Introduction to Many-to-Many Face Reenactment and its Detection

Malte Zietlow, Finn Wellershaus

Nordakademie
Köllner Chaussee 11
25337 Elmshorn
malte.zietlow@nordakademie.de
finn.wellershaus@nordakademie.de

1. Introduction

In late 2017, redditor *deepfake* began publishing pornographic footage on `https://www.reddit.com/r/deepfakes`, where he replaced the actors' faces with those of celebrities — without either consent. As a consequence, he received increased media coverage in early 2018, after the release of a first article and interview on vice-motherboard [1]. While he coined the term *DeepFake* for face-*replacement* and face-*reenactment* technologies, both were already previously known in academia with research dating back to at least 1997 [2]. Following this however, there was an increased awareness for *DeepFakes* which caused more research to be conducted, from 3 papers published in 2017 to more than 150 in 2018 – 2019 [3] as well as resulting in first government action [4].

The aim of this paper is a.) to give the reader an introduction to the process of creating manipulated images via deep learning, i.e. creating *DeepFakes* and b.) to give an overview on different approaches to the detection of such manipulated footage.

1.1. Structure

In the remainder of this section, the limitations to the scope of this paper are presented. Then, in section 2, a distinction between *DeepFakes* and manual video manipulation is made, followed by a discussion of ethical challenges. This is followed by a summary of the creation of many-to-many *DeepFakes* in section 3. Finally, different methods for detecting *DeepFakes* can be found in section 4.

1.2. Limitations

While there are multiple *flavors* of *DeepFakes* (see section 2.2), this study explicitly works within the scope of many-to-many face reenactment and its detection. This means that neither *DeepFakes* of text or audio signals, face-replacement or body reenactment are part of this work. This is mainly due to spatial constraints and considering that, in their core principles, the techniques are mostly equal. Beyond that, one-to-one and one-to-many *DeepFakes* (see section 2.3) are not examined. This restriction is made because we consider many-to-many *DeepFakes* to be of greater risk due to their more generic nature and thus larger applicability.

2. Background

2.1. DeepFakes and CheapFakes

In this work, a distinction between so called CheapFakes and *DeepFakes* is made. On the one hand, CheapFakes are produced by using contemporary video and picture editing software. The name CheapFake should under no circumstance be interpreted as an assessment of their effectiveness since they can still be used to effectively fool users. An example of this usage might be a video of US House Speaker Nancy Pelosi which was slowed down for her to appear drunk [5]. On the other hand, *DeepFakes* describe the results of machine-learning algorithms, which are most often a form of or combination of Neural Networks (NNs). The names simply refer to the different levels of sophistication of the different approaches. Arguably, at some point, they might even merge together with editing software utilizing machine learning more and *DeepFake* technology becoming accessible and integrated into existing tooling.

2.2. Flavors of DeepFakes

Broadly, *DeepFakes* can be sorted into two categories, as described by Mirsky and Lee [3]: There is a source s and a target t

Face Reenacment Here, the expression of s is used to drive the expression of t [3]. The source is therefore also called driver d.

Replacement In this case, some part of t is replaced by the corresponding part of s. An example would be the swapping of t's face with the one from s, preserving t's mimic.

2.3. Generalizability of Generative Models for DeepFakes

In the design of NNs for the creation of *DeepFakes*, one must choose the range of targets to which it is applicable [cf. 3]:

one-to-one after training, the approach is applicable only to a single target and source/driver;

many-to-one after training, the approach is applicable only to a single target, but arbitrary sources/drivers;

many-to-many after training, the approach is applicable both to arbitrary targets and sources/drivers.

As mentioned in section 1.2, only **many-to-many** approaches are presented in this paper because they are more generic than **one-to-one** or **many-to-one**.

