Deep Learning for Deepfakes Creation and Detection: A Survey

Thanh Thi Nyugen, Cuong M. Nguyen, Dung Tien Nguyen, Duc Thanh Nguyen, Saeid Nahavandi, Fellow, IEEE

Published: 28 July 2020

Baran Deniz Korkmaz

August 2020

Abstract

This paper presents a survey of algorithms used to create deepfakes and, more importantly, methods proposed to detect deepfakes in the literature to date. By reviewing the background of deep-fakes and state-of-the-art deepfake detection methods, this study provides a comprehensive overview of deepfake techniques and facilitates the development of new and more robust methods to deal with the increasingly challenging deepfakes.

Index Terms (Authors' Selection): deep learning, computer vision, autoencoders, forensics, GAN

1 Introduction

The underlying mechanism for deepfake creation is deep learning models such as autoencoders and generative adversarial networks, which have been applied widely in the computer vision domain. **[1]-[7]** These models are used to examine facial expressions and movements of a person and synthesize facial images of another person making analogous expressions and movements.

There have been numerous methods proposed to detect deepfakes **[19]-[23]**. Data obtained from https://app.dimensions.ai at the end of July 2020 show that the number of deepfake papers has increased significantly in recent years.

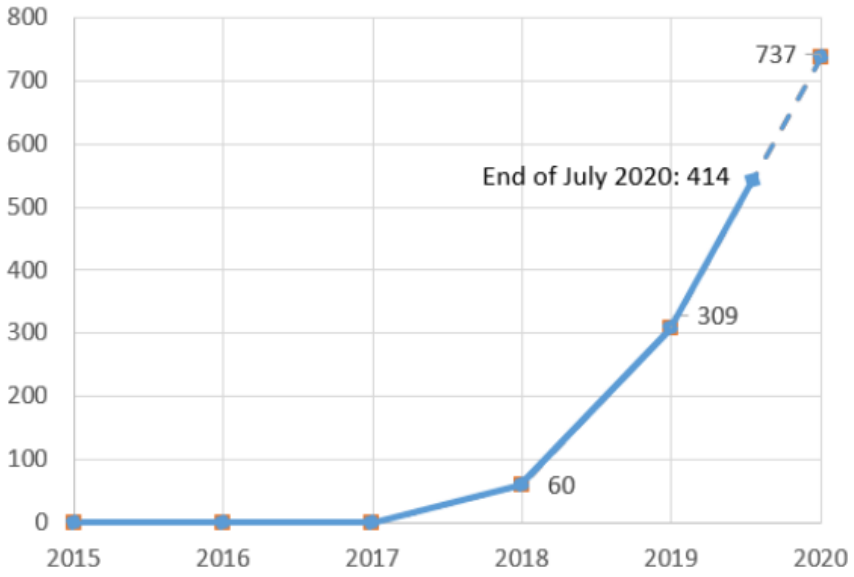

Figure 1: Number of papers related to deepfakes in years from 2015 to 2020, obtained from https://app.dimensions.ai on 24 July 2020 with the search keyword "deepfake" applied to full text of scholarly papers.

The overview of discussions presented in the sections can be listed as below:

- **Section 2** presents the principles of deepfake algorithms and how deep learning has been used to enable such disruptive technologies.
- **Section 3** reviews different methods for detecting deepfakes as well as their advantages and disadvantages.
- In **Section 4** challenges, research trends and directions on deepfake detection and multimedia forensics problems were discussed.

2 Deepfake Creation

Deep learning is well known for its capability of representing complex and high-dimensional data. One variant of the deep networks with that capability is deep autoencoders, which have been widely applied for dimensionality reduction and image compression [26]–[28].

Below, you will find a summary of techniques developed for deepfake creation:

1. The first attempt of deepfake creation was FakeApp, developed by a Reddit user using autoencoder-decoder pairing structure [29], [30]. In that

method, the autoencoder extracts latent features of face images and the decoder is used to reconstruct the face images. To swap faces between source images and target images, there is a need of two encoder- decoder pairs where each pair is used to train on an image set, and the encoder's parameters are shared between two network pairs. In other words, two pairs have the same encoder network. This strategy enables the common encoder to find and learn the similarity between two sets of face images, which are relatively unchallenging because faces normally have similar features such as eyes, nose, mouth positions.

