read till pg 4

Do Deepfakes Feel Emotions? A Semantic Approach to Detecting Deepfakes Via Emotional Inconsistencies

Brian Hosler[1], Davide Salvi[2], Anthony Murray[1], Fabio Antonacci[2],
Paolo Bestagini[2], Stefano Tubaro[2], Matthew C. Stamm[1]

[1]Department of Electrical and Computer Engineering, Drexel University, Philadelphia, PA, USA
{brian, mstamm}@drexel.edu

[2]Dipartimento di Elettronica, Informazione e Bioingegneria, Politecnico di Milano, Milan, Italy
{davide.salvi, fabio.antonacci, paolo.bestagini, stefano.tubaro}@polimi.it

Abstract

Recent advances in deep learning and computer vision have spawned a new class of media forgeries known as deepfakes, which typically consist of artificially generated human faces or voices. The creation and distribution of deepfakes raise many legal and ethical concerns. As a result, the ability to distinguish between deepfakes and authentic media is vital. While deepfakes can create plausible video and audio, it may be challenging for them to to generate content that is consistent in terms of high-level semantic features, such as emotions. Unnatural displays of emotion, measured by features such as valence and arousal, can provide significant evidence that a video has been synthesized. In this paper, we propose a novel method for detecting deepfakes of a human speaker using the emotion predicted from the speaker's face and voice. The proposed technique leverages Long Short-Term Memory (LSTM) networks that predict emotion from audio and video Low-Level Descriptors (LLDs). Predicted emotion in time is used to classify videos as authentic or deepfakes through an additional supervised classifier.

1. Introduction

Deepfake technology has made the creation of realistic media forgeries much more accessible to the general public. This technology allows its users to counterfeit the identity of a person in a video by falsifying their face or voice [58]. Deepfakes have already been used for several malicious purposes, including the creation of fake pornographic videos of celebrities without their consent [11], the theft of over £200,000 from an energy company [54], the creation of fake videos of politicians making controversial claims [22], and the creation of fake videos of private citizens aimed to damage their reputation [53]. Furthermore, deepfake videos can rapidly spread across the internet, creating significant damage to both individuals, organizations, and society at large. As a result, there is a significant need for techniques to automatically detect deepfake videos.

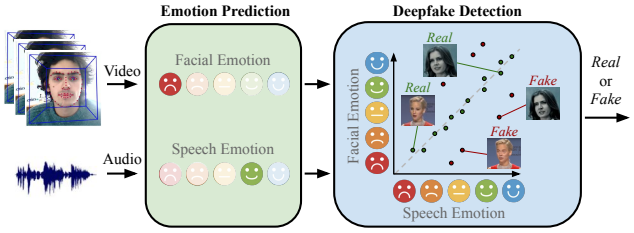


Figure 1: Proposed deepfake detection method exploiting audio-visual emotion analysis.

To combat this growing threat, the research community has recently focused significant effort on developing algorithms to detect deepfakes. Several approaches have been proposed to identify both video [25, 36, 32, 3, 12] and audio [34, 5] deepfakes, and multiple deepfake databases have been created to support this research [46, 33, 19, 56]. Since deepfake technology continues to evolve, developing a wide variety of detection methods is essential to address this problem. Furthermore, this helps provide protection against the evolution of anti-forensic countermeasures [10, 58].

One higher-level semantic concept that may provide clues for detecting deepfakes is emotion. Though deepfake technology can convincingly manipulate low-level visual and audio features, it may be difficult for deepfakes to recreate the subtle emotional cues typically present in humans. In particular, we hypothesize that deepfake videos of a human speaker will likely display irregular or inconsistent emotion in the face and voice, especially when they are not explicitly constrained to do so during creation.

This paper proposes a deepfake detection method that

aims at detecting non natural and inconsistent emotions conveyed by audio and video, as described in Figure 1. To do so, we use the valence-arousal model of emotion. This is a continuous two-dimensional model, where valence represents positivity or negativity of the emotion and arousal represents excitement or calmness. Given a video under analysis, the proposed method first estimates how valence and arousal change in time from audio and facial LLDs. Then, we feed these estimates to a supervised classifier that analyzes valence and arousal behavior to detect whether the video is authentic or a deepfake.

The proposed method is tested on videos from the DeepFake Detection Dataset (DFDC) dataset [19] for which both audio and video have been altered [17]. Results show that valence and arousal behavior is actually different in deepfake audio and video tracks when compared to real speakers. More specifically, results suggest that deepfake speech generation methods are not able to correctly synthesize natural emotion per se. By jointly analyzing audio and video emotional behavior, it is then possible to detect fake videos very accurately, thus increasing the number of tools an investigator can use to run forensic analysis.

2. Background

2.1. Emotion Classification and Recognition

Emotion recognition from audio-visual representations of a speaker refers to the problem of automatically determining the emotion felt by the speaker. This can be done by analyzing both the speaker's speech (i.e., audio recording) and facial expression (i.e., video recording) [35]. In recent years, emotion recognition has become increasingly important in many areas, including social media [39]. Thanks to the growing availability of mobile cameras and sharing platforms, the number of user-generated videos is growing more and more rapidly, and companies are often interested in automatically extracting opinions expressed in user-generated content [13]. However, this task is a very challenging one, as emotion is both open to subjective interpretation and difficult to uniquely define.

Emotions are typically modeled according to two different strategies depending on whether the passage from one emotion to another is considered categorical or dimensional. In the first scenario, emotions are split into discrete and well-defined classes (e.g., happy, sad, neutral, etc.) [9]. In the second case, emotions are described by means of the values of quantitative features, such as valence and arousal and organized in a two-dimensional circumplex space [48]. Valence represents emotional affect ranging from positive to neutral to negative, whereas arousal refers to emotional intensity.

In this work, we utilize the continuous arousal-valence model of emotion as opposed to discrete emotional classes. This allows us to evaluate more nuanced emotions, account for utterances that cannot be uniquely categorized, clearly examine how emotion evolves over time, and avoid issues associated with misclassifying emotions.

