Detecting Deep-Fake Videos from Appearance and Behavior

Shruti Agarwal and Hany Farid
Univeristy of California, Berkeley,
Berkeley, CA, USA
Email: {shruti_agarwal,hfarid}@berkeley.edu

Tarek El-Gaaly and Ser-Nam Lim
Facebook Research, New York,
NY, USA
Email: {tgaaly, sernamlim}@fb.com

2020 IEEE International Workshop on Information Forensics and Security (WIFS) | 978-1-7281-9930-6/20/$31.00 ©2020 IEEE | DOI: 10.1109/WIFS49906.2020.9360904

***Abstract*—Synthetically-generated audios and videos – so-called deep fakes – continue to capture the imagination of the computer-graphics and computer-vision communities. At the same time, the democratization of access to technology that can create a sophisticated manipulated video of anybody saying anything continues to be of concern because of its power to disrupt democratic elections, commit small to large-scale fraud, fuel disinformation campaigns, and create non-consensual pornography. We describe a biometric-based forensic technique for detecting face-swap deep fakes. This technique combines a static biometric based on facial recognition with a temporal, behavioral biometric based on facial expressions and head movements, where the behavioral embedding is learned using a CNN with a metric-learning objective function. We show the efficacy of this approach across several large-scale video datasets, as well as in-the-wild deep fakes.**

I. Introduction

Recent advances in computer graphics, computer vision, and machine learning have made it increasingly easier to synthesize compelling fake audio, image, and video. In the audio domain, highly realistic audio synthesis is now possible in which a neural network, with enough sample recordings, can learn to synthesize speech in your voice [1]. In the static image domain, highly realistic images of people can now be synthesized using generative adversarial networks (GANs) [2], [3]. And, in the video domain, realistic videos can be created of anybody saying anything that its creator wants [4].

These so-called deep-fake videos can be highly entertaining but can also be easily weaponized. The creation of non-consensual pornography, for example, was the first use of deep fakes, and continues to pose a threat particularly to women, ranging from celebrities to journalists, and those that simply attract unwanted attention [5]. We describe a forensic technique to authenticate face-swap deep fake videos in which a person's facial identity is replaced with another's.

There is a significant literature in the general area of digital forensics [6]. Here we focus only on techniques for detecting deep-fake videos, categorized as either low-level or high-level approaches.

Low-level forensic techniques detect pixel-level artifacts introduced by the synthesis process. Some of these techniques detect generic artifacts [7]–[10], while others detect explicit artifacts that result from, for example, image warping [11], image blending [12] and inconsistencies between the image and metadata [13]. Although these techniques detect a variety of fakes with relatively high accuracy, the drawback is that they can be sensitive to unintentional laundering (e.g., transcoding or resizing) or intentional adversarial attacks (e.g., [14]).

High-level approaches, in contrast, tend to generalize and be more resilient to laundering and adversarial attacks. These techniques focus on semantically meaningful features including, for example, inconsistencies in eye blinks [15], head-pose [16], physiological signals [17], and distinct mannerisms [18]. We believe that, because current synthesis techniques are frame-based, incorporating these types of semantic and temporal dynamics is essential to staying slightly ahead of the cat-and-mouse game of synthesis and detection.

The work of [18] is most similar to ours. In their work, the authors analyzed hours of video of specific individuals (in their case, various world leaders and presidential candidates) in order to extract distinct and predictable patterns of facial expressions and head movements. Specifically, from each 10-second clip of an individual, the authors extracted the frame-by-frame facial expressions (parameterized as 18 action units [19] and 3-D head rotation about two axes). The correlation between all pairs of these 20 features yielded a 190-D feature vector capturing an individual's temporal mannerisms. A one-class SVM [20] was employed to classify each 10-second video clip as being consistent or not with the learned mannerisms of an individual. The benefit of this approach is that it captures temporal mannerisms that current frame-based, deep-fake synthesis techniques are not (yet) able to synthesize. The other benefit is that this approach, unlike low-level detection schemes, is more robust to laundering attacks and is more able to generalize to a large class of deep fakes from face-swap to lip-sync, and puppet-master. The drawback of this approach is that it can require significant effort to create models for each individual and it is almost certainly the case that the hand-crafted correlation-based features are not optimal, nor are they capturing all of the distinct properties that might distinguish a real from a fake video.

Building on this earlier work by [18], we employ a convolutional neural network (CNN) with a metric-learning objective function to learn a more discriminating behavioral biometric.

