Audio-Visual Person-of-Interest DeepFake Detection

Davide Cozzolino[1] Alessandro Pianese[1] Matthias Nießner[2] Luisa Verdoliva[1]
[1]University Federico II of Naples [2]Technical University of Munich

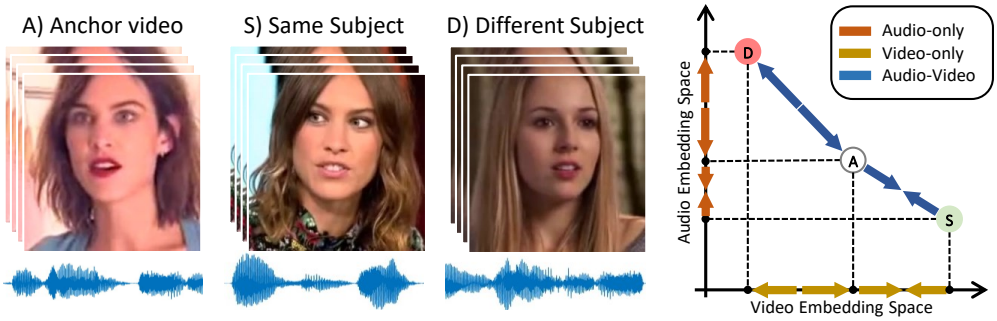

Figure 1. We propose POI-Forensics, a method to detect Deepfakes based on audio-visual identity verification. Given a set of real identities, we propose a contrastive learning method that encourages embedded vectors of a reference video (A) to be close to embedded vectors of the same subject (S), but far from those of different subjects (D).

Abstract

Face manipulation technology is advancing very rapidly, and new methods are being proposed day by day. The aim of this work is to propose a deepfake detector that can cope with the wide variety of manipulation methods and scenarios encountered in the real world. Our key insight is that each person has specific characteristics that a synthetic generator likely cannot reproduce. Accordingly, we extract audio-visual features which characterize the identity of a person, and use them to create a person-of-interest (POI) deepfake detector. We leverage a contrastive learning paradigm to learn the moving-face and audio segment embeddings that are most discriminative for each identity. As a result, when the video and/or audio of a person is manipulated, its representation in the embedding space becomes inconsistent with the real identity, allowing reliable detection. Training is carried out exclusively on real *talking-face video; thus, the detector does not depend on any specific manipulation method and yields the highest generalization ability. In addition, our method can detect both single-modality (audio-only, video-only) and multi-modality (audio-video) attacks, and is robust to low-quality or corrupted videos. Experiments on a wide variety of datasets confirm that our method ensures a SOTA performance, especially on low quality videos. Code is publicly available on-line at* https://github.com/grip-unina/poi-forensics.

1. Introduction

Synthetic media generation has become a key technology in many industrial applications, from film production to the video game industry. Facial manipulations, however, also pose a serious and growing threat to our society, of which financial fraud and disinformation campaigns are just a few examples. With the advancement of such technology, there is a steady increase in the level of photorealism, as more and more methods of video manipulation emerge. In particular, the term deepfake, which is often associated with face-swapping, has now become associated with negative implications. Currently, the deepfake term has taken on an even broader meaning, including a variety of possible video manipulations: speech can be synthesized in anyone's voice, face expression can be modified, the identity of a person can be swapped with another, even altering what they are saying. Some examples of different manipulations of the same identity are shown in Fig. 2: a video where the audio has been manipulated and lip movements have been perfectly synchronized with it [43] and two different types of face-swapping [32, 42].