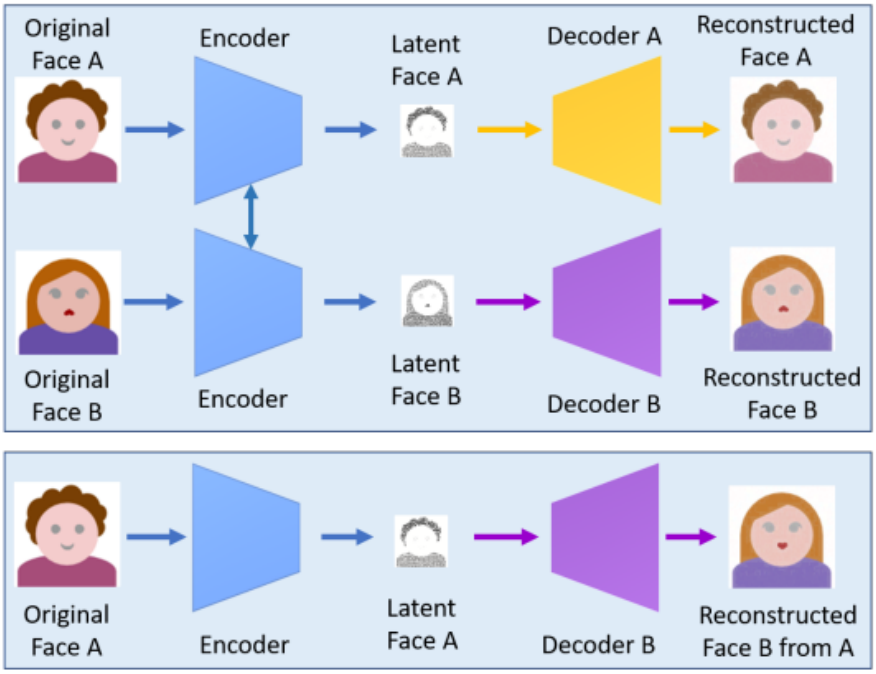

Figure 2: A deepfake creation model using two encoder-decoder pairs. Two networks use the same encoder but different decoders for training process (top). An image of face A is encoded with the common encoder and decoded with decoder B to create a deepfake (bottom).

2. By adding adversarial loss and perceptual loss implemented in VGGFace [34] to the encoder-decoder architecture, an improved version of deepfakes based on the generative adver- sarial network (GAN) [35], i.e. faceswap-GAN, was proposed in [36]. The VGGFace perceptual loss is added to make eye movements to be more realistic and consistent with input faces and help to smooth out artifacts in segmentation mask, leading to higher quality output videos.

3. The multi-task convolutional neural network (CNN) from the FaceNet implementation [37] is introduced to make face detection more stable and face alignment more reliable.

Popular deepfake tools and their features are summarized in the table below:

Tools	Key Features
Faceswap	-Using two encode-decoder pairs. -Parameters of the encoder are shared.
Faceswap-GAN	Adversarial and perceptual loss (VGGface) are added to an auto-encoder architecture.
Few-Shot Face Translation GAN	- Use a pre-trained face recognition model to extract latent embeddings for GAN processing. - Incorporate semantic priors obtained by modules from FUNIT[39] and SPADE[40].
DeepFaceLab	- Expand from the Faceswap method with new models. - Support multiple face extraction modes.
DFaker	- DSSIM loss function [42] is used to reconstruct face. - Implemented by Keras.
DeepFake_tf	Similar to DFaker but implemented based on TensorFlow.
Deepfakes web Beta	Commercial website for face swapping using deep learning algorithms.

3 Deepfake Detection

Early attempts were based on handcrafted features obtained from artifacts and inconsistencies of the fake video synthesis process.

Recent methods, on the other hand, applied deep learning to automatically extract salient and discriminative features to detect deepfakes [44], [45].

The deepfake detection methods can be classified based on the type of source.

3.1 Fake Image Detection

The use of deep learning such as CNN and GAN has made swapped face images more challenging for forensics models as it can preserve pose, facial expression and lighting of the photographs.

1. Zhang et al. [56] used the bag of words method to extract a set of compact features and fed it into various classifiers such as SVM [57], random forest (RF) [58] and multi-layer perceptrons (MLP) [59] for discriminating swapped face images from the genuine.

2. Most works on detection of GAN generated images however do not consider the generalization capability of the detection models although the devel-

opment of GAN is ongoing, and many new extensions of GAN are frequently introduced. Xuan et al. [60] used an image preprocessing step, e.g. Gaussian blur and Gaussian noise, to remove low level high frequency clues of GAN images. This increases the pixel level statistical similarity between real images and fake images and requires the forensic classifier to learn more intrinsic and meaningful features, which has better generalization capability than previ- ous image forensics methods [61], [62] or image steganalysis networks [63].

