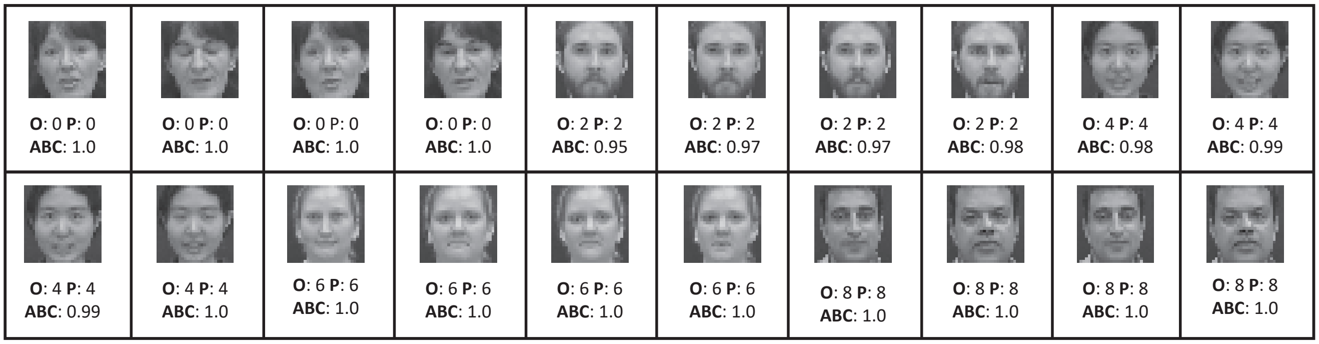


Figure 7: ABC metric values obtained for original videos from VidTIMIT database is greater than 0.94

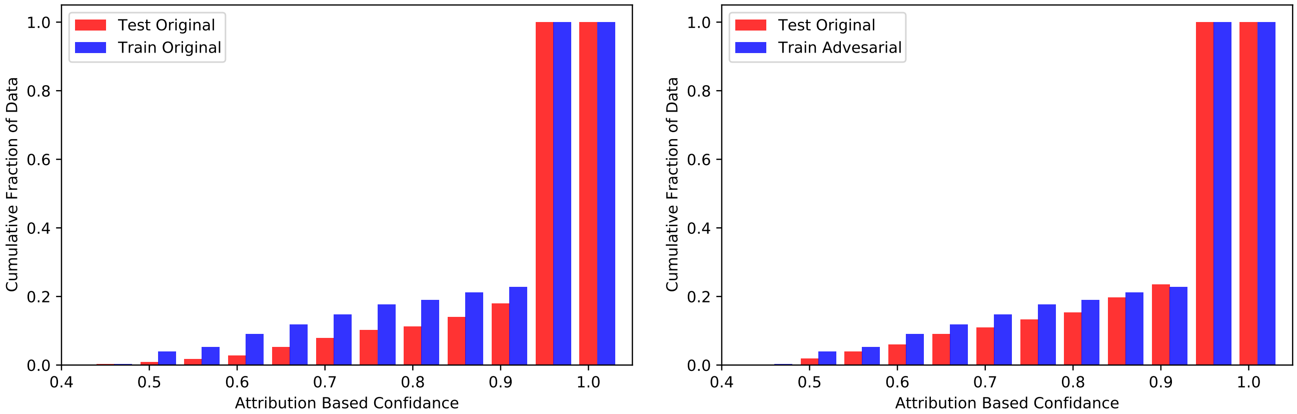


Figure 8: Cumulative data fraction vs. ABC for the test original video from the COHFACE dataset and test adversarial video compared with the trained original video model

4. Results and Discussions

The existing datasets used for the entire face synthesis include FFHQ [23] , CASIA-WebFace [57], VGGFace2 [5] , CelebA [31], and Face Forensics [43]. The datasets generated for face swapping include VidTIMIT [26], Celeb-DF [29]], and FaceForensics++ [44], which is an extension of FaceForensics [43]. Stehouwer et al. [46] generated a face manipulation database with a collection of 300,000 fake images using ProGAN [22] and named it as the Diverse Fake Face Dataset. Neves et al. [37] developed the Face Synthetic Removal database. This database has a collection of more than 150,000 synthetic face images, which are created using StyleGAN. Other GAN based face databases have been presented in [27, 16, 30, 40, 28, 45, 55]. Most of the existing deepfake datasets are not generated using commercial website [1]. Hence, we have generated two deepfake datasets using the commercial website [1].