2.4. Ethical Challenges

The importance of *DeepFake* detection increases in conjunction with the ethical challenges which arise from the use of this technology. The ability to create arbitrary fake images bears multiple risks for misuse. Firstly, there is the aspect of misinformation and propaganda. For example by showing images or videos of political candidates saying things they have not said and that can incriminate them. Moreover, another aspect to take into consideration would be pornography, since it is possible to change faces to other people's bodies, which can lead to non-consensual pornography. Following these dimensions, there are two types of attacks: to public or political figures and to private individuals. The first category is defined by attackers having more resources at hand for an attack, e.g. an organized action against a specific political candidate. The second category, in contrast, is defined by the attackers' lack of adequate resources, e.g. an angry ex-boyfriend. Even though the second category might pose a real risk since not that much effort might be needed to spread a basic amount of false information, in the end, the attacker's capabilities are somewhat limited. However, changes in the ease of use and availability of *DeepFake* technologies can affect its uses towards targeting private individuals.

3. Creation of DeepFakes

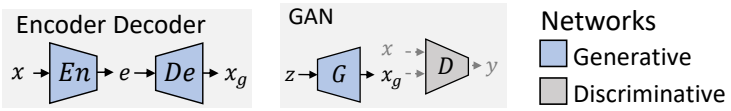
Statistical models come in multiple classes with divergent properties. Two of such classes are discriminative and generative models. Generative models learn the statistical properties of their input domain [cf. 6, 651 sqq.]. For example, they are able to generate unique images of human faces, based on previously observed ones [7].

Generative models are thus the focus of techniques for creating *DeepFakes*. As this serves as an introduction to the creation of *DeepFakes*, we limit the scope to:

Encoder-Decoders (EDs) EDs consist of at least two networks, Encoder (En) and Decoder (De). The En-part is trained to extract useful features $\mathrm{En}(x) = e$ from input image x. This is often done by narrowing layer-width towards its center (see fig. 3.1a) or some other regularization criterion [cf. 6, pp. 499–505]. After encoding $x \to e$, e is fed into De to arrive at the final result $\mathrm{De}(e) = x_g$. Autoencoders (AEs) are a special form

of ED, where the network is trained to reproduce its inputs, i.e. $\mathrm{De}(\mathrm{En}(x)) \approx x$.

Generative Adversarial Networks (GANs) GANs, like ED, consist of at least two networks, Generator (G) and Discriminator (D) (see fig. 3.1b). In a sense, G can be interpreted similar to the De part of the ED. But, instead of relying on En to extract features $\mathrm{En}(x) = e$, the input z is sampled from a so called prior-distribution $z \sim p_z$, e.g. the normal distribution $p_z = \mathcal{N}(0, 1)$. Then, using z, G generates images $\mathrm{G}(z) = x_g$. D, on the other hand, tries to discern whether an input is from the identity of the target $x \in X_{\text{target-id}}$ or artifically generated by G, i.e. x_g. G and D are then trained in alternation, such that G tries to fool D into *thinking* that x_g is an image from the target $X_{\text{target-id}}$ and D trying to discern correctly between $x \sim X$ and $x_g \sim \mathrm{G}(p_z)$. That way, we are able to generate realistically looking faces similar to that of the target [8].



(a) (b) (c)

Figure 3.1: Selection of generative models, adopted from [3]

When trying to create believable face reenactment footage, there are three **goals** we try to reach in the final generated image x_g:

1.) to preserve the mimic of driver x_d in x_g, s.t. $\mathrm{mimic}(x_d) = \mathrm{mimic}(x_g)$

2.) to preserve the identity target x_t in x_g, s.t. $\mathrm{id}(x_t) = \mathrm{id}(x_g)$

3.) to generate a realistically looking image

We can take these goals and construct a model after them. In order to generate an image we utilize the ED-model from fig. 3.1a

3.1. Preserving Mimic

To tackle goal 1.), one must define how to numerically represent the mimic of a face, $\mathrm{mimic}(x) = \textit{undefined}$. Two such representations are a.) via facial landmarks or b.) via the Facial Action Coding System (FACS) For simplicity, only FACS is considered. A discussion of facial landmarks can be found in e.g. [9].

