Detecting Deepfakes

Team members: Sangseok Lee, Shota Takeshima, Jaryl Ngan

Abstract

The volume and quality of fake images and videos, in particular on human faces, generated by neural network based forgery techniques has grown rapidly, resulting in the rampant spread of fakes new and disinformation. In this paper, we will attempt to reproduce some state-of-the-art models, assess the performance of these models and identify their challenges and limitations.

1 Introduction

1.1 Motivation and importance of the problem

Fake news has existed for centuries. However, the rise of social media has resulted in fake news being spread further and wider than ever. Fake news or allegations of fake news are now everywhere.[11] Differentiating between authentic and fake news is often non-trivial, however this task has been made harder with the advent of fake images and videos generated by digital manipulation, in particular, involving facial information. Forgery techniques have gotten so advanced that for a majority of the population it is virtually impossible for them to distinguish between real and fake. These facial manipulation techniques utilize deep learning methods to change the expression on a person's face or even swap the face of one person on to the face of another person. Malicious actors can easily get access to this technology and commit nefarious acts. Some potential threats include fake videos of world leaders making callous comments or company executives releasing information which could send shock waves to the equity markets. The possibilities for abuse are wide-ranging. [12]

Traditionally, media forensics methods were based on factors such as fingerprints from the capturing device and the editing software. [13] However, most of these features are highly specific and are therefore not robust to different conditions. Nowadays, a great volume of media is shared on social media platforms. [14] Most of these platforms automatically modify the media through compression and other types of operations, rendering many of the traditional techniques useless. [15] Although digital forensics experts can still analyze videos for evidence of manipulation, reviewing the droves of information online is impossible. It is therefore necessary to leverage machine learning models with computer vision to create scalable detection methods. However, these models require huge amounts of training data and computation power. Fortunately, over the past few years, interest in this area has been growing rapidly and cloud computation resources have also expanded. Many new datasets have been collected and released to the public, with each release being much bigger than the previous ones. Examples include Faceforensics [16], Celeb-df [17] and DeepFake Detection Challenge (DFDC) dataset [18] just to name a few.

1.2 Types of forgery techniques

Most manipulation techniques fall under two broad categories. The first is identity swapping and the second is facial reenactment. [19] Identity Swapping involves generating a video of the target with the face replaced by a synthesized face of the source. On the other hand, facial reenactments allow the source to change the target expressions while preserving the identity of the target.

The two manipulation techniques that will be used to create the fakes in our dataset are Deepfakes and Neural Textures. Originally, the FaceForensics Dataset [16] that we will be using in this paper

included two other methods. These are Faceswap and Face2Face, however, due to limited computation power and storage space, it was infeasible to include the other two methods.

FaceSwap [20] is an identity swapping technique which transfers the facial region from one image to another. It extracts the facial region by detecting facial landmarks. A 3-D model is then fitted. Using this model, the extracted facial region is projected onto the target image. This is optimized by minimizing the distance between the localized landmarks and the projected shape based on the input image. Finally, the rendered model is blended with the target image.

Figure 1: Faceswapping Ted Cruz's face to fit Donald Trump by Satya Mallick

DeepFakes has become a term widely synonymous with fake facial images, however it is also a specific identity swapping technique. There are various public implementations of DeepFakes available, most notably FakeApp [21]. The method uses two autoencoders with one encoder shared between the source and target images. To create a fake image, the trained encoder and decoder of the source face are applied to the target face. [2]

Figure 2: Nicholas Cage DeepFakes with Amy Adams using FakeApp

Face2Face [22] is a facial reenactment technique. The method uses two video input streams. Manually selected frames are then used to generate a dense reconstruction of the facial images which is then capable of synthesizing the face with different expressions.

