You’re not acting like yourself: Deepfake Detection Based on Facial Behavior

Alexandre Libourel, Jean-Luc Dugelay
Digital Security
Eurecom
Biot, France
{libourel, dugelay}@eurecom.fr

***Abstract*—Politicians and government leaders are critical targets for deepfake attacks. A single deepfake involving these individuals can severely damage their careers or, in extreme cases, pose a national security threat. Attackers can leverage vast amounts of publicly available audio and video recordings to train their models, making this threat even more pressing. In response, specialized deepfake detectors have been developed to focus on detecting deepfakes targeting a specific Person of Interest (POI). By learning facial expressions and movements unique to the POI, these detectors can identify inconsistencies in deepfakes where these authentic attributes are absent. However, previous methods relied on Facial Action Units, which offer an incomplete representation of the POI’s behavior. In this paper, we propose a novel approach to learning POI-specific movements without requiring deepfake samples during training, making it independent of any deepfake generation methods. Although our technique is speaker-dependent, it provides a robust solution for protecting high-profile individuals who are particularly exposed to deepfake threats.**

***Index Terms*—deepfake detection, biometrics, media forensics, behavioral analysis, POI recognition**

I. Introduction

The recent advancements in generative AI, particularly with GANs [7] and diffusion models [9], have made it relatively straightforward for anyone with standard computational resources and a basic understanding of computer science to create a highly realistic deepfake of a public figure. The most common deepfake generation techniques are Face-Swap and Face-Reenactment. The first method consists of ”pasting” the face of person A on the body of person B, while the second consists of animating a picture of person A with the facial expression of person B. Both methods result in a deepfake where person A is depicted moving and saying things they never said (herited from person B).

The growth of mass media and social networks has led to an increased risk of fake multimedia content being used to spread misinformation or discredit public figures. This is even more true during election campaigns. Deepfake detectors are essential for filtering out forged content before it spreads online, thereby preventing as much as possible the dissemination of misinformation.

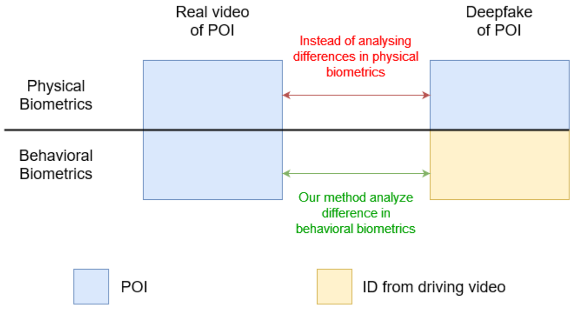
Real video of POI
Deepfake of POI
Instead of analysing differences in physical biometrics
Physical Biometrics
Behavioral Biometrics
Our method analyze difference in behavioral biometrics
POI
ID from driving video

Fig. 1. Principle of behavioral biometrics comparison in the context of deepfake detection.

Most deepfake detectors are trained on datasets to find traces of manipulation left by generators. However, the detectors struggle to accurately classify deepfakes created by previously unseen generators, due to the introduction of unique artifacts with each generator. While deepfake detectors usually reach around 99% of video-level AUC on train datasets, they usually don’t get better performances than 70% of video-level AUC with different datasets [16]. There is a deep asymmetry between the research on generation and detection. On one hand, generation relies on minimizing the number of modifications required to change one face into another. On the other hand, detection aims to detect any trace of manipulation on images to be robust to any deepfake generator. However, today’s generators no longer leave crude traces of tampering, making deepfake detectors not robust to in-the-wild deepfake detection. Indeed, harmless traces of manipulation like compression or social media filters can be interpreted as potential traces of malicious manipulation [13], [15].

To improve the generalization to unknown deepfake generators, some research was conducted to create deepfake detectors conditioned by the identity of a Person Of Interest (POI). The objective of this change of paradigm is to enable the detection of any deepfake from the POI.