For our analysis, we have used three original and three deepfake datasets. Among the three original video datasets, two were obtained from the COHFACE, and VidTIMIT datasets, and the third was obtained from YouTube. We have used videos from the COHFACE dataset and YouTube to generate deepfake videos using a commercial website [1]. The third deepfake videos dataset is generated using GANs [26]. The deepfake video generation commercial website [1] requires us to purchase a minimum of 300 minutes of GPU usage to generate one fake video. Each deepfake video was generated by considering an original video and a donor video. The commercial deepfake video generation website [1] generates realistic fake videos. The videos are converted to frames, and after the face detection is performed, the ResNet50 model, pre-trained on VGGFace2, is trained only on the first 80% of original faces. The trained network is tested using the remaining 20% of the original and deepfake faces obtained by applying a face detection algorithm to deepfake videos.

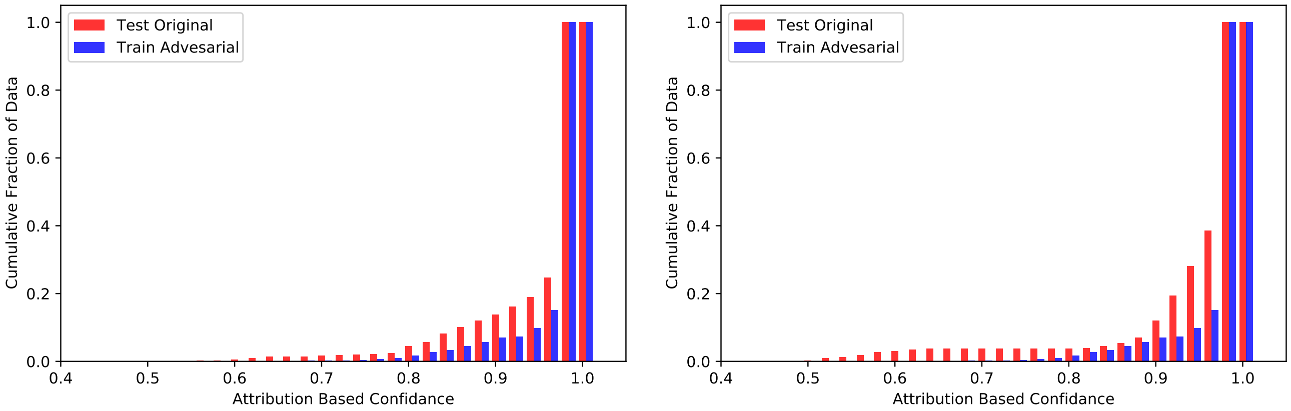


Figure 9: Cumulative data fraction vs. ABC for the test original video from the YouTube dataset and test adversarial video compared with the train original video model

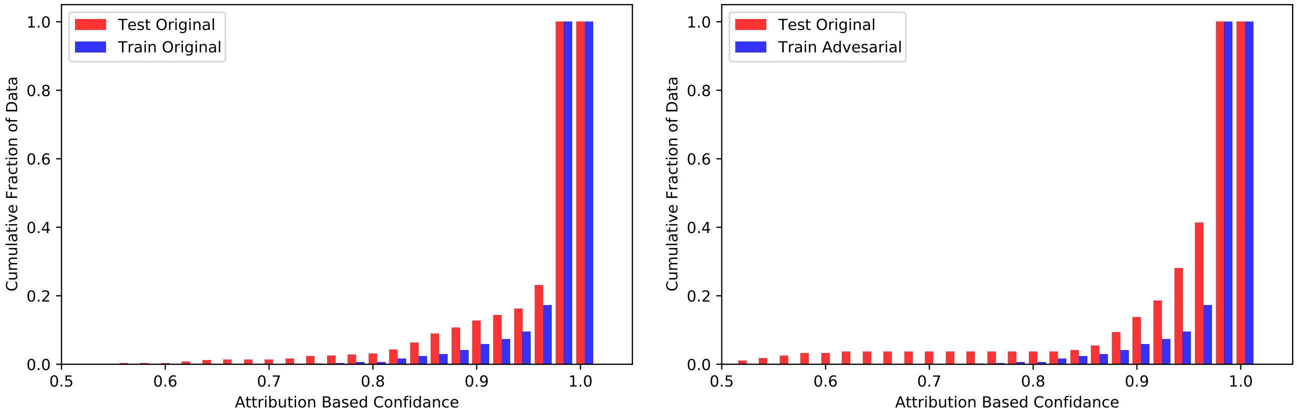


Figure 10: Cumulative data fraction vs. ABC for the test original video from the VidTIMIT dataset and test adversarial video compared with the train original video model