3. Agarwal and Varshney [64] cast the GAN-based deepfake detection as a hypothesis testing prob- lem where a statistical framework was introduced using the information-theoretic study of authentication [65]. The mini- mum distance between distributions of legitimate images and images generated by a particular GAN is defined, namely the oracle error. The analytic results show that this distance increases when the GAN is less accurate, and in this case, it is easier to detect deepfakes.

4. Hsu et al. [66] introduced a two-phase deep learn- ing method for detection of deepfake images. The first phase is a feature extractor based on the common fake feature network (CFFN) where the Siamese network architecture presented in [67] is used.

3.2 Fake Video Detection

Most image detection methods cannot be used for videos because of the strong degradation of the frame data after video compression [83]. Furthermore, videos have temporal characteristics that are varied among sets of frames and thus challenging for methods designed to detect only still fake images.

This subsection focuses on deepfake video detection methods and categorizes them into two groups: methods that employ temporal features and those that explore visual artifacts within frames.

3.2.1 Temporal Features Across Video Frames

1. Sabir et al. [84] leveraged the use of spatio-temporal features of video streams to detect deepfakes. Video manipulation is carried out on a frame-by-frame basis so that low level artifacts produced by face manipulations are believed to further manifest themselves as temporal artifacts with inconsistencies across frames. A recurrent convolutional model (RCN) was proposed based on the integration of the convolutional network DenseNet [68] and the gated recurrent unit cells [85] to exploit temporal discrepancies across frames.

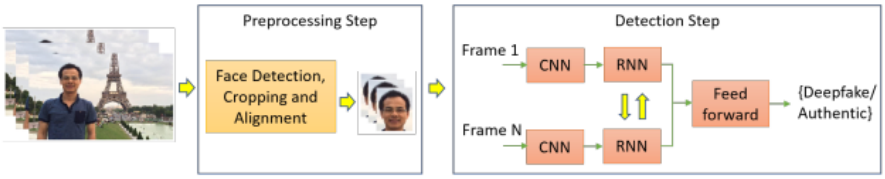

Figure 3: A two-step process for face manipulation detection where the preprocessing step aims to detect, crop and align faces on a sequence of frames and the second step distinguishes manipulated and authentic face images by combining convolutional neural network (CNN) and recurrent neural network (RNN) [84].

2. Guera and Delp [87] highlighted that deepfake videos contain intra-frame inconsistencies and temporal incon- sistencies between frames. They then proposed the temporal- aware pipeline method that uses CNN and long short term memory (LSTM) to detect deepfake videos.CNN is employed to extract frame-level features, which are then fed into the LSTM to create a temporal sequence descriptor. A fully- connected network is finally used for classifying doctored videos from real ones based on the sequence descriptor.

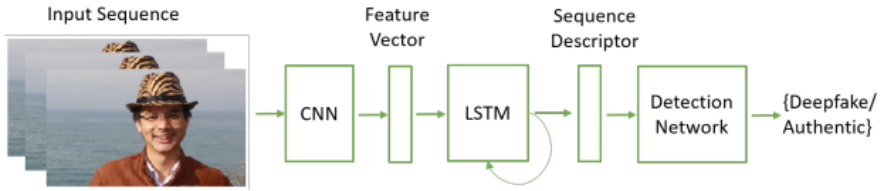

Figure 4: A deepfake detection method using convolutional neural network (CNN) and long short term memory (LSTM) to extract temporal features of a given video sequence, which are represented via the sequence descriptor. The detection network consisting of fully-connected layers is employed to take the sequence descriptor as input and calculate probabilities of the frame sequence belonging to either authentic or deepfake class [87].

3. The use of a physiological signal, eye blinking, to detect deepfakes was proposed in [88] based on the observation that a person in deepfakes has a lot less frequent blinking than that in untampered videos. Blinking rates in deepfakes are much lower than those in normal videos. To discriminate real and fake videos, Li et al. [88] first decompose the videos into frames where face regions and then eye areas are extracted based on six eye landmarks. After few steps of pre-processing such as aligning faces, extracting and scaling the bounding boxes5 of eye landmark points to create new sequences of frames, these cropped eye area sequences are distributed into long-term recurrent convolutional networks (LRCN) [89] for dynamic state prediction.