In comparison to works on the generalization of deepfake detectors, few studies have tried to create deepfake detectors specialized to a specific POI. The main issue is that such a method is not easily scalable to protect everyone from deepfake generation. It can be used to protect critical POIs such as world leaders. However, focusing on the detection of deepfakes for a POI enables us to leverage the consistency of biometric information. This data can be divided into two categories:

This work is part of the DeTOX project. It has been founded by ”Astrid: Guerre Cognitive”, a French defense program initiative.

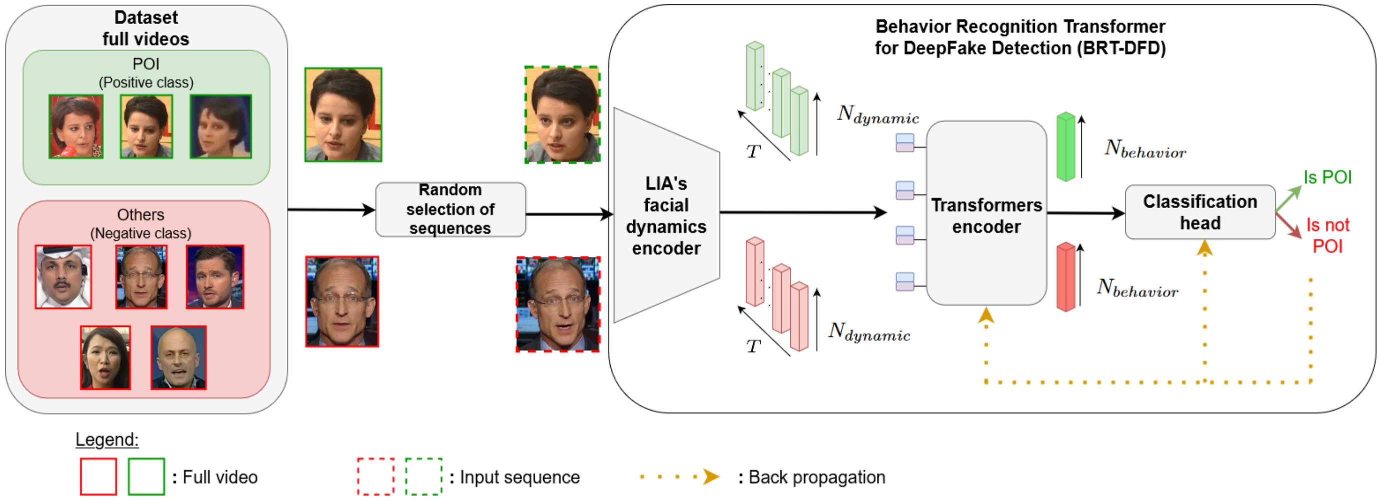

Fig. 2. Training Pipeline. We use head-pose + facial expression description from LIA's encoder for all frames of the sequence to obtain a $h \in \mathbb{R}^{T \times N_{dynamic}}$. h is then fed to a Transformers encoder learning to produce facial behavior embeddings in $\mathbb{R}^{N_{behavior}}$ for POI recognition and a classification head gives a prediction ($T = 40$, $N_{dynamic} = 512$, $N_{behavior} = 512$).

physical and behavioral [10]. The deepfakes created by the current generators are videos where the physical biometrics are close to the POI whereas the behavioral biometrics come from the identity of the driving video, as illustrated in Figure 1. Some methods were developed to capture this behavioral information and used in the context of deepfake detection [1], [4] but these methods rely on a combination of multiple hand-crafted features to monitor facial movement which might not be relevant for describing accurately all the different facial behaviors of different POIs. One of these methods [4] employed supervised learning to model the facial behavior of a POI against a wide range of deepfakes. However, this approach requires significant computational resources and time, and its effectiveness may depend on the specific manipulation techniques used to create the deepfakes.

In this work, we develop a new method for extracting facial movements, to enhance the detection of deepfakes of a POI. An encoder of facial dynamics from a state-of-the-art face reenactment generator is employed to provide a comprehensive representation of the face's expression, pose, gaze direction,... minimizing the inclusion of unintended physical appearance information (i.e. physical biometrics data). This approach ensures that the prediction is primarily based on the observed facial behavior (i.e. behavioral biometrics data). The contributions of our article are the following.