We have applied the state-of-the-art ABC metric for the detection of deepfake videos. The ABC metric is based on two assumptions. First, a linear term dominates the attribution. The same assumption was also made by other attribution methods that use Shapley values [47]. Second, the axiom states that when the difference in the baseline and input is one only feature, but their predictions are different, the difference feature should be assigned an attribution that is non-zero. This axiom was stated in DeepShap [32] and Integrated Gradient [47]. The ABC metric uses the trained model to generate the confidence score. The scores generated by the ABC metric are greater than 0.94 for the original videos.

To validate the results obtained, cumulative data fraction and ABC plots are obtained for COHFACE, YouTube, and VidTIMIT datasets. The cumulative data fraction and ABC for the original test video from the COHFACE dataset and adversarial test video as compared with the train original video model shown in Fig. 8. Fig. 9 shows the plots for the cumulative data fraction and ABC for the original test video from the YouTube dataset and test adversarial video as compared with the train original video model. Fig. 10 shows the plots for the cumulative data fraction and ABC for the original test video from the VidTIMIT dataset and test adversarial video as compared with the train original video model.

5. Conclusion and Future Work

In this study, we apply the state-of-the-art ABC metric to detect deepfake videos. In ABC metric, access to the training data, and training the calibration model on separate validation data are not needed. We have validated our approach on the DeepfakeTIMIT dataset, and two of our deepfake datasets, generated using a commercial website [1]. The loss values obtained from the commercial deepfake website for one of the deepfake datasets are tabulated in Table 1.

Table 1: Loss values obtained for our deepfake dataset

Subject	Original video loss	Donor video loss
1	0.0.01671	0.0.01508
2	0.01507	0.01166
3	0.01405	0.02012
4	0.00925	0.02157
5	0.01326	0.01557
6	0.01524	0.01375
7	0.0172	0.01544
8	0.00748	0.00778
9	0.00988	0.00879
10	0.00704	0.00768

The training loss, and validation loss obtained for the three datasets are tabulated in Table 2.

Table 2: Values for the training loss, and validation loss

Database	Training loss	Validation loss
COHFACE	1.1082	1.1037
YOUTUBE	1.4176	1.4114
VidTIMIT	1.4699	1.4574

In this study, we found that the ABC values obtained for original videos are greater than 0.94. The deepfake videos have low ABC values, so they are easily detected. We obtained the average validation accuracy of greater than 96% on the three datasets.

The significant contributions of our study are listed.

- A new database of deepfake videos was created using a commercial website [1].
- The ABC metric was applied, and a new approach to detect deepfake videos was developed.

To the best of our knowledge, we are the first to generate two deepfake datasets using a commercial website [1] and to use the state-of-the-art ABC metric [19] to detect deepfake videos without accessing training data. In future, we will extract physiological signals from the human face and apply the ABC metric to detect video manipulations.

Acknowledgements.
We acknowledge support from NSF Awards #1822976 and #1422257, an award from the Florida Cybersecurity Center, Royal Bank of Canada, and the Air Force Young Investigator Award to Sumit Jha. Steven Fernandes acknowledges support from the University of Central Florida Preeminent Post-doctoral Fellowship Program.

References

[1] Deepfakes web. https://deepfakesweb.com/. (Accessed on 07/30/2019). 1, 2, 3, 4, 6, 8

[2] Darius Afchar, Vincent Nozick, Junichi Yamagishi, and Isao Echizen. Mesonet: a compact facial video forgery detection network. In *2018 IEEE International Workshop on Information Forensics and Security (WIFS)*, pages 1–7. IEEE, 2018. 2

[3] Akshay Agarwal, Richa Singh, Mayank Vatsa, and Afzel Noore. Swapped! digital face presentation attack detection via weighted local magnitude pattern. In *2017 IEEE International Joint Conference on Biometrics (IJCB)*, pages 659–665. IEEE, 2017. 1