There are many emotion recognition and detection approaches proposed in literature. The majority of recent techniques are based on machine learning and neural networks [29, 4]. State-of-the-art methods can be roughly split into two broad categories that differ in the way networks are fed. Some methods adopt an end-to-end data-driven approach [26, 7], performing the analysis directly on raw data (i.e., video frames and audio samples). Other methods exploit a pre-processing stage that extracts LLDs as handcrafted features from the input signals [52, 16]. This is done to reduce the dimensionality of the input signals to a more compact set of data, thus allowing for simpler classifier architectures that are easier to train.

Advanced emotion recognition systems are often based on multi-modal approaches, where audio and video are jointly exploited to improve the algorithm's performance [40, 50]. Furthermore, these systems can be improved by taking into account also the temporal evolution of emotion, as it provides crucial information that can be used to perform this task. When doing this, certain types of networks such as LSTMs can be used that are capable of identifying feature changes over time [41, 14].

In this work, we leverage a LLD-based multi-modal emotion recognition system exploiting LSTM layers as reported in [45].

2.2. Deepfake Detection

In the last several years, the development of accurate and easy-to-use facial manipulation techniques [61, 55] has become an important information security threat. This has caught the attention of the multimedia forensics community [3], which has begun to develop algorithms to detect and limit this kind of video forgery [58].

Several detection techniques have been proposed in the forensic literature to detect clues left by deepfake videos. Some techniques search for specific semantic inconsistencies present in deepfakes, such as inconsistent eye blinking [31] or head pose [60]. In [32], the authors search for face warping traces. Furthermore, [2] proposes a method that checks the coherence between the movements of the mouth and the pronounced phonemes. Another category of video deepfake detection methods is completely data-driven. The works in [1, 38, 12] utilize neural networks to detect deepfakes by analyzing videos on a frame-by-frame basis. Other data-driven techniques exploit the temporal dimension of video [30, 25].

Deepfake technology is not limited to video editing. Several methods for synthetic speech generation have been proposed through the years [57, 51, 27]. This has motivated

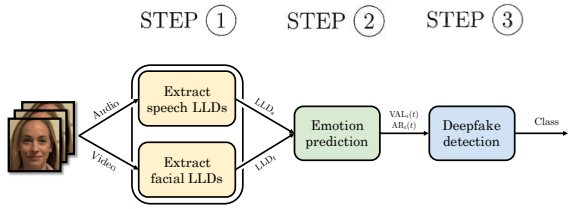


Figure 2: Proposed pipeline for deepfake detection exploiting audio-visual emotion analysis.

forensic research also towards detecting fake speech. Traditional approaches to synthetic speech detection focus on extracting sets of meaningful model-based features from speech samples. This is the case of [49] that relies on classic audio features inherited from the music information retrieval community. Alternatively, in [28] the authors show that the use of long-term audio features has benefits over short-term ones. In [5], audio bicoherence is used instead. More recent methods explore fully data-driven approaches [18]. As an example, in [34] a time frequency representation of speech is fed to a shallow neural network.

3. Method

Deepfake videos typically depict a front-facing person, who is speaking. Given such a video for analysis, we would like to determine if it is *Real* or *Fake*, where

- *Real* - Both the face and the speech of the video subject are authentic and not artificially generated.
- *Fake* - Both the face and the speech of the video subject have been generated through deepfake technologies.

Videos which do not fall into one of these two classes are outside the scope of this work.

In this work, we propose the use of emotions to detect deepfake videos. We hypothesize that deepfake generators may not be able to synthesize all high-level aspects of human emotion. Therefore, we can analyze emotional cues extracted from a person's face and speech to detect whether a video is fake.

Our proposed method is composed of three stages, as shown in Figure 2. In the first stage, we leverage established emotion recognition research to extract low level descriptors (LLDs) of a subject speech and face. In the second stage, these LLDs are used to produce a representation of the subject emotion over time in the visual (i.e., face) and audio (i.e., speech) domain by means of valence and arousal estimation, exploiting well-known findings in emotion recognition literature [45]. The third stage performs deepfake detection on the basis of these emotion signals.

3.1. Low-Level Feature Extraction

The first stage of our proposed system extracts LLDs from the video that describe the speakers face and voice. This approach is commonly used in emotion recognition research to both reduce input dimensionality and only track important features that are linked to emotion. We separately extract features from the visual track (i.e., facial features) and the audio track (i.e., speech features). Given a video under analysis, we can split it into two components: the time-series $f(t)$ representing the temporal evolution of video frames showing the person's face, and the time-series $s(t)$ representing the temporal evolution of the audio track capturing the person's speech.

Facial Features. Previous works identify a subject's emotion by using their expression, which is measured by the location, distance, and motion of specific keypoints on the subject's face [43, 20]. A well-established system to measure facial behavior is the the Facial Action Coding System (FACS), which models different facial movements as elementary muscular activities, including descriptors such as inner and outer brow raising of one or two eyes, jaw drop, and lip tightening [21].

In this work, we use OpenFace [8] features based on FACS extracted on a per-frame basis. Given a video frame f containing a face, we extract our facial features as,

$$\mathrm{LLD_f} = \mathrm{FACS}(f), \tag{1}$$

where $\mathrm{FACS}(\cdot)$ is the facial action unit extractor of OpenFace, and $\mathrm{LLD_f}$ is the resulting 17-element facial feature vector. By applying the feature extractor to multiple frames, we create a 17-dimensional time-series $\mathrm{LLD_f}(t)$, whose length depends on the number of frames used.

Speech Features. Existing work has also provided methods for determining a subjects emotion using the subjects voice [59, 15]. For this purpose several LLD feature sets have been proposed to highlight different aspects of the audio sequence under analysis. For audio analysis, each LLD feature vector is typically extracted from a short time window (i.e., a few millisecond) of the considered speech signal $s(t)$. Time windows are chosen to be short enough so that emotion over this time period can be modeled as a wide-sense stationary process.

In this paper we use a combination of three different feature sets described in OpenSmile [24]. These are the first 13 Mel-Frequency Cepstrum Coefficients (MFCC) [42] as well as their 13 first and 13 second statistical moments, 18 features from the Geneva Minimalistic Acoustic Parameter Set (GeMAPS) [23] and 23 additional ones from the Extended GeMAPS (eGeMAPS), for a total of 80 features extracted from each considered time window.