- We build training datasets containing hundreds of real audio/video recordings of 3 different politicians. We also create deepfakes of them for the evaluation process.
- We propose a new ID recognition algorithm trained to recognize the facial behavior of a POI against real videos containing different IDs.
- We show that this face recognition model can be used to detect deepfakes of a POI without using any deepfake during the training phase.

The paper is organized as follows: Section II proposes a short overview of deepfake detection. In Section III we describe our method and pipeline to extract relevant temporal information on facial movement. Finally, Section IV shows our results on our carefully crafted database.

II. Related Works

Recent works in deepfake detection focus on creating a universal deepfake detector [2], [6], [8], [11], [23], i.e. a detector capable of detecting any deepfakes. To generalize deepfake detection to new generators, these detectors rely on the detection of artifacts. These artifacts are traces of manipulation - often invisible to the human eye - introduced by the deepfake generator. Each generation technique produces unique artifacts. As a result, deepfake detectors often struggle to accurately identify deepfakes created by different techniques. [3], [16], [19], [24].

Another stumbling block to the generalization of deepfake detection is the introduction of artifacts from compression or social media upload pipelines [13], [15]. The artifacts introduced by the genuine image processing pipelines are detected as malicious traces of manipulation, impacting severely the detection accuracy of the deepfakes.

To improve deepfake detection of critical POIs, some work was made to obtain robust deepfake detection against unseen manipulation methods by adding prior knowledge from the POI [5], [17]. Finally, some studies investigate the facial mannerisms of POIs to improve deepfake detection. Boháček et. al [1] designed a one-class SVM to detect fake videos of the Ukrainian president Zelenskyy using head pose and facial action units using only real videos from the POI to train their network. However, the set of action units used for the study does not give a complete description of all possible movements of the face, as demonstrated by Chu et. al [4] where they created a custom dataset of American politicians and failed to correctly classify the different POIs based on their Facial

Action Units. In order to enhance the classification process, a set of 42 novel facial parameters was devised and calculated using three-dimensional facial landmarks. However, since the network was trained on deepfakes of the POIs created with only a few specific generators, this approach may inadvertently learn their specific artifacts, which can limit its ability to generalize to other generation methods.

In this work, we propose a new method to extract a full description of facial movements and a new training strategy that does not require fake data during training, making our method independent of unseen manipulation methods.

III. Method

We introduce the Behavior Recognition Transformers for Deepfake Detection (BRT-DFD), a novel pipeline that is built upon a unique combination of existing techniques to learn facial behaviors specific to a POI. Once trained, BRT-DFD can effectively detect deepfakes of the POI by identifying deviations in these personalized behaviors. The full pipeline is depicted in Figure 2.

From a sequence of frames, BRT-DFD extracts features that describe face dynamic (head pose, gaze direction, facial expression, emotion,...) on all the frames of the sequence thanks to the movement encoder of the LIA deepfake generator [22]. We obtain a sequence of facial dynamic encodings for all the frames. These features are fed to a simple Transformers encoder [21] to learn the movements between the POI and a set of other identities. Finally, a classification head is used to perform the final prediction, i.e. ”This movement belongs to the POI” (positive class) or ”This movement does not belong to the POI” (negative class). A Cross-Entropy loss is used during the training to learn the facial behaviors associated with the POI.

A. Pre-processing and Movement features extraction

First, all the videos are cropped around the head following LIA's preprocessing [20], [22] and the videos are converted to sequences. The video-to-sequence process is fully detailed in Section IV-A.

The pose features are obtained after passing the sequence through LIA's head pose+facial expression descriptor. Here is a brief description of how this encoder works. LIA is an autoencoder trained to reconstruct a face image x_s (source image) in another pose given by the image x_d (driving image). Two encoders are used in LIA's network.

- The identity encoder E_I: It captures the distinct facial attributes that define the physical appearance of image x_s. $E_I(x_s)$ returns $z_{s\to r}$ a latent representation of x_s identity in a facial dynamic of reference.
- The facial dynamic encoder E_D: It captures the facial expression, the head-pose, and other information relative to facial dynamic present in image x_d. $E_D(x_d)$ returns $w_{r\to d}$ a latent representation of the wrapping operation to perform to transform a face with a facial dynamic of reference to the head-pose of x_d.

