A video is worth more than 1000 lies. Comparing 3DCNN approaches for detecting deepfakes

Yaohui Wang, Antitza Dantcheva

A video is worth more than 1000 lies. Comparing 3DCNN approaches for detecting deepfakes

Yaohui Wang and Antitza Dantcheva
Inria, Université Côte d'Azur, France
{yaohui.wang, antitza.dantcheva}@inria.fr

***Abstract*— Manipulated images and videos have become increasingly realistic due to the tremendous progress of deep convolutional neural networks (CNNs). While technically intriguing, such progress raises a number of social concerns related to the advent and spread of fake information and fake news. Such concerns necessitate the introduction of robust and reliable methods for fake image and video detection. Towards this in this work, we study the ability of state of the art *video* CNNs including 3D ResNet, 3D ResNeXt, and I3D in detecting manipulated videos. We present related experimental results on videos tampered by four manipulation techniques, as included in the FaceForensics++ dataset. We investigate three scenarios, where the networks are trained to detect (a) *all* manipulated videos, as well as (b) separately *each* manipulation technique individually. Finally and deviating from previous works, we conduct cross-manipulation results, where we (c) detect the veracity of videos pertaining to manipulation-techniques not included in the train set. Our findings clearly indicate the need for a better understanding of manipulation methods and the importance of designing algorithms that can successfully generalize onto unknown manipulations.**

I. Introduction

Manipulated images have noticeably advanced hand in hand with photography, dating back to the creation of the first photograph in the year 1825 [1]. Related manipulation techniques have been widely driven by profit stemming from identity theft, age deception, illegal immigration, organized crime, and espionage, inflicting negative consequences on businesses, individuals, and political entities.

Currently, we are entering new levels of manipulation of images and videos. While forgery was associated with a slow, painstaking process usually reserved for experts, deep learning based *manipulation-technologies* are streamlined to reduce costs, time and skill needed.

The *manipulation scenario* of interest in this work has to do with a face image of a *target person* being superimposed to a video of a *source* person, widely accepted and referred to as *deepfake*. Therefore deepfakes coerce the target person in a video to reenact the dynamics of the source person.

Most recent research on deepfakes proposed approaches, where forged videos were created based on a *short video* of the source person [2], [3], as well as from a *single ID photo* [4] of the source person. In addition, fully synthesized *audio-video* images were able to replicate synchronous speech and lip movement [5] of a target person.

This work was funded by the French Government (National Research Agency, ANR), under Grant ANR-17-CE39-0002.

Automated generation and manipulation of audio, image and video bare highly exciting perspectives for science, art and video productions, video animation, special effects, reliving already passed actors. While highly intriguing from computer vision perspective, *deepfakes* entail a number of challenges and threats, given that (a) such manipulations can fabricate animations of subjects involved in actions that have not taken place and (b) such manipulated data can be circumvented nowadays rapidly via social media. Particularly, we cannot trust anymore, what we see or hear on video, as deepfakes betray sight and sound, the two predominantly trusted human innate senses [6]. Given that (i) our society relies heavily on the ability to produce and exchange legitimate and trustworthy documents, (ii) sound and images have recorded our history, as well as informed and shaped our perception of reality, axioms and truths such as "I'll believe it when I see it." "Out of sight, out of mind." "A picture is worth a thousand words.", as well as (iii) social media has catapulted online videos as a mainstream source of information; deepfakes pose a threat of distorting what is perceived as reality. To further fuel concern, deepfake techniques have become open to the public via phone applications such as FaceApp[1] and ZAO[2].

We differentiate two cases of concern: the first one has to do with *deepfakes being perceived as real*, and the second relates to *real videos being misdetected for fake*, the latter referred to as "liar's dividend". Given such considerations, video evidence becomes highly questionable.

Additional social threats [7], [8] can affect domains such as journalism, education, individual rights, democratic systems and have intrigued already a set of journalists[3][4][5].