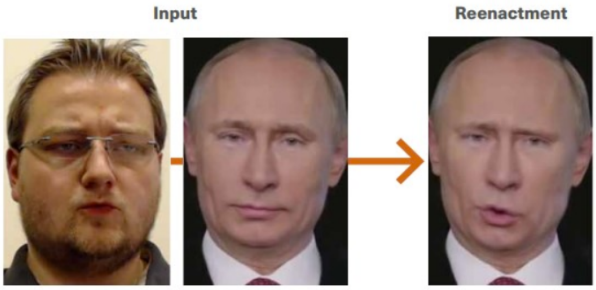

Figure 3: Vladimir Putin pouting using Face2Face

NeuralTextures [23] is another facial reenactment method. It is uses a Generative Adversarial Network to learn from the original image data a neural texture for the target person. [3] The model uses two loss functions, the first is the photo-metric reconstruction loss and the other is an adversarial

loss. Neural textures has the ability to change facial expressions of any facial region, however we concentrate this study to only the mouth region as there is insufficient input the modify the other regions.

Figure 4: Changing Obama's Expressions using NeuralTextures

1.3 Scope of work

Although many effective detection methods based on deep learning have been developed in the past few years, there are still many critical problems that exist which we will discuss in greater depth. The generation of DeepFake videos will continue evolving and improving. As such, detection methods will need to adapt to this ever-changing landscape. This paper will attempt to reproduce some state-of-the-art models, assess the performance of these models and identify their challenges and limitations. [24]

2 Related works

2.1 Human Benchmark

FaceForensics++ [16] carried out a study to evaluate the performance of humans in detecting manipulated images. The participants consisted of 204 participants who were predominantly university computer science students. Participants were given a random time limit of 2,4 or 6 seconds to determine if a randomly selected image from a set of images was authentic or fake. This set of images were equally weighted in terms of real and fake images. The images were also varied in quality and manipulation method.

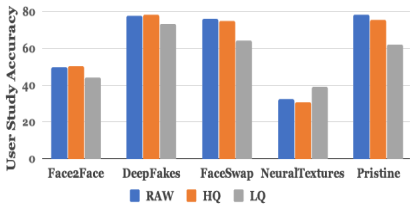

Figure 5: Human Bench on forgery detection by FaceForensics++

The team found that there was a correlation between video quality and participants' ability to detect fakes. The participants managed to get 68.69% in average on raw videos, 66.57% on high quality, and 58.73% on low quality videos. [16] It is worth noting that the partcipants only managed a score of around 40% on compressed Neural Textures manipulated images. It was also found that the differing times to inspect the image did not result in any significant difference in the ability for the candidates to decipher an image's authenticity.

2.2 Traditional Detection Methods

Traditional media detection utilized low-level feature analysis, relying on compression artifacts caused by lossy compression of JPEG files and double image compression. These methods were

improved with Discrete Cosine Transform coefficient analysis. [25] Other traditional methods study the traces left by camera processes such as photo response non-uniformity noise patterns and the camera response function. Statistical properties of natural photographic images and Color Filter Array based solutions have also been successful in the past. [26]

2.3 Convolutional Neural Networks(CNN)

Researchers applying deep learning in the media forensic field have achieved great success. Chen JianSheng utilized convolutional neural networks (CNN) with median filtering in image forensics. [27] Xinyi Ding also utilized CNNs to detect swapped faces.[28] The authors in [9] highlighted that CNNs would only be able to detect manipulations done by techniques that they were trained on. They therefore proposed using a Hierarchical Memory Network (HMN) architecture to combat this problem. HMN considers the contents of the face and previously seen faces. The network encodes the face region and compares it to recently seen encodings.

2.4 Autoencoders

Autoencoders have also been very effective in forgery detection tasks. Zhang et al. [29] used a staked autoencoder to learn complex features for each individual image patch, and then integrated the contextual information of each patch to perform the final results. Cozzolino D and Verdoliva L [30] made use of autoencoders to sieve out implicit information from noisy data, this enabled them to detect anomalies better

2.5 Long short-term memory (LSTM)

Long short-term memory (LSTM) network has been another tool applied in this field. Jawadul H and Amit K. et.al utilized a hybrid of CNNs and LSTM which was able to detect areas in the images that were altered.

2.6 Steganalysis

Development in employing steganalysis rich models (SRM) has also been notable. Cozzolino D and Verdoliva L [30] were able to filter out noisy features using SRM filters. The remaining features are then fed in the CNN. Instead, Peng Zhou et al. [31] made use of a steganalysis feature extractor to pre-process images before inputting it into a CNN.