Dealing with such a large spectrum of manipulations is the main challenge for current video deepfake detectors. In fact, it is particularly difficult to develop a method that can detect multiple known manipulation methods at the same time; this is only exacerbated when targeting unknown methods that were not part of any training samples. As a result, current SOTA detectors, trained over large datasets of deepfake and pristine videos, often show an unsatisfactory

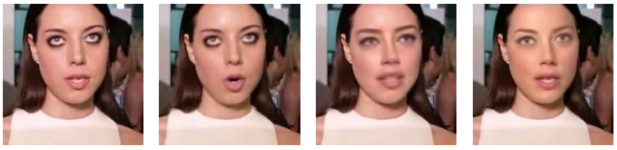

Figure 2. Different manipulations applied to the same original video, from left to right: original video, facial reenactment manipulation using Wav2Lip [43] and two faces-swapping-manipulation using Faceswap [32] and Faceswap GAN (FSGAN) [42].

cross-dataset performance, clearly highlighting the limitations of the supervised deep learning based approach. In addition, performance often drops dramatically under a more challenging scenario, such as low-quality videos. Such conditions, however, are commonplace in the real world where videos are mostly disseminated through social networks where they are further compressed with the loss of relevant audio-visual information. It is also worth pointing out that deep learning-based methods are vulnerable to adversarial attacks with detection performance that degrades sharply even in a black-box scenario [11, 24, 41].

A possible solution to gain generalization and robustness is to shift to a completely different paradigm, training models only on real videos, with the goal to detect manipulated videos based on their anomalous behavior [9]. This approach turns out to be particularly effective if the characterization of pristine faces is based on semantic features, such as soft biometrics, leading to Person-of-Interest (POI) based detection [1, 3–5, 10, 18]. First papers on this topic exploited specific face and head movement patterns [1, 3, 10], inconsistencies between mouth shape dynamics and spoken phonemes [2] or cues related to the specific words uttered by the identity [4] or inconsistencies between inner and outer face regions [18]. These methods present some limitations: in particular, they rely on video-only features, sometimes complemented by categorical information, neglecting precious audio information. Moreover, they often need several hours of videos of the identity under test.

In this work, we propose a new person-of-interest (POI) deepfake detector, called POI-Forensics. The key feature of our approach is the use of a multi-modal analysis. More specifically we rely on a contrastive learning approach and train an audio and video network so that the learned representation characterizes temporal segments of the same identity close to one-another but far from each-other for different identities, as can be seen in Figure 1. At test time, we compute similarity indices between the features extracted from the video under analysis and those extracted by a set of POI-based reference videos. We also include joint audio-video similarity indices that have been shown to improve the discrimination ability of the detector. Overall the proposed method ensures a number of important benefits:

Generalization. Since training does not rely on fake videos, our detector works equally well on known and unknown manipulations. This ensures good generalization to attacks not seen during training (as none are seen during training). The detector can deal with any manipulation method (face swapping, facial reenactment or anything else) with performance that depends only on the fidelity of the video, not on specific inconsistencies or artifacts of the manipulation method.

Flexibility. Due to our multi-modal approach, we can detect also video-only and speech-only manipulations, and even the swapping of a real audio track on the real video of the original identity. When the manipulation involves video and audio jointly, the joint-modality analysis improves performance.

Robustness. The detector is robust to many challenging conditions frequently encountered in real scenarios, where videos are compressed or even maliciously attacked.

No need of re-training. Training does not require videos of the POI under test, hence re-training is not needed when testing new identities, and only a few short POI videos (around 10 minutes) are necessary at test time.

To summarize, our main contributions are the following:

- We propose POI-Forensics, an audio-visual deepfake detection approach which learns a person-of-interest based representation and exploits single and multi-modality consistencies on video temporal segments.
- Experiments show that our method beats SOTA approaches of a significant margin, especially in the most challenging scenarios of compressed and adversarially attacked videos with AUC and accuracy both improving in the range 7%-14%.

