COMPUTER VISION EVALUATION
REPORT ON THE FACEFORENSICS++ PAPER

Quentin Le Roux
MSc in Data Science & Artificial Intelligence
Université Côte d'Azur
quentin.leroux@edhec.com

ABSTRACT

FaceForensics++ is a seminal paper in computer vision and specifically facial forgery detection. The present report offers an overview of the paper, and the namesake dataset it was published with, as well as a re-implementation of its best model based on the XceptionNet architecture. We also cover more recent developments as well as try to outperform its results.

1 INTRODUCTION

Image manipulation is an old field of research with roots hearkening back to the the birth of photography itself. One of the earliest and most well-known image manipulations is that of US President Abraham Lincoln's lithography portrait where the US politician John Calhoun body's was doctored in to replace his (See Figure 1).

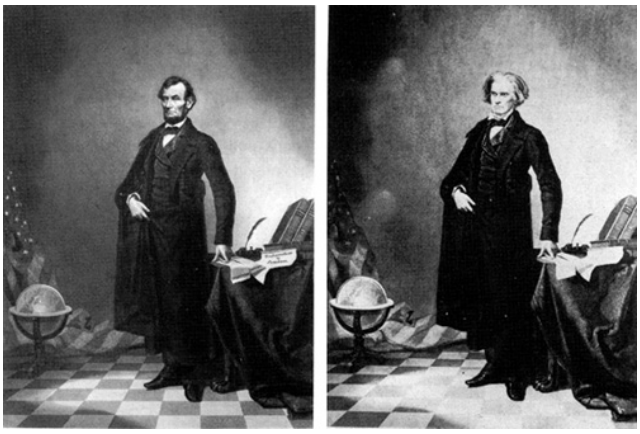


Figure 1: Lithography manipulation of US President A. Lincoln circa 1860 (PRI, 2008)

Over the past decade, the field of computer vision underwent a revolution with the development of new methods thanks to Deep Learning. The computational barrier to entry of image detection as well as manipulation lowered, making such models and methods ubiquitous both as products and a tools. This has raised concerns, however, notably with regards to their potential social, political, and economic impact. The paper notes loss of trust and broader harm to society as an example risk (Rössler et al., 2018).

The paper outlines two main types of image manipulations, specifically regarding facial features (Rössler et al., 2018):

- Expression manipulation
- Identity manipulation

Expression manipulation targets the modification of facial expressions, or their transfer between two persons. Meanwhile, identity manipulation aims at replacing a face entirely either by pasting over, replacing, or replicating another.

Those two types of methods can be applied to images or videos (Korshunova et al., 2016). As such, the reliable and automatic detection of forgeries (the paper specifically mentions the DeepFakes project(deepfakes, 2018)) has become a prominent area of research as newer doctoring methods must be met with more efficient detection models.

2 THE STATE-OF-THE-ART AT THE TIME OF PUBLICATION (OCTOBER 2019)

2.1 FORGERY METHODS

Face manipulation techniques can be broadly sorted into two families (See Table 1). The oldest relies on non deep-learning methods while the other, more recent, group of techniques heavily relies on Generative Adversarial Networks.

Non deep-learning face manipulation methods		
Type	**Name or use**	**Papers**
Video generation	Video Rewrite	Bondi et al. (2017)
Face replacement	Video Face Replacement	Dale et al. (2011)
	VDub	Garrido et al. (2015)
	Real-time expression transfer for facial reenactment	Thies et al. (2015)
	Face2Face	Thies et al. (2018b)
Virtual Reality	FaceVR	Thies et al. (2018a)
	Headon	Thies et al. (2018c)
Reenactment	Deep Video Portraits	Kim et al. (2018)
	NeuralTextures	Thies et al. (2019)
	Bringing Portraits To Life	Averbuch-Elor et al. (2017)
Audio-video syncing	Synthesizing Obama	Suwajanakorn et al. (2017)
Deep-learning face manipulation methods		
Type	**Name or use**	**Papers**
GANs	Face Aging	Antipov et al. (2017)
	Frontal View Synthesis	Huang et al. (2017)
	Face attribute alteration	Lu et al. (2017)
	Deep Feature Interpolation	Upchurch et al. (2017)
Fader Networks	Fader Networks	Lample et al. (2018)

Table 1: State-of-the-art of face manipulation techniques

2.1.1 GRAPHICS-BASED METHODS: FACE2FACE

Non deep-learning methods rely on first tracking the location of a face in two monocular videos (i.e. there are no editing cuts, the video is a single shot without change of angle): the source and target videos. Once the face tracking model is at play, a face transfer model of the tracked features is performed.

Such a method type is exemplified by the Face2Face process (Thies et al., 2018b), one of the main sources of data of the paper at hand here.

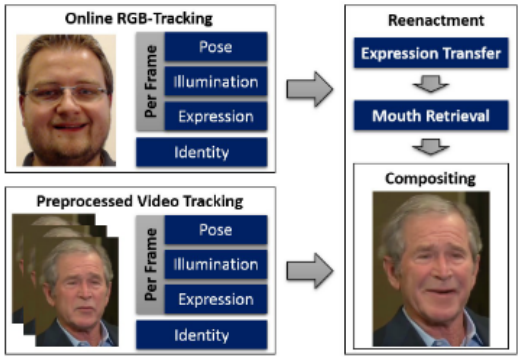


Figure 2: Method of the Face2Face algorithm (Thies et al., 2018b)

Face2Face relies on a so-called parametric face model, (a parameter vector P, see Figure 3, that captures the rigid pose of the face (6 degrees of freedom), its shape (80 dof), its albedo/identity (80 dof), its expression (76 dof), and its illumination (27 dof), totalling 269 parameters).

