Who is driving this deepfake? Beyond Deepfake Detection with Driver Identification

1st Alexandre Libourel
Digital Security
Eurecom
Biot, France
libourel@eurecom.fr

2nd Jean-Luc Dugelay
Digital Security
Eurecom
Biot, France
dugelay@eurecom.fr

***Abstract*—The rapid advancement of deepfake technology has raised significant concerns about the authenticity of digital media and its potential misuse. While much progress has been made in developing methods to detect whether a video is fake or not, a critical question remains: Can we go one step further? What additional information can be derived once a deepfake is identified? Beyond merely flagging manipulated content, understanding the source of the manipulation holds significant value for forensics and investigation. This paper addresses one aspect of this challenge by demonstrating how to recover information from the driving video, i.e., the input video guiding the deepfake generation, to identify the person acting in the driving video (suspected driver). By learning facial expressions and movements unique to a suspected driver, we can identify which deepfake has been generated using videos of the suspected driver in a pool of deepfakes. While the current limitation of this work implies having a large quantity of data concerning your suspected identity, this work proves the feasibility of deducing information on driving videos directly from the deepfakes. Code available at: https://github.com/Thiresias/BRT-driver-identification**

***Index Terms*—Deepfake, biometrics, media forensics, behavioral analysis**

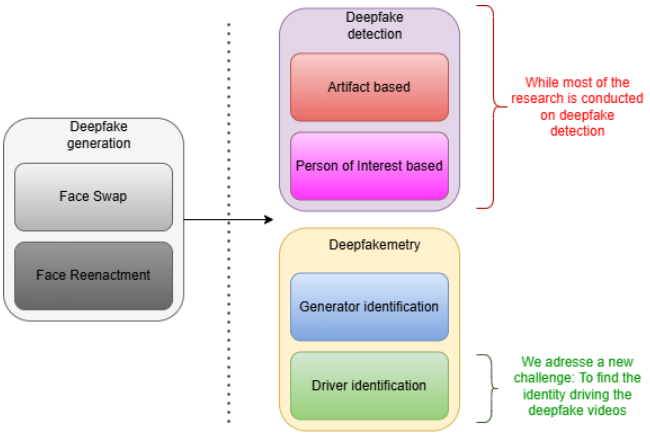

Fig. 1. In the cat-and-mouse game between the generation and detection of deepfakes, we present a new task of deepfake forensics: Identifying the person acting in the driving video.

I. Introduction

Research into synthetic media generation has witnessed unprecedented growth in recent years. Since the advent of Generative Adversarial Networks (GANs) [13] and, more recently, with Diffusion Models (DMs) [16], the number of fake media online has skyrocketed. Websites and platforms specializing in synthetic content generation have emerged, offering tutorials, tools, and software that enable even non-experts to create convincing deepfakes. Despite their benign applications in the entertainment industry, they have increasingly been associated with malicious purposes. They can facilitate misinformation campaigns, damage personal and professional reputations, and enable sophisticated scams, as evidenced by recent high-profile incidents [27]. The potential for deepfakes to be weaponized in criminal activities, such as identity theft or fraud, remains a pressing issue that is still underestimated in its full scope. In particular, facial deepfakes (e.g., Face Swap/Face Reenactment) are more likely to be used with malicious intent because they allow the attacker to "steal" someone else's identity, creating authenticity issues.

Deepfake videos become harder and harder to differentiate from real content. Many efforts are being made to develop tools to differentiate AI-generated content from real content. On the one hand, new laws are promulgated, asking content providers to label fake content as generated using AI (e.g., with the European AI Act). On the other hand, research is conducted to develop tools to automatically detect and label videos as fake or not. Such methods rely on learning any traces of manipulations found in the fake videos. Despite the recent progress in forgery detection, state-of-the-art deepfake detectors struggle to obtain good generalization performances to unseen generators. Generation always seems to be one step ahead of detection. However, generation and detection do not exactly address the same objectives. While generation focuses on learning the biometric features of the target to create realistic deepfakes, detection aims to identify any type of manipulated content, regardless of the target's identity.

To fill this gap, a part of the research community focuses its effort on Person Of Interest (POI) Deepfake detection. Unlike previous methods, the goal is to find inconsistencies in the biometric information of a POI. Compared to traces

This work is part of the DeTOX project. It has been founded by "Astrid: Guerre Cognitive", a French defense program initiative. Project website: https://detox.eurecom.fr/

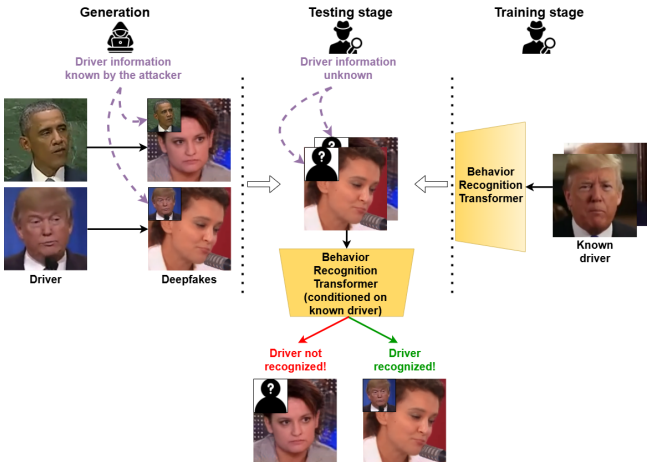

Fig. 2. Summary of our approach: From a set of deepfakes using different driver masters, we determine if a suspected identity is effectively driving the deepfake based on facial behavior.

of forgery, the analysis of the physical attributes and the behavioral signature is not entirely removed by common image processing post-processing, such as compression.