WIFS'2020, December, 6-11, 2020, New York, USA. 978-1-7281-9930-6/20/$31.00

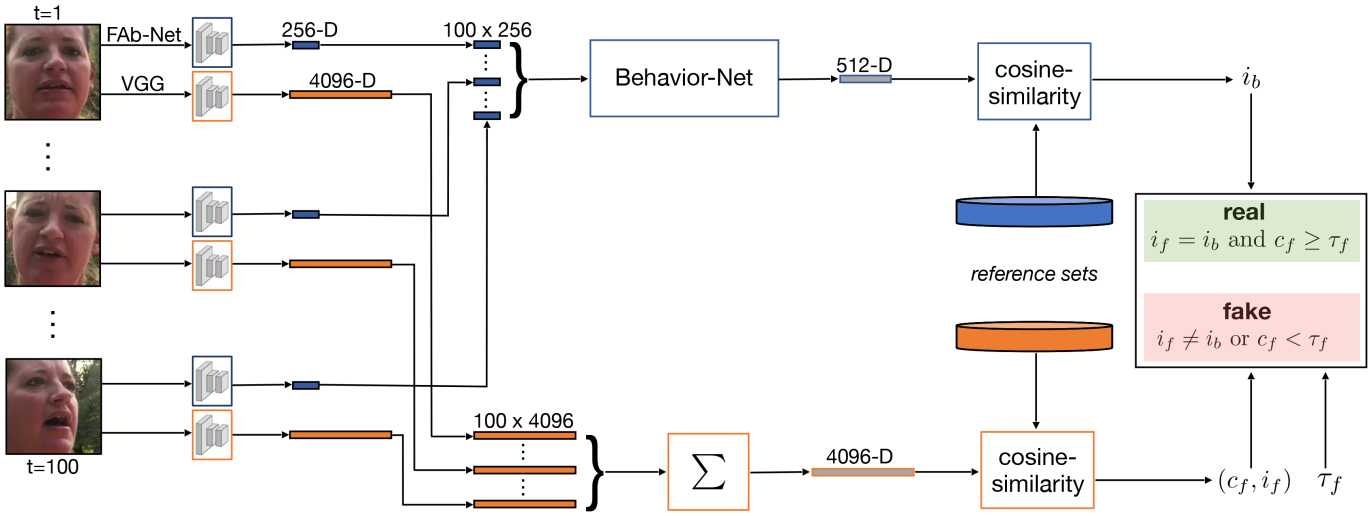

Fig. 1. An overview of our authentication pipeline (see Section II-C).

We pair this learned biometric with a facial biometric in order to determine if a person's identity in video clips as short as four seconds is consistent with the facial and behavioral properties extracted from reference videos. This approach is specifically targeted towards face-swap deep fakes in which the face of one person has been replaced with another.

II. Biometrics

We next describe two biometric measurements that underlie our forensic detection scheme. These include a biometric based on temporal behavioral (facial expressions and head movements) and a biometric based on static facial features.

A. Behavior

In [21], the authors proposed a self-supervised, encoder-decoder network (Facial Attributes-Net, FAb-Net) trained to embed the movement between video frames into a common 256-D space. The authors showed that the network, in turn, learns an embedding space that represents head pose, facial landmarks, and facial expression. We use these 256-D FAb-Net features as building blocks to measure spatiotemporal biometric behavior. A t-frame video clip of a person talking is first reduced to a feature matrix $X \in \mathbb{R}^{256\times t}$, where each matrix column corresponds to each frame's FAb-Net feature.

FAb-Net nicely captures the frame-based facial movements and expressions but is, by design, identity-agnostic. We seek to learn a modified embedding that both captures facial movements and expressions, but also distinguishes these features across individuals. That is, starting with the static FAb-Net features, we learn a low-dimensional mapping that encodes identity-specific spatiotemporal behavior, Fig. 1.

Given FAb-Net feature matrices for n, t-frame video clips $X_1, \ldots, X_n$ with identity labels $y_1, \ldots, y_n$, we learn a mapping $f(\cdot) : \mathbb{R}^{256\times t} \rightarrow \mathbb{R}^d$, that projects X_i to an embedding space such that the similarity S_{ij} between $f(X_i)$ and $f(X_j)$ is high if $y_i = y_j$ (positive sample) and S_{ij} is low if $y_i \neq y_j$ (negative sample). Because, the output $f(X_i)$ is normalized to lie on a unit sphere, a cosine similarity, between two vector-based representations, is used to compute S_{ij}.