$z_{s\to r}$ and $w_{r\to d}$ are combined together to obtain $z_{s\to d}$

$$z_{s\to d} = E_I(x_s) + E_D(x_d) = z_{s\to r} + w_{r\to d} \tag{1}$$

where $z_{s\to d}$ is the latent representation of the face in x_s with the facial dynamic of x_d. $z_{s\to d}$ is used by LIA's decoder to reconstruct the face of x_s with the pose and expression of x_d.

In this work, we focus on LIA's pose encoder E_D. The use of this motion decomposition highlights a strong property of disentanglement between the facial features (i.e. physical biometrics) and the facial pose and expression (i.e. behavioral biometrics) in the latent vector $w_{r\to d}$. LIA's authors emphasized on the importance of their Linear Motion Decomposition (LMD) to disentangle facial dynamic and physical attributes. In Section IV-D, we show that the use of LMD allows better to detect deepfakes more accurately. For more information on how the LIA motion decomposition works, please refer to their publication [22].

Here, for a sequence of frames $s = (s_1, ..., s_T)$ all the frames are passed through LIA's pose encoder to obtain $h = (h_1, ..., h_T) \in \mathbb{R}^{T\times 512}$ a latent encoding of the facial position and expression of the video sequence.

B. Transformer encoder and classifier

With h we only have a description of the facial dynamic on all the frames of the sequence. We need to exploit the temporal information to learn the behaviors associated with the POI. To find relevant temporal information in our encodings we rely on a Transformers encoder. Thanks to their attention mechanisms, Transformers capture dependencies across all time steps simultaneously. Unlike traditional recurrent models, which process sequences step-by-step and can struggle with long-range dependencies, Transformers process the entire sequence in parallel. This allows them to model complex temporal relationships more efficiently and accurately, making them particularly powerful for tasks involving sequential data. Following Zheng et. al training guidelines [24] we add a [CLS] token before the Transformers encoder. The encoding of the [CLS] token is meant to embed the general context of the sequence. This embedding is then passed through a Multi-Layer Perceptron classifier to perform the final classification, i.e. ”*Does this movement belong to the POI, or not?*”.

The weights are updated using the Cross-Entropy loss.

$$L_{cls}(y, \hat{y}) = -\sum_{c=0}^{C} \mathbb{1}_{y=c} log(\hat{y}_c) \tag{2}$$

where in our case, $C = 1$ for binary classification.

IV. Experiments and Results

A. Database

To recognize the video of a POI we gathered 191 videos from the former French minister Najat Vallaud-Belkacem from her YouTube channel. We filtered out all the videos where she did not speak in front of the camera for more than 10 seconds without any cuts, zooms, or other forms of video editing. We kept 2 173 clips from 79 videos of her, speaking in different

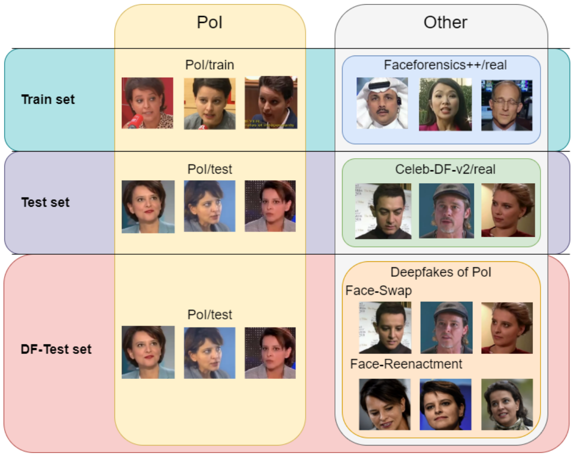

Fig. 3. Visual description of the database for one POI. **Train set** contains real videos of POI and real videos of FF++. No deepfakes were seen during training process. **Test set** has real videos of POI and real videos from Celeb-DF-v2. This serves as validation to check if we effectively manage to recognize the POI with facial dynamics. The **DF-Test set** includes the same real videos of the POI as the Test set, along with deepfake videos of the POI generated using the same real videos from the Celeb-DF-v2. Two generators were used.

places (radio studio, TV set, conference stage, National Assembly, Zoom meetings, ...) and in different contexts (political debate, indoor/outdoor interviews, conferences, promotional videos, ...) for a total duration of 4 hours, 43 minutes, and 21 seconds of videos.

