DEEPFAKES CREATION AND DETECTION USING DEEP LEARNING

Shaik Ismail[1], Shaik Saif Ali[2], Zulqharnain Mustafa Sheriff[3], Dr. K Naggi Reddy[4]
B.E Student,Department of IT, Lords Institute of Engineering and Technology, Hyderabad
Professor, Department & HoD of IT, Lords Institute of Engineering and Technology, Hyderabad
hodit@lords.ac.in

Abstract— Computer vision, NLP, and image identification are just a few of the many areas where deep learning has been put to use. Deepfakes are a byproduct of the development of deep learning algorithms for image recognition and modification; these deepfakes employ these algorithms to generate fake pictures that may be difficult to tell apart from the genuine thing. In this paper, we explore the use of deep learning for both creating and detecting deepfakes, and we propose the use of a deep learning image enhancement method to improve the quality of deepfakes created in response to growing concerns about personal privacy and security.

***Keywords*—deepfake, deep Learning, Artificial intelligence, machine learning, tensor flow.**

I. 1NTRODtlCTION

The use of machine vision is expanding rapidly in a wide variety of industries, from standard image-detection programs to the auto and robotics sectors [1][2]. Deepfake is a product of machine vision, which is one of many such applications. Using deep learning algorithms, a method known as "deepfake" may be used to generate fake photos that can be difficult to spot. Typically, this is accomplished by replacing one person's face in a source image with another's face in a target image. The generation of deepfakes relies on deep learning encoders and decoders, which have seen substantial application in the field of machine vision [4]. [5]. The encoders pull out all the characteristics of a picture, and then the decoders make the forgery. While training deep Learning models for use in deepfake techniques required a vast amount of photos and videos, this was a far more difficult undertaking before the advent of social media. The proliferation of data has facilitated the development of sophisticated deepfake techniques. Tensorflow is used in the creation of several of the algorithms used in deerfake [6]. TensorFlow is a free library for data-flow graph-based numerical computation. While the system was designed for internal use by Gonfle in its research and development of machine learning and deep neural neiworks, its generalizability and low entry barrier to entry have made it a popular choice for machine learning applications.

Since TensorFlow's APIs are compatible with Python, we can rapidly alter the CNN architecture and experiment with other architectures without having to make extensive changes to the source code.

You could be watching a video of a prominent public figure or president giving a speech and have no idea if what you're seeing is real or fake [7] due to the prevalence of deepfakes. Creating these fake images and videos is much easier today than it was in the past; all you need is an image or video of the target individual to generate the fake contour.

The proliferation of deepfakes on the internet has prompted major tech firms to begin investigating ways to detect and expose them. As of late, facebook. To promote greater research and development towards identifying and preventing deepfakes, Amazon, Microsoft, and the Partnership on AI's Media Integrity Steering Committee have announced the Deepfake Detection Challenge [8]. In addition, Google has contributed to the deepfake detection cballenge [9] by making available to the public a free dataset. The fact that major players in the IT industry like Google and Microsoft are investigating deepfake demonstrates the magnitude of the problem.

Deepfakes are digital forgeries that swap out the intended victim's visage with that of another person of a different race. Developers and user groups on the internet have refined this method to provide user-friendly tools like FakeApp and FaceSwap that can be found on the internet. [l0J [11].

The autoencoder-decoder pipeline is crucial to Deepfake since en- coders are often used for image compression, rely on deep neural networks, and cause the network to produce a compressed version of the original input by imposing a bottleneck. because of the advent of better encoders Less processing power is needed for deepfake tasks because to the possibility of high-quality picture compression (13). [14]. Two autoencoders are trained to create convincing fakes. The characteristics of the original picture are learned by a single autoenccctler.