Currently manipulations include subtle imperfections that can be detected by humans and, if trained well, by computer vision algorithms. Towards thwarting such adversarial attacks, early multimedia forensics based detection strategies have been proposed [9], [10], [11], [12]. Similar strategies, although essential, cannot provide a comprehensive solution against manipulated audio, images and video. Specifically,

[1] https://apps.apple.com/gb/app/faceapp-ai-face-editor/id1180884341
[2] https://apps.apple.come/cn/app/id146519927
[3] https://edition.cnn.com/interactive/2019/01/business/pentagons-race-against-deepfakes/
[4] https://www.nytimes.com/2019/11/24/technology/tech-companies-deepfakes.html
[5] https://www.theguardian.com/commentisfree/2018/jul/22/deep-fake-news-donald-trump-vladimir-putin

the detection of deepfakes is challenging for several reasons: (a) it evolves a "cat-and-mouse-game" between the adversary and the system designer, (b) deep models are highly domain-specific and likely yield big performance degradation in cross-domain deployments, especially with large train-test domain gap.

*a) **Contributions**:* Motivated by the above, this work makes following two contributions:

(i) We compare state-of-the art *video* based techniques in detecting deepfakes. Our intuition is that current state of the art forgery detection techniques [13], [14], [15], [16], [17], [18] omit a pertinent clue, namely *motion*, by investigating only spatial information. We found out that generative models exhibit difficulties in preserving appearance throughout generated videos, as well as motion consistency [19], [20], [21]. Hence, we here show that using 3DCNNs indeed outperforms state of the art image-based techniques.

(ii) We show that such models trained on known manipulation techniques generalise poorly to tampering methods outside of the training set. Towards this, we provide an evaluation, where train and test sets do not intersect with respect to manipulation techniques.

A. Related work

A very recent survey has revisited image and video manipulation approaches and early detection efforts [22]. An additional comprehensive survey paper [23] reviews manipulations of images, graphs and text.

1) Generation of images and videos: Generative adversarial networks (GANs) [24] have enabled a set of face manipulations including identity [25], [26], facial attributes [27], as well as facial expressions [28], [29].

2) Deepfake detection: While a number of *manipulation-detection-approaches* are *image-based* [18], [14], others are targeted towards *videos* [30], [9], [10] or jointly on audio-videos [31]. We note that although some **video-based approaches** might perform better than image-based ones, such approaches are only applicable to particular kinds of attacks. For example, many of them [30], [9] may fail, if the quality of the eye area is sufficiently good or the synchronization between video and audio is sufficiently natural [32].

Image-based approaches are general-purpose detectors, for instance, the algorithms proposed by Fridrich and Kodovsky [13] is applicable to both steganalysis and facial reenactment video detection. Rahmouni et al.' [17] presented an algorithm to detect computer-generated images, which was later extended to detecting computer-manipulated images. However, performance of such approaches on new tasks is limited compared with that of task-specific algorithms [14], [33].

What we show in this work is that, such algorithms are indeed challenged, if confronted with videos outside of the training data.

3) Adversarial training and detection: An **adversarial example** is an instance with small, intentional feature perturbations that cause a machine learning model to make a false prediction. Adversarial examples are inputs to machine learning models that an attacker intentionally designed to cause the model to make mistakes. In our context, the manipulated videos aim to be detected as real.

Adversarial training Currently defending against adversarial attacks is brute-force, where given a known network, a number of adversarial examples are generated against such a network, and then the network is explicitly trained not be fooled by them. However, this refers to the above named "cat-and-mouse-game" without a clear winner.

Adversary detection approaches attempt to verify the truthfulness of samples.

Rössler [14] presented a comparison of existing hand-crafted, as well as deep neural networks (DNNs), which analyzed the **FaceForensics++** dataset and proceeded to detect adversarial examples in an *image-based* manner. This was done for (i) raw data, (ii) high quality videos compressed by a constant rate quantization parameter equal to 23 (denoted as HQ), as well as (iii) low quality videos compressed by a quantization rate of 40 (denoted as LQ). There were two trainings-settings used: (a) training on all manipulation methods concurrently, (b) individual training on each manipulation method separately. These two settings refer to two scenarios of interest in this work.

We summarize for training setting (a), which is the more challenging setting (as indicated by lower related detection rates).

1) **Raw data**: It is interesting to note that the correct detection rates for all seven compared algorithms ranged between 97.03% and 99.26%. The highest score was obtained by XceptionNet [34].
2) **HQ**: High quality compressed data was detected with rates ranging between 70.97% and 95.73% (XceptionNet).
3) **LQ**: Intuitively low quality compressed data had the lowest detection rates with 55.98% to 81% (XceptionNet).

Therefore we here focus on the LQ-compression as the most challenging setting.

We note that reported detection rates pertained to the analysis of a facial area with the dimension 1.3 times the cropped face. Analyzing the full frame obtained lower accuracy.