2.7 Robust Methods

Building a detection system robust to new images is of utmost importance and has been a great challenge. Marra et al. [32] proposed a multi-task incremental learning detection method in order to detect and classify new types of Generative Adversarial Network (GAN) generated images, without worsening the performance on the previous ones. Their proposed detection approach, based on the XceptionNet model, achieved promising results being able to correctly detect new GAN generated images. Attention mechanisms were also applied to further improve the training process of the detection systems. [33]

3 Experimental results

3.1 Computation Resources

For this project, we utilized Google Drive to store the data and had a capacity of 100 GB. We used Google Colab for our computing resources. This meant the computing power was often limited. The notebook would time out when the load was too heavy or after 12-hour intervals. A certain amount of time would have to pass before these resources would then become available again. The computation restrictions derailed our efforts in attempting more models and utilizing more data.

Although Duke Virtual Machines (VM) GPUScavenger had a much higher capacity with around 500 GB storage, after checking with the IT team, we found that it was not possible for all the team

members to access the same VM, thus hindering collaboration. We therefore, ultimately decided to use the Google Suite instead.

3.2 Data

In this project, we use the FaceForensics++ dataset [1] of pristine and manipulated videos. The dataset consists of 1000 original videos from YouTube. We apply two manipulation methods resulting in a total of 2000 manipulated videos. The two methods which were earlier discussed are DeepFakes and Neural Textures. In addition, these 3000 videos are formatted into three different compression rates: raw (uncompressed), c23 (medium compression), c40 (high compression). In this paper, we focus on c40 for two main reasons. First, existing literature has indicated that facial forgery detection becomes more difficult when the video quality is lower. Second, due to the lack of storage and computing resources, it was not feasible to include the other video qualities.

3.3 Preprocessing

Two datasets were used, one contains the original and DeepFake manipulated videos and the other contains the original and Neural Textures manipulated videos. Each of the datasets therefore contains 2000 videos, with 1000 being original and the other 1000 being manipulated. Both the datasets were split into 3 parts, training, validation and test set with an 72-14-14 split. The videos were then cut by frames, with 10 frames from each video so that the input data would be images instead of videos.

We extracted faces from all the frames using Multi-task Cascaded Convolutional Networks(MTCNN developed by Zhang et al.[4]. MTCNN not only detects the face but also five key points (left eye, right eye, nose, mouth left and mouth right) . This method achieves superior accuracy over the state-of-the-art techniques.

3.4 Model

Ultimately, we decided to compare the following three models: Resnet-18, Xception and Inception Resnet V1.

3.4.1 Resnet-18

Resnet-18 utilizes a common convolutional architecture, we therefore use it as the baseline model. This model was pre-trained on ImageNet, a dataset that contains of over 15 million labeled high-resolution images in over 22,000 categories. We then fine-tuned this model by training it on our data. By overcoming the issue of vanishing gradient, Resnet-18 demonstrated that extremely deep networks can be trained using standard SGD through the use of residual modules, as shown in Figure 6.

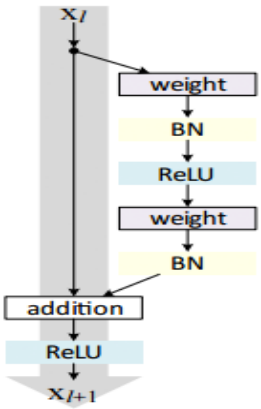

Figure 6: The residual module in ResNet as originally proposed by He et al. in 2015.

3.4.2 Xception

Introduced by Google, Xception [6] stands for Extreme version of Inception-v3 [7]. Xception is an extension of the Inception architecture which replaces the standard Inception modules with depthwise separable convolutions. A complete description of the specifications of the network is given in Figure 7. The Xception architecture has 36 convolutional layers forming the feature extraction base of the network. In our experimental evaluation we will exclusively investigate image classification and therefore our convolutional base will be followed by a logistic regression layer. The 36 convolutional layers are structured into 14 modules, all of which have linear residual connections around them, except for the first and last modules.