Facial Action Coding System (FACS) The FACS was introduced in 1978 by Ekman and Friesen [10]. They grouped muscle regions of the face to 58 so called Action Units (AUs), which can be represented by vectors. A major advantage of AUs over facial landmarks is that they are (more or less) invariant (i.e. the same expression is encoded similarly) between faces, head angle and scale [11].

Following this discussion, FACS is used. AUs might be computed using a generic Action Unit Estimator (AUE), e.g. [12]. The AU-vectors are then concatenated to the encoded

representation ($\text{En}(x_t) = e$) and used in the De-step. To enforce goal 1.), the difference in AUs between final *DeepFake* $\text{AUE}(x_g) = a_g$ and driver $\text{AUE}(x_d) = a_d$ are minimized in training: $\min(|a_g - a_d|)$. A visualization of this is given in fig. 3.2.

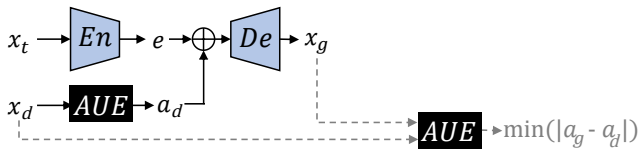


Figure 3.2: ED-architecture enriched with AUs, where $\bigoplus$ denotes concatenation. Grey paths are used solely in training. Inspired by [3], [11]

3.2. Preserving Identity

Training the model on preserving the identity of the target requires a training set of known targets, e.g. [13]–[15] which together consist of about 2000 identities [11]. One can then learn an image classifier I, e.g. MobileNet [16] (for its simplicity and speed) to assign the correct identity to a given image. Once the classifier I converges, it can be added to the training procedure of the model. The ED model can then be trained to also minimize the cross-entropy identity loss $\mathcal{L}_{id}$ (see eq. (1)). A visualization of this is given in fig. 3.3.

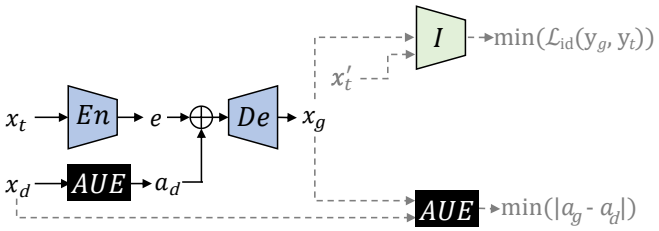


Figure 3.3: Architecture from fig. 3.2, enriched with an identity constraint, where x'_t is a randomly chosen image with identity of the target t. Inspired by [3], [11]

3.3. Increasing Realism

Finally, to increase realism (goal 3.)), a GAN-inspired discriminator D (see fig. 3.1b) is adopted and trained alternatingly to discern between real and generated images, minimizing the adversarial loss $\mathcal{L}_{adv}$ (see eq. (2)). A visualization of this is given in fig. 3.4.

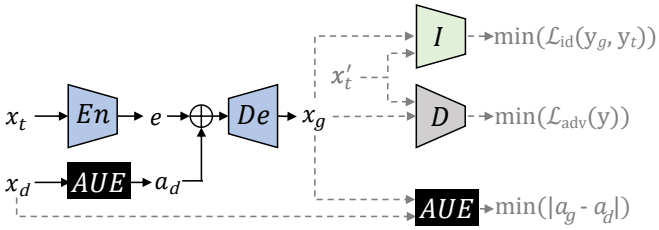


Figure 3.4: Architecture from fig. 3.3, enriched with a realism constraint. Inspired by [3], [11]

4. Detection of DeepFakes

The methods of detecting *DeepFakes* usually do not inhibit major differences between cases like face swapping or face reenactment. Instead, they focus on more general concepts like artifacts which are introduced while using GANs or more general aspects of *DeepFakes*. The area of *DeepFake* detection can be separated in different approaches. Hereafter, there will be a broad introduction to each approach with an example; however, there are many more different shapes of these approaches.