2. Related Work

Single-modality methods. In the large and rapidly growing literature on deepfake detection, most methods rely on supervised training, leveraging large datasets of real and fake videos. For the majority, they analyze only video and exploit low-level features, artifacts due to some imperfections of the generation method. These methods are typically very effective when the target video has been generated with a manipulation present in training, and almost useless otherwise [49, 50]. Since new methods for generating synthetic data are proposed by the day, this latter case may occur quite frequently. Even assuming that examples of new types of manipulation were available, re-training the models on datasets that grow without bounds becomes intractable. On the other hand, fine-tuning them on the new data would likely disrupt performance of old

ones (catastrophic forgetting). To face this problem, specific solutions have been proposed in the literature, resorting to few-shot learning [12,25] or incremental learning [27,39] approaches. In any case, the problem remains of readily acquiring examples of new manipulations. Another simple, yet effective, resource to improve generalization is augmentation. For forensic applications, beyond the standard operations considered in computer vision, it is important to include compression and resizing, so as to gain robustness against the typical impairments caused by social networks. In addition, some dedicated forms of cut-out appear to be useful for deepfake detection [14]. Ensembling is also helpful to boost the performance, and to gain robustness against possible misalignments [6].

Aside from limited performance, another significant limitation of the above approaches is the lack of interpretability. This problem is partially addressed by methods that aim at detecting some specific cues related to the generation process. One direction is to rely on low-level artifacts caused by the up-sampling operation, clearly visible in the Fourier domain. Accordingly, frequency-based analyses are carried out in [34, 36, 38]. Another direction is to learn the specific artifacts introduced by blending, which is a necessary processing step in many different manipulation approaches [35]. More in general, attention mechanisms can be used to guide the network to focus on low-level and/or high-level inconsistencies, both in the spatial and in the temporal domain [13,51,54–56,58]. While these solutions provide some more insight on the type of manipulation performed, they are also very vulnerable to all quality impairment actions that disguise the low-level features they rely upon. These include not only casual image impairment processes, but also voluntary perturbations and adversarial attacks [24, 41] which are becoming more and more commonplace. To gain robustness, the method proposed in [22] is based on semantic features by focusing on inconsistencies in the mouth movements. The spatio-temporal architecture is pre-trained on the visual speech recognition task and then fine-tuned on mouth embeddings of real and forged videos.

Multimodal analysis. In recent years, a few pioneering works have began analyzing audio and video jointly to perform deepfake detection. Some works look for inconsistencies between the audio and video content. The method developed in [29, 31], for example, relies on the inability of some generation methods to correctly synchronize the audio stream with the video content. Likewise, the key idea of [57] is to learn and exploit the intrinsic synchronization between video and audio. However, as technology has advanced quickly, there are now several methods that can generate realistic deepfakes where speech and lips movements are accurately synchronized [43], making audio-visual synchronization analysis extremely challenging. The approach proposed by [40] focuses on the extraction of the emotion features from both modalities together with a similarity analysis within the same audio and the same video. In [52], a multi-modal and multi-scale transformer is designed to exploit both spatial and frequency domain artifacts, while in [7] the idea is to look for inconsistencies between audio and visual streams by training a modality dissonance score. While these methods show promising results, they require both fake and real videos for training, which could impair their generalization ability.

POI-based detection. To avoid being polarized by the type of manipulated videos included in the training set, one may choose to learn only an intrinsic model of pristine videos, and interpret all deviations from this model as an anomaly and, therefore, a manipulation [9, 21]. In particular, the biometrics of an individual allow for the construction of expressive and reliable models [3]. Under this perspective, deepfake video detection can be interpreted as a POI-based attribution task. In fact, the question to answer is thus not, *real or fake?*, but rather *is this the POI or not?*. This identity verification approach is very general as it can handle both cheapfakes and deepfakes at the same time.

Several works in this area rely on facial and head movement features [3, 5, 10]. Specifically, in [1] a method is proposed that includes a face recognition module to capture facial features and another module to learn frame-based facial movements and expressions, that can better measure spatio-temporal biometric behaviors. This approach turns out to be very effective on face-swapping manipulations, much less on facial reenactment ones. To handle both types of attacks, in [10] a low-dimensional 3D morphable model of the person is extracted from the target video so as to analyze the face motion by means of an adversarial learning strategy. In [2], inconsistencies between the mouth shape dynamics and the spoken phonemes are exploited, and the idea is further developed in [4] where inconsistencies between facial movements and spoken words are targeted. This approach can handle also sophisticated speech manipulations where facial characteristics have not been altered. Its main drawback is the need to train on several hours of videos of the POI, which may not be always available and so reduces flexibility. More recently, in [18] it is proposed an identity-based method that detects if the inner and outer face regions belong to the same individual.