To assess that our method also works with English speakers, we created 2 additional databases of POI with real videos from former American presidents Barack Obama and Donald Trump scrapped on YouTube. Real videos of a POI represent the positive class

Additionally, we have videos of people who are different from the PoI. We use 997 real videos from FaceForensics++ [18] and 649 from Celeb-DF-v2 [12] for a total of 1646 videos for a total duration of 6 hours, 39 minutes, and 13 seconds. For the rest of the paper, we will refer to this subset as the negative class. 3 videos of Faceforensics++ contained videos of Obama. We suppressed them during the training of Obama's model. A summary is given in Table I.

TABLE I
DESCRIPTION OF THE DIFFERENT REAL VIDEO DATABASES

Databases	**Number of real videos**	**Total duration**
FF++ and Celeb-DF-v2	1528	6h 39min
Najat Vallaud-Belkacem	79	4h 43min
Barack Obama	72	2h 54min
Donald Trump	80	1h 32min

Finally, we split the dataset between train and test. In the train set, the negative class is represented by the real videos from FF++, and the real videos of Celeb-DF-v2 are put in the test set. For the positive class, we split the videos so that 80% of the real videos are in the **Train set**, and the 20% remaining in the **Test set**. Finally, we also create the **DF-Test set**, composed of the POI's real video from the test set and deepfake of the POI generated with the videos of the test set to asses if our model is robust to deepfake manipulation. Figure 3 gives a visual description of the different sets .

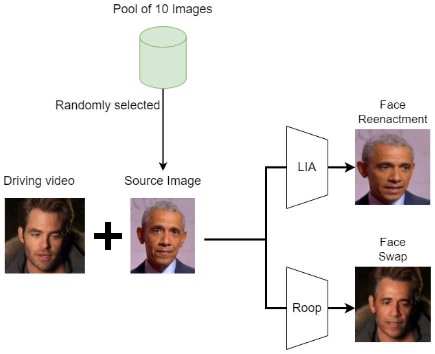

Fig. 4. Pipeline for the deepfake generation. Face Reenactment consists of animating a portrait of the POI with the movements of the POI. Face Swapping consists of swapping the face of the driving video with the face of the POI.

For each POI we generated around 500 Face Swaps and 500 Face Reenactments using real videos from Celeb-DF-v2 as driving video and one image randomly chosen from a pool of 10 images of the POI as source image. The pipeline of the generation of deepfake is presented in Figure 4. We used LIA to generate face reenactment deepfakes and Roop[1] for face swapping. Roop's face swap model comes from the Insightface[2] project. It is worth noting that our test set is composed of videos generated with LIA and our method rely on the encoder of LIA. One may argue that our result on LIA-generated video might be biased. However, since we only use the facial expression encoder and not the decoder, we think that there must not be a strong bias on the detection of LIA-generated videos as seen in Table V where we test our model on different generators.

B. Training details

First, because each video has a different frame rate, we define a sequence as a succession of frames spaced by a fixed time interval. For example, to extract a sequence from a 30-FPS (*resp. 25-FPS*) video, we take one frame every 6 frames (*resp. 5 frames*) so that there is a 0.20-second interval between each successive frame. We show an example of the transformation in Figure 5. All the videos that do not have a frame rate of 25 or 30 were discarded before training and testing.

Second, to extract long-range facial behaviors, we select the minimal time duration of a video to be 8 seconds. Videos lasting less than 8 seconds have been discarded. During each epoch, and because one 8-second sequence is not representative of the full video, we randomly pick K 8-second sequences from each video.