4.1. Physical Approach

A human on average blinks every two to ten seconds with a duration of $0.1 - 0.4$ seconds [17]. Based on this, spontaneous eye blinking would be expected in *DeepFake* videos. This, however, is not always the case. An explanation for this circumstance lies in the training data for GANs. As Li, Chang and Lyu [17] describe, with an exposure time of 1/30 of a second, the likelihood of being captured on a picture while blinking is around 7.5% [17]. In combination with the fact that most people might prefer pictures of them when they do not blink online, is commonly thought to be one of the major reasons for this discrepancy [18]. Based on this observation, measuring the frequency of blinking can serve as a method of detection of *DeepFake* videos.

In order to determine the state of the eye, it its necessary to first determine where the eye is located. This is done in the preprocessing-phase. Here, the facial landmarks of the pictures are extracted from each face area [17]. To deal with changes in the orientation of the face, there is first an alignment of the face in a unified coordinate space which moves the face towards the center of the image [17], provided it is not already centered.

The next step is the prediction of the eye's new state. For this, a Long-Term Recurrent Convolutional Neural Network (LRCN) [19] is used. This is because the blinking of human eyes displays strong dependencies to previous and next images. The LRCN-model enables capturing this by using a combination of Convolutional Neural Networks (CNNs) and Long Short-Term Memorys (LSTMs) [19]. The CNN is responsible for the feature detection, in this case the eye region, into discriminative features [17], [19]. These are fed into the sequence learning which is implemented as a LSTM. This enables considering information from previous eye states. The result of this step is then passed to another NN containing a fully connected layer which has generated the probability of an eye being closed or open [17].

This is only an example of using physical human characteristics to determine a video or pictures' authenticity. Li, Chang and Lyu [17] describe in their paper critically how this might be circumvented by improving the *DeepFake* generation models to also incorporate eye blinking. Here there are possibilities to a.) also check for frequency of blinking [17] or b.) change to completely different approaches also used in medicine for checking skin color to check for heart rate or similar aspects [18].

4.2. Artifact-Based Approach

The algorithms for creating *DeepFakes* might introduce artifacts in the resulting images. These artifacts can be used to discover *DeepFakes* and lay the foundation for this type of detection.

Amerini, Galteri, Caldelli *et al.* [20] describe one of these approaches utilizing optical flow between different frames. Optical flow, also called motion estimation, aims at the discovery of motion between an observed object and the observer in two consecutive frames [21]. This is used to find discrepancies between real images and *DeepFakes* in regard to the movement of the face [20]. Amerini, Galteri, Caldelli *et al.* [20] tested their models on the FaceForensics++ data with a resulting accuracy of 81.61%. Li, Yang, Sun *et al.* [22] found that on this data set also had the comparatively highest Area under the Curve (AUC) of 82.3%. This hints to the model not being that successful, verifying its performance on a more challenging data set like CELEB-DF [22] poses an interesting direction for future research.

Another approach which showed in [22] the overall highest average AUC was Dual Spatial Pyramid-Face Warp Artifacts (DSP-FWA), which was introduced by [23] and focuses on discovering face warping artifacts which are introduced by the resizing operations in *DeepFake* creation algorithms [22], [23].

Similar to the physical approach, this one also requires constant adaptation to improvements in the generative models which only need to improve to incorporate dealing with the artifacts which are used for *DeepFake* detection[3].

4.3. Undirected Approaches

In these approaches, the focus is not on the artifacts, but on training generic NNs, which can select relevant features, informing about whether or not an image is fake. Xuan, Peng, Wang *et al.* [24] introduced blur and noise to the training phase to force their model to focus on more intrinsic aspects which in turn should improve the generalization abilities. This is contrary to other approaches which focus on e.g. pixel density as described by Mirsky and Lee [3].

5. Conclusion

There are different approaches to dealing with *DeepFakes*, some more promising than others. We consider methods of detection to be only one part of a possible solution since, on their own, they would possibly only lead to a race between NNs for the creation and for the detection of *DeepFakes*, both getting more and more sophisticated. In this aspect, we agree with Mirsky and Lee [3]'s ideas that out-of-band methods for signatures of multimedia content and other prevention mechanisms are required for a better solution. However, a sole focus on technical solutions might also be futile. More effective might

be to, at the same time, try to adapt on a legal and society level to this new situation, being aware of the existence of possibly false media content.