Finding the truth in the digital realm is therefore more important than ever. The difficulty increases when dealing with deepfakes, which are often used for nefarious ends and can be created by almost anybody with the help of currently available deepfake tools. Numerous techniques have been presented so far [25-29] to identify deepfakes. Most of them rely on deep learning, which has led to a conflict between malevolent and beneficial applications of the technique. To counter the danger posed by biometric face-swapping technologies Figure 1: The total number of articles published on deepfakes from 2016 to 2021, as found by searching the complete texts of academic papers using the term "deepfake" on the website https://app.dimensions.ai at the end of 2021. In order to speed up the development of technologies for detecting fraudulent digital visual material, the United States Defense Advanced Research Projects Agency (DARPA) launched a research program in media forensics (called material Forensics or MediFor) [30]. To encourage greater study into identifying and preventing the use of deepfakes to deceive viewers, Facebook Inc., Microsoft Corp., and the Partnership on AI consortium have announced the Deepfake Detection Challenge [21]. The number of deepfake publications has skyrocketed in recent years, as seen by data collected by https://app.dimensions.ai by the end of 2021 (Fig. 1). Even though the counted deepfake publications is less than the true quantity, there is a clear upward trend in the study of this phenomenon.

The publications [19, 20, 20] provide overviews of the current state of the art in deepfake creation and detection. For instance, Mirsky and Lee [19] zeroed in on replacement procedures (such as swapping or transferring a face) and reenactment approaches (which include altering a target's expression, lips, attitude, gaze, or body). Conventional techniques (such as blind methods without utilizing any external data for training, one-class sensor-based and model-based methods, and supervised methods using handmade features) and deep learning-based approaches (such as CNN models) were distinguished by Verdoliva [20]. Based on how deepfakes are made, Tolosana et al. [22] classified production and detection approaches as whole-face synthesis, identity swap, attribute manipulation, and expression swap, respectively. On the other hand, we conduct the survey using a novel classification scheme.

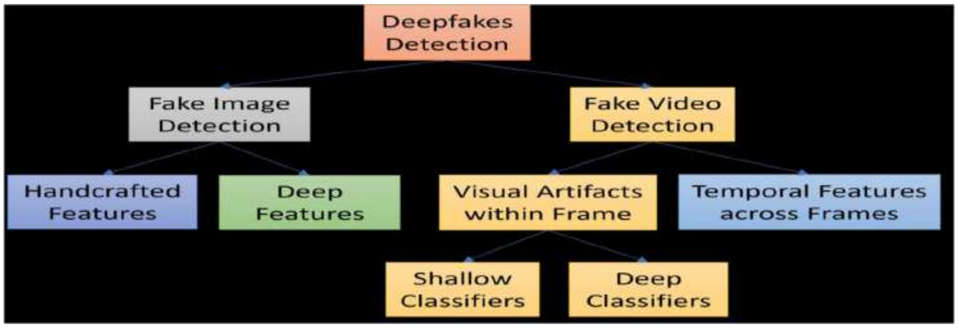

Fig. 2. Categories of reviewed papers relevant to deepfake detection methods

2. Deepfake Creation

The high quality of the spoofed films and the accessibility of the related software have contributed to deepfakes' meteoric rise in popularity. Most of these programs are created using deep learning methods. It is well knowledge that deep learning can effectively represent high-dimensional data sets. Deep autoencoders are a subset of deep networks with this feature that has found widespread use in a variety of contexts, including dimensionality reduction and picture compression [29–30]. In the beginning, there existed FakeApp, an autoencoder-decoder pairing structure devised by a Reddit user [31, 32]. The autoencoder is responsible for extracting latent characteristics from facial pictures, while the decoder is responsible for reconstructing them. Two encoder-decoder pairs, one for each picture set used in training, and a common set of encoder parameters, are required to perform the face swap between source and target images. This means that the encoder network is shared by the two pairs. Using this method, the generic encoder may discover and master the resemblance between two groups of face photos.- Using two sets of encoders and decoders, as shown in Fig. 3, we can create deepfakes. The training procedure (above) for two networks that share an encoder but utilize distinct decoders. A deepfake (bottom) is generated by taking a picture of face A, encoding it using the standard encoder, and decoding it with the alternative decoder. Face B, with the mouth geometry of face A, is what we see in the rebuilt picture (down below). Originally, Face B's mouth was inverted like a heart, but after reconstruction, it looked normal.