3.2.2 Visual Artifacts within Video Frame

This subsection investigates the other approach that normally decomposes videos into frames and explores visual artifacts within single frames to obtain discriminant features.

1. Deep classifiers:

   (a) A deep learning method to detect deepfakes based on the artifacts observed during the face warping step of the deepfake generation algorithms was proposed in [92].

   (b) Nguyen et al. [95] proposed the use of capsule networks for detecting manipulated images and videos. The recent development of capsule network based on dynamic routing algorithm [97] demonstrates its ability to describe the hierarchical pose relationships between object parts. This development is employed as a component in a pipeline for detecting fabricated images and videos.

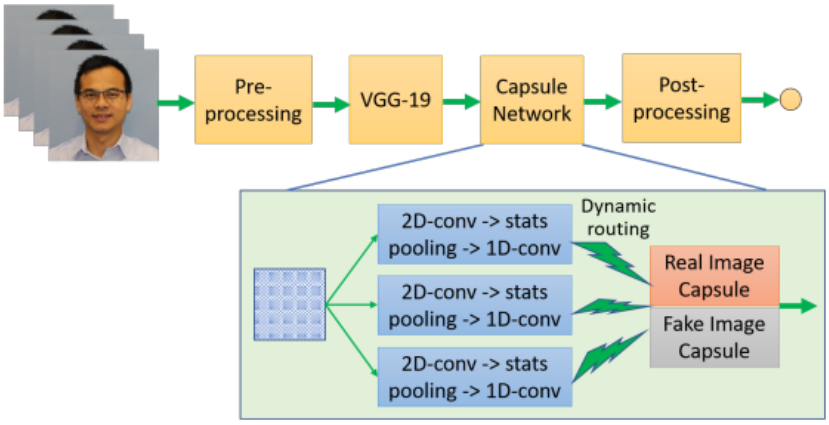

Figure 5: Capsule network takes features obtained from the VGG-19 network [90] to distinguish fake images or videos from the real ones.

2. Shallow Classifiers:

   (a) Yang et al. [93] proposed a detection method by observing the differences between 3D head poses comprising head orientation and position, which are estimated based on 68 facial landmarks of the central face region. The 3D head poses are examined because there is a shortcoming in the deepfake face generation pipeline. The extracted

features are fed into an SVM classifier to obtain the detection results. Experiments on two data sets show the great performance of the proposed approach against6 its competing methods.

(b) A method to exploit artifacts of deepfakes and face manipulations based on visual features of eyes, teeth and facial contours was studied in [103]. **The visual artifacts arise from lacking global consistency, wrong or imprecise estimation of the incident illumination, or imprecise estimation of the underlying geometry.**

(c) The use of photo response non uniformity (PRNU) analysis was proposed in [104] to detect deepfakes from authentic ones. PRNU is a component of sensor pattern noise, which is attributed to the manufacturing imperfection of silicon wafers and the inconsistent sensitivity of pixels to light because of the variation of the physical characteristics of the silicon wafers.

(d) Hasan and Salah [111] proposed the use of blockchain and smart contracts to help users detect deepfake videos based on the assumption that videos are only real when their sources are traceable.

4 Source

- Deep Learning for Deepfakes Creation and Detection: A Survey; Thanh Thi Nguyen, Cuong M. Nguyen, Dung Tien Nguyen, Duc Thanh Nguyen, Saeid Nahavandi

5 Referenced Sources

1. **Autoencoders & GANs for Deepfake Creation**

   - [1] - Badrinarayanan, V., Kendall, A., and Cipolla, R. (2017). SegNet: A deep convolutional encoder-decoder architecture for image segmenta- tion. IEEE Transactions on Pattern Analysis and Machine Intelligence, 39(12), 2481-2495.
   - [2] - Yang, W., Hui, C., Chen, Z., Xue, J. H., and Liao, Q. (2019). FV-GAN: Finger vein representation using generative adversarial networks. IEEE Transactions on Information Forensics and Security, 14(9), 2512-2524.
   - [3] - Tewari, A., Zollhoefer, M., Bernard, F., Garrido, P., Kim, H., Perez, P., and Theobalt, C. (2020). High-fidelity monocular face reconstruc- tion based on an unsupervised model-based face autoencoder. IEEE Transactions on Pattern Analysis and Machine Intelligence. DOI: 10.1109/TPAMI.2018.2876842.