3. Method

We propose a new method for deepfake detection which exploits audio-visual features of the portrayed individual. Our key assumption is that any manipulation method, however accurate, will eventually perturb some of these features, which will be thus identified as anomalous, allowing for the detection of the attack. So, our method does not

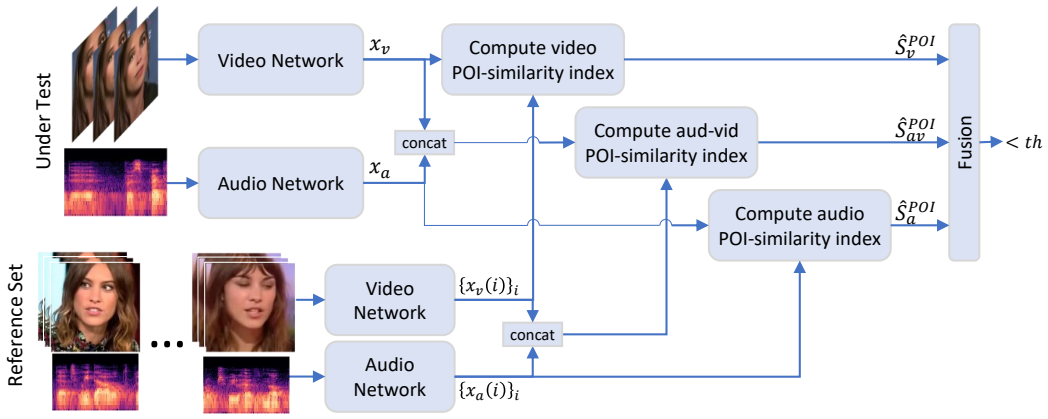

Figure 3. POI-Forensics testing scheme. We extract from the audio and video segments the embedded vectors and compare them with those extracted from a set of pristine videos of person of interest by means of the POI similarity indices (audio-only, video-only, audio-video). Finally, a fusion POI similarity index is computed.

look for direct traces of manipulation, such as generation artifacts, but rather for the indirect but compelling clue that the individual present in the video is not the person who is claimed to be. These features are learned in a suitable embedding space by means of a contrastive learning procedure, looking for consistencies/inconsistencies among embedded vectors associated with different identities. Our method trains exclusively on real videos, a large dataset comprising more than 5,000 identities with the associated audio [8], thereby maintaining independence from any specific manipulation and an intrinsic high generalization ability. The dataset presents around 61% males and spans a wide range of different ethnicities and ages. It varies in terms of environment (indoor/outdoor), hence including different lighting and variations in pose.

Fig. 3 shows the testing scheme. From each 3-second segment of the video at test time, two 256-component embedded vectors are extracted by means of dedicated neural networks (in both cases, ResNet-50 with Group-Normalization [53]), one for the video modality and one for the audio modality. The input of the audio network is the 300×257 amplitude spectrogram extracted from the audio using a window length of 25 ms and a stride of 10 ms. The input of video network is the cropped faces from 25 frames evenly distributed along the segment. The output vectors are denoted as $x_m(c)$, with c indicating the video-segment and $m \in \{a, v\}$ the modality (a for audio and v for video).

During training, performed by the contrastive learning approach described in Section 3.1, the audio and video networks learn to extract features that are generally close to one-another for segments of the same identity and far from each-other for segments of different identities. Accordingly, in the testing procedure, described in Section 3.2, we compute POI similarity indices between the features extracted from the target video and those extracted by a set of reference videos of the same POI. In addition, joint audio-video similarity indices are computed as they are found experimentally to improve performance. In the following, we will describe in detail both train and test phases.