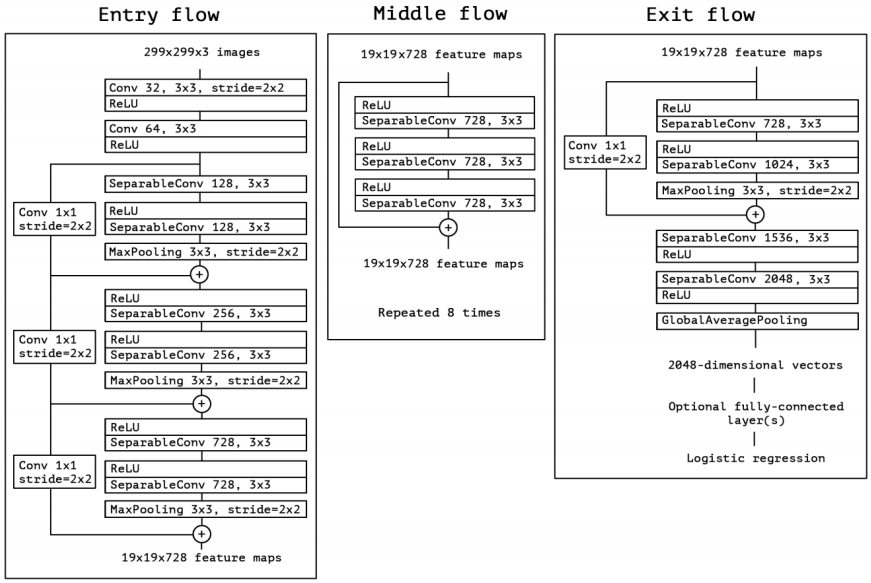

Figure 7: The Xception architecture: the data first goes through the entry flow, then through the middle flow which is repeated eight times, and finally through the exit flow.

3.4.3 Inception Resnet V1

Inspired by the performance of the ResNet, a hybrid inception module was proposed. Inception Resnet V1 [8] is a hybrid Inception version that has a similar computational cost to Inception-v3 [7]. Empirical research has shown that Inception-ResNet model is able to achieve higher accuracies at a lower epoch. We use a Inception Resnet V1 model pretrained on the VGGFace2 dataset[9], instead of Imagenet that was we used for Resnet-18. We made this choice as the VGGFace2 dataset is more widely used for facial recognition tasks. Since our objective was to detect facial image forgeries, it seems like a reasonable choice. Figure 8 is the schema for Inception Resnet V1. This model consists of several blocks including the modules A, B, C and the reduction blocks as described in Figures 9, 10 and 11.

3.5 Evaluation of Results

We report the results of the three models discussed in the previous sections.

3.5.1 Model Comparison

Table 1 taken from [10] summarizes the accuracy results on the test set for various models trained on the full Faceforensics dataset. The accuracy scores given are for classifying original and Neural-Textures frames at the highest compression level (c40). To avoid overfitting and poor generalization, pretrained models were used. The Resnet Model is pretrained on ImageNet classification. Inception Resnet V1 is pretrained on VGGFace2 face recognition. Transfer learning improves the performance significantly, especially when the model is pretrained for facial recognition tasks.

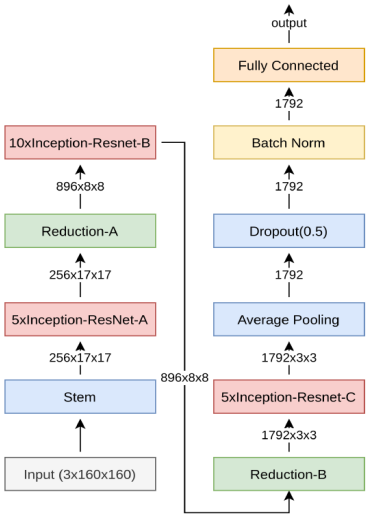

Figure 8: The schema for Inception-ResNet-v1 network.

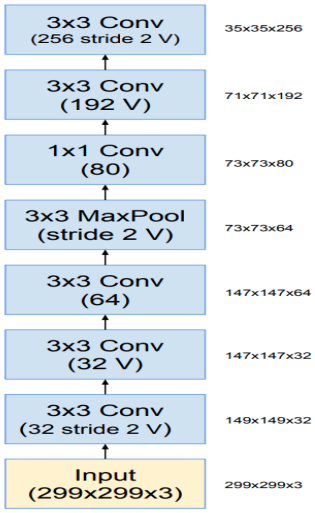

Figure 9: The stem of the Inception-ResNet-v1 network.