Figure 3: Parametric face model of the Face2Face method (Thies, 2016)

The process relies on graphics card processing to model/rasterize the image and perform a face capture from a source and target videos. Face capture relies on an energy formulation per frame (see Figure 4) that estimates the different parameters of the face to be modified, smoothed over several keyframes using a technique called non-rigid model-based bundling (see Figure 5), which relies on iterative reweighted least squares. The end result is the capture of the geometry and texture of a person's face and the ability to track its movement throughout a video.

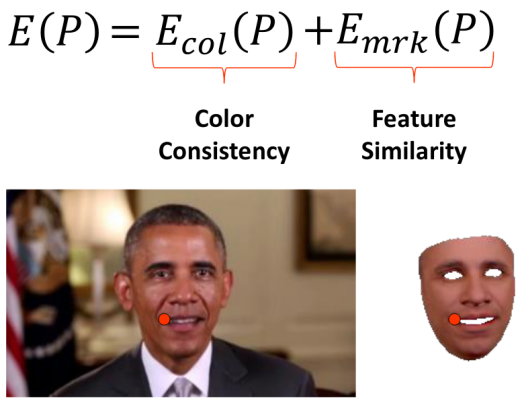


Figure 4: Energy formulation of the Face2Face method (Thies, 2016)

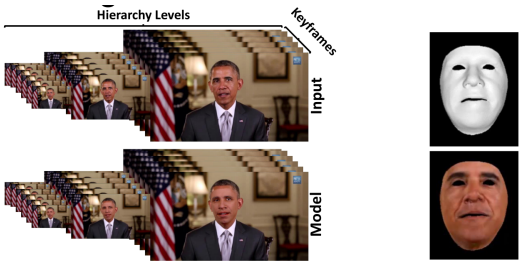


Figure 5: Use of sets of keyframes to compute a face with the Face2Face method (Thies, 2016)

Once the tracked face is modeled, the parameters of the tracking can be modified in real-time with imported parameters from a source face, i.e. to perform facial feature transfers and reenactments. This phase relies on two steps: an expression transfer, and a mouth retrieval (see Figure 2). This is done per frame in a video. The geometry of the target face is modified to fit the source's face parameters while the target's face texture is kept.

The mouth retrieval phase is used to fill the inside of a mouth during reenactment and relies on the use of a database computed from the target's video (See Figure 6). The new face features and the interior of the mouth are composed together into an end result (the mouth interior is retrieved via k-nearest neighbors between the source's mouth and the created database).

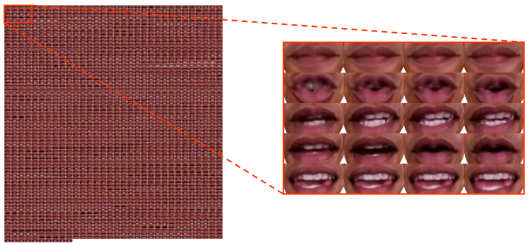


Figure 6: Mouth database being produced with the Face2Face method (Thies, 2016)

The method is limited in some regards, assuming a smooth illumination, the absence of face occlusion (e.g. a hand, mask, etc.), and a training phase with regards to the computing of a mouth database. Furthermore, personal details are lost in the process, such as wrinkles.

2.1.2 Graphics-based methods: FaceSwap

Similar to Face2Face, the FaceSwap methods relies on detecting facial features, or landmarks, to identify the face region in a source video. Given a pre-constructed three-dimensional facial template, the source faces is blended in and projected onto a target image where facial features have also been detected.

The projection relies on minimizing the difference between the source template and the target. Once the transfer has been performed and blended in the target video, the image is color-corrected.

2.1.3 Learning-based methods: DeepFakes

The DeepFakes method relies on learning the shape of a face from a source video and pasting it in a target video using an autoencoder setup. Two autoencoders are typically used on source and target images or videos, where feature maps of the corresponding cropped faces are learned.

With such a setup (see Figure 7), and once the training phase is done, the encoder of the source-processing autoencoder and the decoder of the target-processing autoencoder are combined. As such, the context (feature map) of the source video or image is used to reconstruct the original data of the target video or image.

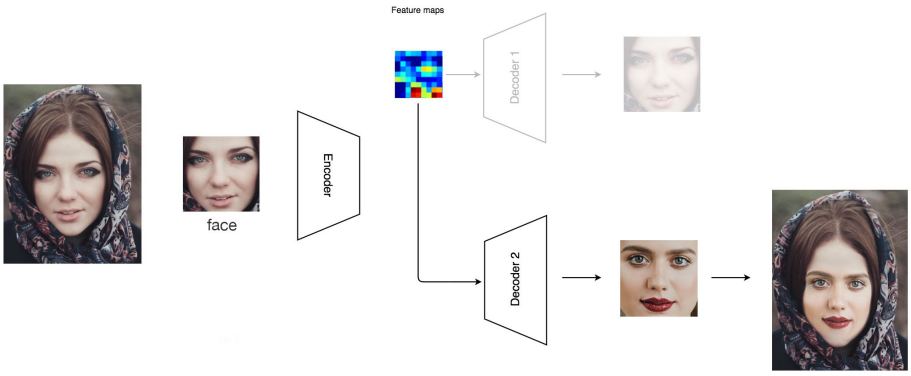


Figure 7: Autoencoding process to produce DeepFakes images (Hui, 2018)

A blending/cleaning step is then performed, yielding a DeepFake content.

2.1.4 Learning-based methods: NeuralTextures

NeuralTextures is a learning-based methods where a neural network is used to learn facial renderings in a generative adversarial setting. Generative Adversarial Neural networks are a type of unsupervised machine learning process. GAN methodology relies on two competing models that competes against each other in order to retrieve and copy variations in a dataset. They are called discriminator and generator networks.