Given an short time window $s_{\mathrm{win}}(t)$ of speech signal, we extract the LLD feature vector as

$$\mathrm{LLD_s} = \mathrm{SMILE}(s_{\mathrm{win}}(t)), \tag{2}$$

where $\mathrm{SMILE}(\cdot)$ is the speech-based feature extraction process described in OpenSmile [24], and $\mathrm{LLD_s}$ is the 80-

element speech feature vector. Considering multiple time windows extracted from the speech time-series $s(t)$, we obtain the time-series $\text{LLD}_\text{s}(t)$, whose length depends on how many windows are selected.

3.2. Emotion Recognition

The second stage of our system estimates the speaker's emotion over time for both the audio and visual tracks of the input video. This is done on the basis of the sequence of LLDs produced by the first stage. As discussed in Section 2, we capture the speaker's emotion using the continuous valence-arousal model of emotion commonly used in psychology [44]. Under this model, valence represents the subject's emotional affect (i.e. positive to negative), while arousal represents the magnitude of that emotion. Discrete emotional classes (e.g. happy, sad, angry, calm) can all be represented in the two dimensional space created by arousal and valence [47], however this continuous emotional model also allows capture subtle and ambiguous emotional behavior. However, the continuous arousal-valence emotional model also allows capture subtle and ambiguous emotional behavior that may be important to distinguish real emotions from synthesized ones.

To capture the evolution of emotion over time, we produce time varying arousal and valence signals as opposed to single measurements for each video. To accomplish this, we use an LSTM-based neural network based on the work presented in [45]. The input of this network is a multi-dimensional time-series of length T, consisting of either speech or facial LLDs. The network architecture is composed of a bi-directional LSTM layer of 64 cells, followed by another layer of 32 cells, a fully connected layer with 16 neurons, and finally a dense layer with two output neurons. The network outputs two time series of length T. One represents valence and the other one represents arousal evolution over time. The network is trained on each modality (i.e., face and speech), thus producing four time-series, describing valence and arousal behavior for both the subject's face and speech. More formally, we compute the speech-derived valence and arousal signals as

$$[\text{VAL}_\text{s}(t),\ \text{AR}_\text{s}(t)] = \text{LSTM}_\text{s}(\text{LLD}_\text{s}(t)), \tag{3}$$

where $\text{LSTM}_\text{s}(\cdot)$ is the trained emotion recognition model, $\text{LLD}_\text{s}(t)$ is the speech-derived LLD time-series, while $\text{VAL}_\text{s}(t)$ and $\text{AR}_\text{s}(t)$ are the speech-based valence and arousal time-series. Likewise, our facial-derived features can be expressed as,

$$[\text{VAL}_\text{f}(t),\ \text{AR}_\text{f}(t)] = \text{LSTM}_\text{f}(\text{LLD}_\text{f}(t)), \tag{4}$$

where $\text{LSTM}_\text{f}(\cdot)$ is the emotion recognition network, $\text{LLD}_\text{f}(t)$ is the facial-derived LLD time-series, while $\text{VAL}_\text{f}(t)$ and $\text{AR}_\text{f}(t)$ are the face-based valence and arousal time-series.

In order to capture fine-grained face and speech emotional temporal behavior and to satisfy the wide-sense stationary emotional process assumption over the analyzed time windows, all the considered time-series are sampled at 10Hz. As most modern video is recorded with 30 or more frames per second, this is done by selecting 10 evenly spaced frames within a second, and extracting LLDs from each frame. To extract speech-based features, 10ms time windows are extracted every 100ms, and LLDs are extracted from each window. Thus the sequences $\text{VAL}_\text{s}(t)$, $\text{VAL}_\text{f}(t)$, $\text{AR}_\text{s}(t)$, and $\text{AR}_\text{f}(t)$ capture the subjects emotion at a rate of 10 samples per second.

3.3. Deepfake Detection

The third stage of the proposed method performs deepfake detection by analyzing the arousal and valence signals extracted by the second stage. To capture real and synthetic visual and audio emotional behavior, we propose different sets of features and classification methods.

Statistical Features. Because they are not explicitly constrained to do so, deepfake technologies may struggle to produce emotionally consistent videos. Specifically, they may fail to produce the same coherent emotion across multiple modalities, such as facial expression and speech. To capture this inconsistency, we propose measuring the correlation between facial and speech valence signals, as well as the correlation between facial and speech arousal signals. We do this using Lin's Concordance Correlation Coefficient (CCC), a commonly-used metric in this field [6]. Formally let us define valence and arousal concordance features as

$$\text{C}_\text{VAL} = \text{CCC}(\text{VAL}_\text{s}(t),\ \text{VAL}_\text{f}(t)), \tag{5}$$

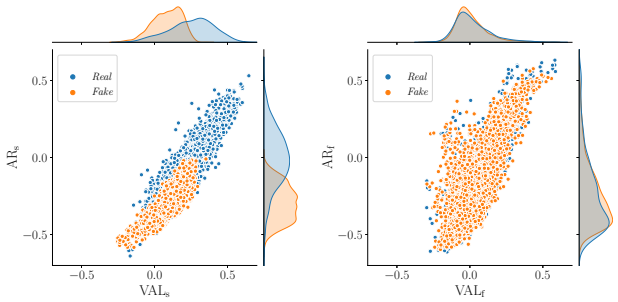
$$\text{C}_\text{AR} = \text{CCC}(\text{AR}_\text{s}(t),\ \text{AR}_\text{f}(t)), \tag{6}$$

where the function $\text{CCC}(X, Y)$ computes the concordance correlation coefficient between signals X and Y. Furthermore, deepfakes may fail to produce the intensity and range of emotion that an authentic subject can convey. To capture this behavior, we consider eight additional features. First, we consider the mean of speech and facial valence and arousal. Mean speech valence is defined as

$$\mu_{\text{VAL, s}} = \frac{1}{T}\sum_t \text{VAL}_\text{s}(t), \tag{7}$$

where T is the length of the time series. The other values $\mu_{\text{AR, s}}$, $\mu_{\text{VAL, f}}$ and $\mu_{\text{AR, f}}$ are computed similarly. Then, to capture the range of emotion for each modality we consider the standard deviations of all the four time-series. Speech valence standard deviation is defined as

$$\sigma_{\text{VAL, s}} = \sqrt{\frac{1}{T}\sum_t (\text{VAL}_\text{s}(t) - \mu_{\text{VAL, s}})^2}. \tag{8}$$



(a) Speech emotion predictions (b) Facial emotion predictions

Figure 3: Mean valence and arousal predictions for both speech (a) and facial (b) features in *Real* (●) and *Fake* (●) sequences.