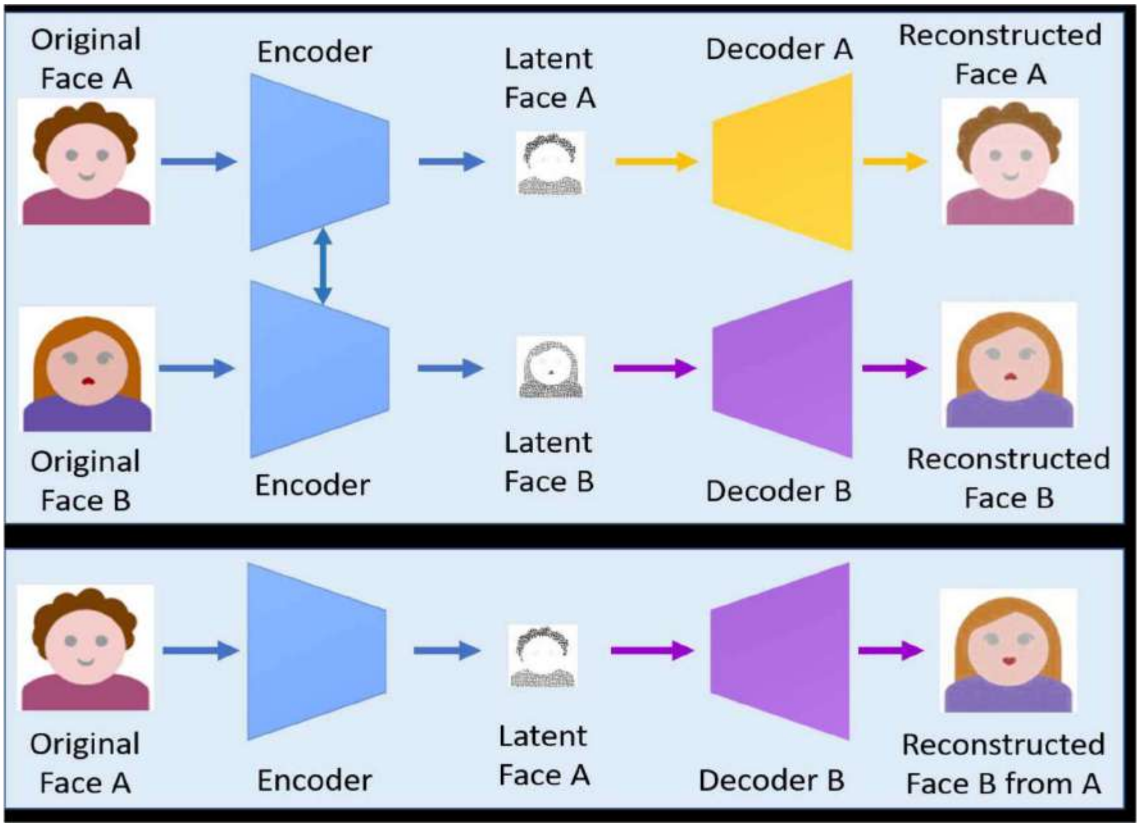

Fig. 3. A deepfake creation model using two encoder-decoder pairs

3. Deepfake Detection

Typically, classifiers are employed to distinguish genuine movies from those that have been tampered with, making deepfake detection a binary classification issue. In order to train classification algorithms, this approach needs a large collection of both authentic and spoofed films. Although there is a growing supply of fraudulent movies, this does not provide a sufficient sample size against which to test existing detection strategies. Korshunov and Marcel [63] created a noteworthy deepfake dataset of 620 videos using the Faceswap-GAN [58] open source code to solve this problem. Low and high grade deepfake films, which can effectively simulate the facial expressions, lip movements, and eye blinking, were generated using videos from the publicly accessible VidTIMIT collection . Different approaches to identifying deepfakes were put to the test with these films. The widely used VGG and Facenet based facial recognition algorithms fail to identify deepfakes in experiments. When applied to this newly created dataset, other techniques, such as lip-syncing approaches and picture quality measures using support vector machine (SVM) , result in a very high mistake rate when attempting to identify deepfake films. As a result, there is a pressing need for research into more reliable techniques of distinguishing deepfakes from the real thing. In this article, we provide a comprehensive overview of deepfake detection techniques, classifying them into two broad categories: techniques for detecting fake images and techniques for detecting false videos (Fig. 2). The latter are split into two subsets, one dealing with visual artifacts unique to single video frames and the other with temporal aspects shared by several frames. Though deep learning recurrent classification models are used in most temporal feature-based approaches, both deep and shallow classifiers may be used to analyze visual artifacts inside a video frame.