To learn the mapping $f(\cdot)$, a CNN is trained with a multi-similarity metric-learning objective function [22]. Following the approach in [22], the loss for a mini-batch is computed as follows. First, for every input X_i, hard positive and negative samples are selected. For hard negative samples (where $y_i \neq y_j$), a sample X_j is selected if $S_{ij} > \min\{S_{ik} - \epsilon\}$, for all k such that $y_i = y_k$, and where ϵ is a small margin. This formulation selects the most confusing negative samples whose similarity with the input is larger than the minimum similarity between the input and all positive samples. Similarly, for hard positive samples (where $y_i = y_j$), a sample X_j is selected if $S_{ij} < \max\{S_{ik} + \epsilon\}$, for all k such that $y_i \neq y_k$. Here, the most meaningful positive samples are selected by comparing to the negative samples most similar to the input.

A soft weighting is then applied to rank these selected samples according to their importance for learning the desired embedding space. For a given input X_i, let $\mathcal{N}_i$ and $\mathcal{P}_i$ represents the selected negative and positive samples that are weighted as follows:

$$w_{ij}^{-} = \frac{e^{\beta(S_{ij}-\lambda)}}{1+\sum_{k\in\mathcal{N}_i} e^{\beta(S_{ik}-\lambda)}} \tag{1}$$

and

$$w_{ij}^{+} = \frac{e^{-\alpha(S_{ij}-\lambda)}}{1+\sum_{k\in\mathcal{P}_i} e^{-\alpha(S_{ik}-\lambda)}}, \tag{2}$$

where α, β, and λ are hyper-parameters. Finally, the loss $\mathcal{L}$ over a mini-batch of size m is:

$$\mathcal{L} = \frac{1}{m}\sum_{i=1}^{m}\left\{\frac{1}{\alpha}\log\left[1+\sum_{k\in\mathcal{P}_i} e^{-\alpha(S_{ik}-\lambda)}\right] + \frac{1}{\beta}\log\left[1+\sum_{k\in\mathcal{N}_i} e^{\beta(S_{ik}-\lambda)}\right]\right\}. \tag{3}$$

TABLE I
CLASSIFICATION PERFORMANCE AS AUC (TOP) AND DETECTION ACCURACIES AT A FIXED THRESHOLD OF $\tau_f = 0.86$ (BOTTOM).

	Average	WLDR	FF	DFD	DFDC-P	CDF
AUC	0.97	0.99	0.99	0.95	0.92	0.99
real	96.5%	99.6%	99.2%	93.1%	93.1%	97.6%
fake	91.8%	95.8%	98.7%	93.2%	71.7%	99.4%
average	94.2%	97.7%	98.9%	93.2%	82.4%	98.5%

By performing supervised training using the identity labels, the network is encouraged to learn an embedding space that clusters the biometric signatures by identity.

Our model is trained on the VoxCeleb2 dataset [23], containing over a million utterances from $5,994$ unique identities. The size of the input feature matrix is fixed to $t = 100$, corresponding to a 4-second video clip at 25 frames/second (this clip size was selected as it was the minimum clip size of the VoxCeleb2 utterances). We used the ResNet-101 network architecture [24], where the input layer of the network is modified to the size of our feature matrix (256×100). A fully-connected output layer of size $d = 512$ is added on top of this network, forming our final feature vector, which is normalized to be zero-mean and unit-length before computing the loss. We name this network Behavior-Net.

The CNN training is performed for $10,000$ iterations with a mini-batch of size 256. Following [22], in each mini-batch, 32 identities are randomly selected, for which 8 utterance videos (each of variable length) are randomly selected, from which a randomly selected 100-frame sequence is extracted. All other optimization hyper-parameters are the same as in [22].

Even though the Behavior-Net features are trained only on the VoxCeleb2 dataset, these features will be used to classify different identities across different datasets. This generalizability is both practically useful and suggests that the underlying Behavior-Net captures intrinsic properties of people.

B. Appearance

Rapid advances in deep learning and access to large datasets have led to a revolution in face recognition. We leverage one such fairly straight-forward approach, VGG [25], a 16-layer CNN trained to perform face recognition on a dataset consisting of 2622 identities. VGG yields a distinct 4096-D face descriptor per face, per video frame. These descriptors are averaged over the 100 frames of the 4-second video clip to yield a single facial descriptor, Fig. 1.