Usually the discriminator and generator are both convolutional neural networks. The problem set by a GAN is that the generator tries to maximize the discriminator's error rate while the discriminator tries to minimize it. This result in an adversarial setup that gives the method its namesake.

In the case of image forgery, a generator networks will try to learn to produce fake images (based on a random noise vector) which will be tested by the discriminator network for authenticity. The discriminator is trained on a set of real images plus the fake images created by the generator while the generator is trained on the binary feedback of the discriminator (See Figure 8, Silva (2018)).

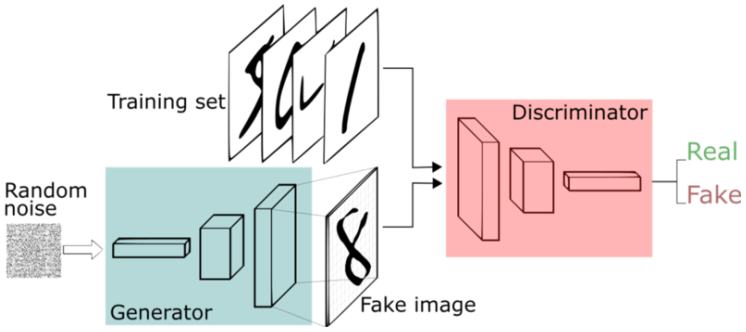


Figure 8: Generic representation of a GAN network trying to generate fake digit pictures based on the MNIST dataset (Silva, 2018)

In the case of NeuralTextures used in the paper, the method relies on the graphics-based method Face2Face in order to track and generate facial features besides the mouth area. The mouth area is generated using the GAN approach instead. The goal is to reduce the computational needs of the method.

2.2 FORGERY DETECTION METHODS

The team identified two types of forgery detection methods:

- Expert-crafted steganographic features (hypothetical-deductive process)
- Learned features (inductive process)

However, both may target the same sets of evidence and hints in order to distinguish a forgery from a real image (see Table 2).

Generic artifacts	
Type	**Papers**
Dropped, duplicated frames	Long et al. (2017)
Interpolation	Ding et al. (2018)
Copy-move (part of an image is reused/copied to another part)	D'Amiano et al. (2019)
Chroma-key (Blending of different videos together such as foreground and background)	Mullan et al. (2017)
Face artifacts	
Type	**Papers**
Computer graphics	Rahmouni et al. (2017)
Morphed face	Raghavendra et al. (2017)
Face splicing	Carvalho et al. (2016)
Face swapping	Zhou et al. (2017)
Eye blinks	Li et al. (2018)
color/texture/shape cues	Carvalho et al. (2016)
DeepFakes	Güera & Delp (2018)

Table 2: State-of-the-art of face manipulation techniques

Most modern forensics methods rely on convolutional neural networks to perform detection. They show "impressive results" but may be lacking robustness in the presence of common, practical cases such as resizing and compression – common operations on online platforms.

2.3 DATASETS AVAILABLE

Historically, image forensics datasets were the result of "significant manual efforts under controlled conditions". The goal was to generate sets of images where specific forgery artifacts would manifest. Video footage was rarely addressed. Examples are displayed in Table 3.

Dataset name	**Content**	**Paper**
MICC_F2000	700 images	Amerini et al. (2011)
Wild Web Dataset	90 real cases	Zampoglou et al. (2015)
Realistic Tampering	220 images	Korus & Huang (2017)
DeepFakes	620 videos from 43 subjects	Korshunov & Marcel (2018)
NIST	50,000 images, 500 videos	Guan et al. (2019)

Table 3: Examples of classical image forensics datasets

3 THE AUTHORS' MOTIVATIONS

In this context, the paper's team (Andreas Rössler, Davide Cozzolino, Luisa Verdoliva, Christian Riess, Justus Thies, and Matthias Nießner) wished to push the literature on the topic of forgery detection and computer vision overall. As such, the authors wanted to contribute in four different aspects to the field:

- A realistic, automated benchmark for future forgery detection techniques
- A robust assessment of the currently available detection methods

- A large scale, industry-shaping dataset
- A new, state-of-the-art method for detecting facial feature doctoring

4 THE PAPER'S PROPOSED APPROACH

4.1 PROPOSING A NEW STATE-OF-THE-ART DATASET: FACEFORENSICS++

The FaceForensics++ datasets aims to cover realistic scenarios: "videos from the wild, manipulated and compressed with different quality levels". To this end, the paper's author proposed a new, state-of-the-art dataset that is an order of magnitude bigger than previous iterations (more than 1.8 million images drawn from c. 4000 fake videos, see Table 4).

Methods	**Train**	**Validation**	**Test**
Pristine	366,847	68,511	73,770
Face2Face	366,843	68,511	73,770
FaceSwap	291,434	54,618	59,640
DeepFakes	366,835	68,506	73,768
NeuralTextures	291,834	54,630	59,672

Table 4: Number of images per manipulation method

The goal of such a dataset is to enable the training of newer, better forensic tools to apprehend the new forgery methods that deep-learning has notably enabled.

In order to construct this dataset, the authors relied on videos from the YouTube platform (YouTube-8m dataset, videos with the tags "face", "newscaster", "newsprogram", "interview", "[video] blog"), manually screened to ensure the current forgery method will work (see Figure 9).