- [4] - Guo, Y., Jiao, L., Wang, S., Wang, S., and Liu, F. (2018). Fuzzy sparse autoencoder framework for single image per person face recognition. IEEE Transactions on Cybernetics, 48(8), 2402-2415.
- [5] - Liu, F., Jiao, L., and Tang, X. (2019). Task-oriented GAN for PolSAR image classification and clustering. IEEE Transactions on Neural Networks and Learning Systems, 30(9), 2707-2719.
- [6] - Cao, J., Hu, Y., Yu, B., He, R., and Sun, Z. (2019). 3D aided duet GANs for multi-view face image synthesis. IEEE Transactions on Information Forensics and Security, 14(8), 2028-2042.
- [7] - Zhang, H., Xu, T., Li, H., Zhang, S., Wang, X., Huang, X., and Metaxas, D. N. (2019). StackGAN++: Realistic image synthesis with stacked generative adversarial networks. IEEE Transactions on Pattern Analysis and Machine Intelligence, 41(8), 1947-1962.
- [8] - Lyu, S. (2018, August 29). Detecting deepfake videos in the blink of an eye.

2. **Pros of Deepfake Creation** - [14] - Marr, B. (2019, July 22). The best (and scari- est) examples of AI-enabled deepfakes.

3. **Deepfake Detection**

   - [19] - Lyu, S. (2020, July). Deepfake detection: current challenges and next steps. In IEEE International Conference on Multimedia and Expo Workshops (ICMEW) (pp. 1-6). IEEE.
   - [20] - Guarnera, L., Giudice, O., Nastasi, C., and Battiato, S. (2020). Preliminary forensics analysis of deepfake images. arXiv preprint arXiv:2004.12626.
   - [21] - Jafar, M. T., Ababneh, M., Al-Zoube, M., and Elhassan, A. (2020, April). Forensics and analysis of deepfake videos. In The 11th Interna- tional Conference on Information and Communication Systems (ICICS) (pp. 053-058). IEEE.
   - [22] - Trinh, L., Tsang, M., Rambhatla, S., and Liu, Y. (2020). Interpretable deepfake detection via dynamic prototypes. arXiv preprint arXiv:2006.15473
   - [23] - Younus, M. A., and Hasan, T. M. (2020, April). Effective and fast deepfake detection method based on Haar wavelet transform. In 2020 International Conference on Computer Science and Software Engineer- ing (CSASE) (pp. 186-190). IEEE.

4. **Deepfake Creation in Practice**

   - [29] - Faceswap: Deepfakes software for all. `https://github.com/deepfakes/faceswap`
   - [30] - FakeApp 2.2.0. `https://www.malavida.com/en/soft/fakeapp/`

- [31] - DeepFaceLab. `https://github.com/iperov/DeepFaceLab`
- [32] - DFaker. `https://github.com/dfaker/df`
- [33] - DeepFake tf: Deepfake based on tensorflow. `https://github.com/StromWine/DeepFaketf`
- [34] - Keras-VGGFace: VGGFace implementation with Keras framework. `https://github.com/rcmalli/keras-vggface`
- [35] - Goodfellow, I., Pouget-Abadie, J., Mirza, M., Xu, B., Warde-Farley, D., Ozair, S., ... and Bengio, Y. (2014). Generative adversarial nets. In Advances in Neural Information Processing Systems (pp. 2672-2680).
- [36] - Faceswap-GAN `https://github.com/shaoanlu/faceswap-GAN`
- [37] - FaceNet `https://github.com/davidsandberg/facenet`
- [38] - CycleGAN `https://github.com/junyanz/pytorch-CycleGAN-and-pix2pix`

5. **Some Other Recent Methods for Deepface Detection**

   - [44] - de Lima, O., Franklin, S., Basu, S., Karwoski, B., and George, A. (2020). Deepfake detection using spatiotemporal convolutional net- works. arXiv preprint arXiv:2006.14749.
   - [45] - Amerini, I., and Caldelli, R. (2020, June). Exploiting prediction error inconsistencies through LSTM-based classifiers to detect deepfake videos. In Proceedings of the 2020 ACM Workshop on Information Hiding and Multimedia Security (pp. 97-102).

6. **Fake Image Detection**

   (a) **Bag of Words** - [56] - Zhang, Y., Zheng, L., and Thing, V. L. (2017, August). Automated face swapping and its detection. In 2017 IEEE 2nd International Conference on Signal and Image Processing (ICSIP) (pp. 15-19). IEEE.

   (b) **Image Preprocesing Proposals** - [60] - Xuan, X., Peng, B., Dong, J., and Wang, W. (2019). On the general- ization of GAN image forensics. arXiv preprint arXiv:1902.11153.