Faces for this facial biometric and the behavioral biometric are extracted using OpenFace [26]. Once localized and extracted from a video frame, each face is aligned and re-scaled to a size of 256×256 pixels.

C. Authentication

Given authentic, 4-second video clips for all unique identities, two reference sets are created with the VGG facial and Behavior-Net features. Define F_i to be the $4096 \times m_i$ real-valued matrix consisting of the VGG features for m_i video clips of identity i. Similarly, define B_i to be the $512 \times m_i$ real-valued matrix consisting of the Behavior-Net features for the same m_i video clips, also of identity i. Each column of the matrices F_i and B_i contains the VGG and Behavior-Net features for a single video clip.

Fig. 2. Shown is an example frame of a face-swap deep fake (middle) from the DFDC-P dataset, in which the source identity (left) should be mapped onto the target (right), which is clearly not the case in this example.

Given these reference sets, an unseen 4-second video clip is authenticated as follows. First, extract the facial and Behavior-Net features, $\vec{f} \in \mathbb{R}^{4096}$ and $\vec{b} \in \mathbb{R}^{512}$. Next, find the identities, i_f and i_b in the reference sets with the most similar features using a cosine-similarity metric: $i_f = \arg\max_i\{\max(\vec{f}^t \cdot F_i)\}$ and $i_b = \arg\max_i\{\max(\vec{b}^t \cdot B_i)\}$.

With these matched identities, a video clip is classified as real or fake following two simple rules (see also Fig. 1):

1) A video clip is classified as real if the facial and Behavior-Net identities are the same, $i_f = i_b$, and if the facial similarity is above a specified threshold, $c_f >= \tau_f$, where $c_f = \max(\vec{f}^t \cdot F_{i_f})$ (i.e., a close facial match is found).
2) A video clip is classified as fake if either:
   a) the matched identities are different, $i_f \neq i_b$, or
   b) the facial similarity is below threshold, $c_f < \tau_f$.

The rationale for the asymmetric treatment of the facial and Behavior-Net similarities is that in a face-swap deep fake, the facial identity of a person is modified but typically not the behavior. As a result, it is possible for a person's facial identity to be significantly different in a test video than in their reference videos, in which case, we should not be confident of the facial identity match.

III. RESULTS

We begin by describing the five datasets used for validation and analysis. We then describe the overall accuracy of detection followed by an analysis of robustness and relative importance of the appearance and behavioral features.

A. Datasets

The world leaders dataset (WLDR) [18] consists of several hours of real videos of five U.S. political figures, their political impersonators, and face-swap deep fakes between each political figure and their corresponding impersonator. We augmented this dataset with five new U.S. political figures.

The FaceForensics++ dataset (FF) [27] consists of 1000 YouTube videos of 1000 different people, mostly news anchors

and video bloggers. Each video was used to create four types of deep fakes: DeepFake, FaceSwap, Face2Face, and Neural Textures. We only use the first two categories of fakes as only these are face-swap deep fakes. After removing videos with multiple people or with identities overlapping to other datasets, we were left with 990 real videos and the corresponding 1980 deep fake videos.

The DeepFake Detection dataset (DFD) [28] by Google/Jigsaw consists of 363 real and 3068 face-swap deep fakes of 28 paid and consenting actors. Each individual was made to perform tasks like walking, hugging, talking, etc. in different expressions ranging from happy, to angry, neutral, or disgust. For our analysis, we selected only those videos where the individual was talking, resulting in 185 real and 1577 deep fake videos.

The Deep Fake Detection Challenge Preview dataset (DFDC-P) [29] consists of 1131 real and 4113 face-swap deep fakes videos of 66 individuals of various genders, ages and ethnicities.

The Celeb-DF (Ver. 2) dataset (CDF) [30] is currently the largest publicly available deep-fake dataset. It is reported as containing 5639 face-swap deep fakes generated from 590 YouTube videos of 61 celebrities speaking in different settings ranging from interviews, to TV-shows, and award functions (we, however, only identified 59 unique identities in the downloaded dataset).