The values $\sigma_{\text{AR, s}}$, $\sigma_{\text{VAL, f}}$ and $\sigma_{\text{AR, f}}$ are computed similarly.

Our final 10-element detection feature vector consists of the concatenation of all these statistical features. This can be fed to any supervised classifier to detect deepfakes.

Learned Features. Emotion statistics captured through the previously defined features can provide strong cues about a video being fake. However, deepfake generation techniques may produce temporal emotional changes that differ from the emotional behavior of a real human. Because of this, we propose an alternative deepfake detection technique that exploits the analysis of valence and arousal temporal evolution.

In particular, we propose the use of a learning based detection approach. Due to the low feature dimensionality (i.e., audio and video valence and arousal), we use a lightweight LSTM to detect fake videos. The proposed LSTM architecture consists of a single bidirectional LSTM layer with 8 neurons, followed by an output layer with two neurons. The network is trained with MSE loss. The network input consists of the four time series $\text{VAL}_\text{f}(t)$, $\text{VAL}_\text{s}(t)$, $\text{AR}_\text{f}(t)$ and $\text{AR}_\text{s}(t)$. The softmaxed output of the network can be used to detect if a video is fake.

4. Experimental Setup

4.1. Datasets

To evaluate our system, we used two different datasets. The first was used to train the emotion prediction module, while the second was used to train the deepfake detection module.

Emotion Recognition. The LSTM model used to predict valence and arousal values starting from speech and facial LLDs was trained using the SEMAINE database [37]. This database is widely used by the research community to train and benchmark emotion recognition algorithms. It contains 95 videos of 22 talking human subjects with ground truth valence and arousal annotations. In total, there are approximately 427 minutes of video, which we split into 10-second windows to train our emotion recognition network. The dataset focuses on interactions between subjects and a Sensitive Artificial Listener (SAL), i.e., an operator who responds according to a predetermined set of rules. The subjects are facing the camera and speaking to the off-screen operator. The videos in the database were collected in a laboratory environment, so factors such as lighting, background noise, and subject movement are controlled. The subjects in the database are all Caucasian adults with British accents.

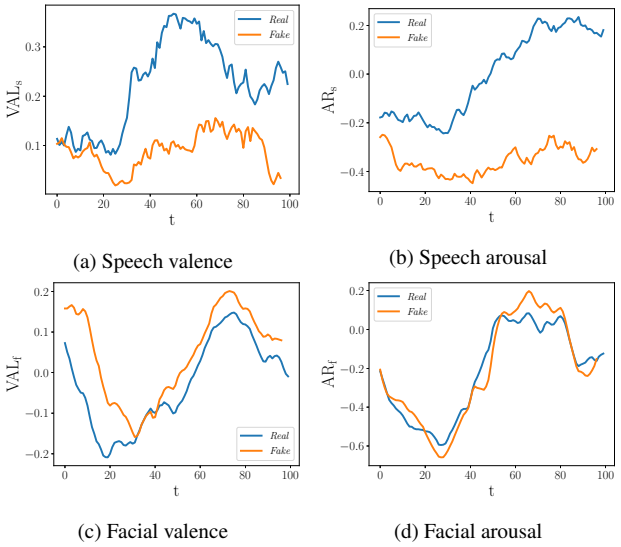
Deepfake Detection. In the following experiments, we used a subset of the Deepfake Detection Challenge training dataset (DFDC) [19]. The DFDC contains nearly 120,000 videos, of which, 100,000 are labeled as *Fake*, and the rest as *Real*. The videos are divided into 50 folders, numbered from 0 to 49, where each subset contains a set of *Real* videos, along with all derivative *Fake* videos. While the videos are largely visual-based fakes, some of the videos in divisions 45 to 49 contain falsified audio in addition to possible falsified video [19]. A given subject may appear multiple times within the DFDC, however, only within the same division. By separating the divisions, we can ensure that our system is not over-fitting to specific identities.

To create a dataset for our experiments, we considered the videos within folders 45 to 49 that contained both *Fake* video and audio and extracted them with the corresponding *Real* source videos. Videos with fake audio were detected by computing the difference between the audio tracks of pairs of *Real* and *Fake* videos, and determining those that differed significantly. The collected videos were then divided based on their original partition within the DFDC. This ended in 5 folders, each with between 1,230 and 1,749 videos, corresponding to the 5 divisions of the DFDC dataset. Our resulting dataset contained 5,248 fake videos, created from 1,844 real videos, totaling 7,092 videos.

4.2. Emotion Recognition Setup

The first stage of our proposed system determines the emotions of a subject that appears in a video. To do this, we extract visual and speech LLDs and we feed them to an LSTM model to predict the subject's emotional valence and arousal. We used the SEMAINE database, as described in Section 4.1, to train this system.

LLDs Extraction. To extract facial and speech LLDs, we used all the videos in the SEMAINE dataset and divided them into 10-second windows. From each window we extracted audio and video features at a sampling rate of 10Hz, resulting in time-series LLD_f and LLD_s with length $T = 100$ time samples. As described in Section 3.1, we used the openSMILE [24] and openFace [8] libraries to extract the speech and the facial LLDs respectively. At the end of this process, we obtained 80×100 samples LLD_s



(a) Speech valence (b) Speech arousal

(c) Facial valence (d) Facial arousal

Figure 4: Temporal evolution of speech valence VAL_s (a), speech arousal AR_s (b), facial valence VAL_f (c) and facial arousal AR_f (d) for *Real* and *Fake* videos.

representation and a 17×100 samples LLD_f representation for each 10-second window of video.

Emotion Recognition. To train the emotion recognition network starting from the extracted LLDs, we divide the dataset into a training, validation, and testing set, based on the video's subject. The training set contains 328 minutes of video from the first 16 subjects of the SEMAINE dataset. The validation set contains 46 minutes of video from subjects 17-19 of the dataset. The model is trained on the training set of LLD sequences for up to 50 epochs, using the validation set to determine the best-performing checkpoint. We trained using an RMSprop optimizer with an initial learning rate of 0.001, and a batch size of 256. The loss function we use aims to maximize the CCC between the predicted and the true emotion. In this stage we produce a set of four identical LSTM models, each one trained to better predict one of the speech/facial values of valence and arousal.