3.1. Fake Image Detection

The threats posed by deepfakes to individual safety, public order, and democratic processes are growing. As soon as this danger was made public, methods were offered to identify and counteract deepfakes. Earlier approaches relied on characteristics manually created from false picture synthesis faults and discrepancies. In order to automatically extract prominent and discriminative characteristics to identify deepfakes, recent approaches have often employed deep learning.

3.1.1. Handcrafted Features-based Methods

While GAN is constantly evolving and new extensions are often developed, the generalization potential of detection models is often overlooked in existing publications. Low-level high-frequency hints of GAN pictures were removed by Xuan et al. using an image preparation phase, such as using Gaussian blur and Gaussian noise. This improves the forensic classifier's generalization capability over earlier image forensics methods or image steganalysis networks by increasing the statistical similarity between real and fake images at the pixel level. For example, Zhang et al. employed the bag of words technique to extract a collection of compact features, which they then put into classifiers such support vector machines, random forests, and multi-layer perceptrons . pictures synthesized by GAN models are realistic and high-quality due to GAN's capacity to learn distribution of the complicated input data and produce new outputs with comparable input distribution, making them among the most difficult to spot deep learning-generated pictures. The GAN-based deepfake detection was reframed by Agarwal and Varshney as a hypothesis testing issue, and a statistical framework was presented via an examination of authentication from an information-theoretic perspective. The oracle error is the smallest allowable discrepancy between the distributions of authentic pictures and those produced by a certain GAN. Analytic studies reveal that this gap grows in the presence of a less-than-perfect GAN, making it simpler to spot deep fakes. When dealing with high-resolution picture inputs, a GAN has to be very precise to produce convincing false images.

3.1.2. Deep Features-based Methods

Since it is possible to substitute faces in pictures with ones from a library of stock images, face swapping has several appealing applications in video compositing, transfiguration in portraiture, and notably in identity protection. However, it is also one of the methods used by cybercriminals to obtain unauthorized entry into identity or authentication systems. Because deep learning techniques like CNN and GAN can maintain picture details like stance, facial expression, and lighting, swapped face photos provide a new challenge for forensics models .

3.2. Fake Video Detection

Due to the severe loss of frame data following video compression, most image identification algorithms cannot be used to movies. Furthermore, videos present difficulties for algorithms developed to identify solely still false pictures due to temporal properties that vary between sets of frames. This section analyzes the techniques used to detect deepfake videos and divides them into two subsets: those that rely on temporal characteristics and those that investigate visual artifacts inside individual frames.

3.2.1. Temporal Features across Video Frames

Sabir et al. [103] used spatiotemporal aspects of video streams to identify deepfakes on the basis of the finding that temporal coherence is not maintained effectively in the synthesis process of deepfakes.

Low-level distortions from face modifications are thought to also show as temporal artifacts with irregularities between frames due to the frame-by-frame nature of video editing. To take advantage of temporal differences between frames, a recurrent convolutional model (RCN) was suggested by combining the convolutional network DenseNet [61] with the gated recurrent unit cells. The suggested technique is evaluated using the Face Forensics dataset [105], which consists of one thousand movies.

Deepfake movies, as pointed out by G uera and Delp [112], include both internal (between frames) and external (between frames) irregularities. Next, they suggested a temporally-aware pipeline approach to detecting deepfake films by combining convolutional neural networks (CNNs) with long short-term memories (LSTMs). The LSTM is used to generate a temporal sequence descriptor from the frame-level data extracted by the CNN. To differentiate edited from unedited movies using the sequence descriptor, we deploy a fully-connected network, as shown in Fig. 7. In , researchers used a dataset of 600 films, composed of 300 deepfake movies culled from various video hosting websites and 300 clean videos chosen at random from the Hollywood human activities dataset, and achieved an accuracy of better than 97%.