Finally, one question remains: What can we do after labeling fake videos? Just as biometry refers to the elements that allow the identification of a biological person, or hardwaremetry [12] refers to the identification of the camera that took a certain image, we use deepfakemetry to denote all the clues that characterize a deepfake video (source image, driving video, architecture of the neural network, ...).

In this paper, we address the problem of **driver identification** to demonstrate that we can extract information concerning the driving video used to generate a deepfake. Because Face Swap and Face Reenactment rely on the use of a driving video to generate their deepfakes, the head-pose, the eye gaze, and the facial expressions are out of the distribution of the behavioral signature of the identity portrayed in the video. A visual explanation of the task is provided in Figure 2. Therefore, if we have prior knowledge of the puppeteer, are we able to detect and unmask deepfakes that have been generated by this person?

POI's deepfake detectors, based on facial behaviour analysis, can recognise when a person's behavioural signature is or is not present in a video. Therefore, these models should perfectly fit this novel task. Instead of learning the behavior of the target of the deepfake, we aim to learn the behavior of the attacker to recognize the deepfakes made by this identity.

The contributions are the following:

- We improved an existing pipeline to learn the facial behavior of a deepfake impersonator so that we learn only the facial dynamics without leaking information from the physical appearance.
- We proved that this upgraded pipeline allows us to recognize the hidden driver behind a deepfake video.

The paper is divided as follows: In Section II, we will present the state of the art of face swap and face reenactment, alongside recent progress in deepfake detection and other research that started to deal with going beyond the simple detection of deepfakes. In Section III, we will present our new approach by upgrading an existing pipeline, and we will present the task of driver identification. In Section IV, we will show that our approach can identify an attacker inside a deepfake video. Finally, we will discuss the obtained results and the future works in Section V.

II. Related Works

Generation: Face Swap and Face Reenactment technologies have rapidly evolved with advancements in deep learning and computer vision. Face swap [19]–[21] focuses on replacing one person's face with another's by identifying facial landmarks and ensuring natural alignment. In contrast, face reenactment manipulates a target face to mimic the expressions or movements of a source [33], [36], [39], [42].

Tools like DeepFaceLab [29], SimSwap [4], or DeepLive-Cam [11] enable high-fidelity face swapping. Face reenactment frameworks such as First Order Motion Model [33] or LIA [39] allow expressive control of a target face based on a single input image.

Finally, more advanced techniques have emerged that use diffusion models to create fully synthetic motion for a face driven by audio input [10], [32], [34], [40]. The movements in the deepfake are completely artificial and do not come from any real person. Therefore, we can not expect to find the movement of the attacker within the deepfake video. So we will not consider these generators for our study.

Detection: The cornerstone of a good deepfake detector is the ability to stay robust even with unseen generation techniques. This underscores why significant research efforts are devoted to designing detectors capable of identifying all types of forgeries, regardless of the generation technique used [2], [3], [8], [35], [41]. However, artifact-based detection suffers from difficulties in generalizing to real-world scenarios. Indeed, traces of manipulations can be caused by genuine algorithms, from lossy compression to social networks pipelines [17], [23], [24], [38].

Therefore, POI deepfake detection allows us to rely on information other than artifacts for labeling videos: Biometrics. It is defined by the National Institute of Standards and Technology as "*A measurable physical characteristic or personal behavioral trait used to recognize the identity or verify the claimed identity, of an applicant* " [28]. They can be physical (facial attributes, iris, fingerprint) or behavioral (facial mimics, specific head gestures, eye-blinking rhythm) [18]. The process of deepfake creation involves the copying of physical biometric attributes to match with expressions and facial movements that do not belong to the target of the deepfake. Analyzing mismatch in biometrics leverages discriminative information between fake and real faces. For example, in [9], they analyzed the mismatch between the inner face and outer face to improve the detection of face swaps. However, behavioral biometrics is more suitable for detection because deepfakes do not contain the behavioral