4.3. Deepfake Detection Setup

To perform deepfake detection based on emotional features, we trained our proposed classifiers on the DFDC dataset. Here we want to ensure that none of the speakers considered in the training phase also appears during the test. This is to avoid any unwanted speaker recognition and ensure that the obtained results depend only on the emotional features extracted. To do so, we exploit the division of the dataset into folders in building the subsets we use. In particular, we use folders 45 and 46 as the training set, folder 47 as validation, and folders 48 and 49 as the test set. All DFDC videos are 10 seconds long. When these videos are sampled at a rate of 10Hz, our LLD time-series are $T = 100$ samples long. These LLDs were then fed to the emotion recognition LSTM to produce valence and arousal facial and speech signals.

Statistical Features. From the obtained series VAL_f, VAL_s, AR_s and AR_f we computed a statistical feature vector for each 10-second video clip as explained in Section 3.3. These features were then used to train and test a supervised classifier and estimate the binary label (*Real* vs. *Fake*). We performed an extensive analysis training different types of classifiers and comparing their performances with each other. In particular, the classifiers we considered are Random Forest (RF), XGBoost (XGB), Logistic Regression (LR), and K-Nearest Neighbors (k-NN). In addition to doing a multi-modal study (audio and video), we want to assess each modality's ability to detect deepfakes. To do so, we repeated this experiment considering two subsets of the feature vector. First, we removed the visual and correlation features, feeding the classifiers only with the emotional features derived from speech (audio only). Then, we did the opposite by keeping only the facial ones (video only).

Learned Features. Statistical features may not fully exploit the temporal evolution differences in valence and arousal, which could be crucial to detect deepfakes. Therefore, we propose using an LSTM-based classifier on top of valence and arousal. This stage's input is the set of VAL_f, VAL_s, AR_s and AR_f time-series. This results in a 4×100 matrix. We used the same train-validation-test subdivision based on folders previously introduced in this section, and trained the classifier for 100 epochs with an early-stopping patience of 15 epochs. We used the Adam optimizer, Mean Squared Error (MSE) as loss function, an initial learning rate of 0.01, and a batch size of 64. The validation set was used to select the best-performing checkpoint, and the testing set was used to compute the final performance metrics. As we did in the statistical approach study, to assess each modality's ability to detect deepfakes, we repeated the experiment by taking two subsets of the full feature set. First, we removed the facial-derived signals, and we fed the classifier considering only the two speech emotion channels (audio only). Then, we removed the speech-derived features, and we trained and tested the network only with the two facial emotion channels (video only).

5. Results and Discussion

5.1. Emotional Features in Deepfakes

Our first experiment was to train a system to predict emotion in a video or audio signal, in the form of valence and arousal, based on LLDs. Once this system was trained, we used it to predict emotions in an independent dataset, the DFDC [19]. Figure 3 shows the prediction values, where valence and arousal values are divided between *Real* and

Table 1: Deepfake detection performance of the tested classifiers and feature sets including accuracy, AUC, and detection rate at 5% false alarm rate.

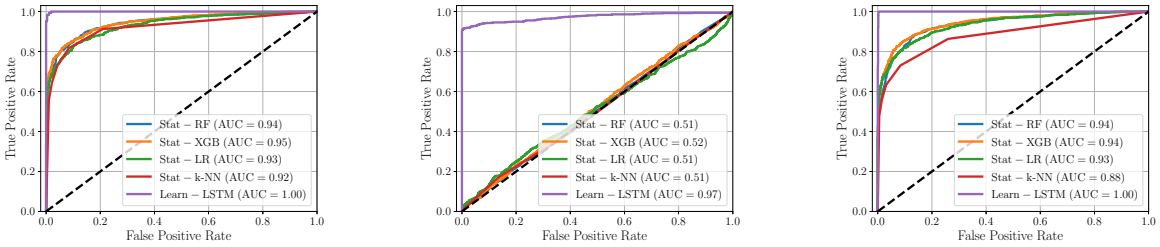
Model	Scenario	Balanced Acc.	AUC	TPR@5%
Stat - RF	Audio	87.1%	0.937	83.2%
	Video	50.2%	0.509	12.0%
	A+V	84.9%	0.945	84.9%
Stat - XGB	Audio	87.8%	0.944	94.7%
	Video	51.1%	0.519	51.9%
	A+V	87.4%	0.947	94.4%
Stat - LR	Audio	84.7%	0.930	81.8%
	Video	50.4%	0.508	13.2%
	A+V	85.3%	0.933	82.7%
Stat - k-NN	Audio	84.5%	0.882	86.3%
	Video	51.8%	0.507	16.8%
	A+V	80.1%	0.921	91.3%
Learn - LSTM	Audio	98.9%	1.000	100.0%
	Video	95.7%	0.973	94.3%
	A+V	99.5%	1.000	100.0%

Fake sequences. The plotted values represent the mean values of valence and arousal for each time window of 10s we consider in the prediction stage. Figure 3a shows that the range spanned by valence and arousal for real audio is much larger than that of fake audio. This suggests that audio faking techniques fail in synthesizing the original speaker's range and intensity of emotion. On the other hand, Figure 3b shows little difference between the emotion distributions represented in real and fake visual sequences. One possible explanation for this similarity is that deepfakes are designed to transfer a target's facial features onto a subject's facial expression. In this process, the subject's expression, and therefore instantaneous emotion, is largely preserved. This hypothesis can be tested by examining the time series of valence and arousal. Figure 4 shows the predicted values of valence and arousal (i.e., VAL_f, VAL_s, AR_f, AR_s) in one *Fake* video, and its associated *Real* source video. Fake audio exhibits valence and arousal values that are consistently lower than those of its authentic counterpart. The facial emotion however, is highly correlatedThis re-enforces our previous hypothesis that, while synthetic audio is not as expressive as authentic audio, fake video emotions are much more easily synthesized.