[4] Aparna Bharati, Richa Singh, Mayank Vatsa, and Kevin W Bowyer. Detecting facial retouching using supervised deep learning. *IEEE Transactions on Information Forensics and Security*, 11(9):1903–1913, 2016. 2

[5] Qiong Cao, Li Shen, Weidi Xie, Omkar M Parkhi, and Andrew Zisserman. Vggface2: A dataset for recognising faces across pose and age. In *2018 13th IEEE International Conference on Automatic Face & Gesture Recognition (FG 2018)*, pages 67–74. IEEE, 2018. 1, 6

[6] Yunjey Choi, Minje Choi, Munyoung Kim, Jung-Woo Ha, Sunghun Kim, and Jaegul Choo. Stargan: Unified generative adversarial networks for multi-domain image-to-image translation. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 8789–8797, 2018. 2

[7] John S Denker and Yann Lecun. Transforming neural-net output levels to probability distributions. In *Advances in Neural Information Processing Systems*, pages 853–859, 1991. 3

[8] Brian Dolhansky, Russ Howes, Ben Pflaum, Nicole Baram, and Cristian Canton Ferrer. The deepfake detection challenge (dfdc) preview dataset. *arXiv preprint arXiv:1910.08854*, 2019. 2

[9] Jonathan St BT Evans and Keith Frankish. *In two minds: Dual processes and beyond*, volume 10. Oxford University Press Oxford, 2009. 3

[10] Steven Fernandes, Sunny Raj, Eddy Ortiz, Iustina Vintila, Margaret Salter, Gordana Urosevic, and Sumit Jha. Predicting heart rate variations of deepfake videos using neural ode. In *Proceedings of the IEEE International Conference on Computer Vision Workshops*, pages 0–0, 2019. 2

[11] Ester Gonzalez-Sosa, Julian Fierrez, Ruben Vera-Rodriguez, and Fernando Alonso-Fernandez. Facial soft biometrics for recognition in the wild: Recent works, annotation, and cots evaluation. *IEEE Transactions on Information Forensics and Security*, 13(8):2001–2014, 2018. 2

[12] Ian Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. Generative adversarial nets. In *Advances in neural information processing systems*, pages 2672–2680, 2014. 1

[13] Philip M Groves and Richard F Thompson. Habituation: a dual-process theory. *Psychological review*, 77(5):419, 1970. 3

[14] David Güera and Edward J Delp. Deepfake video detection using recurrent neural networks. In *2018 15th IEEE International Conference on Advanced Video and Signal Based Surveillance (AVSS)*, pages 1–6. IEEE, 2018. 2

[15] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 770–778, 2016. 1

[16] Zhenliang He, Wangmeng Zuo, Meina Kan, Shiguang Shan, and Xilin Chen. Attgan: Facial attribute editing by only changing what you want. *IEEE Transactions on Image Processing*, 28(11):5464–5478, 2019. 2, 6

[17] Phillip Isola, Jun-Yan Zhu, Tinghui Zhou, and Alexei A Efros. Image-to-image translation with conditional adversarial networks. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 1125–1134, 2017. 2

[18] Anubhav Jain, Richa Singh, and Mayank Vatsa. On detecting gans and retouching based synthetic alterations. In *2018 IEEE 9th International Conference on Biometrics Theory, Applications and Systems (BTAS)*, pages 1–7. IEEE, 2018. 2

[19] Susmit Jha, Sunny Raj, Steven Fernandes, Sumit K Jha, Somesh Jha, Brian Jalaian, Gunjan Verma, and Ananthram Swami. Attribution-based confidence metric for deep neural networks. In *Advances in Neural Information Processing Systems*, pages 11826–11837, 2019. 1, 2, 3, 5, 8

[20] Heinrich Jiang, Been Kim, Melody Guan, and Maya Gupta. To trust or not to trust a classifier. In *Advances in Neural Information Processing Systems*, pages 5541–5552, 2018. 3

[21] Daniel Kahneman. *Thinking, fast and slow*. Macmillan, 2011. 3

[22] Tero Karras, Timo Aila, Samuli Laine, and Jaakko Lehtinen. Progressive growing of gans for improved quality, stability, and variation. *arXiv preprint arXiv:1710.10196*, 2017. 2, 6

[23] Tero Karras, Samuli Laine, and Timo Aila. A style-based generator architecture for generative adversarial networks. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 4401–4410, 2019. 2, 6