For each identity in the WLDR, DFD, and DFDC-P datsets, a random 80% of the real videos are used for the reference set and the remaining 20% are used for testing. In these three datasets there were sufficient videos of each individual in similar contexts. In contrast, the FF and CDF datasets had either only a small number of videos per individual or the context for each individual varied drastically. For these two datasets, therefore, we take a different approach to creating the reference/testing sets. In particular, each real video is divided in half, the first half of which is used for reference, and the second half used for testing. Similarly, we split each fake video in half, discard the first half and subject the second half to testing. The first half is discarded because the real counterpart of this video is used for reference, thus avoiding any overlap in utterances between the reference and testing. We recognize that this split is not ideal as video halves are not independent, but as we will see below, there is little difference in the results between the $80/20$ splits and these $50/50$ splits.

Each reference and testing video is re-saved at a frame-rate of 25fps (and a ffmpeg quality of 20). This consistent frame-rate allows us to partition each video into overlapping 4-second clips, each of 100 frames, with a 5-frame sliding window.

B. Identification

Shown in Table I is the measure of efficacy for our algorithm for each of five datasets, along with the average across all datasets. The performance is reported as the area under the curve (AUC) of the receiver operating characteristic curve (top) and the detection accuracies of real and fake videos (bottom) for a fixed threshold of $\tau_f = 0.86$.

TABLE II
COMPARISON OF OUR APPROACH WITH PREVIOUS WORK OVER MULTIPLE BENCHMARKS [30]. THE REPORTED VALUES CORRESPOND TO THE AUC. ALTHOUGH NOT A PERFECT COMPARISON DUE TO SIGNIFICANTLY DIFFERENT UNDERLYING METHODOLOGIES, OUR APPROACH DOES PERFORM WELL. THE FF DATASET IN THIS COMPARISON CONSISTS OF THE *F*ACE*S*WAP AND *D*EEPFAKE CATEGORIES.

	WLDR	FF	DFD	DFDC-P	CDF
Protecting World Leaders [18]	0.93	–	–	–	–
2-stream [7]	–	0.70	0.52	0.61	0.53
XceptionNet-c23 [30]	–	**0.99**	0.85	0.72	0.65
Head Pose [16]	–	0.47	0.56	0.55	0.54
MesoNet [31]	–	0.84	0.76	0.75	0.54
Face Warping [11]	–	0.80	0.74	0.72	0.56
Ours: Appearance and Behavior	**0.99**	**0.99**	**0.95**	**0.93**	**0.99**

Note that the accuracy for the DFDC-P is unusually low. This is because many of the fake videos in this dataset failed to convincingly map the facial appearance of the desired source identity into the target video. Shown in Fig. 2 is a representative example of this problem. Shown is one frame from the source video, one frame from the target video, and the corresponding frame from the face-swap deep fake video in which the source identity should be mapped into the target video. In this example drawn from the DFDC-P dataset, we can clearly see that the source identity was not mapped into the target video, but rather continues to look like the target. This difference accounts for the low accuracy on the DFDC-P dataset as both behavior and appearance of the fakes correspond to the target identity and are thus classified as real by our algorithm. Although this effect is most pronounced in the DFDC-P dataset, the DFD dataset also suffers from a similar problem, failing to convincingly map the source to the target identity.

We next evaluate our detection algorithm against three in-the-wild, face-swap deep fake videos downloaded from YouTube. These three deep fakes were created using the following source and target combinations: 1) Steve Buscemi mapped onto Jennifer Lawrence [1]; 2) Tom Cruise mapped onto Bill Hader [2]; and 3) Billie Eilish mapped onto Angela Martin [3]. Because, only Jennifer Lawerence was already in our reference set (CDF), real videos for the other five identities were downloaded from YouTube to augment our reference set. This included three minutes of videos of Angela Martin from The Office and 20 minutes of interview videos for each of Billie Eilish, Steve Buscemi, Bill Hader, and Tom Cruise. The accuracy rate for each of these face-swap deep fakes is 100%.