3.1. Contrastive Learning

Our model is trained using a contrastive learning formulation to embedded vectors of a video to be close to embedded vectors of the same identity, but far from those of different identities. Overall the loss is defined as:

$$\mathcal{L}_{tot} = \mathcal{L}_v + \mathcal{L}_a + \lambda \mathcal{L}_{av}, \tag{1}$$

where $\mathcal{L}_v$ and $\mathcal{L}_a$ are the contrastive losses applied to the single modality, while $\mathcal{L}_{av}$ is a joint contrastive loss that takes into account both modalities. Note that the joint contrastive loss is used to push a modality to make up for the shortcomings of the other modality, this gives an improvement when both the modalities are used in testing.

For all three terms, we adopt a multi-way matching loss that compares the positive matches with all the negative matches, leading to a more stable learning as compared with popular pairwise losses, such as the triplet loss, that compare each positive match with only one negative match [47]. We adopt the contrastive loss function used in [10, 28], where the similarity between feature vectors is evaluated as the negative of squared Euclidean distance normalized by a factor τ. The contrastive loss function is expressed as:

$$\mathcal{L}_m = -\sum_c \log \frac{\sum_{k \in \mathcal{N}_c} e^{S_m(c,k)}}{\sum_{k \neq c} e^{S_m(c,k)}} \qquad m \in \{a, v, av\}, \tag{2}$$

where c and k are 3-second segments extracted from different videos, and $S_m(c,k)$ measures their similarity according to modality $m \in \{a, v, av\}$. Note that the summation at the numerator is restricted to the set $\mathcal{N}_c$, comprising only

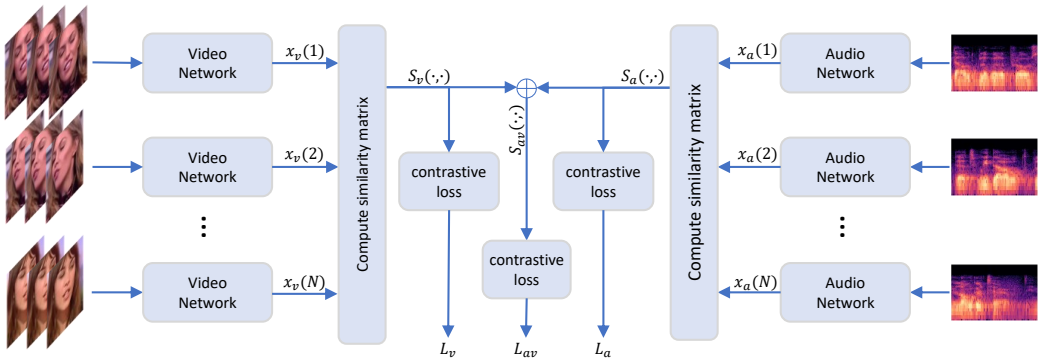

Figure 4. POI-Forensics training scheme. At each training iteration, we analyze N video-segments and extract the embedded vectors from the audio and video signals. All the embedded vectors are compared computing three $N \times N$ matrices of similarity measures for only-video, only-audio and audio-video, respectively. For each matrix, a contrastive loss is evaluated to push closer the embedded vectors of the same individual but move farther from those of different individuals.

segments k that portray the same identity as c. The single-modality similarity measures are defined as:

$$S_m(c,k) = -\frac{1}{\tau}\left\|x_m(c) - x_m(k)\right\|^2 \qquad m \in \{a, v\} \quad (3)$$

while the joint similarity measure $S_{av}(c,k)$ is the sum of the two single-modality measures.

We divide the dataset on a per-individual basis, using 4608 identities for training and 512 for validation. Note that we excluded the 500 identities that were used to build the deepfake dataset FakeAVCelebV2 [23] in order to avoid any possible polarization in our analysis. The networks are trained for 12 epochs with 2304 batches per epoch, and each batch is formed by 8×8 video-segments, with 8 segments each for 8 different individuals. We use the Adam optimizer with decoupled weight decay [37], a learning rate of 10^{-4}, weight decay of 0.01, and the $\tau = 0.01$. To increase robustness we use a wide variety of augmentations, including geometric operations (rescaling, rotation, flipping and shifting), point-wise operations (random variation of brightness, contrast, saturation, and hue) and image impairments (blurring, compression, noise adding and patch removal).