However, in Li Fig. 7, it was suggested that eye blinking may be used as a physiological indication to identify deepfakes. The sequence descriptor is used in this deepfake detection technique that use a convolutional neural network (CNN) and a long short term memory (LSTM) to extract temporal information from a video sequence. Using the sequence descriptor as input, a detection network made up of fully connected layers determines the likelihood that a given frame sequence is either legitimate or deepfake . et al. found that compared to unaltered movies, people in deepfakes blink far less often. The average human adult blinks between two and ten times in a two-minute period, with each blink lasting between 0.2 and 0.6 seconds. However, deepfake algorithms often train on publicly accessible face photos from the internet, and these images almost always depict humans with wide eyes. Therefore, deepfake algorithms can't make convincing fake faces that can blink naturally unless they have access to photos of individuals blinking. Blinking rates in deepfakes are, however, far lower than those in authentic videos. Li et al. crop eye regions in the films and distribute them into long term recurrent convolutional networks (LRCN) for dynamic state prediction, which they use to distinguish between genuine and false movies. The LRCN uses a CNN-based feature extractor, an LSTM-based sequence learner, and a fully connected layer to estimate the likelihood that a user's eyes are open or closed. Eye blinking has robust temporal relationships, which may be captured electively by using LSTM.

3.2.2. Visual Artifacts within Video Frame

As was said in the preceding section, deep recurrent network models provide the basis of most approaches that use temporal patterns across video frames to identify deepfake films. This section digs into the alternative method, which often breaks down films into frames and analyses visual abnormalities inside individual frames to extract discriminant characteristics. To determine genuine from false films, these attributes are fed into a classifier, either a deep one or a shallow one. In this section, we categorize techniques according on whether they use deep or shallow classifiers. Intelligent, in-depth classifiers. Due to the low quality of most deepfake films, one face warping (scaling, rotation, and shearing) is usually necessary to get them to seem like the real thing. This method leaves artifacts that may be recognized by CNN models like VGG16, ResNet50, ResNet101, and ResNet152 due to the resolution discrepancy between the warped face region and the surrounding environment. In, the authors suggest using deep learning to identify deepfakes by analyzing the artifacts produced by deepfake generating algorithms during the face warping stage. Two deepfake datasets, UADFV and DeepfakeTIMIT, are used to test the effectiveness of the proposed approach.

The UADFV dataset includes a total of 32,752 frames spread over 49 authentic and 49 synthetic videos. The DeepfakeTIMIT dataset has 320 films divided into two quality categories: low quality (64 × 64) and high quality (128 x 128), with a total of 10,537 authentic pictures and 34,023 fake images. Two deepfake detection MesoNet approaches, namely Meso-4 and MesoInception-4, the HeadPose method and the twostream NN method for detecting faces that have been tampered with are used to evaluate how well the suggested system performs. The suggested technique has the advantage that deepfake videos are not required for negative example generation prior to training the detection models. Instead, negative instances are created on-the-fly by isolating the face area, aligning it across several scales, adding a Gaussian blur to a randomly selected scaled picture, and then warping back to the original image. When compared to approaches that need pre-generating deepfakes, this saves a significant amount of time and computing resources. The use of capsule networks to identify digitally altered media was suggested by Nguyen et al.. To overcome the shortcomings of conventional neural network architectures (CNNs), the capsule network was first used in inverted graphics applications. In order to express the hierarchical pose connections between object pieces, the capsule network based on dynamic routing algorithm was recently developed. This innovation is used as a part of a pipeline for detecting fake images and videos, as shown in Fig. 4. The pipeline comprises three capsules, each of which performs a distinct task in the detection of fake images and videos. Four datasets with examples of various types of picture and video forgeries are used to test the approach. Some examples are the Face2Face method facial reenactment dataset Face

Forensics and the fully computer-generated image dataset created by Rahmouni et al.. The well-known Idiap Research Institute replay-attack dataset is another example. In each of these data sets, the suggested technique outperforms the alternatives. This demonstrates the capsule network's promise as a platform for developing a universal detection system that can function effectively against a wide range of fraudulent image and video assaults.