A challenge, not being addressed by Rössler has to do with the generalization of such methods. When detection methods, as the presented ones are confronted with adversarial attacks, outside of the training set, such networks are challenged.

The rest of the paper is organized as follows. In Section II we present the dataset, in Section III the compared detection algorithms. Experiments and related results are summarized in Section IV. Finally, Section V concludes the paper.

II. Dataset

The FaceForensics++ dataset [14] comprises of 1000 talking subjects, represented in 1000 real videos. Further, based

Fig. 1. **Sample frames from the Faceforensics++ dataset** From left to right: original source (large) and target (small) images, deepfakes, face2face, faceswap, neuraltextures.

on these 1000 real videos, 4x1000 adversarial examples have been generated by following four manipulation schemes:

1) **face-swap**[6] represents a graphic approach transferring a face region from a source video to a target video. Using facial landmarks, a 3D template model employs blendshapes to fit the transferred face.
2) **deepfakes** has become the synonym for face manipulations of all kind, however it origins to FakeApp[7] and faceswap github[8].
3) **face2face** [2] is a facial reenactment system that transfers the expressions of a source video to a target video while maintaining the identity of the target person.
4) **neuraltextures** [35] incorporates facial reenactment as an example for a *NeuralTextures*-based rendering approach. It uses the original video data to learn a neural texture of the target person, including a rendering network.

III. Algorithms

We select following three state-of-the-art 3D CNN methods, which have excelled in action recognition.

- **I3D** [36] incorporates sets of RGB frames as input. It replaces 2D convolutional layers of the original Inception model by 3D convolutions for spatio-temporal modeling and inflates pre-trained weights of the Inception model on ImageNet as its initial weights. Results showed that such inflation has the ability to improve 3D models.
- **3D ResNet** [37] and **3D ResNeXt** are inspired by I3D, both extending initial 2D ResNet and 2D ResNeXt to spatio-temporal dimension for action recognition. We note that deviating from the original ResNet-bottleneck block, the ResNeXt-block introduces group convolutions, which divide the feature maps into small groups.

Given the binary classification problem in this work, we replace the prediction layer in all networks by a single neuron layer, which outputs one scalar value. All three networks have been pre-trained on the large-scale human action dataset Kinetics-400. We inherit the weights in the neural network models and further fine-tune the networks on the Faceforensics++ dataset in all our experiments.

Experimental Setup We use PyTorch to implement our models. The three entire networks are trained end-to-end on 4 NVIDIA V100 GPUs. We set the learning rates to $1e^{-3}$. For training, I3D accepts videos of 64 frames with spatial dimension 224×224 as input. The size of input of 3D ResNet and 3D ResNeXt are 16 frames of spatial resolution 112×112. For testing, we split each video into short trunks, each of temporal size of 250 frames. The final score assigned to each test video is the average value of the scores of all trunks.

We detect and crop the face region based on facial landmarks, which we detect in each frame using the method from Bulat and Tzimiropoulos [38]. Next, we enlarge the detected region by a factor of 1.3, in order to include more pixels around the face region.

IV. Experiments

We conduct three experiments on the above enlisted manipulation techniques with I3D, 3D ResNet and 3D ResNext aiming at training and detecting (a) all manipulation techniques, (b) each manipulation technique separately and (c) cross-manipulation techniques. Towards this, we split train, test and validation sets according to the protocol provided in the Faceforensics++ dataset. We report in all experiments the true classification rates (TCR).

A. All manipulation techniques

Firstly we evaluate TCR of the three *video* CNNs in comparison to *image*-forgery detection algorithms. For the latter we have in particular the state of the art XceptionNet [14], learning-based methods used in the forensic community for generic manipulation detection [15], [16], computer-generated vs. natural image detection [17] and face tampering detection [18]. Given the unbalanced classification problem in this experiment (number of fake videos being nearly

[6] htpps://github.com/MarekKowalski/FaceSwap/
[7] https://www.fakeapp.com
[8] https://github.com/deepfakes/faceswap

TABLE I

DETECTION OF ALL FOUR MANIPULATION METHODS, LQ. TCR...TRUE CLASSIFICATION RATE, DF...DEEPFAKES, F2F...FACE2FACE, FS...FACE-SWAP, NT...NEURALTEXTURES.