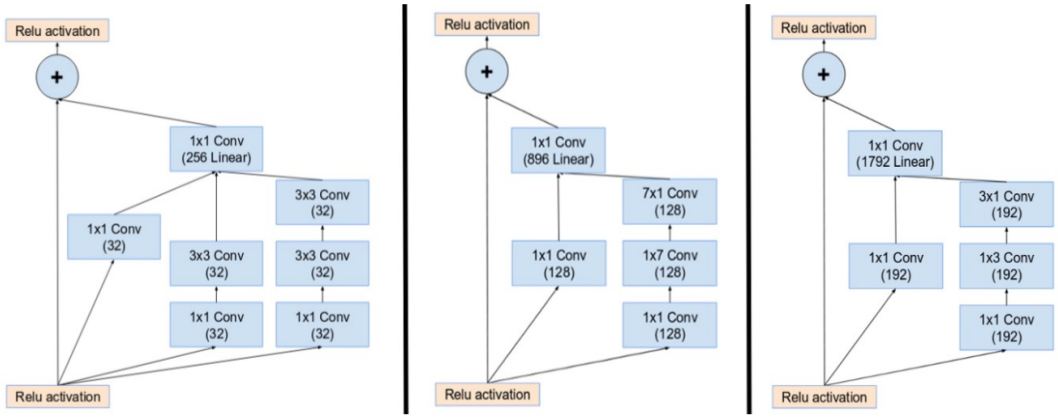

Figure 10: Inception modules A,B,C in an Inception ResNet.

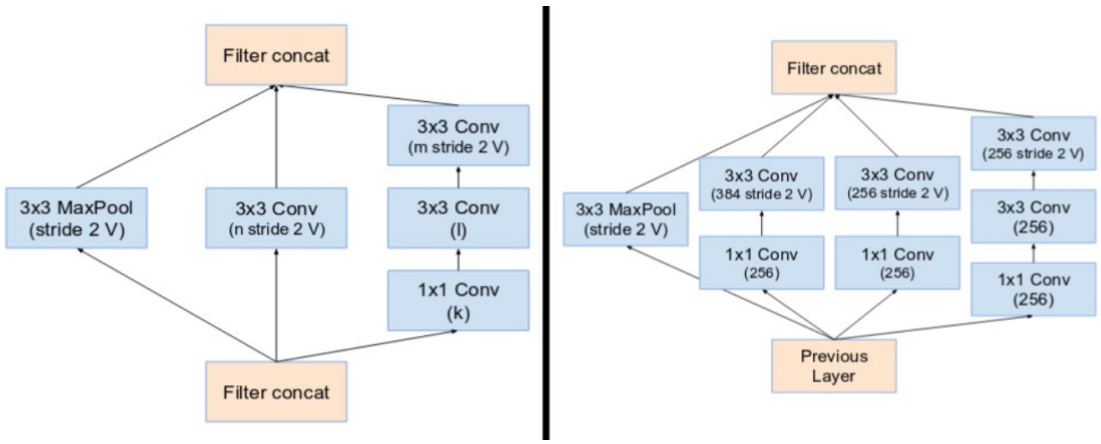

Figure 11: Reduction Block A (35x35 to 17x17 size reduction) and Reduction Block B (17x17 to 8x8 size reduction)

Model	Original vs NT
Resnet-18	66.4
Resnet-18 pretrained	73.6
XceptionNet	66.5
Inception ResNet pretrained	74.8

Table 1: Accuracy percentages for various model trained and evaluated on classifying original vs. NeuralTextures frames at compression level c40. The Resnet Model is pretrained on ImageNet classification. Inception Resnet V1 is pretrained on VGGFace2 face recognition.