5.2. Deepfake Detection

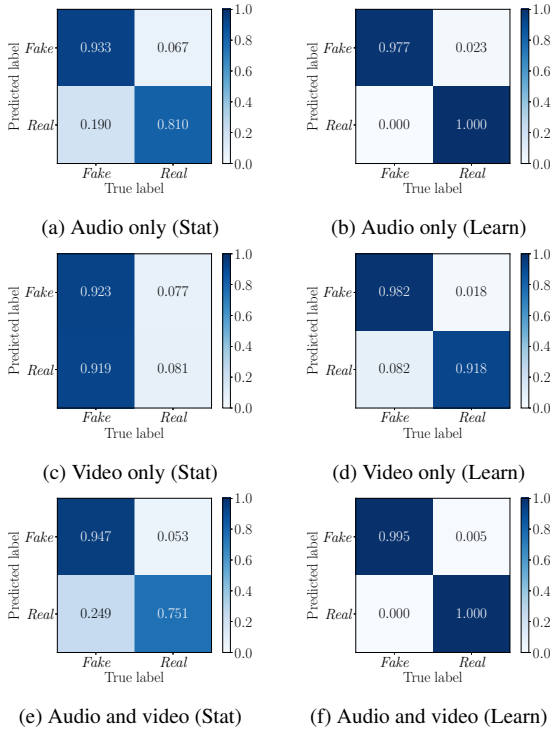
Statistical Features. To test our statistical features' discriminative performance, we trained multiple classifiers using the same feature sets and compared their performance. Table 1 shows the performance of each tested classifier. From Table 1, we can see that all of them can achieve an accuracy of over 80%. These classifiers also maintain their accuracy when operating at low false-alarm rates. Interestingly, Table 1 shows that these classifiers consistently perform better using speech-derived features than when using facial features. Classification based on the statistical facial features achieves a maximum of 51.8% accuracy, little better than random chance, whereas the same classifiers trained only on speech features achieved roughly 85% accuracy. The same trend is confirmed from RF confusion matrices reported in Figures 6a, 6c and 6e as well as from all classifiers ROC curves shown in Figure 5. This result is consistent with the considerations based on the analysis of Figures 3 and 4. We believe that the LLDs used to create these facial emotion signals are closely tied to the features used by deepfake algorithms. Many deepfake algorithms focus not just on changing the identity of the subject in a video, but on preserving its original instantaneous expression and movement. This allows them to become more convincing fakes. However, the subject's expression is what our proposed system uses to predict emotion, meaning that the facial features we use for discrimination are the same that are being faked. The end result is that real and fake faces both express the same instantaneous emotion, thus our system cannot distinguish between them on the basis of this.

Learned Features. In addition to our handcrafted statistical features, we propose the use of learned features that capture the temporal evolution of emotion for performing deepfake detection. Table 1 shows the results of this approach (LSTM). With an accuracy of 99.5%, this method strongly outperforms the statistical feature approach. The same trend is confirmed by the confusion matrices in Figures 6b, 6d and 6f. Figure 5 compares ROC curves for the examined classifiers. Most notably, Figure 5b shows that the LSTM gives impressive low-false-alarm rate accuracy when classifying only based on facial features, while all other classifiers fail to outperform random chance consistently. This supports our hypothesis that deepfakes struggle to create semantic consistencies within a medium in the form of emotion. Our previous experiments demonstrated, primarily, that synthetic speech does not achieve the same emotional range as authentic speech, and that this property is useful for detection. Contrarily, synthetic faces can recreate the same instantaneous emotions as authentic faces. However, this does not mean that deepfake algorithms can create *temporally consistent* emotion. For example, many deepfake algorithms are stateless. They perform image-to-image or face-to-face translation on a single image at a time. These methods do not impose strict consistency between multiple executions, which may result in a choppy or inconsistent sequence of pictures, or video which is smooth, but contains emotional changes over time that differ from a real human's behavior. Algorithms that enforce consistency in one domain, such as the visual domain, may struggle from this same issue in another domain. As described in this paper, we believe that the emotion domain is one such



(a) Audio only (b) Video only (c) Audio and video

Figure 5: Receiver Operating Characteristic (ROC) curves obtained with statistical features (Stat) and learned features (Learn) considering audio only (a), video only (b), and audio and video jointly (c).



(a) Audio only (Stat) (b) Audio only (Learn)

(c) Video only (Stat) (d) Video only (Learn)

(e) Audio and video (Stat) (f) Audio and video (Learn)

Figure 6: Confusion matrices obtained using RF on statistical features (Stat) on audio (a), video (c), and audio and video (e) compared to confusion matrices obtained using LSTM on learned features (Learn) on audio (b), video (d), and audio and video (f).

example that can be exploited for detection. Our LSTM approach supports this hypothesis by exploiting the subjects' temporal evolution to produce accurate detection results.

5.3. Discussion of Bias

As researchers and authors in the field of automated learning, we find it necessary to discuss the potential sources of bias in our work. In this work we present *emotion* to detect deepfakes and synthetic-audio in videos. This is done by first training an emotion recognition model. The dataset used to train this model consists of Caucasian adults with British accents. This dataset is **not** representative of all peoples. We expect that those cultures and languages which are not represented in this dataset express emotions differently, both visually and audibly, than those are represented. Furthermore, in order to validate our proposed technique we used a dataset containing primarily US American residents. As such we do not purport that our trained classifiers will generalize to other cultures. The extension of our proposed method to arbitrary populations may be non-trivial, and is a subject of future research.

6. Conclusions

In this work we have proposed a system for performing deepfake detection using semantic consistency in emotion. Our proposed system builds upon existing emotion recognition work to extract emotions over time from a subject's speech and face separately. These emotion signals are then analyzed to detect the presence of synthesized speech or faces. We show experimentally that our system is able to discriminate between real and deepfake videos, achieving accuracy of up to 99.5%. Additionally, our proposed technique achieves 100% detection accuracy on our test set at a false alarm rate of only 5%.

Acknolwedgement

Research was sponsored by the Army Research Office and was accomplished under Cooperative Agreement Number W911NF-20-2-0111, the Defense Advanced Research Projects Agency (DARPA) and the Air Force Research Laboratory (AFRL) under agreement numbers FA8750-20-2-1004 and HR001120C0126, and by the National Science Foundation under Grant No. 1553610. The U.S. Government is authorized to reproduce and distribute reprints for Governmental purposes notwithstanding any copyright notation thereon. The views and conclusions contained herein are those of the authors and should not be interpreted as necessarily representing the official policies or endorsements, either expressed or implied, of DARPA and AFRL, the Army Research Office, the National Science Foundation, or the U.S. Government.