Lastly, shown in Table II is a comparison of our detection accuracy, measured using area under the curve (AUC), to six previous deep-fake detection schemes. Our scheme outperforms or is equal to previous approaches across all datasets. Note, however, that this is not a perfect comparison because our approach has access to a reference set of only real videos to

[1] https://www.youtube.com/watch?v=r1jng79a5xc
[2] https://www.youtube.com/watch?v=VWrhRBb-1Ig
[3] https://www.instagram.com/p/B6lXvJlIU92/

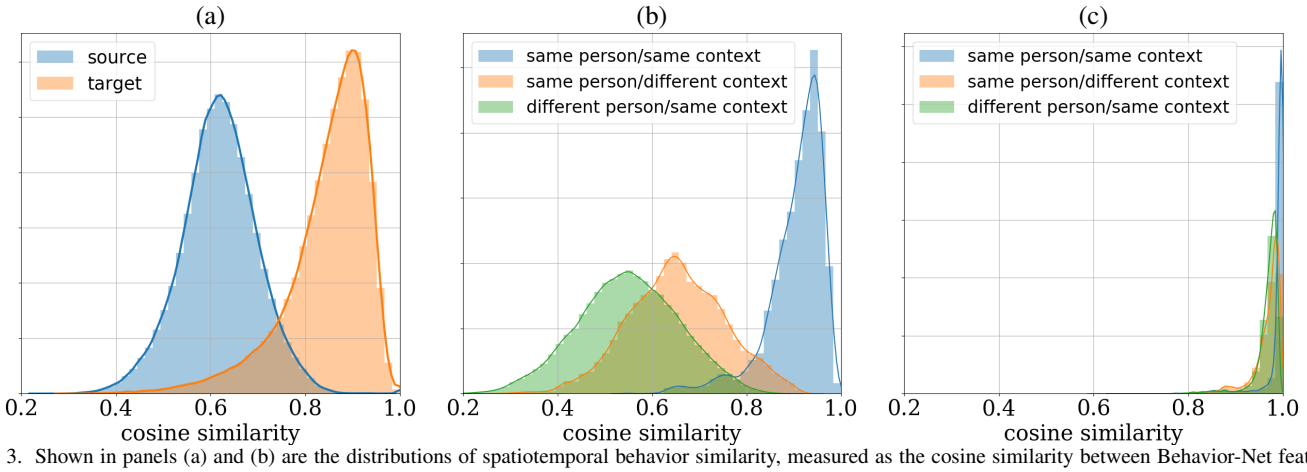

Fig. 3. Shown in panels (a) and (b) are the distributions of spatiotemporal behavior similarity, measured as the cosine similarity between Behavior-Net feature vectors. Shown in panel (c) is the distribution of spatial FAb-net similarity. See Section III-C for a detailed explanation of each panel.

compare against, as compared to these other fully-supervised approaches with access to real and fake reference videos.

C. Analysis

Our Behavior-Net feature was designed to capture spatiotemporal behavior, while the VGG feature captures facial identity. Here we analyze our results in more detail to ensure that these two features are not entangled and that the Behavior-Net does in fact capture temporal properties not captured by the static FAb-Net features.

In the first analysis, we show that Behavior-Net does in fact capture behavior and not just a person's facial identity. Shown in Fig. 3(a) are the distributions of Behavior-Net similarities between source (blue)/target (orange) identities relative to their face-swap deep fakes (recall that a face-swap deep fake is created by mapping an identity in a source video to a target video). The similarity of the target behavior relative to the face-swap deep fakes is much higher than the source, meaning that even though the facial identity in the deep fake matches the source, the behavioral identity still matches the target. This indicates that the Behavior-Net is capturing more information than just facial identity.

In the second analysis, we show that Behavior-Net captures identity-specific behaviors and not just identity-agnostic expressions or behaviors. This analysis is based on the real videos in the DFD dataset, where each of the 28 actors were recorded talking in different contexts ranging from a casual conversation sitting on a couch to a speech at a podium. Each of these contexts captured a specific facial expression ranging from neutral, to angry, happy, and laughing. And, each of these contexts were recorded twice, once with a still camera and once with moving camera. Shown in Fig. 3(b) are the distributions of Behavior-Net similarities between the same person in the same context (blue), the same person in different contexts (orange), and different people in the same context (green). When different people are recorded in the same context, we see that their Behavior-Net features are not similar, indicating that Behavior-Net captures identity-specific behaviors and not just specific contexts. At the same time, however, we see, that context can change an individual behavior (the orange vs. blue distributions). For example, a person is likely to have a different behavior when they are speaking casually to their friends as opposed to giving a formal speech to a large crowd. Nevertheless, our Behavior-Net captures identity-specific behaviors, albeit somewhat context dependent. Shown in Fig. 3(c) are the same distributions as in panel (b) but for only the static FAb-Net features. The distributions for the same person in the same context (blue), the same person in different contexts (orange), and different people in the same context (green) are all nearly identical, revealing that the static FAb-Net features does not capture identity-specific information.