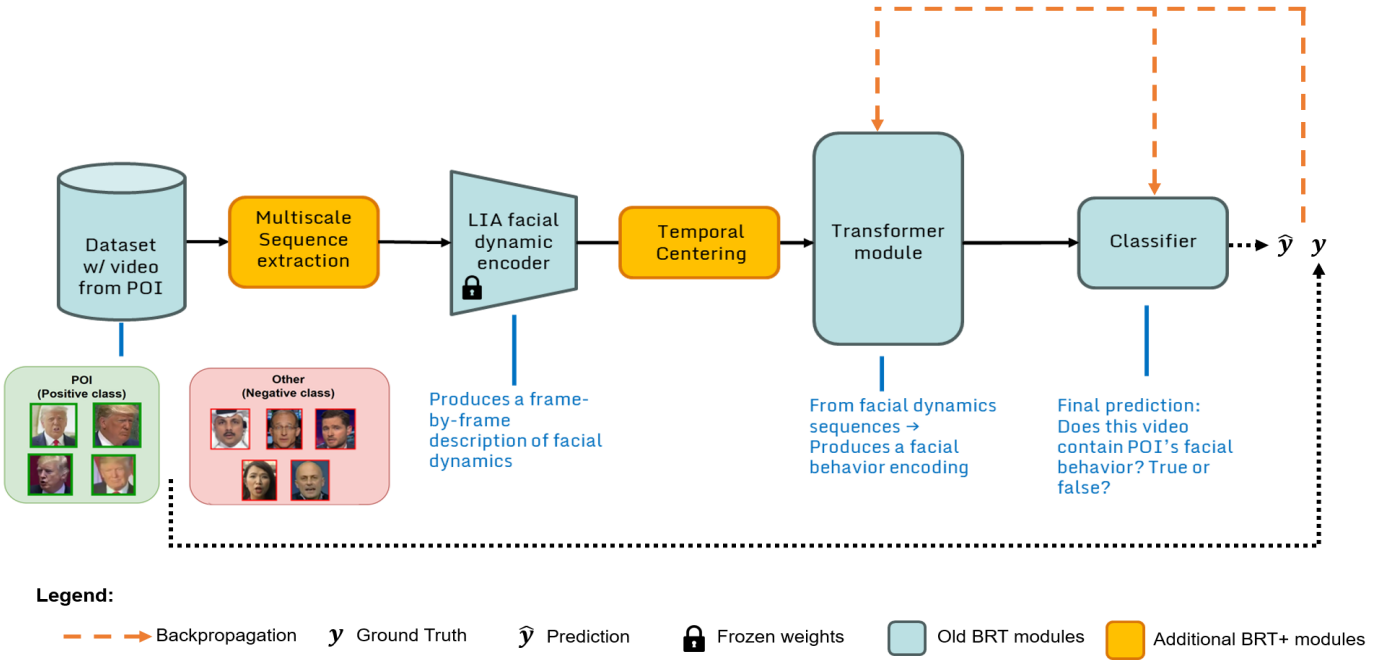

Fig. 3. Pipeline of BRT+. We introduce a *multiscale sequence extraction* module and a *Temporal centering* layer to the existing pipeline BRT from [22]. BRT used a facial dynamic encoder from [39] on all frames of a sequence to obtain a frame-by-frame description of the head-pose, gaze direction, facial expression, etc... Then, we feed this sequence of descriptions to a Transformers encoder that learns to produce facial behavior embeddings at the video level. Finally, a classification head gives the final prediction (i.e., *Does this behavioral signature belong to our suspected driver?).*

signature of the identity displayed in the video. Furthermore, [26] suggested that the behavioral signature could be exploited to determine if a video is fake or not. Few works relying on behavior analysis already exist, and they have two distinct approaches to tackle the problem. The first strategy consists of training a model to learn similarities between an input video and a set of real POI reference videos [6], [7]. The second strategy consists of training a model to learn the POI's behavioral signature directly from videos of the POI [1], [5], [22] against a set of videos from other identities.

Beyond the detection of deepfakes: As stated in the Introduction, an emerging field of research is dedicated to identifying elements to determine the origin of a deepfake. Many works have already been conducted in the identification of the generator used to create deepfake images. Indeed, CNN-generated images leave unique fingerprints based on their architecture [37]. These digital fingerprints can be exploited to identify the generator at the source of the deepfake [14], [15], [25].

Regarding driver identification, no study is fully dedicated to this task. However, Cozzolino et al. [7] have already demonstrated the potential for forming clusters of individuals based on the similarity of their facial movements. They propose a deepfake detection model specifically designed to detect deepfakes of a POI based on facial movements. Their approach analyzes facial motions and expressions to detect inconsistencies in how a person behaves when talking. The model is based on a Temporal ID Network (3D-CNN) trained using an adversarial strategy, where a 3D Morphable Model is used to challenge and improve the detector's robustness. In a nutshell, their model is able to compute a similarity score between an input video and a set of reference videos. This similarity indicates how the facial movements of the input video are close from those inside in the reference videos. Therefore, their model should provide results that are better than random in the task of driver identification.

Problem of Identity Leakage: Studies have shown that neural networks are likely to learn the physical attributes of people present in their training set [8]. Therefore, specific architectures are required to disentangle information related to physical and behavioral biometrics. A solution to this problem already exists if we look at works on deepfake generation. In [39], the authors propose an extra module to encode facial dynamics disentangled from the physical attributes.

In this paper, experiments with ID-Reveal showed that despite great performances in POI-deepfake detection and driver identification for Face Swapping, ID-Reveal failed at accurately identifying the driver in Face Reenactment deepfakes (see Section IV). These tests suggest that the drop in performance is related to a leakage of physical attributes in the encoding of the movements.

In the next section, we describe our approach for the task of driver indentification.

III. METHOD

In this section, we present the methodology used to recognize the driver behind deepfake videos. First, we detail the model used to learn the facial behavior of a single individual, leveraging gaze patterns, facial mannerisms, and head pose signatures. Second, we will explain how this learned behavior allows us to identify the driver in manipulated videos.