We trained our model for 100 epochs on a GeForce RTX 3090 with 24 GB of VRAM. As in [24], we first apply warm-

[1] https://github.com/s0md3v/roop
[2] https://github.com/deepinsight/insightface

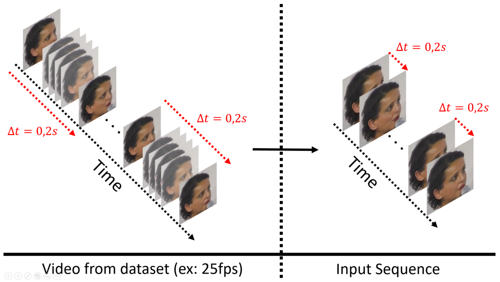

Fig. 5. Video-to-sequence conversion. In this example, we show how a 25-fps video is transformed into a sequence by extracting one frame every 5 frames.

up on the 10 first and then a cosine decay on the learning rate with an initial value of 10^{-4}. However, we use AdamW optimizer [14] instead of Stochastic Gradient Descent (SGD).

C. Results

To evaluate the performance of our model, we compute the Video-Level AUC: Area Under the Curve of the ROC curve at the video level. The score of each video is the average score obtained for K sequences of 8 seconds that come from the same video.

$$score(video_j) = \frac{1}{K}\sum_{k=1}^{K} score(sequence_{j,k}) \qquad (3)$$

In Table IV, we report the result obtained after training the network on multiple epochs and different classifiers.

We trained three models with real videos of French former minister Najat Vallaud-Belkacem, and former American presidents: Barack Obama and Donald Trump. The results in Table II show similar results in video-level AUC for different POIs. It proves that BRT-DFD extracts an embedding that contains enough movement description to learn behaviors from different POIs.

TABLE II
VIDEO LEVEL AUC (%) OBTAINED ON TRAIN, TEST, AND DF-TEST DATASET FOR FACIAL BEHAVIOR RECOGNITION WITH DIFFERENT POIS.

POI	**Train set**	**Test set**	**DF-Test set**
Najat Vallaud-Belkacem	99.99	96.57	91.47
Barack Obama	99.94	98.80	93.09
Donald Trump	100.0	98.80	93.77

We compare the performance of BRT-DFD with state-of-the-art deepfake detectors RECCE [2] and SBI [19], each trained using FF++. RECCE uses an auto-encoder to reconstruct real face images, along with a multi-scale graph reasoning module to detect generator-specific artifacts. In contrast, SBI was trained solely on real images versus self-blended images, without any deepfakes. Results on our custom databases are shown in Table III. For RECCE and SBI, the video-level AUC calculation is adjusted, with $sequence_{j,k}$ replaced by $frame_{j,k}$ in Equation 3.

We observe that RECCE and SBI struggle to maintain high performance across different POIs, whereas our method provides more consistent predictions. Additionally, both state-of-the-art detectors show better detection performance on the Obama subset. The driving videos for generating deepfakes were sourced from Celeb-DF-v2, a dataset containing real video clips of American actors. Due to the predominance of white actors in this dataset, manipulation traces were more noticeable, which resulted in better classification when evaluating the Obama subset.

TABLE III
VIDEO LEVEL AUC (%) OBTAINED ON DF-TEST SET OF DIFFERENT POIS WITH SOTA DETECTORS. IN BOLD, THE BEST-PERFORMING DETECTOR.

POI	RECCE	SBI	BRT-DFD (Ours)
Najat Vallaud-Belkacem	66.79	91.20	**91.47**
Barack Obama	82.41	**95.83**	93.09
Donald Trump	81.83	78.50	**93.77**

D. Ablation study

In this section, we show the importance of the disentanglement of physical biometrics and behavioral biometrics. We retrain our model after removing the Linear Motion Decomposition module from LIA's encoder. We obtain a classification score with and without LMD that we report in Table IV.