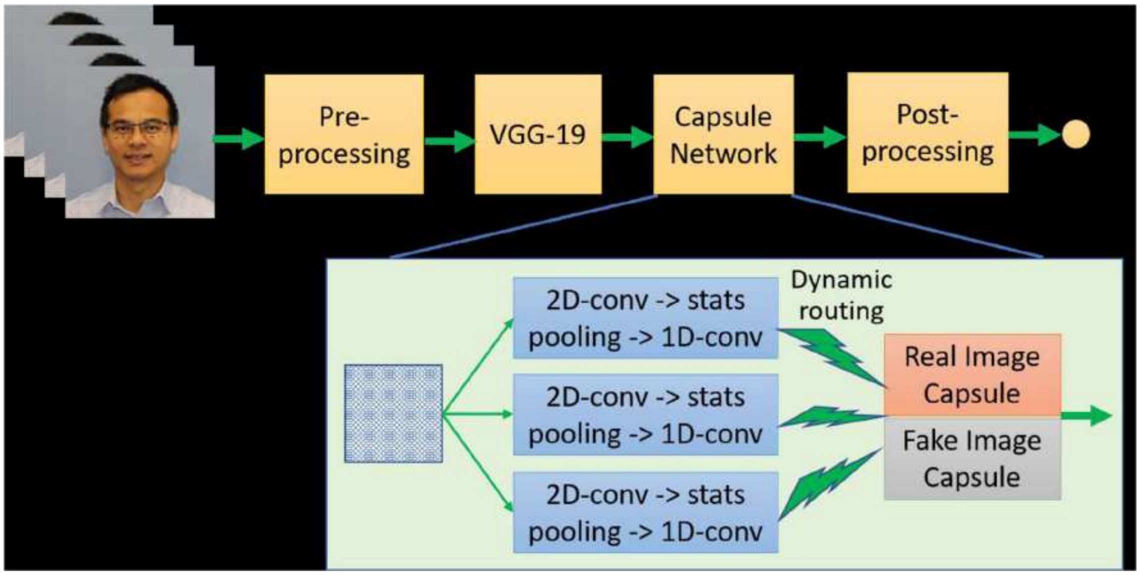
Pre-processing
VGG-19
Capsule Network
Post-processing
2D-conv -> stats pooling -> 1D-conv
2D-conv -> stats pooling -> 1D-conv
2D-conv -> stats pooling -> 1D-conv
Dynamic routing
Real Image Capsule
Fake Image Capsule

Fig. 4. Capsule network

4. Discussions and Future Research Directions

The use of deep learning has made the production of deepfakes more simpler. This proliferation of phony social media sites . It is not always necessary for a deepfake to reach a large audience to have an adverse effect. Without the use of social media, saboteurs whose goal is to build deepfakes for malevolent purposes need to send them to target audiences. Threats to national and international security might result, for instance, when intelligence agencies use this strategy to exert influence on key decision-makers like politicians. Researchers have been hard at work building deepfake detection algorithms, and several positive findings have been published. provides an overview of common techniques used, and the study as a whole provides a review of the state-of-the-art in this area. It's becoming more clear that individuals who employ sophisticated machine learning to construct deepfakes are at odds with those who make the effort to spot them. The quality of deepfakes has been on the rise, thus detection systems need to step up their game. The idea behind it is that if AI can break something, AI can also repair it. Various approaches for detection have been presented and assessed, but only on limited datasets. Developing a regularly updated benchmark dataset of deepfakes to verify the continual improvement of detection techniques is one strategy for improving detection performance. Since deep learning detection models benefit most from a large training set, this is good news for everyone involved in the training process.

5. Conclusions

With the advent of deepfakes, people's faith in the veracity of media reports has started to diminish. They may aggravate political tension, incite public unrest, violence, or even war, as well as inflict grief and unpleasant effects for the targeted individuals. This is particularly important today since deepfake creation tools are easily accessible, and because fake news may spread rapidly on social media. This study contains a wide-ranging debate on difficulties, possible trends, and possible future approaches in the field of deepfake production and detection. Researchers in the field of artificial intelligence may use the findings of this study to create more efficient strategies for countering deepfakes.

REFERENCES

[1] Wiley, Victor and Lucas, Thomas. (2018). Computer Vision and Image Processing: A Paper Review. International Journal of Artificial Intelli- gence Research.

[2] Bharathi, S.Shankar and N.Radhakrishnan, and Prasad, Pinnamaneni. (2013). Machine Vision Solutions in Automotive Industry.

[3] M. H. Wagdy, H. A. KhaIi1 and S. A. Maged, “Swarm Robotics Pattern Formation Algorithms,” 2020 8th International Conference on Control, Mechatronics and Automation (ICCMA), 2020, pp. 12-17.

[4] Badrinarayanan, V., Kendall, A., and Cipolla, R. (2017). SegNet: A deep convolutional encoder-decoder architecture for image segmentation. IEEE Transactions on Pattern Analysis and Machine Intelligence,39(12), 2481-2495.