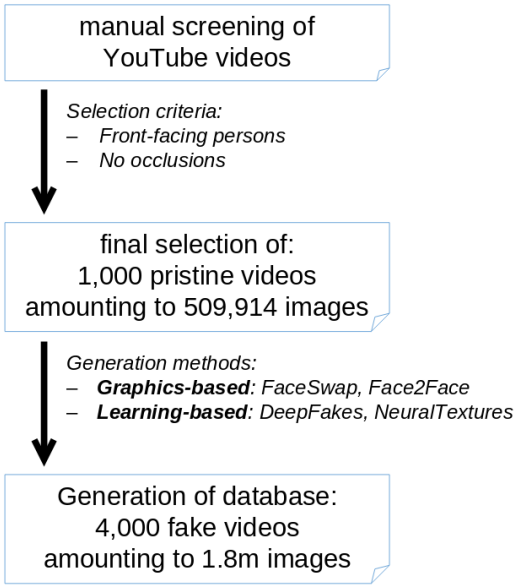
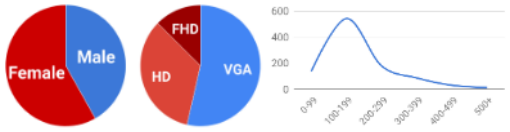


Figure 9: Dataset generation process

A further selection criteria focuses on diversity of format, for example relating to video quality, to emulate the varied content that can be found online (see Figure 10). The dataset construction relies on the 4 state-of-the-art methods previously presented: FaceSwap, Face2Face, DeepFakes, and NeuralTextures.

Whereas the DeepFakes manipulates each frame in a target video, the FaceSwap and NeuralTextures methods doctor the minimum amount of frames across both source and target videos. Finally, Face2Face maps all expressions of the source video onto the target video.



(a) Gender (b) Resolution (c) Pixel Coverage of Faces

Figure 10: Statistics of the FaceForensics++ sequences (Rössler et al., 2018). VGA denotes 480p, HD denotes 720p, and FHD denotes 1080p resolution of the videos. The graph (c) shows the number of sequences (y-axis) with given bounding box pixel height (x-axis).

4.2 Formalizing the problem of forgrey detection

The authors represent the problem of forgery detection as a per-frame binary classification with regards to video forensics. Based on the produced database, the authors split the data between training, validation and test sets with the following ratios: 72%, 14%, 14%.

In order to formalize the problem, the authors first performed a human trial with computer science students in order to construct a human baseline for the detection of forgeries. Given a set times to classify a random set of fake and pristine images, humans struggled to classify images produced by the Face2Face and NeuralTexture methods.

Once the human baseline is constructed, the authors tested a thorough selection of learning-based and steganalytic (hand-crafted features followed by a SVM) methods. Compared to the humans, who were presented full pictures, the automatic forgery detection methods were presented images that underwent a first step of processing, called "domain-specific knowledge", where the images have been conservatively cropped to the face region using the face tracking method proposed by Thies et al. (2020).

Overall, the automatic methods covered by the paper are listed in Table 5. Each was trained using the hyperparameters listed in Table 6 using the Adam optimizer with default values. The transfer learning process used for the XceptionNet model followed two phases: a first 3-epoch training where only the last, newly-inserted fully connected layer is trained, then a 15-epoch training phase where all the weights are trained.

Furthermore, the model were trained and tested in two given situation, either on one set of data, or on all four at once.

Type	**Method**	**Paper**
Steganalysis	Hand-crafted processed followed by SVM	Fridrich & Kodovsky (2012)
Learned	Steganalysis features of the previous method as a starting point to a CNN	Cozzolino et al. (2017)
	Constrained CNN	Bayar & Stamm (2018)
	Stats-2L CNN	Rahmouni et al. (2017)
	MesoInception-4 (based on InceptionNet)	Afchar et al. (2018)
	XceptionNet (Transfer Learning)	Chollet (2017)

Table 5: Methods covered in the paper

For computing accuracy of the learning-based methods, the authors implemented an averaging process over 100 images per video, during validation and testing, and 270 images per video during training. Finally, they implemented an early-stopping heuristic where accuracy is evaluated 10 times during validation. If no improvement is recorded, the training stops. The imbalance in the dataset (more fakes than pristine videos) is dealt by weight-balancing the training images.

Model	Learning rate	Batch size
Cozzolino et al. (2017)	10^{-5}	16
Bayar & Stamm (2018)	10^{-5}	64
Rahmouni et al. (2017)	10^{-4}	64
Afchar et al. (2018)	10^{-3}	76
Chollet (2017)	0.0002	32

Table 6: Method hyperparameters

4.3 Proposing a benchmark

Using a further 1,000 pristine videos from YouTube as a baseline, the authors created a public benchmark for researchers to use to test their forgery detection solution.

Each participant in the benchmark (limited to once every two weeks) is provided one thousand images yielded from the dataset (either pristine or forged using one of the four previously presented methods) to analyze.

The goal of the benchmark is to push the research in forgery detection, but also offer a comparison metric between future research. In that sense, the FaceForensics++ paper is a seminal one.