References

[1] 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 (WIFS)*, 2018. 2

[2] S. Agarwal, H. Farid, O. Fried, and M. Agrawala. Detecting deep-fake videos from phoneme-viseme mismatches. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*, 2020. 2

[3] 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 Workshops (CVPRW)*, 2019. 1, 2

[4] M. B. Akçay and K. Oğuz. Speech emotion recognition: Emotional models, databases, features, preprocessing methods, supporting modalities, and classifiers. *Speech Communication*, 116:56–76, 2020. 2

[5] E. A. AlBadawy, S. Lyu, and H. Farid. Detecting AI-synthesized speech using bispectral analysis. In *IEEE Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*, 2019. 1, 3

[6] B. T. Atmaja and M. Akagi. Evaluation of error and correlation-based loss functions for multitask learning dimensional speech emotion recognition. *CoRR*, abs/2003.10724, 2020. 4

[7] T. Baltrušaitis, P. Robinson, and L.-P. Morency. 3d constrained local model for rigid and non-rigid facial tracking. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. IEEE, 2012. 2

[8] 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 (WACV)*. IEEE, 2016. 3, 5

[9] E. Y. Bann and J. J. Bryson. The conceptualisation of emotion qualia: Semantic clustering of emotional tweets. In *Neural Computation and Psychology Workshop (NCPW)*, 2014. 2

[10] M. Barni, M. C. Stamm, and B. Tondi. Adversarial multimedia forensics: Overview and challenges ahead. In *European Signal Processing Conference (EUSIPCO)*, 2018. 1

[11] BBC News. Deepfakes porn has serious consequences. https://www.bbc.com/news/technology-42912529, 2018. 1

[12] N. Bonettini, E. D. Cannas, S. Mandelli, L. Bondi, P. Bestagini, and S. Tubaro. Video Face Manipulation Detection Through Ensemble of CNNs. In *International Conference on Pattern Recognition (ICPR)*, 2020. 1, 2

[13] E. Cambria, D. Das, S. Bandyopadhyay, and A. Feraco. Affective computing and sentiment analysis. In *A practical guide to sentiment analysis*. Springer, 2017. 2

[14] S. Chen and Q. Jin. Multi-modal conditional attention fusion for dimensional emotion prediction. In *ACM International Conference on Multimedia*, 2016. 2

[15] M. Day. Emotion recognition with boosted tree classifiers. In *Proceedings of the 15th ACM on International conference on multimodal interaction*, pages 531–534, 2013. 3

[16] D. Deng, Y. Zhou, J. Pi, and B. E. Shi. Multimodal utterance-level affect analysis using visual, audio and text features. *CoRR*, abs/1805.00625, 2018. 2

[17] DeepFake Detection Challenge Results. https://ai.facebook.com/blog/deepfake-detection-challenge-results-an-open-initiative-to-advance-ai. 2

[18] H. Dinkel, N. Chen, Y. Qian, and K. Yu. End-to-end spoofing detection with raw waveform CLDNNS. In *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2017. 3

[19] B. Dolhansky, J. Bitton, B. Pflaum, R. Lu, Jikuo ans Howes, M. Wang, and C. Canton Ferrer. The deepfake detection challenge dataset. *CoRR*, abs/2006.07397, 2020. 1, 2, 5, 6

[20] M. Egger, M. Ley, and S. Hanke. Emotion recognition from physiological signal analysis: a review. *Electronic Notes in Theoretical Computer Science*, 343:35–55, 2019. 3

[21] R. Ekman. *What the face reveals: Basic and applied studies of spontaneous expression using the Facial Action Coding System (FACS)*. Oxford University Press, USA, 1997. 3

[22] Euronews. French charity publishes deepfake of Trump saying 'AIDS is over'. https://www.euronews.com/2019/10/09/french-charity-publishes-deepfake-of-trump-saying-aids-is-over, 2019. 1

[23] F. Eyben, K. R. Scherer, B. W. Schuller, J. Sundberg, E. André, C. Busso, L. Y. Devillers, J. Epps, P. Laukka, S. S. Narayanan, et al. The geneva minimalistic acoustic parameter set (GeMAPS) for voice research and affective computing. *IEEE transactions on affective computing*, 7:190–202, 2015. 3

[24] F. Eyben, M. Wöllmer, and B. Schuller. Opensmile: the munich versatile and fast open-source audio feature extractor. In *ACM International Conference on Multimedia*, 2010. 3, 5

[25] D. Güera and E. J. Delp. Deepfake Video Detection Using Recurrent Neural Networks. *IEEE International Conference on Advanced Video and Signal-Based Surveillance (AVSS)*, 2019. 1, 2

[26] S. Ji, W. Xu, M. Yang, and K. Yu. 3D convolutional neural networks for human action recognition. *IEEE Transactions on Pattern Analysis and Machine Intelligence (TPAMI)*, 35:221–231, 2012. 2

[27] N. Kalchbrenner, E. Elsen, K. Simonyan, S. Noury, N. Casagrande, E. Lockhart, F. Stimberg, A. van den Oord, S. Dieleman, and K. Kavukcuoglu. Efficient neural audio synthesis. *CoRR*, abs/1802.08435, 2018. 2

[28] M. R. Kamble, H. B. Sailor, H. A. Patil, and H. Li. Advances in anti-spoofing: from the perspective of ASVspoof challenges. *APSIPA Transactions on Signal and Information Processing*, 9, 2020. 3

[29] R. Khan and O. Sharif. A literature review on emotion recognition using various methods. *Global Journal of Computer Science and Technology*, 2017. 2

[30] P. Korshunov and S. Marcel. DeepFakes: a New Threat to Face Recognition? Assessment and Detection. *CoRR*, abs/1812.08685, 2018. 2

[31] Y. Li, M. 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 (WIFS)*, 2018. 2

[32] Y. Li and S. Lyu. Exposing deepfake videos by detecting face warping artifacts. In *IEEE Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*, 2019. 1, 2