In the third analysis, we analyse the amount of data required to build a reference set for an individual. For this analysis, the same reference set as before was used for the identities in FF, DFD, DFDC-P, and CDF. For the identities in the WLDR dataset (the only one with hours of video per person), the reference sets consists of between 1 and 2000 randomly selected 4-second clips. With 2, 30, 50, 100, 1000, and 2000 video clips, the average detection accuracy for identities in the WLDR dataset are 65.4%, 92.2%, 93.2%, 94.0%, 97.3%, and 97.7%, respectively. This rapid increase in accuracy and leveling off shows that large reference sets are not needed, assuming, again, that the context in which the individual is depicted is similar.

In this fourth, and final, analysis, we analyse the robustness of classification against a simple compression laundering operation. The video clips in our reference and testing sets, Section III-B, are each encoded at a relatively high ffmpeg quality of qp=20 (the lower this value, the higher the quality). Each testing video clip was recompressed at a lower quality of qp= 40 and classified against the original reference set. For the same threshold ($\tau_f = 0.86$), the average detection accuracy remains high at 94.5% (WLDR), 98.1% (FF), 93.2% (DFD), 80.9% (DFDC-P), and 93.3% (CDF). These results are almost identical to the high-quality videos in Table I.

IV. Discussion

We have developed a novel technique for detecting face-swap deep fakes. This technique leverages a fundamental flaw in these deep fakes in that the person depicted in the video is simply not the person that it purports to be. We have shown that a combination of a facial and behavioral biometric is highly effective at detecting these face-swap deep fakes. Unlike many other techniques, this approach is less vulnerable to counter attack and generalizes well to previously unseen deep fakes with previously unseen people.

Our forensic technique should generalize to so-called puppet-master deep fakes in which one person's facial expressions and head movements are mapped onto another person. These deep fakes suffer from the same basic problem as face-swap deep fakes in that the underlying behavior of the person is not that who it purports to be. As such, our combined facial and behavioral biometric should be able to detect these deep fakes.

We will, however, likely struggle to classify so-called lip-sync deep fakes in which only the mouth has been modified to be consistent with a new audio track. The facial identity and the vast majority of the behavior in these deep fakes will be consistent with the person depicted. To overcome this limitation, we seek to customize our behavioral model to learn explicit inconsistencies between the mouth and the rest of the face and/or underlying audio signal.

Acknowledgment

The PI's research group (Farid) is partially supported with funding from the Defense Advanced Research Projects Agency (DARPA FA8750-16-C-0166). The views, opinions, and findings expressed are those of the authors and should not be interpreted as representing the official views or policies of the Department of Defense or the U.S. Government. The PI's research group is also partially supported by Facebook. There is no collaboration between Facebook and DARPA. We thank Yipin Zhou for her help in data collection.

References

[1] A. v. d. Oord, S. Dieleman, H. Zen, K. Simonyan, O. Vinyals, A. Graves, N. Kalchbrenner, A. Senior, and K. Kavukcuoglu, "Wavenet: A generative model for raw audio," 2016. 1

[2] 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*, 2019, pp. 4401–4410. 1

[3] T. Karras, S. Laine, M. Aittala, J. Hellsten, J. Lehtinen, and T. Aila, "Analyzing and improving the image quality of stylegan," 2019. 1

[4] R. Tolosana, R. Vera-Rodriguez, J. Fierrez, A. Morales, and J. Ortega-Garcia, "Deepfakes and beyond: A survey of face manipulation and fake detection," 2020. 1

[5] B. Chesney and D. Citron, "Deep fakes: A looming challenge for privacy, democracy, and national security," *California Law Review*, vol. 107, p. 1753, 2019. 1

[6] H. Farid, *Photo Forensics*. MIT Press, 2016. 1

[7] P. Zhou, X. Han, V. I. Morariu, and L. S. Davis, "Two-stream neural networks for tampered face detection," in *IEEE Conference on Computer Vision and Pattern Recognition Workshops*, 2017. 1, 4

[8] N. Yu, L. Davis, and M. Fritz, "Attributing Fake Images to GANs: Learning and Analyzing GAN Fingerprints," in *IEEE International Conference on Computer Vision*, 2018. 1

[9] X. Zhang, S. Karaman, and S.-F. Chang, "Detecting and simulating artifacts in GAN fake images," arxiv: 1907.06515, 2019. 1