A. Behavior Recognition Transformer +

For our work, we design the improved pipeline BRT+ to learn the facial behavior (facial expression, eye gaze, head pose) of one single identity as in [22]. It depends on a facial dynamic encoder from the Face Reenactment tool LIA [39], followed by a transformer encoder that transforms the frame-level information to a video-level description of the behavior from the input video. The full pipeline is depicted in Figure 3.

In the following sub-section, we will detail the full architecture. The new modules and layers introduced by BRT+ are flagged by the "*(BRT+)*" mention.

1) Multiscale input sequences (BRT+): Facial behavior encompasses the unique set of expressions, micro-expressions, and mannerisms that characterize an individual. These behaviors are inherently distinctive and can serve as valuable features for building classifiers as they provide discriminative information about a person's identity. However, accurately capturing and representing facial behavior poses a challenge due to the varying temporal dynamics of different expressions. While some behaviors, such as eye blinks and muscle contraction, occur within milliseconds (high-frequency behaviors), others, like nodding or frowning, unfold over a longer duration (low-frequency behaviors).

To detect as many behavioral signatures as possible, we designed a multiscale frame-to-sequence extraction module. Let $V = (f_{t_0}, ...f_{t_T})$ be a full video clip, where t_0 (*resp.* t_T) is the time of the first (*resp.* last) frame of the video clip. We extract a sequence of N frames $I = (f_{t_k}, f_{t_k+\delta t}, ..., f_{t_k+(N-1)\delta t})$ where δt is a time interval chosen randomly according to the duration and the frame rate of the video, such that $N\delta t < t_T - t_0$. For the rest of the paper, we define t_k and $t_l = t_k + (N-1)\delta t$, which are respectively the first and the last frames of the extracted sequence.

2) Encoding face dynamic: Once the input sequence is obtained, we feed it to the facial dynamic LIA's encoder frame by frame.

LIA is a face reenactment algorithm that is able to reconstruct a face with a different facial expression extracted from another picture. It suggests that LIA includes a highly advanced encoder of facial dynamics. Furthermore, the authors have shown that the encoder effectively captures facial dynamics independently of the subject's physical identity. This property is particularly valuable in the context of driver identification, as it enables the encoding of all the facial movements without being influenced by the individual's visual appearance in the video.

In summary, for an input sequence $I = (f_{t_k}, ..., f_{t_l})$, we obtain $H = (h_{t_k}, ..., h_{t_l})$, a frame-by-frame description of the facial dynamics.

3) Temporal centering layer (BRT+): However, some information regarding the physical appearance might remain in each embedding. To ensure we get rid of the maximum of information linked to the physical appearance, we center the embedding along the time axis. That way, we force the model to learn temporal variation instead of average information.

From a sequence of facial dynamics H we remove the average temporal information to get $\bar{H} = (\bar{h}_{t_k}, ..., \bar{h}_{t_l}) = (h_{t_k} - \bar{h}, ..., h_{t_l} - \bar{h})$ where

$$\bar{h} = \frac{1}{T} \sum_{j=0}^{N-1} h_{t_k + j\delta t} \tag{1}$$

This is a crucial step to remove as many physical attributes as possible. Not doing this step induces an important decrease in performance for the task of driver identification for Face-Reenactment deepfakes. The weights of the facial dynamic encoder are frozen during the training.

4) Gathering video-level information: The sequence of embeddings h provides only a description of facial dynamics across all frames in the sequence, without explicitly leveraging temporal information. We must incorporate temporal dependencies to effectively learn the behaviors associated with one identity, For this purpose, we use a Transformer encoder, which excels at capturing relevant temporal patterns within our sequence of embeddings.

Outperforming traditional recurrent models that process sequences step-by-step and often struggle with long-range dependencies, Transformers leverage attention mechanisms to analyze all time steps simultaneously. It leads to a better representation of the relationships between the different dynamics of the face, making them well-suited to learn behaviors specific to an individual.

We encode the video-level information as Zheng et al. did in [43]. They used a learnable *[CLS]* token to encode the class of the sequence. After the training of the temporal transformer encoder, the encoding of the *[CLS]* token by the Transformer will embed the facial behavior of the sequence.

5) Classification, loss, and score at the video-level.: The last part of the pipeline consists of a simple Multi-Layer Perceptron, composed of ReLU activation functions and dropout layers to improve classification generalization for unseen videos.

For the classification task, we use a simple Binary Cross-Entropy loss. For a single observation/prediction pair $(y/\hat{y})$, the binary Cross-Entropy loss is defined as:

$$L_{BCE}(y, \hat{y}) = -(\ y.log(\hat{y}) + (1-y).log(1-\hat{y})\ ) \tag{2}$$

The backpropagation will update the weights of both the Transformer encoder and the classifier.

Finally, to evaluate our model, we use the following guidelines: For one video clip V, we randomly extract K sequences.