3.2. Testing

We assume to have a reference set of at least 10 pristine videos of the person of interest, each one lasting about 30 seconds, for a total of 100 disjoint segments, that is a reference set $\mathcal{R}$. At test time, for each segment, c, of the target video and each modality we compute an index, which measures the similarity between the test segment and the closest reference segment:

$$S_m^{\text{POI}}(c) = \max_{i \in \mathcal{R}} S_m(c,i) \qquad m \in \{a, v, av\}\,, \quad (4)$$

These statistics will guide our decision, as they provide an indication of how plausible it is that the segment portrays the intended POI. Indeed we rely on the normalized metrics defines as

$$\widehat{S}_m^{\text{POI}}(c) = \frac{S_m^{\text{POI}}(c) - \mu_m^{\text{POI}}}{\sigma_m^{\text{POI}}} \quad (5)$$

where μ_m^{POI} and σ_m^{POI} are, for each modality and person, the mean and standard deviation of the similarity index estimated from the pristine videos of the reference set. In detail, these are estimated on the set $\left\{S_m^{\text{POI}}(i)\right\}_{i \in \mathcal{R}}$ computed by avoiding to evaluate the similarity measures between two segments of the same video. For our assumptions on the reference set, there is a pretty large number of similarity indices to average upon, so this estimates can be considered reliable. We also consider a fusion POI-similarity index, motivated by preliminary experiments, as the minimum of the three normalized POI-similarity indices. Then, under the hypothesis that the video under test is real, that is, it portrays the claimed POI, the normalized similarity index can be regarded as a standard Gaussian random variable $\widehat{S}_m^{\text{POI}}(c) \sim \mathcal{N}(0,1)$. This modeling allows one to compute the probability of false alarm, P_{fa}, in closed form for any decision threshold or, conversely, to set the decision threshold so as to obtain the desired false alarm rate. In the experimental part, we will fix the threshold considering a desired false alarm rate of 10%. Considering a whole target video of, say, 30 seconds, we have 10 disjoint segments with 10 POI-similarity indices for each modality, certainly not independent of one another, to be taken into account jointly. Therefore, we calculate the overall decision statistic by averaging the POI-similarity indices of all the disjoint segments in the target video.

4. Experimental Results

4.1. Datasets and Metrics

To evaluate our method and reference state-of-the-art approaches, we consider several deepfake datasets that also

	v	a	ai	$\lambda=0$ video index	$\lambda=0$ audio index	$\lambda=0$ AV index	$\lambda=0$ Fusion	$\lambda=1$ video index	$\lambda=1$ audio index	$\lambda=1$ AV index	$\lambda=1$ Fusion
AUC (%)	✓			79.0	49.5	70.6	73.4	78.3	49.1	68.8	71.4
	✓	✓		74.4	97.0	93.5	95.5	73.1	96.0	94.6	96.0
		✓	✓	53.6	93.3	82.0	88.7	49.5	95.2	83.2	90.9
	✓	✓	✓	72.0	99.0	95.4	96.5	70.8	98.8	96.4	97.5
	AVG			69.8	84.7	85.4	88.5	67.9	84.8	85.8	**88.9**
Pd@10% (%)	✓			50.1	15.1	29.9	39.4	40.9	14.6	34.1	37.3
	✓	✓		49.6	94.1	87.9	90.9	46.5	94.7	91.6	94.5
		✓	✓	15.6	89.1	62.5	65.6	10.9	90.6	64.1	76.6
	✓	✓	✓	42.3	98.8	90.0	95.6	37.0	97.6	94.3	95.6
	AVG			39.4	74.3	67.6	72.9	33.8	74.4	71.0	**76.0**

Table 1. Results in terms of AUC and Pd%. We compare the four similarity indices and evaluate the effect to include the joint contrastive loss term during training.

provide identity information:

pDFDC, preview DeepFake Detection Challenge dataset [17]. It includes realistic face-swapping manipulations relative to 68 identities.