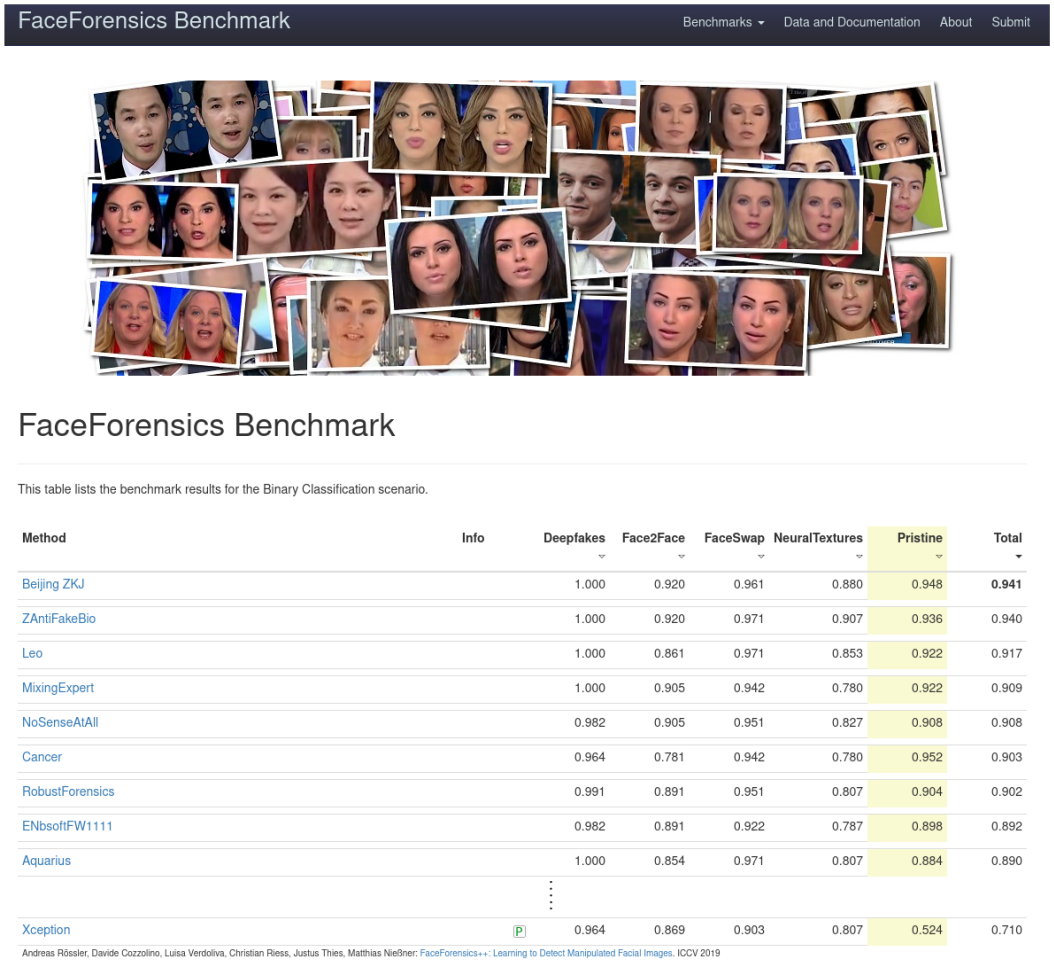


Figure 11: FaceForensics++ Benchmark Leaderboard

5 RESULTS, PERFORMANCE, AND LIMITATIONS

The second result from the paper is that forgery detection methods perform best on raw input – increasing levels of compression lead to downgraded detection performance. Neural networks are more resilient to compression.

When combined with domain-specific knowledge (face isolation), the neural network XceptionNet yields the best performance on all tests.

NeuralTextures is the most challenging doctoring method to detect. DeepFakes, despite being a neural network method is relatively easily detectable as the method relies on a post-processing step that yields detectable artifacts.

Results per architecture/methods are listed in Figure 13 and 14. It is interesting to note that computer detection methods perform better than human at detecting forgeries (see Figure 12).

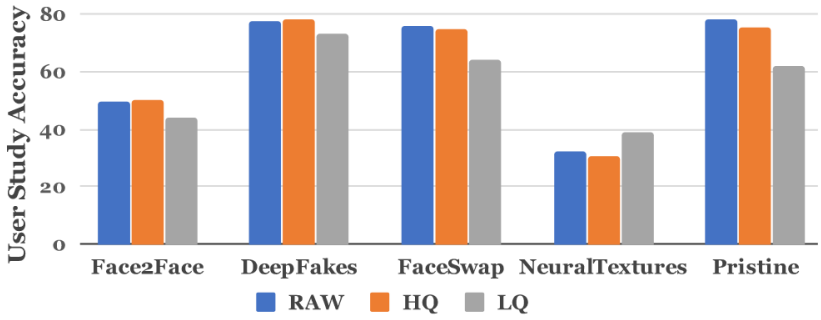


Figure 12: Human baseline, accuracy of forged pictures from the FaceForensics++ dataset

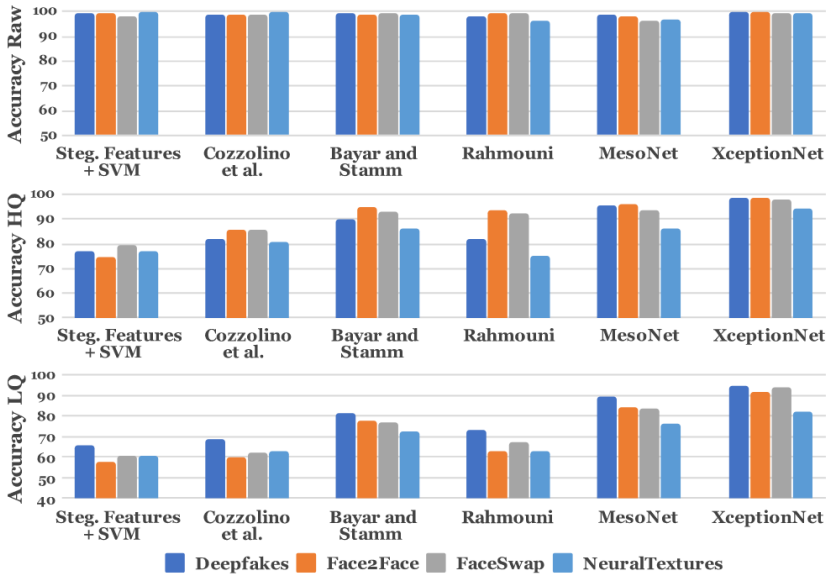


Figure 13: Architecture results, average accuracy of forged pictures from the FaceForensics++ dataset when trained on each manipulation methods separately