Algorithm	Train and Test	TCR
Steg. Features + SVM [13]	FS, DF, F2F, NT	55.98
Cozzolino et al. [15]	FS, DF, F2F, NT	58.69
Bayar and Stamm [16]	FS, DF, F2F, NT	66.84
Rahmouniet al. [17]	FS, DF, F2F, NT	61.18
MesoNet [18]	FS, DF, F2F, NT	70.47
XceptionNet [34]	FS, DF, F2F, NT	81.0
3D ResNet	FS, DF, F2F, NT	83.86
3D ResNeXt	FS, DF, F2F, NT	85.14
I3D	FS, DF, F2F, NT	87.43

TABLE II

DETECTION OF EACH MANIPULATION METHOD INDIVIDUALLY, LQ. TCR...TRUE CLASSIFICATION RATE, DF...DEEPFAKES, F2F...FACE2FACE, FS...FACE-SWAP, NT...NEURALTEXTURES.

Algorithm	DF	F2F	FS	NT
Steg. Features + SVM [13]	73.64	73.72	68.93	63.33
Cozzolino et al. [15]	85.45	67.88	73.79	78.00
Bayar and Stamm [16]	84.55	73.72	82.52	70.67
Rahmouniet al. [17]	85.45	64.23	56.31	60.07
MesoNet [18]	87.27	56.20	61.17	40.67
XceptionNet [34]	96.36	86.86	90.29	80.67
3D ResNet	91.81	89.6	88.75	73.5
3D ResNeXt	93.36	86.06	92.5	80.5
I3D	95.13	90.27	92.25	80.5

four times the number of real videos), we use weighted cross-entropy loss, in order to reduce the effects of unbalanced data. Related results are depicted in Table I.

B. Single manipulation techniques

Next, we investigate the performances of all algorithms when trained and tested on single manipulation techniques. We report the TCRs in Table II. Interestingly, here the video based algorithms perform similarly as the best image-based algorithm. This can be due to the reduced data-size pertaining to videos of a single manipulation.

The experiments show that all detection approaches are consistently utmost challenged by the GAN-based *neuraltextures*-approach. We note that *neuraltextures* trains a unique model for each video, which results in a higher variation of possible artifacts. While *deepfakes* similarly trains one model per video, a fixed post-processing pipeline is used, which is similar to the computer-based manipulation methods and thus has consistent artifacts.

C. Cross-manipulation techniques

In our third experiment, we train the 3D CNNs with videos pertained to 3 manipulation techniques, as well as the original (real) videos and proceed to test with the remaining manipulation technique, as well as original videos. We show related results in Table III. Naturally, this is the most challenging setting. At the same time, it is the most realistic one, because it is unlikely that knowledge on whether and how videos have been manipulated will be provided. Similar to the first experiment, we use weighted cross-entropy loss, in order to solve the unbalanced classification problem. The significantly utmost challenging setting in this experiment is when *faceswap* is the manipulation technique to be detected.

TABLE III

DETECTION OF CROSS-MANIPULATION METHODS, LQ. TRUE CLASSIFICATION RATES REPORTED. DF...DEEPFAKES, F2F...FACE2FACE, FS...FACE-SWAP, NT...NEURALTEXTURES.

Train	Test	3D ResNet	3D ResNeXt	I3D
FS, DF, F2F	NT	64.29	68.57	66.79
FS, DF, NT	F2F	74.29	70.71	68.93
FS, F2F, NT	DF	75.36	75.00	72.50
F2F, NT, DF	FS	59.64	57.14	55.71

While *face2face* and *faceswap* represent graphics-based approaches, *deepfakes* and *neuraltextures* are learning-based approaches. However, *faceswap* replaces the largest facial region in the target image and involves advanced blending and color correction algorithms to seamlessly superimpose source onto target. Hence the challenge might be due to the inherent dissimilarity of *faceswap* and learning-based approaches, as well as due to the seamless blending between source and target, different from *face2face*. We note that *humans* easily detected manipulations affected by *faceswap* and *deepfakes* and were more challenged by *face2face* and ultimately *neuraltextures* [14]. This ranking corresponds to the CNNs-results in experiment 2 (see Table II).

V. CONCLUSIONS

In this work we compared three state-of-the-art 3D CNN methods in detecting four deepfake-manipulation-techniques. The three tested methods included 3D ResNet, 3D ResNeXt and I3D, which we adapted from action recognition. Despite the pre-training of mentioned methods on the action recognition dataset Kinetics-400, the methods generalized very well to deepfake detection. Experimental results showed that 3D / video CNNs outperformed or performed at least similarly to image-based forgery detection algorithms.