   (c) **Oracle Error** - [64] - Agarwal, S., and Varshney, L. R. (2019). Limits of deepfake detection: A robust estimation viewpoint. arXiv preprint arXiv:1905.03493.

   (d) **Deep Fake Image Detection Based on Pairwise Learning**

   - [66] - Hsu, C. C., Zhuang, Y. X., and Lee, C. Y. (2020). Deep fake image detection based on pairwise learning. Applied Sciences, 10(1), 370.
   - [67] - Chopra, S. (2005). Learning a similarity metric discriminatively, with application to face verification. In IEEE Conference on Compter Vision and Pattern Recognition (pp. 539-546).

- [68] - Huang, G., Liu, Z., Van Der Maaten, L., and Weinberger, K. Q. (2017). Densely connected convolutional networks. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (pp. 4700- 4708).

7. **Some Other Detection Methods**

   - [79] - Farid, H. (2009). Image forgery detection. IEEE Signal Processing Magazine, 26(2), 16-25.
   - [80] - Mo, H., Chen, B., and Luo, W. (2018, June). Fake faces identification via convolutional neural network. In Proceedings of the 6th ACM Workshop on Information Hiding and Multimedia Security (pp. 43-47).
   - [81] - Marra, F., Gragnaniello, D., Cozzolino, D., and Verdoliva, L. (2018, April). Detection of GAN-generated fake images over social networks. In 2018 IEEE Conference on Multimedia Information Processing and Retrieval (MIPR) (pp. 384-389). IEEE.
   - [82] - Hsu, C. C., Lee, C. Y., and Zhuang, Y. X. (2018, December). Learning to detect fake face images in the wild. In 2018 International Symposium on Computer, Consumer and Control (IS3C) (pp. 388-391). IEEE.

8. **Fake Video Detection**

   (a) **Temporal Features across Video Frames**

   - **Spatio-Temporal Features** - [84] - Sabir, E., Cheng, J., Jaiswal, A., AbdAlmageed, W., Masi, I., and Natarajan, P. (2019). Recurrent convolutional strategies for face ma- nipulation detection in videos. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops (pp. 80-87).
   - **Intra-Frame Inconsistencies by LSTM** - [87] - Guera, D., and Delp, E. J. (2018, November). Deepfake video detection using recurrent neural networks. In 2018 15th IEEE International Conference on Advanced Video and Signal Based Surveillance (AVSS) (pp. 1-6). IEEE.
   - **Eye Blinking** - [88] - Li, Y., Chang, M. C., and Lyu, S. (2018, December). In ictu oculi: Exposing AI created fake videos by detecting eye blinking. In 2018 IEEE International Workshop on Information Forensics and Security (WIFS) (pp. 1-7). IEEE.

   (b) **Visual Artifacts within Video Frame**

   - **Capsule Networks** - [95] - Nguyen, H. H., Yamagishi, J., and Echizen, I. (2019, May). Capsule- forensics: Using capsule networks to detect forged images and videos. In 2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP) (pp. 2307-2311). IEEE.

- **Dynamic Routing** - [97] - Sabour, S., Frosst, N., and Hinton, G. E. (2017). Dynamic routing between capsules. In Advances in Neural Information Processing Systems (pp. 3856-3866).
- **3D Head Poses** - [93] - Yang, X., Li, Y., and Lyu, S. (2019, May). Exposing deep fakes using inconsistent head poses. In 2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP) (pp. 8261-8265). IEEE.
- **Facial Artifacts** - [103] - Matern, F., Riess, C., and Stamminger, M. (2019, January). Exploiting visual artifacts to expose deepfakes and face manipulations. In 2019 IEEE Winter Applications of Computer Vision Workshops (WACVW) (pp. 83-92). IEEE.
- **PRNU** - [104] - Koopman, M., Rodriguez, A. M., and Geradts, Z. (2018). Detection of deepfake video manipulation. In The 20th Irish Machine Vision and Image Processing Conference (IMVIP) (pp. 133-136). [105] Lukas, J., Fridrich, J., and Goljan, M. (2006). Digital c
- **Blockchain** - [111] - Hasan, H. R., and Salah, K. (2019). Combating deepfake videos using blockchain and smart contracts. IEEE Access, 7, 41596-41606

6 Notes

1. **Data Sets for Deepfake Detection:**

   (a) UADFV

   (b) DeepfakeTIMIT

   (c) Data Set used in DARPA MediFor GAN Image/Video Challenge