TABLE IV
VIDEO LEVEL AUC (%) OBTAINED ON DF-TEST FOR FACIAL BEHAVIOR RECOGNITION WITH AND WITHOUT LINEAR MOTION DECOMPOSITION. IN BOLD, THE BEST PERFORMANCES ARE OBTAINED WITH THE LMD MODULE

Method	DF-Test set		
	N. Vallaud-Belkacem	B. Obama	D. Trump
BRT w/o LMD	84.8	87.0	86.8
BRT w/ LMD	**91.47**	**93.09**	**93.77**

The removal of the LMD module results in a more important leakage of physical information. We observe a 5 to 7% decrease in video-level AUC. Because physical biometrics are more discriminating than behavioral biometrics, the networks tend to focus on appearance rather than facial dynamics. Consequently, the attenuation of physical information allows the model to focus on facial dynamics for the prediction.

E. Results per generator

We compared the video-level AUC on two subsets of DF-test to evaluate the behavior of BRT-DFD on Face Swap and Face Reenactment separately. We create **FS-Test set**, a subset of **DF-Test set** with the real videos of POI and FaceSwaps of the POI generated with Roop. Similarly, we create **FR-Test set**, a subset of **DF-Test set** with the real videos of the POI and the Face Reenactments of the POI generated with LIA. We observe that Face-Reenactment deepfakes are more difficult to distinguish from real videos than Face-Swap deepfakes. While Face-Swaps preserve the facial dynamics of the driving video, Face-Reenactments often produce deepfakes focused on a single, specific expression of the POI, which is subtly altered by the driving video's dynamics. As a result, Face-Reenactments maintain more of the POI's original facial dynamics than Face-Swaps.

TABLE V
VIDEO LEVEL AUC (%) OBTAINED PER DEEPFAKE GENERATORS. FS AND FR STAND RESPECTIVELY FOR FACE SWAP AND FACE REENACTMENT

POI	FS-Test set	FR-Test set	**DF-Test set**
Najat Vallaud-Belkacem	91.37	91.57	91.47
Barack Obama	98.31	87.87	93.09
Donald Trump	98.00	89.54	93.77

V. DISCUSSION AND CONCLUSION

Previous methods [1], [4] failed in providing solutions that combined both, a full description of facial behaviors and training on real videos only. To detect deepfakes of a POI based on the facial behaviors of the video, we introduced BRT-DFD a Behavior Recognition Transformers for Deepfake Detection. This network was designed to learn a POI's facial behaviors against those of decoys. We emphasized that to achieve better results in POI deepfake detection, it is important to detect movement without leaking information from facial attributes to be sure that our prediction is mainly based on behavioral biometrics. LIA's motion descriptor works well at providing information about the facial dynamic (head pose, gaze direction, expression, etc...) without relying on physical biometrics data.

BRT-DFD is speaker-dependent; When one wants a deepfake detector for a new POI, one has to train a new model using new data. However, it provides a reliable solution for critical individuals such as world leaders, as there are no universal deepfake detectors today to protect all from deepfakes.

Further studies are necessary to determine the optimal ratio of videos featuring POIs to those featuring other identities during the training phase. An additional enhancement would be to incorporate vocal behavior into the detection process and develop a fusion model that integrates audio, video, and their synchronization. Finally, creating a public benchmark is necessary to properly evaluate the existing POI deepfake detectors.

ACKNOWLEDGMENT

This work was founded by the French ASTRID program, facilitated by the Defense Innovation Agency and the ANR under the DeTOX project (ANR-22-ASGC-0005).

REFERENCES

[1] M. Boháček and H. Farid, "Protecting world leaders against deep fakes using facial, gestural, and vocal mannerisms," *Proceedings of the National Academy of Sciences*, vol. 119, no. 48, p. e2216035119, 2022.

[2] J. Cao, C. Ma, T. Yao, S. Chen, S. Ding, and X. Yang, "End-to-end reconstruction-classification learning for face forgery detection," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2022, pp. 4113–4122.

[3] L. Chen, Y. Zhang, Y. Song, L. Liu, and J. Wang, "Self-supervised learning of adversarial example: Towards good generalizations for deepfake detection," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2022, pp. 18 710–18 719.

[4] B. Chu, W. You, Z. Yang, L. Zhou, and R. Wang, "Protecting world leader using facial speaking pattern against deepfakes," *IEEE Signal Processing Letters*, vol. 29, pp. 2078–2082, 2022.