[10] S.-Y. Wang, O. Wang, R. Zhang, A. Owens, and A. A. Efros, "CNN-generated images are surprisingly easy to spot...for now," in *IEEE Conference on Computer Vision and Pattern Recognition*, 2020. 1

[11] Y. Li and S. Lyu, "Exposing deepfake videos by detecting face warping artifacts," arXiv: 1811.00656, 2018. 1, 4

[12] L. Li, J. Bao, T. Zhang, H. Yang, D. Chen, F. Wen, and B. Guo, "Face X-ray for more general face forgery detection," *arXiv preprint arXiv:1912.13458*, 2019. 1

[13] M. Huh, A. Liu, A. Owens, and A. A. Efros, "Fighting fake news: Image splice detection via learned self-consistency," in *European Conference on Computer Vision*, 2018. 1

[14] N. Carlini and D. Wagner, "Towards evaluating the robustness of neural networks," arXiv: 1608.04644, 2016. 1

[15] Y. Li, M.-C. Chang, and S. Lyu, "In ictu oculi: Exposing AI created fake videos by detecting eye blinking," in *IEEE International Workshop on Information Forensics and Security*, 2018, pp. 1–7. 1

[16] X. Yang, Y. Li, and S. Lyu, "Exposing deep fakes using inconsistent head poses," in *IEEE International Conference on Acoustics, Speech and Signal Processing*, 2019, pp. 8261–8265. 1, 4

[17] U. A. Ciftci and I. Demir, "Fakecatcher: Detection of synthetic portrait videos using biological signals," arXiv: 1901.02212, 2019. 1

[18] S. Agarwal, H. Farid, Y. Gu, M. He, K. Nagano, and H. Li, "Protecting world leaders against deep fakes," in *IEEE Conference on Computer Vision and Pattern Recognition, Workshop on Media Forensics*, 2019, pp. 38–45. 1, 3, 4

[19] P. Ekman and W. V. Friesen, "Measuring facial movement," *Environmental Psychology and Nonverbal Behavior*, vol. 1, no. 1, pp. 56–75, 1976. 1

[20] B. Schölkopf, J. C. Platt, J. C. Shawe-Taylor, A. J. Smola, and R. C. Williamson, "Estimating the support of a high-dimensional distribution," *Neural Computation*, vol. 13, no. 7, pp. 1443–1471. 1

[21] O. Wiles, A. S. Koepke, and A. Zisserman, "Self-supervised learning of a facial attribute embedding from video," arXiv: 1808.06882, 2018. 2

[22] X. Wang, X. Han, W. Huang, D. Dong, and M. R. Scott, "Multi-similarity loss with general pair weighting for deep metric learning," in *IEEE Conference on Computer Vision and Pattern Recognition*, 2019. 2, 3

[23] J. S. Chung, A. Nagrani, and A. Zisserman, "VoxCeleb2: Deep speaker recognition," arXiv: 1806.05622, 2018. 3

[24] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *IEEE Conference on Computer Vision and Pattern Recognition*, 2016, pp. 770–778. 3

[25] O. M. Parkhi, A. Vedaldi, and A. Zisserman, "Deep face recognition," in *British Machine Vision Conference*, 2015. 3

[26] T. Baltrušaitis, P. Robinson, and L.-P. Morency, "OpenFace: an open source facial behavior analysis toolkit," in *IEEE Winter Conference on Applications of Computer Vision*, 2016, pp. 1–10. 3

[27] A. Rössler, D. Cozzolino, L. Verdoliva, C. Riess, J. Thies, and M. N. ner, "FaceForensics++: Learning to detect manipulated facial images," arXiv: 1901.08971, 2019. 3

[28] N. Dufour and A. Gully, "Contributing Data to Deepfake Detection Research," https://ai.googleblog.com/2019/09/contributing-data-to-deepfake-detection.html. 4

[29] B. Dolhansky, R. Howes, B. Pflaum, N. Baram, and C. C. Ferrer, "The deepfake detection challenge (DFDC) preview dataset," arXiv: 1910.08854, 2019. 4

[30] Y. Li, X. Yang, P. Sun, H. Qi, and S. Lyu, "Celeb-DF: A new dataset for deepfake forensics," arXiv: 1909.12962, 2019. 4

[31] D. Afchar, V. Nozick, J. Yamagishi, and I. Echizen, "Mesonet: A compact facial video forgery detection network," in *IEEE International Workshop on Information Forensics and Security*, 2018. 4