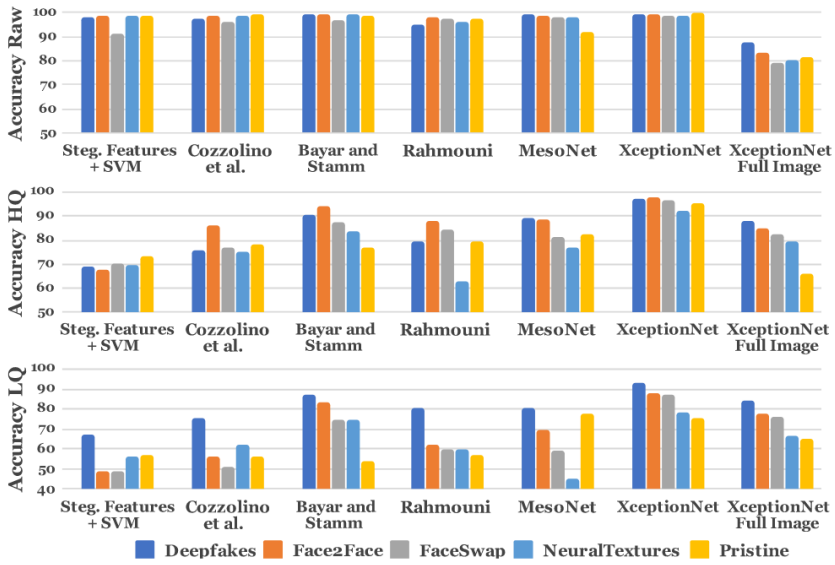


Figure 14: Architecture results, average accuracy of forged pictures from the FaceForensics++ dataset when trained on all manipulation methods at once

The main limitations of the current image forgery methods as at the publication of the paper and still today is the need for source and target videos to be front-facing, single-shot, without occlusions. The benchmark and most doctoring and detection methods revolves around producing and detecting videos and images that follow those principle. Furthermore, the detection relies on first singling out faces within a pictures.

The development of doctoring methods that can overcome that limitation would put into question existing papers, datasets, and benchmarks, including the one proposed by the paper.

6 Future directions

styleGAN https://en.wikipedia.org/wiki/StyleGAN

https://paperswithcode.com/paper/video-face-manipulation-detection-through

https://paperswithcode.com/paper/mantra-net-manipulation-tracing-network-for

https://paperswithcode.com/dataset/dfdc

7 Reimplementing the XceptionNet

8 Conclusion

Though limited in some extent, the ability to produce image forgeries has exploded over the past few years. The social, economic and political risks forgeries raise leads to an ever-increasing risk to detect them. As such, FaceForensics++ is a seminal paper and dataset in the area of computer vision and specifically the detection of forgeries produced by the current state-of-the-art.

REFERENCES

Tampered photos. *QPRI*, 88, 06 2008. URL https://archive.ph/20130704233149/http://www.pri.org/theworld/#selection-249.1-259.1.

Darius Afchar, Vincent Nozick, Junichi Yamagishi, and Isao Echizen. Mesonet: a compact facial video forgery detection network. In *2018 IEEE International Workshop on Information Forensics and Security (WIFS)*, pp. 1–7, 2018. doi: 10.1109/WIFS.2018.8630761.

Irene Amerini, Lamberto Ballan, Roberto Caldelli, Alberto Del Bimbo, and Giuseppe Serra. A sift-based forensic method for copy–move attack detection and transformation recovery. *IEEE Transactions on Information Forensics and Security*, 6(3):1099–1110, 2011. doi: 10.1109/TIFS.2011.2129512.

Grigory Antipov, Moez Baccouche, and Jean-Luc Dugelay. Face aging with conditional generative adversarial networks, 2017.

Hadar Averbuch-Elor, Daniel Cohen-Or, Johannes Kopf, and Michael F. Cohen. Bringing portraits to life. *ACM Transactions on Graphics (Proceeding of SIGGRAPH Asia 2017)*, 36(6):196, 2017.

Belhassen Bayar and Matthew C. Stamm. Constrained convolutional neural networks: A new approach towards general purpose image manipulation detection. *IEEE Transactions on Information Forensics and Security*, 13(11):2691–2706, 2018. doi: 10.1109/TIFS.2018.2825953.

Luca Bondi, Silvia Lameri, David Güera, Paolo Bestagini, Edward J. Delp, and Stefano Tubaro. Tampering detection and localization through clustering of camera-based cnn features. In *2017 IEEE Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*, pp. 1855–1864, 2017. doi: 10.1109/CVPRW.2017.232.

Tiago Carvalho, Fábio A. Faria, Hélio Pedrini, Ricardo da S. Torres, and Anderson Rocha. Illuminant-based transformed spaces for image forensics. *IEEE Transactions on Information Forensics and Security*, 11(4):720–733, 2016. doi: 10.1109/TIFS.2015.2506548.

François Chollet. Xception: Deep learning with depthwise separable convolutions. In *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 1800–1807, 2017. doi: 10.1109/CVPR.2017.195.

Davide Cozzolino, Giovanni Poggi, and Luisa Verdoliva. Recasting residual-based local descriptors as convolutional neural networks: An application to image forgery detection. In *Proceedings of the 5th ACM Workshop on Information Hiding and Multimedia Security*, pp. 159–164, New York, NY, USA, 2017. Association for Computing Machinery. ISBN 9781450350617. doi: 10.1145/3082031.3083247. URL https://doi.org/10.1145/3082031.3083247.

Kevin Dale, Kalyan Sunkavalli, Micah K. Johnson, Daniel Vlasic, Wojciech Matusik, and Hanspeter Pfister. Video face replacement. *ACM Trans. Graph.*, 30(6):1–130, dec 2011. ISSN 0730-0301. doi: 10.1145/2070781.2024164. URL http://doi.acm.org/10.1145/2070781.2024164.

deepfakes. Faceswap. https://github.com/deepfakes/faceswap/network, 2018.