Further, we noted a significant decrease in true classification rates in when detecting manipulated videos pertained to manipulation techniques not represented in the training set. One reason relates to the fact that networks lack an adaptation-ability to transfer learned knowledge from one domain (trained manipulation methods) to another domain (tested manipulation method). It is known that current machine learning models exhibit unpredictable and overly confident behaviour outside of the training distribution.

Future work will involve the consideration of additional deepfake manipulation-techniques. Further, we plan to develop novel deepfake-detection approaches, which place emphasis on appearance, motion, as well as pixel-level based generated noise, targeted to outsmart ever-evolving generation and manipulation algorithms.

VI. Acknowledgment

This work is supported by the French Government (National Research Agency, ANR), under Grant ANR-17-CE39-0002.

References

[1] H. Farid, "Photo tampering throughout history," *online] http://www. cs. dartmouth. edu/farid/research/digitaltampering*, 2011.

[2] J. Thies, M. Zollhofer, M. Stamminger, C. Theobalt, and M. Nießner, "Face2face: Real-time face capture and reenactment of rgb videos," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2016, pp. 2387–2395.

[3] H. Kim, P. Carrido, A. Tewari, W. Xu, J. Thies, M. Niessner, P. Pérez, C. Richardt, M. Zollhöfer, and C. Theobalt, "Deep video portraits," *ACM Transactions on Graphics (TOG)*, vol. 37, no. 4, p. 163, 2018.

[4] H. Averbuch-Elor, D. Cohen-Or, J. Kopf, and M. F. Cohen, "Bringing portraits to life," *ACM Transactions on Graphics (TOG)*, vol. 36, no. 6, p. 196, 2017.

[5] S. Suwajanakorn, S. M. Seitz, and I. Kemelmacher-Shlizerman, "Synthesizing obama: learning lip sync from audio," *ACM Transactions on Graphics (TOG)*, vol. 36, no. 4, p. 95, 2017.

[6] J. Silbey and W. Hartzog, "The upside of deep fakes," *Md. L. Rev.*, vol. 78, p. 960, 2018.

[7] R. Chesney and D. K. Citron, "Deep fakes: A looming challenge for privacy, democracy, and national security. 107 california law review (2019, forthcoming); u of texas law," *Public Law Research Paper*, no. 692, pp. 2018–21, 2018.

[8] K. Eichensehr, "Don't believe it if you see it: Deep fakes and distrust," *Jotwell: J. Things We Like*, p. 1, 2018.

[9] S. Agarwal, H. Farid, Y. Gu, M. He, K. Nagano, and H. Li, "Protecting world leaders against deep fakes," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops*, 2019, pp. 38–45.

[10] E. Sabir, J. Cheng, A. Jaiswal, W. AbdAlmageed, I. Masi, and P. Natarajan, "Recurrent convolutional strategies for face manipulation detection in videos," *Interfaces (GUI)*, vol. 3, p. 1, 2019.

[11] I. Amerini, L. Galteri, R. Caldelli, and A. Del Bimbo, "Deepfake video detection through optical flow based cnn," in *Proceedings of the IEEE International Conference on Computer Vision Workshops*, 2019, pp. 0–0.

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

[13] J. Fridrich and J. Kodovsky, "Rich models for steganalysis of digital images," *IEEE Transactions on Information Forensics and Security*, vol. 7, no. 3, pp. 868–882, 2012.

[14] A. Rössler, D. Cozzolino, L. Verdoliva, C. Riess, J. Thies, and M. Nießner, "Faceforensics++: Learning to detect manipulated facial images," *arXiv preprint arXiv:1901.08971*, 2019.

[15] D. Cozzolino, G. Poggi, and L. 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*. ACM, 2017, pp. 159–164.

[16] B. Bayar and M. C. Stamm, "A deep learning approach to universal image manipulation detection using a new convolutional layer," in *Proceedings of the 4th ACM Workshop on Information Hiding and Multimedia Security*. ACM, 2016, pp. 5–10.

[17] N. Rahmouni, V. Nozick, J. Yamagishi, and I. Echizen, "Distinguishing computer graphics from natural images using convolution neural networks," in *2017 IEEE Workshop on Information Forensics and Security (WIFS)*. IEEE, 2017, pp. 1–6.