3.5.2 Generalization from NerualTextures and DeepFakes

We trained our models on highly compressed c40 NeuralTextures and DeepFakes videos and cross-tested both datasets to check the generalizability of the models on unseen manipulation techniques. Table 2 shows the accuracy of the Inception ResNet Model trained and tested on both datasets. The models performed well with 89.6% accuracy on the DeepFakes dataset and 65.2% on the Neural-Textures dataset. These results were achieved in a short training time but were able to outperform humans significantly. These models would likely benefit from more computation resources. It is also important to note that the models were able to detect forgeries made by DeepFakes much better than those by NeuralTextures. This result is consistent with human performance. This is possibly because Neural Textures makes only slight amendments to the expression of the target, making it harder to detect. However, the models did not perform well when it was exposed to manipulation methods they were not trained on. The model trained with the DeepFakes dataset only achieved a 52.7% accuracy when evaluated on the test set of the Neural Textures dataset. Whereas, the model trained with the Neural Textures dataset only achieved a 52.6% when evaluated on the test set of the DeepFakes dataset. Both models performed only slightly better than random chance when exposed to an unseen manipulation technique.

Model	Original vs DF	Original vs NT
Inception ResNet pretrained + trained by DF	89.6	52.7
Inception ResNet pretrained + trained by NT	52.6	65.2

Table 2: The Inception ResNet pretrained model is trained by two different manipulation datasets (DF and NT).

4 Contribution of each member of the team

Everyone contributed to implementing the models, doing literature review, writing the report and creating the video.

5 Concluding remarks

We found that deep learning models performed well, much better than humans and much more scalable. Going forward, utilizing such tools will definitely be vital for the media forensics field. The use of transfer learning has also shown to be very useful and can significantly decrease the training time and the amount of data required. However, these models performed poorly when exposed to unseen manipulation methods. This remains an important area to tackle as new forgery techniques come up frequently and gathering a dataset of these new techniques will be a big challenge.

Because of limited computational resources and time we were not able to further examine uncompressed (raw) and moderately compressed (c23) videos of the same manipulations such as NeuralTextures or DeepFakes or even other manipulation techniques like Face2Face or Faceswap. A possible extension to this project would be to try out the other models discussed in the past works section.

6 References

1. Faceforensics++ dataset. https://github.com/ondyari/FaceForensics, Accessed: 2020-11-10.

2. Deepfakes github. https://github.com/deepfakes/faceswap, Accessed: 2020-11-10.

3. Justus Thies, Michael Zollh¨ofer, and Matthias Nießner. Deferred neural rendering: Image synthesis using neural textures. ACM Transactions on Graphics 2019 (TOG), 2019.

4. Zhang, K., Zhang, Z., Li, Z., and Qiao, Y. (2016). Joint face detection and alignment using multitask cascaded convolutional networks. IEEE Signal Processing Letters, 23(10):1499–1503.

5. K. He, X. Zhang, S. Ren and J. Sun, "Deep Residual Learning for Image Recognition," 2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Las Vegas, NV, 2016, pp. 770-778, doi: 10.1109/CVPR.2016.90.

6. F. Chollet, "Xception: Deep Learning with Depthwise Separable Convolutions," 2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Honolulu, HI, 2017, pp. 1800-1807, doi: 10.1109/CVPR.2017.195.

7. C. Szegedy, V. Vanhoucke, S. Ioffe, J. Shlens and Z. Wojna, "Rethinking the Inception Architecture for Computer Vision," 2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Las Vegas, NV, 2016, pp. 2818-2826, doi: 10.1109/CVPR.2016.308.

8. Christian Szegedy, Sergey Ioffe, Vincent Vanhoucke, and Alexander A. Alemi. 2017. Inception-v4, inception-ResNet and the impact of residual connections on learning. In Proceedings of the Thirty-First AAAI Conference on Artificial Intelligence (AAAI'17). AAAI Press, 4278–4284.

9. Q. Cao, L. Shen, W. Xie, O. M. Parkhi, and A. Zisserman. Vggface2: A dataset for recognising faces across pose and age. In International Conference on Automatic Face and Gesture Recognition, 2018.

10.Nika Dogonadze and Jana Obernosterer and Ji Hou: Deep Face Forgery Detection, 2020.

11. Cortada, James W. and William Aspray. Fake News Nation: The Long History of Lies and Misinterpretations in America (Rowman Littlefield Publishers, 2019

12. Thanh Thi Nguyen and Cuong M. Nguyen and Dung Tien Nguyen and Duc Thanh Nguyen and Saeid Nahavand. Deep Learning for Deepfakes Creation and Detection: A Survey, 2020. arXiv:1909.11573.