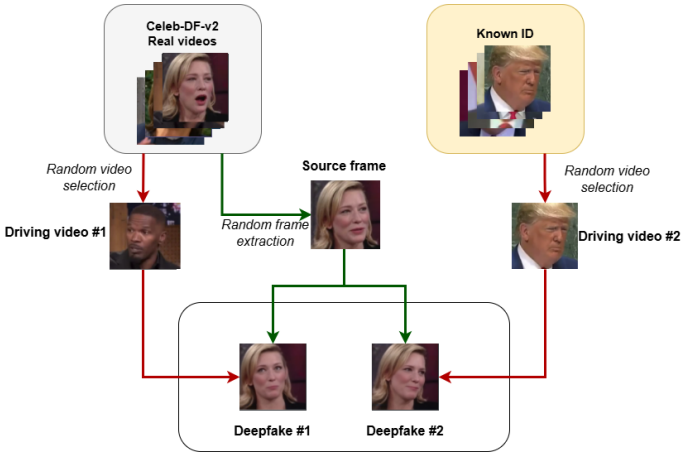

Fig. 4. Scheme of the database for one identity. We first generate a pool of deepfakes made with all the different identities from CDFv2 (on the left). Then, we add deepfakes of CDFv2 identities driven by videos from our list of known driving identities (on the right). For clarity, only Face Reenactment is illustrated in this figure.

In consequence, the network will return K predictions. The likelihood of a video containing the facial behavior of the suspect identity is defined as the average score obtained by the K sequences. Once all videos have their score, we can compute the usual classification metrics as defined in the next subsection.

B. Driver identification task

The task is described in Figure 2. After training one model per suspected identity, we test our approach at accurately identifying the videos driven by the suspected identity. Specifically, we assessed whether each trained model could correctly classify deepfakes in which the suspected identity was used as the driver, discriminating them from other manipulated videos.

Each of our three specialized models was tested on the full pool of deepfake videos generated from the Celeb-DF-v2 dataset. This pool contains a mixture of deepfakes, including those where the suspected identity was the driver and those featuring other random identities. To evaluate performance, we ran each model sequentially on its corresponding test set and measured its ability to correctly classify deepfakes driven by our suspected identity against others.

To quantify our results, we computed the following classification metrics: ROC AUC and Precision-Recall (PR) AUC. ROC AUC evaluates the discriminative power of our models, i.e., their ability to discriminate between deepfakes driven by the suspected ID and deepfakes driven by other people. Unlike accuracy, ROC AUC is better suited for imbalanced datasets. A random (*resp.* perfect) classifier is expected to have a ROC AUC of 50% (*resp.* 100%) PR AUC is a metric that is sensitive to samples predicted positive with a high confidence score; it controls the precision/recall balance for different classification thresholds. It is informative to know if we can trust a score when the confidence score is high. It is also well-suited when there is a class imbalance, which is the case in this study. Also in practice, compared to the number of all existing deepfakes, the number of deepfakes driven by a specific identity is not significant. A random (*resp.* perfect) classifier is expected to have a PR AUC equal to $(100 \times r)\%$ (*resp.* 100%), with r being the ratio of positive samples in the dataset.

TABLE I
NUMBER OF REAL VIDEOS PER SUSPECTED IDENTITY USED TO GENERATE THE TRAIN AND TEST SETS.

	Number of videos	
	Train	Test
Belkacem	571	132
Obama	168	29
Trump	153	30

TABLE II
NUMBER OF FAKE VIDEOS VIDEOS GENERATED FOR THE TEST SET. FS AND FR MEAN RESPECTIVELY FACE SWAP AND FACE REENACTMENT. (-) AND (+) REFER TO THE NEGATIVE AND THE POSITIVE CLASS OF EACH TEST SET.

Test set	Source ID	Driver ID	FS	FR
Belkacem VS CDF	CDF	CDF (-)	132	536
		Belkacem (+)	58	130
Obama VS CDF	CDF	CDF (-)	132	536
		Obama (+)	59	29
Trump VS CDF	CDF	CDF (-)	132	536
		Trump (+)	46	30

IV. RESULTS

In this section, we first describe the construction of our database, which consists of real videos of different suspected identities (only 3) and other individuals. Second, we present and analyze the results of our driver identification experiments, where we evaluated the ability of our models to correctly recognize the driver behind deepfake videos. We report key performance metrics, including ROC Area Under the Curve (ROC AUC) and Precision-Recall AUC (PR AUC) for two different types of deepfake manipulations: face reenactment and face swap. Table II reports the number of videos generated. The face swaps were generated using the Roop FaceSwap generator [31]. The reenacted faces were generated using FOMM [33].

We compared our approach with BRT [22] and ID-Reveal [7]. Even if these last methods have been designed for the deepfake detection task, they are the closest works we can derive for driver identification based on facial movements.

Table III summarizes the classification performances of our models for each suspected ID.