[18] D. Afchar, V. Nozick, J. Yamagishi, and I. Echizen, "Mesonet: a compact facial video forgery detection network," in *2018 IEEE International Workshop on Information Forensics and Security (WIFS)*. IEEE, 2018, pp. 1–7.

[19] Y. Wang, P. Bilinski, F. Bremond, and A. Dantcheva, "G3AN: Disentangling motion and appearance for video generation," in *Conference on Computer Vision and Pattern Recognition (CVPR)*, 2020.

[20] ——, "*G3AN*: This video does not exist. disentangling motion and appearance for video generation," *arXiv preprint arXiv:1912.05523*, 2019.

[21] ——, "Imaginator: Conditional spatio-temporal gan for video generation," in *Winter Conference on Applications of Computer Vision (WACV)*, 2020.

[22] R. Tolosana, R. Vera-Rodriguez, J. Fierrez, A. Morales, and J. Ortega-Garcia, "Deepfakes and beyond: A survey of face manipulation and fake detection," *arXiv preprint arXiv:2001.00179*, 2020.

[23] H. Xu, Y. Ma, H. Liu, D. Deb, H. Liu, J. Tang, and A. Jain, "Adversarial attacks and defenses in images, graphs and text: A review," *arXiv preprint arXiv:1909.08072*, 2019.

[24] I. Goodfellow, J. Pouget-Abadie, M. Mirza, B. Xu, D. Warde-Farley, S. Ozair, A. Courville, and Y. Bengio, "Generative adversarial nets," in *Advances in neural information processing systems*, 2014, pp. 2672–2680.

[25] T. Karras, S. Laine, and T. Aila, "A style-based generator architecture for generative adversarial networks," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2019, pp. 4401–4410.

[26] P. Majumdar, A. Agarwal, R. Singh, and M. Vatsa, "Evading face recognition via partial tampering of faces," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops*, 2019, pp. 0–0.

[27] Y. Wang, A. Dantcheva, and F. Bremond, "From attributes to faces: a conditional generative adversarial network for face generation," in *International Conference of the Biometrics Special Interest Group (BIOSIG)*, vol. 17, 2018.

[28] Z. Liu, G. Song, J. Cai, T.-J. Cham, and J. Zhang, "Conditional adversarial synthesis of 3d facial action units," *Neurocomputing*, vol. 355, pp. 200–208, 2019.

[29] L. Jiang, W. Wu, R. Li, C. Qian, and C. C. Loy, "Deeperforensics-1.0: A large-scale dataset for real-world face forgery detection," *arXiv preprint arXiv:2001.03024*, 2020.

[30] Y. Li, M.-C. Chang, and S. Lyu, "In ictu oculi: Exposing ai generated fake face videos by detecting eye blinking," *arXiv preprint arXiv:1806.02877*, 2018.

[31] P. Korshunov and S. Marcel, "Speaker inconsistency detection in tampered video," in *2018 26th European Signal Processing Conference (EUSIPCO)*. IEEE, 2018, pp. 2375–2379.

[32] ——, "Vulnerability assessment and detection of deepfake videos," in *The 12th IAPR International Conference on Biometrics (ICB)*, 2019, pp. 1–6.

[33] A. Khodabakhsh, R. Ramachandra, K. Raja, P. Wasnik, and C. Busch, "Fake face detection methods: Can they be generalized?" in *2018 International Conference of the Biometrics Special Interest Group (BIOSIG)*. IEEE, 2018, pp. 1–6.

[34] F. Chollet, "Xception: Deep learning with depthwise separable convolutions," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2017, pp. 1251–1258.

[35] J. Thies, M. Zollhöfer, and M. Nießner, "Deferred neural rendering: Image synthesis using neural textures," *arXiv preprint arXiv:1904.12356*, 2019.

[36] J. Carreira and A. Zisserman, "Quo vadis, action recognition? a new model and the kinetics dataset," in *proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2017, pp. 6299–6308.

[37] K. Hara, H. Kataoka, and Y. Satoh, "Can spatiotemporal 3d cnns retrace the history of 2d cnns and imagenet?" in *Proceedings of the IEEE conference on Computer Vision and Pattern Recognition*, 2018, pp. 6546–6555.

[38] A. Bulat and G. Tzimiropoulos, "How far are we from solving the 2d & 3d face alignment problem? (and a dataset of 230,000 3d facial landmarks)," in *International Conference on Computer Vision*, 2017.