Xiangling Ding, Gaobo Yang, Ran Li, Lebing Zhang, Yue Li, and Xingming Sun. Identification of motion-compensated frame rate up-conversion based on residual signals. *IEEE Transactions on Circuits and Systems for Video Technology*, 28(7):1497–1512, 2018. doi: 10.1109/TCSVT.2017.2676162.

Luca D'Amiano, Davide Cozzolino, Giovanni Poggi, and Luisa Verdoliva. A patchmatch-based dense-field algorithm for video copy–move detection and localization. *IEEE Transactions on Circuits and Systems for Video Technology*, 29(3):669–682, 2019. doi: 10.1109/TCSVT.2018.2804768.

Jessica Fridrich and Jan Kodovsky. Rich models for steganalysis of digital images. *IEEE Transactions on Information Forensics and Security*, 7(3):868–882, 2012. doi: 10.1109/TIFS.2012.2190402.

P. Garrido, L. Valgaerts, H. Sarmadi, I. Steiner, K. Varanasi, P. Pérez, and C. Theobalt. Vdub: Modifying face video of actors for plausible visual alignment to a dubbed audio track. *Comput. Graph. Forum*, 34(2):193–204, May 2015. ISSN 0167-7055. doi: 10.1111/cgf.12552. URL https://doi.org/10.1111/cgf.12552.

Haiying Guan, Mark Kozak, Eric Robertson, Yooyoung Lee, Amy N. Yates, Andrew Delgado, Daniel Zhou, Timothee Kheyrkhah, Jeff Smith, and Jonathan Fiscus. Mfc datasets: Large-scale benchmark datasets for media forensic challenge evaluation. In *2019 IEEE Winter Applications of Computer Vision Workshops (WACVW)*, pp. 63–72, 2019. doi: 10.1109/WACVW.2019.00018.

David Güera and Edward J. Delp. Deepfake video detection using recurrent neural networks. In *2018 15th IEEE International Conference on Advanced Video and Signal Based Surveillance (AVSS)*, pp. 1–6, 2018. doi: 10.1109/AVSS.2018.8639163.

Rui Huang, Shu Zhang, Tianyu Li, and Ran He. Beyond face rotation: Global and local perception gan for photorealistic and identity preserving frontal view synthesis. In *2017 IEEE International Conference on Computer Vision (ICCV)*, pp. 2458–2467, 2017. doi: 10.1109/ICCV.2017.267.

Jonathan Hui. How deep learning fakes videos (deepfake) and how to detect it? 2018. URL https://jonathan-hui.medium.com/how-deep-learning-fakes-videos-deepfakes-and-how-to-detect-it-c0b50fbf7cb9.

Hyeongwoo Kim, Pablo Garrido, Ayush Tewari, Weipeng Xu, Justus Thies, Matthias Nießner, Patrick Pérez, Christian Richardt, Michael Zollhöfer, and Christian Theobalt. Deep video portraits, 2018.

Pavel Korshunov and Sebastien Marcel. Deepfakes: a new threat to face recognition? assessment and detection, 2018.

Iryna Korshunova, Wenzhe Shi, Joni Dambre, and Lucas Theis. Fast face-swap using convolutional neural networks. *CoRR*, abs/1611.09577, 2016. URL http://arxiv.org/abs/1611.09577.

Paweł Korus and Jiwu Huang. Multi-scale analysis strategies in prnu-based tampering localization. *IEEE Transactions on Information Forensics and Security*, 12(4):809–824, 2017. doi: 10.1109/TIFS.2016.2636089.

Guillaume Lample, Neil Zeghidour, Nicolas Usunier, Antoine Bordes, Ludovic Denoyer, and Marc'Aurelio Ranzato. Fader networks: Manipulating images by sliding attributes, 2018.

Yuezun Li, Ming-Ching Chang, and Siwei Lyu. 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, 2018. doi: 10.1109/WIFS.2018.8630787.

Chengjiang Long, Eric Smith, Arslan Basharat, and Anthony Hoogs. A c3d-based convolutional neural network for frame dropping detection in a single video shot. In *2017 IEEE Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*, pp. 1898–1906, 2017. doi: 10.1109/CVPRW.2017.237.

Yongyi Lu, Yu-Wing Tai, and Chi-Keung Tang. Conditional cyclegan for attribute guided face image generation. 05 2017.

Patrick Mullan, Davide Cozzolino, Luisa Verdoliva, and Christian Riess. Residual-based forensic comparison of video sequences. In *2017 IEEE International Conference on Image Processing (ICIP)*, pp. 1507–1511, 2017. doi: 10.1109/ICIP.2017.8296533.

R. Raghavendra, Kiran B. Raja, Sushma Venkatesh, and Christoph Busch. Transferable deep-cnn features for detecting digital and print-scanned morphed face images. In *2017 IEEE Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*, pp. 1822–1830, 2017. doi: 10.1109/CVPRW.2017.228.

Nicolas Rahmouni, Vincent Nozick, Junichi Yamagishi, and Isao Echizen. Distinguishing computer graphics from natural images using convolution neural networks. In *2017 IEEE Workshop on Information Forensics and Security (WIFS)*, pp. 1–6, 2017. doi: 10.1109/WIFS.2017.8267647.

Andreas Rössler, Davide Cozzolino, Luisa Verdoliva, Christian Riess, Justus Thies, and Matthias Nießner. Faceforensics: A large-scale video dataset for forgery detection in human faces. *CoRR*, abs/1803.09179, 2018. URL http://arxiv.org/abs/1803.09179.

Thalles Silva. An intuitive introduction to generative adversarial networks (gans). 2018. URL https://www.freecodecamp.org/news/an-intuitive-introduction-to-generative-adversarial-networks-gans-7a2264a81394/.