13. A. Rocha et al., "Authorship Attribution for Social Media Forensics," in IEEE Transactions on Information Forensics and Security, vol. 12, no. 1, pp. 5-33, Jan. 2017, doi: 10.1109/TIFS.2016.2603960.

14. Allcott, H. and M. Gentzkow (2017a). Social Media and Fake News in the 2016 Election. Journal of Economic Perspectives 31, 211–236.

15. Karim, Sajida He, Hui Laghari, Asif Memon, Kamran Khan, Mehak Magsi, Arif. (2020). The Evaluation Video Quality in Social Clouds. Entertainment Computing. 35. 100370. 10.1016/j.entcom.2020.100370.

16. Andreas Rossler, Davide Cozzolino, Luisa Verdoliva, Chris- ¨ tian Riess, Justus Thies, and Matthias Nießner. FaceForensics++: Learning to detect manipulated facial images. In ICCV, 2019.

17. Yuezun Li, Xin Yang, Pu Sun, Honggang Qi, Siwei Lyu; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2020, pp. 3207-3216

18. Nicholas Dufour, Andrew Gully, Per Karlsson, Alexey Victor Vorbyov, Thomas Leung, Jeremiah Childs, and Christoph Bregler. Deepfakes detection dataset by google jigsaw.

19. Pavel Korshunov and Sebastien Marcel. Deepfakes: a new ´ threat to face recognition? assessment and detection. arXiv preprint arXiv:1812.08685, 2018

20. Faceswap github. https://github.com/ deepfakes/faceswap, Accessed Nov 2, 2020.

21. FakeApp. https://www.malavida.com/en/soft/ fakeapp/, Acessed Nov 2, 2020.

22. Justus Thies, Michael Zollhofer, Marc Stamminger, Christian Theobalt, and Matthias Niessner. Face2face: Real-time face capture and reenactment of rgb videos. In CVPR, June 2016.

23. Justus Thies, Michael Zollhofer, and Matthias Nießner. De- ¨ ferred neural rendering: Image synthesis using neural textures. In SIGGRAPH, 2019.

24. Conrad Sanderson and Brian C Lovell. Multi-region probabilistic histograms for robust and scalable identity inference. In International Conference on Biometrics, 2009.

25. Darius Afchar, Vincent Nozick, Junichi Yamagishi, and Isao Echizen. Mesonet: a compact facial video forgery detection network. In IEEE International Workshop on Information Forensics and Security (WIFS), 2018

26. Hany Farid. Digital Image Forensics. MIT Press, 2012.

27. Chen, Jiansheng Kang, Xiangui Liu, Ye Wang, Z.. (2015). Median Filtering Forensics Based on Convolutional Neural Networks. Signal Processing Letters, IEEE. 22. 1849-1853. 10.1109/LSP.2015.2438008.

28. Xinyi Ding, Zohreh Raziei, Eric C Larson, Eli V Olinick, Paul Krueger, and Michael Hahsler. 2019. Swapped Face Detection using Deep Learning and Subjective Assessment. arXiv preprint arXiv:1909.04217 (2019).

29. Zhang Y, Goh J, Win L L, et al. Image Region Forgery Detection: A Deep Learning Approach[C]//SG-CRC. 2016: 1-11.

30. D. Cozzolino and L. Verdoliva, "Single-image splicing localization through autoencoder-based anomaly detection," in IEEE Workshop on Information Forensics and Security, 2016.

31. Zhou P, Han X, Morariu V I, et al. Two-stream neural networks for tampered face detection[C]//2017 IEEE Conference on Computer Vision and Pattern Recognition Workshops (CVPRW). IEEE. 2017: 1831-1839.

32. F. Marra, D. Gragnaniello, L. Verdoliva, and G. Poggi, "Do GANs leave artificial fingerprints?" in MIPR, 2019.

33. J. Stehouwer, H. Dang, F. Liu, X. Liu, and A. Jain, "On the detection of digital face manipulation," arXiv preprint arXiv:1910.01717v1, 2019.