References

[1] S. Cole, *Vice-Motherboard: AI-Assisted Fake Porn Is Here and We're All Fucked*, vice.com, Ed., 2017. [Online]. Available: `https://www.vice.com/en_us/article/gydydm/gal-gadot-fake-ai-porn` (visited on 12/08/2020).

[2] C. Bregler, M. Covell and M. Slaney, 'Video Rewrite: driving visual speech with audio', in *Proceedings of the 24th Annual Conference on Computer Graphics and Interactive Techniques, SIGGRAPH 1997, Los Angeles, CA, USA, August 3-8, 1997*, G. S. Owen, T. Whitted and B. Mones-Hattal, Eds., ACM, 1997, pp. 353–360.

[3] Y. Mirsky and W. Lee, 'The Creation and Detection of Deepfakes: A Survey', *CoRR*, vol. abs/2004.11138, 2020.

[4] Senate - Homeland Security and Governmental Affairs and House - Energy and Commerce, *Deepfake Report Act of 2019*, 2019. [Online]. Available: `https://www.congress.gov/bill/116th-congress/senate-bill/2065/text` (visited on 14/08/2020).

[5] A. Fichera, *Manipulated Video Targeting Pelosi Goes Viral*, 2019. [Online]. Available: `https://www.factcheck.org/2019/05/manipulated-video-targeting-pelosi-goes-viral/` (visited on 17/08/2020).

[6] I. Goodfellow, Y. Bengio and A. Courville, *Deep Learning*. MIT Press, 2016.

[7] T. Karras, S. Laine and T. Aila, 'A Style-Based Generator Architecture for Generative Adversarial Networks', in *IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2019, Long Beach, CA, USA, June 16-20, 2019*, Computer Vision Foundation / IEEE, 2019, pp. 4401–4410.

[8] I. J. Goodfellow, J. Pouget-Abadie, M. Mirza, B. Xu, D. Warde-Farley, S. Ozair, A. C. Courville and Y. Bengio, 'Generative Adversarial Nets', in *Advances in Neural Information Processing Systems 27: Annual Conference on Neural Information Processing Systems 2014, December 8-13 2014, Montreal, Quebec, Canada*, Z. Ghahramani, M. Welling, C. Cortes, N. D. Lawrence and K. Q. Weinberger, Eds., 2014, pp. 2672–2680.

[9] S. Ha, M. Kersner, B. Kim, S. Seo and D. Kim, 'MarioNETte: Few-Shot Face Reenactment Preserving Identity of Unseen Targets', in *The Thirty-Fourth AAAI Conference on Artificial Intelligence, AAAI 2020, The Thirty-Second Innovative Applications of Artificial Intelligence Conference, IAAI 2020, The Tenth AAAI Symposium on Educational Advances in Artificial Intelligence, EAAI 2020, New York, NY, USA, February 7-12, 2020*, AAAI Press, 2020, pp. 10 893–10 900.

[10] P. Ekman and W. Friesen, 'Facial action coding system', 1978.

[11] H. X. Pham, Y. Wang and V. Pavlovic, 'Generative Adversarial Talking Head: Bringing Portraits to Life with a Weakly Supervised Neural Network', *CoRR*, vol. abs/1803.07716, 2018.

[12] T. Senechal, D. J. McDuff and R. E. Kaliouby, 'Facial Action Unit Detection Using Active Learning and an Efficient Non-linear Kernel Approximation', in *2015 IEEE International Conference on Computer Vision Workshop, ICCV Workshops 2015, Santiago, Chile, December 7-13, 2015*, IEEE Computer Society, 2015, pp. 10–18.

[13] B.-C. Chen, C.-S. Chen and W. H. Hsu, 'Face Recognition and Retrieval Using Cross-Age Reference Coding With Cross-Age Celebrity Dataset', *IEEE Trans. Multimedia*, vol. 17, no. 6, pp. 804–815, 2015.