[5] D. Cozzolino, A. Rössler, J. Thies, M. Nießner, and L. Verdoliva, "Id-reveal: Identity-aware deepfake video detection," in *Proceedings of the IEEE/CVF international conference on computer vision*, 2021, pp. 15 108–15 117.

[6] S. Dong, J. Wang, R. Ji, J. Liang, H. Fan, and Z. Ge, "Implicit identity leakage: The stumbling block to improving deepfake detection generalization," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2023, pp. 3994–4004.

[7] I. Goodfellow, J. Pouget-Abadie, M. Mirza, B. Xu, D. Warde-Farley, S. Ozair, A. Courville, and Y. Bengio, "Generative adversarial nets," *Advances in neural information processing systems*, vol. 27, 2014.

[8] Y. Guo, C. Zhen, and P. Yan, "Controllable guide-space for generalizable face forgery detection," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2023, pp. 20 818–20 827.

[9] J. Ho, A. Jain, and P. Abbeel, "Denoising diffusion probabilistic models," *Advances in neural information processing systems*, vol. 33, pp. 6840–6851, 2020.

[10] L. Johnson, "Chapter 11 - security component fundamentals for assessment," in *Security Controls Evaluation, Testing, and Assessment Handbook (Second Edition)*, 2nd ed., L. Johnson, Ed. Academic Press, 2020, pp. 471–536.

[11] D.-K. Kim and K.-S. Kim, "Generalized facial manipulation detection with edge region feature extraction," in *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, 2022, pp. 2828–2838.

[12] Y. Li, X. Yang, P. Sun, H. Qi, and S. Lyu, "Celeb-df: A large-scale challenging dataset for deepfake forensics," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2020, pp. 3207–3216.

[13] A. Libourel, S. Husseini, N. Mirabet-Herranz, and J.-L. Dugelay, "A case study on how beautification filters can fool deepfake detectors," in *IWBF 2024, 12th IEEE International Workshop on Biometrics and Forensics*, 2024.

[14] I. Loshchilov, "Decoupled weight decay regularization," *arXiv preprint arXiv:1711.05101*, 2017.

[15] Y. Lu and T. Ebrahimi, "Impact of video processing operations in deepfake detection," in *2023 24th International Conference on Digital Signal Processing (DSP)*. IEEE, 2023, pp. 1–5.

[16] A. V. Nadimpalli and A. Rattani, "On improving cross-dataset generalization of deepfake detectors," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2022, pp. 91–99.

[17] S. Ramachandran, A. V. Nadimpalli, and A. Rattani, "An experimental evaluation on deepfake detection using deep face recognition," in *2021 International Carnahan Conference on Security Technology (ICCST)*. IEEE, 2021, pp. 1–6.

[18] A. Rossler, D. Cozzolino, L. Verdoliva, C. Riess, J. Thies, and M. Nießner, "Faceforensics++: Learning to detect manipulated facial images," in *Proceedings of the IEEE/CVF international conference on computer vision*, 2019, pp. 1–11.

[19] K. Shiohara and T. Yamasaki, "Detecting deepfakes with self-blended images," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2022, pp. 18 720–18 729.

[20] A. Siarohin, S. Lathuilière, S. Tulyakov, E. Ricci, and N. Sebe, "First order motion model for image animation," *Advances in neural information processing systems*, vol. 32, 2019.

[21] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, Ł. Kaiser, and I. Polosukhin, "Attention is all you need," *Advances in neural information processing systems*, vol. 30, 2017.

[22] Y. Wang, D. Yang, F. Bremond, and A. Dantcheva, "Latent image animator: Learning to animate images via latent space navigation," in *International Conference on Learning Representations*, 2022.

[23] Z. Yan, Y. Luo, S. Lyu, Q. Liu, and B. Wu, "Transcending forgery specificity with latent space augmentation for generalizable deepfake detection," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2024, pp. 8984–8994.

[24] Y. Zheng, J. Bao, D. Chen, M. Zeng, and F. Wen, "Exploring temporal coherence for more general video face forgery detection," in *Proceedings of the IEEE/CVF international conference on computer vision*, 2021, pp. 15 044–15 054.