[33] Y. Li, P. Sun, H. Qi, and S. Lyu. Celeb-DF: A Large-scale Challenging Dataset for DeepFake Forensics. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2020. 1

[34] A. Lieto, D. Moro, F. Devoti, C. Parera, V. Lipari, P. Bestagini, and S. Tubaro. "Hello? Who Am I Talking to?" A Shallow CNN Approach for Human vs. Bot Speech Classification. In *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2019. 1, 3

[35] S. Marsella and J. Gratch. Computationally modeling human emotion. *Communications of the ACM*, 57:56–67, 2014. 2

[36] F. Matern, C. Riess, and M. Stamminger. Exploiting visual artifacts to expose deepfakes and face manipulations. In *IEEE Winter Applications of Computer Vision Workshops (WACVW)*, 2019. 1

[37] G. McKeown, M. Valstar, R. Cowie, M. Pantic, and M. Schroder. The semaine database: Annotated multimodal records of emotionally colored conversations between a person and a limited agent. *IEEE Transactions on Affective Computing (TAC)*, 3:5–17, 2011. 5

[38] H. H. Nguyen, F. Fang, J. Yamagishi, and I. Echizen. Multi-task learning for detecting and segmenting manipulated facial images and videos. *CoRR*, abs/1906.06876, 2019. 2

[39] S. Poria, E. Cambria, R. Bajpai, and A. Hussain. A review of affective computing: From unimodal analysis to multimodal fusion. *Information Fusion*, 37:98–125, 2017. 2

[40] S. Poria, E. Cambria, D. Hazarika, N. Majumder, A. Zadeh, and L.-P. Morency. Context-dependent sentiment analysis in user-generated videos. In *Annual Meeting of the Association for Computational Linguistics*, 2017. 2

[41] S. Poria, E. Cambria, D. Hazarika, N. Mazumder, A. Zadeh, and L.-P. Morency. Multi-level multiple attentions for contextual multimodal sentiment analysis. In *IEEE International Conference on Data Mining (ICDM)*. IEEE, 2017. 2

[42] L. Rabiner and B.-H. Juang. *Fundamental of Speech Recognition*. Prentice-Hall, 1993. 3

[43] P. Rathod, K. George, and N. Shinde. Bio-signal based emotion detection device. In *2016 IEEE 13th International Conference on Wearable and Implantable Body Sensor Networks (BSN)*, pages 105–108. IEEE, 2016. 3

[44] N. A. Remington, L. R. Fabrigar, and P. S. Visser. Reexamining the circumplex model of affect. *Journal of personality and social psychology*, 79(2):286, 2000. 4

[45] F. Ringeval, B. Schuller, M. Valstar, N. Cummins, R. Cowie, and M. Pantic. AVEC'19: Audio/visual emotion challenge and workshop. In *ACM International Conference on Multimedia*, 2019. 2, 3, 4

[46] A. Rössler, D. Cozzolino, L. Verdoliva, C. Riess, J. Thies, and M. Nießner. FaceForensics++: Learning to detect manipulated facial images. In *International Conference on Computer Vision (ICCV)*, 2019. 1

[47] D. C. Rubin and J. M. Talarico. A comparison of dimensional models of emotion: Evidence from emotions, prototypical events, autobiographical memories, and words. *Memory*, 17(8):802–808, 2009. 4

[48] J. A. Russell. A circumplex model of affect. *Journal of personality and social psychology*, 39:1161, 1980. 2

[49] M. Sahidullah, T. Kinnunen, and C. Hanilçi. A comparison of features for synthetic speech detection. In *Interspeech*, 2015. 3

[50] G. Sahu. Multimodal speech emotion recognition and ambiguity resolution. *CoRR*, abs/1904.06022, 2019. 2

[51] J. Shen, R. Pang, R. J. Weiss, M. Schuster, N. Jaitly, Z. Yang, Z. Chen, Y. Zhang, Y. Wang, R. Skerrv-Ryan, R. A. Saurous, Y. Agiomvrgiannakis, and Y. Wu. Natural TTS synthesis by conditioning wavenet on mel spectrogram predictions. In *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2018. 2

[52] L. Sun, Z. Lian, J. Tao, B. Liu, and M. Niu. Multi-modal continuous dimensional emotion recognition using recurrent neural network and self-attention mechanism. In *International Workshop on Multimodal Sentiment Analysis in Real-life Media Challenge*, 2020. 2

[53] The New York Times. Pennsylvania Woman Accused of Using Deepfake Technology to Harass Cheerleaders. https://www.nytimes.com/2021/03/14/us/raffaela-spone-victory-vipers-deepfake.html, 2021. 1

[54] The Wallstreet Journal. Fraudsters Used AI to Mimic CEO's Voice in Unusual Cybercrime Case. https://www.wsj.com/articles/fraudsters-use-ai-to-mimic-ceos-voice-in-unusual-cybercrime-case-11567157402, 2019. 1

[55] J. Thies, M. Zollhöfer, and M. Nießner. Deferred neural rendering: Image synthesis using neural textures. *ACM Transactions on Graphics (TOG)*, 38:1–12, 2019. 2

[56] M. Todisco, X. Wang, V. Vestman, M. Sahidullah, H. Delgado, A. Nautsch, J. Yamagishi, N. Evans, T. Kinnunen, and K. A. Lee. ASVspoof 2019: Future Horizons in Spoofed and Fake Audio Detection. *CoRR*, abs/1904.05441v2, 2019. 1

[57] A. van den 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. *CoRR*, abs/1609.03499, 2016. 2

[58] L. Verdoliva. Media forensics and deepfakes: an overview. *CoRR*, abs/2001.06564, 2020. 1, 2

[59] R. Xia and Y. Liu. Using i-vector space model for emotion recognition. In *Thirteenth Annual Conference of the International Speech Communication Association*, 2012. 3

[60] X. Yang, Y. Li, and S. Lyu. Exposing deep fakes using inconsistent head poses. In *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2019. 2

[61] M. Zollhöfer, J. Thies, P. Garrido, D. Bradley, T. Beeler, P. Pérez, M. Stamminger, M. Nießner, and C. Theobalt. State of the art on monocular 3D face reconstruction, tracking, and applications. *Computer Graphics Forum*, 37:523–550, 2018. 2