A. Database

To train and evaluate our models, we constructed a dataset comprising both real and deepfake videos. Our goal is to train multiple distinct models, each specialized in recognizing the facial behavior of a specific identity while distinguishing it from general facial behaviors observed in other individuals. We took the real videos of [22], which consists of tv/radio interviews, conferences and public addresses from 3 politicians;

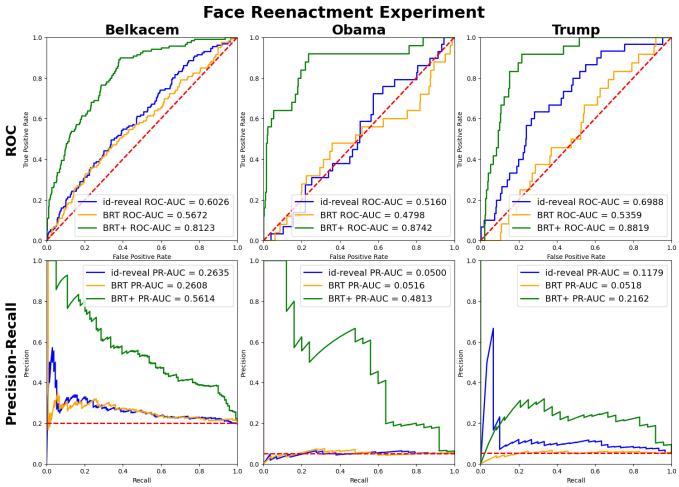

Fig. 5. **ROC curves** (on top) and **Precision-Recall curves** (on bottom) on Face-Reenactment subset only for 3 suspected IDs. The blue, orange, and green curves refer respectively to ID-Reveal, BRT, and BRT+ model performances on **Face Reenactment**. The red dotted line corresponds to the expected curve of a random classifier.

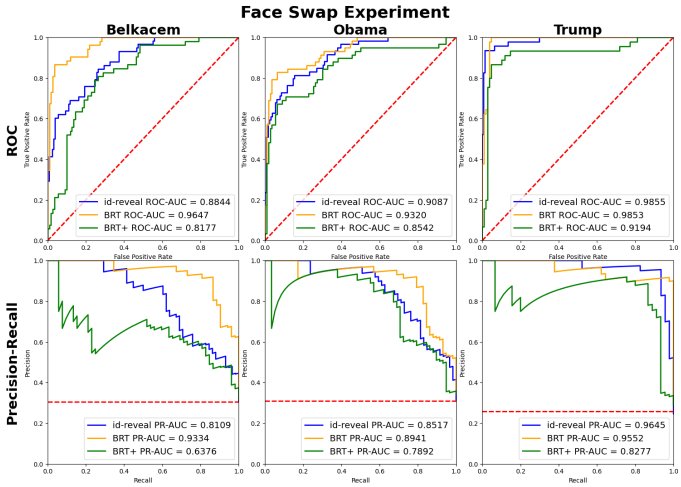

Fig. 6. **ROC curves** (on top) and **Precision-Recall curves** (on bottom) on Face-Reenactment subset only for 3 suspected IDs. The blue, orange, and green curves refer respectively to ID-Reveal, BRT, and BRT+ model performances on **Face Swap**. The red dotted line corresponds to the expected curve of a random classifier.

The former French minister Najat Vallaud-Belacem, and two American presidents; Barack Obama and Donald Trump. Table I shows how many video clips were gathered by suspected drivers.

1) Training Data: For each identity, we collected a set of real videos showcasing their natural facial expressions, gaze behavior, and head pose dynamics. To provide a diverse contrast during training, we supplemented these videos with approximately 1,000 real videos from the FaceForensics++ dataset [30]. This combination enables the model to learn the unique facial behavior of our suspected drivers against a cohort of varied individuals, improving its ability to generalize and differentiate between identities.

2) Testing data: To evaluate our approach, we constructed a test set consisting of a pool of deepfake videos generated from the CelebDFv2 dataset [21]. Using 536 real videos from CelebDFv2, we applied Face Swaps and Face Reenactments techniques to synthesize manipulated videos where different identities were swapped. Within this pool of deepfakes, we carefully embedded deepfakes generated with each suspect as the driving identity.

For testing, we ran our trained model to the pool of deepfakes driven either by Celeb-DF-v2 IDs or our suspected driver, systematically verifying whether the model could correctly classify deepfakes where the suspected driver was indeed the driver versus deepfakes involving other individuals. This process resulted in multiple test sets, one for each identity, allowing us to assess the effectiveness of our approach in identifying the driver across different manipulated videos.

B. Global performances

Overall, BRT+ results demonstrate a strong performance across both face-reenactments and face swaps tasks, with AUC values consistently exceeding 80%.

As mentioned in Section III-B, a random classifier is expected to have a PR AUC equal to the ratio of positive samples. Therefore, for a random classifier, the baselines are the following:

- For Face Reenactment experiment: the baseline of PR AUC is 0.200, 0.051, and 0.053 for N. Vallaud-Belkacem, B. Obama, and D. Trump, respectively. It represents the area under the curve of the red dotted line in Figure 5 PR curves.
- For Face Swap experiment: the baseline of PR AUC is 0.305, 0.309, and 0.258 for N. Vallaud-Belkacem, B. Obama, and D. Trump, respectively. It represents the area under the curve of the red dotted line in Figure 6 PR curves.

In summary, our model is able to differentiate deepfakes made with a driving video of the suspected identity and deepfakes driven by another. However, the low Precision-Recall AUC indicates that driver identification is still far from ready for real-world application, as high-confidence predictions cannot yet be fully trusted.