Supasorn Suwajanakorn, Steven M. Seitz, and Ira Kemelmacher-Shlizerman. Synthesizing obama: Learning lip sync from audio. *ACM Trans. Graph.*, 36(4), July 2017. ISSN 0730-0301. doi: 10.1145/3072959.3073640. URL https://doi.org/10.1145/3072959.3073640.

Justus Thies. Face2face presentation at cvpr 2016. 2016. URL https://on-demand.gputechconf.com/siggraph/2016/presentation/sig1641-justus-thies-matthias-niessner-face-to-face-real-time-capture-reenactment.pdf.

Justus Thies, Michael Zollhöfer, Matthias Nießner, Levi Valgaerts, Marc Stamminger, and Christian Theobalt. Real-time expression transfer for facial reenactment. *ACM Trans. Graph.*, 34(6), October 2015. ISSN 0730-0301. doi: 10.1145/2816795.2818056. URL https://doi.org/10.1145/2816795.2818056.

Justus Thies, Michael Zollhöfer, Marc Stamminger, Christian Theobalt, and Matthias Nießner. Facevr: Real-time gaze-aware facial reenactment in virtual reality. *ACM Trans. Graph.*, 37(2), June 2018a. ISSN 0730-0301. doi: 10.1145/3182644. URL https://doi.org/10.1145/3182644.

Justus Thies, Michael Zollhöfer, Marc Stamminger, Christian Theobalt, and Matthias Nießner. Face2face. *Communications of the ACM*, 62:96 – 104, 2018b.

Justus Thies, Michael Zollhöfer, Christian Theobalt, Marc Stamminger, and Matthias Niessner. Headon. *ACM Transactions on Graphics*, 37(4):1–13, Aug 2018c. ISSN 1557-7368. doi: 10.1145/3197517.3201350. URL http://dx.doi.org/10.1145/3197517.3201350.

Justus Thies, Michael Zollhöfer, and Matthias Nießner. Deferred neural rendering: Image synthesis using neural textures, 2019.

Justus Thies, Michael Zollhöfer, Marc Stamminger, Christian Theobalt, and Matthias Nießner. Face2face: Real-time face capture and reenactment of rgb videos, 2020.

Paul Upchurch, Jacob Gardner, Geoff Pleiss, Robert Pless, Noah Snavely, Kavita Bala, and Kilian Weinberger. Deep feature interpolation for image content changes, 2017.

Markos Zampoglou, Symeon Papadopoulos, and Yiannis Kompatsiaris. Detecting image splicing in the wild (web). In *2015 IEEE International Conference on Multimedia Expo Workshops (ICMEW)*, pp. 1–6, 2015. doi: 10.1109/ICMEW.2015.7169839.

Peng Zhou, Xintong Han, Vlad I. Morariu, and Larry S. Davis. Two-stream neural networks for tampered face detection. In *2017 IEEE Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*, pp. 1831–1839, 2017. doi: 10.1109/CVPRW.2017.229.

APPENDIX A - CONSTRAINED CNN

Presented in 2018 by Bayar & Stamm (2018), a constrained CNN is a CNN method that uses a new type of layer at its start. The goal is to help the network learn not an image's content but manipulation traces. This new layer is trained in a similar fashion as a normal CNN, expect that it undergoes a normalization step where the middle weight of the filter (in a 3x3 instance case) is set to -1 while the surrounding weights are normalized to sum up to 1.

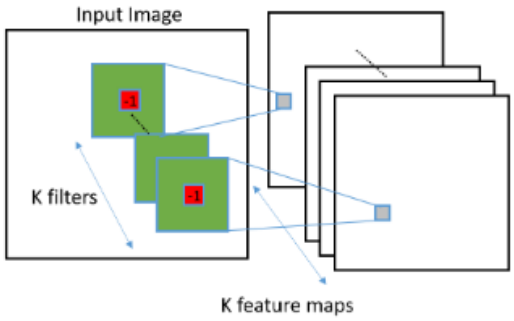


Figure 15: Constrained CNN layer. The middle weight is set to 1, and the surrounding ones normalized to sum up to 1

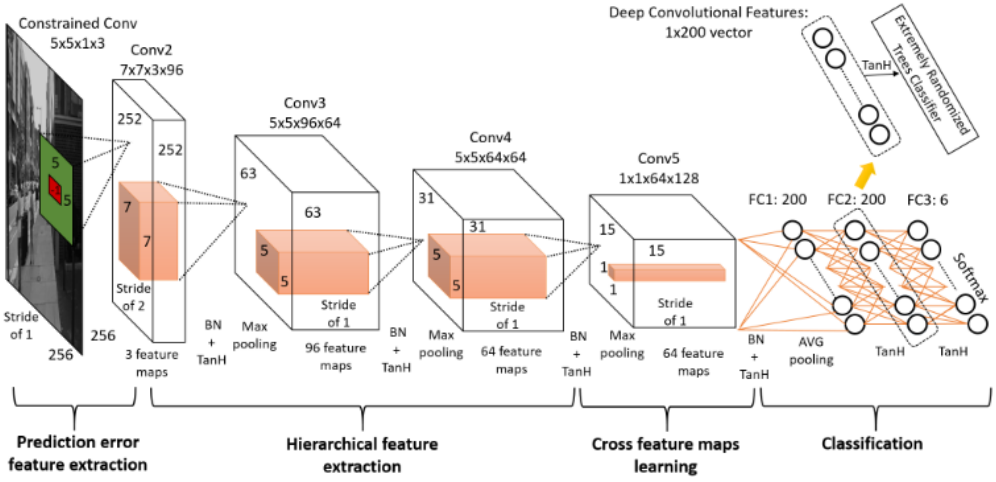


Figure 16: Constrained CNN as proposed by Bayar & Stamm (2018)