[24] Niki Kilbertus, Giambattista Parascandolo, and Bernhard Schölkopf. Generalization in anti-causal learning. *arXiv preprint arXiv:1812.00524*, 2018. 3

[25] Durk P Kingma and Prafulla Dhariwal. Glow: Generative flow with invertible 1x1 convolutions. In *Advances in Neural Information Processing Systems*, pages 10215–10224, 2018. 2

[26] Pavel Korshunov and Sébastien Marcel. Deepfakes: a new threat to face recognition? assessment and detection. *arXiv preprint arXiv:1812.08685*, 2018. 2, 6

[27] Guillaume Lample, Neil Zeghidour, Nicolas Usunier, Antoine Bordes, Ludovic Denoyer, and Marc'Aurelio Ranzato. Fader networks: Manipulating images by sliding attributes. In *Advances in Neural Information Processing Systems*, pages 5967–5976, 2017. 2, 6

[28] Mu Li, Wangmeng Zuo, and David Zhang. Deep identity-aware transfer of facial attributes. *arXiv preprint arXiv:1610.05586*, 2016. 6

[29] Yuezun Li, Xin Yang, Pu Sun, Honggang Qi, and Siwei Lyu. Celeb-df: A new dataset for deepfake forensics. *arXiv preprint arXiv:1909.12962*, 2019. 2, 6

[30] Ming Liu, Yukang Ding, Min Xia, Xiao Liu, Errui Ding, Wangmeng Zuo, and Shilei Wen. Stgan: A unified selective transfer network for arbitrary image attribute editing. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 3673–3682, 2019. 2, 6

[31] Ziwei Liu, Ping Luo, Xiaogang Wang, and Xiaoou Tang. Deep learning face attributes in the wild. In *Proceedings of the IEEE international conference on computer vision*, pages 3730–3738, 2015. 6

[32] Scott M Lundberg and Su-In Lee. A unified approach to interpreting model predictions. In *Advances in Neural Information Processing Systems*, pages 4765–4774, 2017. 7

[33] Puspita Majumdar, Akshay Agarwal, Richa Singh, and Mayank Vatsa. Evading face recognition via partial tampering of faces. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops*, pages 0–0, 2019. 2

[34] Falko Matern, Christian Riess, and Marc Stamminger. Exploiting visual artifacts to expose deepfakes and face manipulations. In *2019 IEEE Winter Applications of Computer Vision Workshops (WACVW)*, pages 83–92. IEEE, 2019. 2

[35] Scott McCloskey and Michael Albright. Detecting gan-generated imagery using color cues. *arXiv preprint arXiv:1812.08247*, 2018. 2

[36] Lakshmanan Nataraj, Tajuddin Manhar Mohammed, BS Manjunath, Shivkumar Chandrasekaran, Arjuna Flenner, Jawadul H Bappy, and Amit K Roy-Chowdhury. Detecting gan generated fake images using co-occurrence matrices. *Electronic Imaging*, 2019(5):532–1, 2019. 2

[37] Joao C Neves, Ruben Tolosana, Ruben Vera-Rodriguez, Vasco Lopes, and Hugo Proença. Real or fake? spoofing state-of-the-art face synthesis detection systems. *arXiv preprint arXiv:1911.05351*, 2019. 2, 6

[38] Huy H Nguyen, Fuming Fang, Junichi Yamagishi, and Isao Echizen. Multi-task learning for detecting and segmenting manipulated facial images and videos. *arXiv preprint arXiv:1906.06876*, 2019. 2

[39] Nicolas Papernot and Patrick McDaniel. Deep k-nearest neighbors: Towards confident, interpretable and robust deep learning. *arXiv preprint arXiv:1803.04765*, 2018. 3

[40] Guim Perarnau, Joost Van De Weijer, Bogdan Raducanu, and Jose M Álvarez. Invertible conditional gans for image editing. *arXiv preprint arXiv:1611.06355*, 2016. 6

[41] Christian Rathgeb, Antitza Dantcheva, and Christoph Busch. Impact and detection of facial beautification in face recognition: An overview. *IEEE Access*, 7:152667–152678, 2019. 2

[42] Sylvestre-Alvise Rebuffi, Alexander Kolesnikov, Georg Sperl, and Christoph H Lampert. icarl: Incremental classifier and representation learning. In *Proceedings of the IEEE conference on Computer Vision and Pattern Recognition*, pages 2001–2010, 2017. 2