C. Performance per generators

Comparing the two deepfake generation techniques, we observe that:

With BRT+, we obtain similar ROC AUC on Face Swap and Face Reenactment, outperforming BRT and ID-Reveal at finding the driver identity on Face Reenactment deepfakes. The performance of BRT and ID-Reveal on face reenactment data is close to that of a random classifier. We hypothesize that this is due to the Identity Leakage phenomenon described in Section II, as face reenactments preserve only the behavioral signature and contain no physical information about the driver. This hypothesis is further supported by the observation that

BRT and ID-Reveal are more effective with Face Swap videos (see Figure 6), surpassing BRT+ in both ROC and PR AUC.

Another explanation could be that face swap better preserves the movements of the driver than face reenactment. Where Face Swap only alters physical attributes of the inner face, Face Reenactment generates a video where all frames are close from the source image, and the semantics of the movement are not fully retained in the final deepfake. This would align with a key finding from [26], which states: "*synthetic videos [...] are seen as less real and less engaging than the original source video.*". The authors conducted subjective tests where participants evaluated the naturalness of a person's talking behavior in real videos versus deepfakes created with FOMM [33]. There is a lack of movement amplitude and liveliness in the deepfakes created with Face Reenactment, suggesting that the facial movement is not perfectly transferred.

In both cases, BRT+ results suggest that our method relies on behavioral biometrics for driver identification, as it achieves similar ROC-AUC performance on both face swap and face reenactment data. Thanks to the temporal centering layer, any remaining traces of physical identity — which can be assumed to remain constant over a few seconds — as well as individuals' default facial expressions, are not learned during the training phase.

TABLE III
VIDEO-LEVEL ROC AUC AND PRECISION-RECALL AUC FOR THE TASK OF DRIVER IDENTIFICATION FOR FACESWAP AND FACE-REENACTMENT. COMPARISION BETWEEN BRT+ AND ID-REVEAL.

	Face reenactment			
	ID-Reveal		BRT+ (our)	
IDs	ROC AUC%	PR AUC%	ROC AUC%	PR AUC%
N.V-B	60.26	26.35	81.23	56.14
B.O	51.60	5.00	87.42	48.13
D.T	69.88	11.79	88.19	21.62
	Face swap			
	ID-Reveal		BRT+ (our)	
IDs	ROC AUC%	PR AUC%	ROC AUC%	PR AUC%
N.V-B	88.44	81.09	81.77	63.76
B.O	90.87	85.17	85.42	78.92
D.T	98.55	96.45	91.94	82.77

D. Performance per IDs

We also noticed differences between the classification performances for the different identities. For example, the model trained on Donald Trump videos always obtains a better AUC, whatever the generator. We suspect that Donald Trump's facial mimics are more expressive than those of Barack Obama or Najat Vallaud-Belkacem. Therefore, an individual with a very specific set of mannerisms and behavioral signatures is more likely to be identified when one of their videos is used to drive a deepfake. This assumption must be verified with a bigger dataset containing more identities.

V. DISCUSSION AND CONCLUSION

In this work, we introduced driver identification, which aim at identifying the driver behind deepfake videos. To do so, we propose a new pipeline based on the analysis of the facial behavior of a suspected person. The main challenge of this task lay in finding a tool to extract the facial movements while preventing the leakage of the physical appearance. Our results demonstrate that the proposed method is able to discriminate deepfakes generated using a specific identity as the driver from those driven by other identities. Our method retrieves information about the driver as well for Face Swap and Face Reenactment

Performance varies across different identities, with strong mannerisms being the easiest to recognize, as evidenced by consistently higher AUC values. This suggests that individuals with highly distinct facial mannerisms are more easily unveiled in deepfake videos, reinforcing the idea that behavioral signatures play a crucial role in driver identity attribution.

Overall, our findings support the feasibility of driver identification as a novel tool for deepfake forensics, providing new insights into understanding and attributing deepfake content based on facial behavior dynamics. While driver identification shows promise, its real-world applicability remains limited due to the identity-dependent nature of our models, as it is impractical to maintain a comprehensive collection of video recordings for all possible identities.

Future research could focus on determining the minimal amount of video recordings needed to effectively learn an individual's facial behavior. This would help optimize driver identification for practical applications by reducing data requirements while maintaining reliable performance. Investigating how different recording conditions (e.g., lighting, camera angle, and emotional expressions) affect the learning process could further refine this approach.

Additionally, another promising direction is the development of a foundation model capable of encoding facial dynamics without requiring separate training for each identity. Such a model could leverage large-scale datasets to learn generalized facial behavior patterns while allowing for efficient adaptation to specific identities. This would significantly enhance the scalability of driver identification, making it more applicable in real-world forensic scenarios where pre-recorded data of all individuals may not be available.

ACKNOWLEDGMENT

This work was founded by the French ASTRID program, facilitated by the Defense Innovation Agency 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] R. Chen, X. Chen, B. Ni, and Y. Ge, "Simswap: An efficient framework for high fidelity face swapping," in *Proceedings of the 28th ACM international conference on multimedia*, 2020, pp. 2003–2011.

[5] 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.

[6] D. Cozzolino, A. Pianese, M. Nießner, and L. Verdoliva, "Audio-visual person-of-interest deepfake detection," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2023, pp. 943–952.

[7] 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.

[8] 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.

[9] X. Dong, J. Bao, D. Chen, T. Zhang, W. Zhang, N. Yu, D. Chen, F. Wen, and B. Guo, "Protecting celebrities from deepfake with identity consistency transformer," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2022, pp. 9468–9478.