[5] 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.

[6] TensorFlow. Accessed on: December 19, 2020. Available at https://www.tensorflow.org/

[7] AKaIiyar, R. K., Goswami, A., and Narang, P. (2020). Deepfake: improving fake news detection using tensor decomposition based deep neural network. Journal of Supercomputing.

[8] Deepfake Detection Challenge Results. Accessed on: January 15, 2021. Available at https://ai.facebook.com/blog/deepfake-detection-challenge- results-an-open-initiative-to-advance-at/

[9] Dolhansky, Brian and Howes, Russ and Pflaum, Ben and Baram, Nicole and Ferrer, Cristian.(2019). The Deepfake Detection Chal- lenge(DFDC)Preview Dataset.

[10] Faceswap: Deepfakes software for all. Accessed on: February 2, 2021. Available at https://github.com/deepfakes/faceswap

[11] FakeApp 2.2.0. Accessed on: February 2, 2021. Available at https://www.ma1avida.com/en/soft/fakeapp/

[12] Ballé, Johannes and Laparra, Valero and Simoncelli, Eero.(2016). End- to-end Optimized Image Compression.

[13] Cheng, Zhengxue and Sun, Heming and Takeuchi, Masaru and Katto, Jiro. (2019). Energy Compaction-B ased Image Compression Using Convolutional AutoEncoder. IEEE Transactions on Multimedia. PP. 1-1.

[14] A. Punnappurath and M. S. Brown, “Learning Raw Image Reconstruction-Aware Deep Image Compressors,” in IEEE Transactions on Pattern Analysis and Machine Intelligence, vol. 42, no. 4, pp. 1013- 1019, 1 April

2020.

[15] Petrov, Ivan, et a1. (2020). DeepFaceLab: A simple, flexible and exten- sible face swapping framework. arXiv preprint arXiv:2005.05535, 2020.

[16] Tewari, Ayush, et a1. (2018). High-Fidelity Monocular Face Reconstruc- tion Based on an Unsupervised Model-Based Face Autoencoder. IEEE Transactions on Pattern Analysis and Machine Intelligence.

[17] Deep Fakes, Fake News, and What Comes Next. Accessed on: January 20, 2021. Available at https://jsis.washington.edu/news/deep-fakes-fake- news-and-what-comes-next/

[18] Korshunova, I., Shi, W., Dambre, J., and Theis, L. (2017). Fast faceswap using convolutional neural networks. In Proceedings of the IEEE Inter- national Conference on Computer Vision (pp. 3677-3685).

[19] Mirsky, Yisroel and W. Lee. “The Creation and Detection of Deepfakes.” ACM Computing Surveys (CSUR). vol. 54 no. 1 , April 2021.

[20] D. Afchar, V. Nozick, J. Yamagishi and I. Echizen, ’MesoNet: a Com- pact Facial Video Forgery Detection Network,” 2018 IEEE International Workshop on Information Forensics and Security (WIFS), 2018, pp. 1-7.

[21] V. Schetinger, M. M. Oliveira, R. da Silva, and T. J. Carvalho. Humans are easily fooled by digital images. arXiv preprint arXiv:1509.05301, 2015.

[22] lofte, Sergey and Szegedy, Christian. (2015). Batch Normalization: Accelerating Deep Network Training by Reducing Internal Covariate Shift. arXiv preprint arXiv:1502.03167, 2015.

[23] Siarohin, Aliaksandr, Stéphane Lathuiliére, S. Tulyakov, E. Ricci and
N. Sebe. “First Order Motion Model for Image Animation.” NeurIPS (2019).

[24] Wang X. et a1. ESRGAN: Enhanced Super-Resolution Generative Ad- versarial Networks. In: Leal-Taixé L., Roth S. (eds) Computer Vision ECCV 2018 Workshops. ECCV 2018. Lecture Notes in Computer
Science, vol 11133. Springer, Cham.

[25] Xiaoming Li. et a1. Blind Face Restoration via Deep Multi-scale Com- ponent Dictionaries. arXiv preprint arXiv:2008.00418, 2020.

[26] Xintao Wang and Ke Yu and Kelvin C.K. Chan and Chao Dong and Chen Change Loy, “BasicSR,” 2020, Accessed on: January 15, 2021. Available at https://github.com/xinntao/BasicSR