[14] C. Cao, Y. Weng, S. Zhou, Y. Tong and K. Zhou, 'FaceWarehouse: A 3D Facial Expression Database for Visual Computing', *IEEE Trans. Vis. Comput. Graph.*, vol. 20, no. 3, pp. 413–425, 2014.

[15] F. Tarres and A. Rama, *GTAV face database*, 2011.

[16] A. G. Howard, M. Zhu, B. Chen, D. Kalenichenko, W. Wang, T. Weyand, M. Andreetto and H. Adam, 'MobileNets: Efficient Convolutional Neural Networks for Mobile Vision Applications', *CoRR*, vol. abs/1704.04861, 2017.

[17] Y. Li, M.-C. Chang and S. Lyu, 'In Ictu Oculi: Exposing AI Created Fake Videos by Detecting Eye Blinking', in *2018 IEEE International Workshop on Information Forensics and Security (WIFS)*, 2018, pp. 1–7.

[18] A. Pishori, B. Rollins, N. v. Houten, N. Chatwani and O. Uraimov, *Detecting Deepfake Videos: An Analysis of Three Techniques*. 2020.

[19] J. Donahue, L. A. Hendricks, M. Rohrbach, S. Venugopalan, S. Guadarrama, K. Saenko and T. Darrell, *Long-term Recurrent Convolutional Networks for Visual Recognition and Description*. 2014.

[20] I. Amerini, L. Galteri, R. Caldelli and A. D. Bimbo, 'Deepfake Video Detection through Optical Flow Based CNN', in *2019 IEEE/CVF International Conference on Computer Vision Workshop (ICCVW)*, 2019, pp. 1205–1207.

[21] S. S. Beauchemin and J. L. Barron, 'The Computation of Optical Flow', *ACM Comput. Surv.*, vol. 27, no. 3, pp. 433–466, Sep. 1995, ISSN: 0360-0300.

[22] Y. Li, X. Yang, P. Sun, H. Qi and S. Lyu, *Celeb-DF: A Large-scale Challenging Dataset for DeepFake Forensics*. 2019.

[23] K. He, X. Zhang, S. Ren and J. Sun, 'Spatial Pyramid Pooling in Deep Convolutional Networks for Visual Recognition', *Lecture Notes in Computer Science*, pp. 346–361, 2014, ISSN: 1611-3349.

[24] X. Xuan, B. Peng, W. Wang and J. Dong, *On the generalization of GAN image forensics*. 2019.

Glossary

redditor a contributor to the webpage `https://www.reddit.com`. 1

Acronyms

AE Autoencoder. 3

AU Action Unit. 4, 5

AUC Area under the Curve. 7

AUE Action Unit Estimator. 4

CNN Convolutional Neural Network. 6

D Discriminator. 4, 5

De Decoder. 3–5, 11

DSP-FWA Dual Spatial Pyramid-Face Warp Artifacts. 7

ED Encoder-Decoder. 3–5, 11

En Encoder. 3, 4, 11

FACS Facial Action Coding System. 4

G Generator. 4

GAN Generative Adversarial Network. 4–6, 11

LRCN Long-Term Recurrent Convolutional Neural Network. 6

LSTM Long Short-Term Memory. 6

NN Neural Network. 2, 6, 7

Appendix

A. Identity loss

$$\mathcal{L}_{id}(\hat{y}_g, y_t) = -\sum y_t, \log(\hat{y}_g) \tag{1}$$

where:

$\hat{y}_g$ is the predicted identity of the generated DeepFake $I(x_g) = \hat{y}_g$
y_t is the correct identity of the target x_t

B. Adversarial loss

$$\mathcal{L}_{adv}(y) = \max_{ED} \min_{D} \begin{cases} (1 - D(y))^2, & \text{if } y = D(x'_t) \\ (D(y))^2, & \text{if } y = D(x_g) \end{cases} \tag{2}$$

where:

ED is the combination of En and De as discussed in section 3
D is the GAN-like discriminator as discussed ibid.
x'_t is a random image of the target identity
x_g is the image generated by the ED