[10] C. Du, Q. Chen, T. He, X. Tan, X. Chen, K. Yu, S. Zhao, and J. Bian, "Dae-talker: High fidelity speech-driven talking face generation with diffusion autoencoder," in *Proceedings of the 31st ACM International Conference on Multimedia*, 2023, pp. 4281–4289.

[11] K. Estanislao, "Deep-live-cam," https://github.com/hacksider/Deep-Live-Cam, 2023.

[12] C. Galdi, M. Nappi, and J.-L. Dugelay, "Combining hardwaremetry and biometry for human authentication via smartphones," in *Image Analysis and Processing—ICIAP 2015: 18th International Conference, Genoa, Italy, September 7-11, 2015, Proceedings, Part II 18*. Springer, 2015, pp. 406–416.

[13] 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.

[14] L. Guarnera, O. Giudice, and S. Battiato, "Level up the deepfake detection: a method to effectively discriminate images generated by gan architectures and diffusion models," in *Intelligent Systems Conference*. Springer, 2024, pp. 615–625.

[15] L. Guarnera, O. Giudice, M. Nießner, and S. Battiato, "On the exploitation of deepfake model recognition," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2022, pp. 61–70.

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

[17] J. Hu, X. Liao, W. Wang, and Z. Qin, "Detecting compressed deepfake videos in social networks using frame-temporality two-stream convolutional network," *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 32, no. 3, pp. 1089–1102, 2021.

[18] 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.

[19] I. Korshunova, W. Shi, J. Dambre, and L. Theis, "Fast face-swap using convolutional neural networks," in *Proceedings of the IEEE international conference on computer vision*, 2017, pp. 3677–3685.

[20] L. Li, J. Bao, H. Yang, D. Chen, and F. Wen, "Faceshifter: Towards high fidelity and occlusion aware face swapping," *arXiv preprint arXiv:1912.13457*, 2019.

[21] 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.

[22] A. Libourel and J.-L. Dugelay, "You're not acting like yourself: Deepfake detection based on facial behavior," in *IPAS 2025, 6th IEEE International Conference on Image Processing Applications and Systems*, 2025.

[23] 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.

[24] 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.

[25] F. Marra, D. Gragnaniello, L. Verdoliva, and G. Poggi, "Do gans leave artificial fingerprints?" in *2019 IEEE conference on multimedia information processing and retrieval (MIPR)*. IEEE, 2019, pp. 506–511.

[26] Q. Miao, S. Kang, S. Marsella, S. DiPaola, C. Wang, and A. Shapiro, "Study of detecting behavioral signatures within deepfake videos," *arXiv preprint arXiv:2208.03561*, 2022.

[27] F. Muhly, E. Chizzonic, and P. Leo, "Ai-deepfake scams and the importance of a holistic communication security strategy," *International Cybersecurity Law Review*, pp. 1–9, 2025.

[28] M. Nieles, K. Dempsey, V. Y. Pillitteri *et al.*, "An introduction to information security," *NIST special publication*, vol. 800, no. 12, p. 101, 2017.

[29] I. Perov, D. Gao, N. Chervoniy, K. Liu, S. Marangonda, C. Umé, C. S. Facenheim, L. RP, J. Jiang, S. Zhang *et al.*, "Deepfacelab: Integrated, flexible and extensible face-swapping framework," *arXiv preprint arXiv:2005.05535*, 2020.

[30] 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.

[31] S. Sangwan, "Roop," https://github.com/s0md3v/roop, 2023.

[32] S. Shen, W. Zhao, Z. Meng, W. Li, Z. Zhu, J. Zhou, and J. Lu, "Difftalk: Crafting diffusion models for generalized audio-driven portraits animation," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2023, pp. 1982–1991.

[33] 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.

[34] M. Stypułkowski, K. Vougioukas, S. He, M. Zięba, S. Petridis, and M. Pantic, "Diffused heads: Diffusion models beat gans on talking-face generation," in *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, 2024, pp. 5091–5100.

[35] C. Tan, Y. Zhao, S. Wei, G. Gu, P. Liu, and Y. Wei, "Rethinking the up-sampling operations in cnn-based generative network for generalizable deepfake detection," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2024, pp. 28 130–28 139.

[36] 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.

[37] S.-Y. Wang, O. Wang, R. Zhang, A. Owens, and A. A. Efros, "Cnn-generated images are surprisingly easy to spot... for now," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2020, pp. 8695–8704.

[38] Y. Wang, Q. Sun, D. Rong, and R. Geng, "Multi-domain awareness for compressed deepfake videos detection over social networks guided by common mechanisms between artifacts," *Computer Vision and Image Understanding*, vol. 247, p. 104072, 2024.

[39] 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.

[40] S. Xu, G. Chen, Y.-X. Guo, J. Yang, C. Li, Z. Zang, Y. Zhang, X. Tong, and B. Guo, "Vasa-1: Lifelike audio-driven talking faces generated in real time," *Advances in Neural Information Processing Systems*, vol. 37, pp. 660–684, 2024.

[41] 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.

[42] J. Zhao and H. Zhang, "Thin-plate spline motion model for image animation," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2022, pp. 3657–3666.

[43] 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.