[43] Andreas Rössler, Davide Cozzolino, Luisa Verdoliva, Christian Riess, Justus Thies, and Matthias Nießner. Faceforensics: A large-scale video dataset for forgery detection in human faces. *arXiv preprint arXiv:1803.09179*, 2018. 2, 6

[44] Andreas Rossler, Davide Cozzolino, Luisa Verdoliva, Christian Riess, Justus Thies, and Matthias Nießner. Faceforensics++: Learning to detect manipulated facial images. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 1–11, 2019. 2, 6

[45] Wei Shen and Rujie Liu. Learning residual images for face attribute manipulation. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 4030–4038, 2017. 6

[46] Joel Stehouwer, Hao Dang, Feng Liu, Xiaoming Liu, and Anil Jain. On the detection of digital face manipulation. *arXiv preprint arXiv:1910.01717*, 2019. 2, 6

[47] Mukund Sundararajan, Ankur Taly, and Qiqi Yan. Axiomatic attribution for deep networks. In *ICML*, pages 3319–3328. JMLR. org, 2017. 3, 7

[48] Christian Szegedy, Wei Liu, Yangqing Jia, Pierre Sermanet, Scott Reed, Dragomir Anguelov, Dumitru Erhan, Vincent Vanhoucke, and Andrew Rabinovich. Going deeper with convolutions. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 1–9, 2015. 2

[49] Shahroz Tariq, Sangyup Lee, Hoyoung Kim, Youjin Shin, and Simon S Woo. Detecting both machine and human created fake face images in the wild. In *Proceedings of the 2nd International Workshop on Multimedia Privacy and Security*, pages 81–87, 2018. 2

[50] Justus Thies, Michael Zollhöfer, and Matthias Nießner. Deferred neural rendering: Image synthesis using neural textures. *ACM Transactions on Graphics (TOG)*, 38(4):1–12, 2019. 2

[51] Justus Thies, Michael Zollhofer, Marc Stamminger, Christian Theobalt, and Matthias Nießner. Face2face: Real-time face capture and reenactment of rgb videos. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 2387–2395, 2016. 2

[52] Ruben Tolosana, Ruben Vera-Rodriguez, Julian Fierrez, Aythami Morales, and Javier Ortega-Garcia. Deepfakes and beyond: A survey of face manipulation and fake detection. *arXiv preprint arXiv:2001.00179*, 2020. 1

[53] Run Wang, Lei Ma, Felix Juefei-Xu, Xiaofei Xie, Jian Wang, and Yang Liu. Fakespotter: A simple baseline for spotting ai-synthesized fake faces. *arXiv preprint arXiv:1909.06122*, 2019. 2

[54] Sheng-Yu Wang, Oliver Wang, Andrew Owens, Richard Zhang, and Alexei A Efros. Detecting photoshopped faces by scripting photoshop. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 10072–10081, 2019. 2

[55] Taihong Xiao, Jiapeng Hong, and Jinwen Ma. Elegant: Exchanging latent encodings with gan for transferring multiple face attributes. In *Proceedings of the European conference on computer vision (ECCV)*, pages 168–184, 2018. 6

[56] Xin Yang, Yuezun Li, and Siwei Lyu. Exposing deep fakes using inconsistent head poses. In *ICASSP 2019-2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 8261–8265. IEEE, 2019. 2

[57] Dong Yi, Zhen Lei, Shengcai Liao, and Stan Z Li. Learning face representation from scratch. *arXiv preprint arXiv:1411.7923*, 2014. 6

[58] 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. 2

[59] Xu Zhang, Svebor Karaman, and Shih-Fu Chang. Detecting and simulating artifacts in gan fake images. *arXiv preprint arXiv:1907.06515*, 2019. 2

[60] Peng Zhou, Xintong Han, Vlad I Morariu, and Larry S Davis. Two-stream neural networks for tampered face detection. In *2017 IEEE Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*, pages 1831–1839. IEEE, 2017. 2

[61] Jun-Yan Zhu, Taesung Park, Phillip Isola, and Alexei A Efros. Unpaired image-to-image translation using cycle-consistent adversarial networks. In *Proceedings of the IEEE international conference on computer vision*, pages 2223–2232, 2017. 2