DF-TIMIT, DeepFake-TIMIT [30]. It contains face-swapping manipulations applied to 320 videos of 32 identities of the Vid-TMIT dataset [45].

FakeAVCelebV2, Audio-Video Deepfake dataset [23]. It comprises both face-swapping and facial reenactment methods. To generate cloned fake audio a transfer learning-based real-time voice cloning tool (SV2TTS [26]) is used.

KoDF, a large-scale Korean DeepFake dataset [33]. It contains 403 identities and is composed of manipulated videos generated by six different techniques, including both face swapping and face-reenactment manipulations.

We report results using the following metrics: AUC, which is very popular in the literature and does not require to set a threshold, Accuracy, with a fixed 0.5 threshold and the probability of detection obtained for a 10% false alarm rate (Pd@10%). All these metrics are evaluated at video-level for each technique included in the analysis.

4.2. Ablation Study

Influence of similarity indexes. First, we assess the importance of each of the similarity indices, $\{a, v, av\}$, contributing the decision statistic and, also, the importance of including or not the joint contrastive loss term: $\lambda = 1$ or 0 in eq. 1. To this end, we experiment on a subset of FakeAVCelebV2 and KoDF datasets comprising a total of more than 140 individuals, and the whole variety of manipulations. Results are presented in Tab. 1 in terms of AUC (top) and Pd@10% (bottom). For this analysis, the deepfakes are grouped according to which parts of the multimedia asset (video+audio) is manipulated: only the video, only the audio, both video and audio. Actually, the first group is further divided in two subgroups depending on whether the non-manipulated audio is consistent with the claimed identity or not (for example, because only the face was swapped). In the first case, the audio similarity index will not help, in the second case it could be important. These four groups are identified by the checkmarks in the first three columns of the table, indicating video manipulation (v), audio manipulation (a), audio inconsistency (ai).

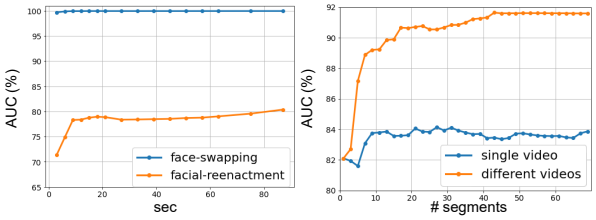

Figure 5. Results in terms of AUC by varying the length of the video under test (left) and by varying the number of video segments of the reference set (right).

Let us focus, for the time being, on the $\lambda = 0$ case and the AUC measure (upper-left part of the table). Several important observations are already possible. First, using the audio similarity index ensures a large performance gain with respect to using video-only features, from 15% to 20% on the average. This is a huge improvement, and hence a very significant result. We believe this is likely due to the limited emphasis that the audio component has received to date with respect to the video component. As expected, the audio index becomes useless in case 1, but the video index makes up for this loss. Notably, in this case, including the audio index in the decision statistic is not immaterial, as it reduces the performance in both the audio-video and fusion cases. On the average, however, the fusion of all three indices grants a significant gain with respect to all other cases. The best performance is observed in case 4, of course, when all components are manipulated. Moving right to the $\lambda = 1$ case, where the joint contrastive loss term is activated, we observe a limited but consistent performance improvement, all other considerations remaining the same. Moving down, all numbers drop significantly, as the Pd@10% is a much less forgiving performance measure, but general behaviors are basically confirmed. Notable differences concern the poor performance obtained when only the video index is used, and the great improvement achieved by including the contrastive loss term, especially with the fusion strategy.

Influence of test video length. We aim at studying the role of the test video length, analyzing how the performance varies as a function of the number of available segments. Results are shown in the left graph of Fig.5, separately for face-swapping (FS) and facial reenactment (FR) manipulations. Interestingly, for face-swapping, even a sin-

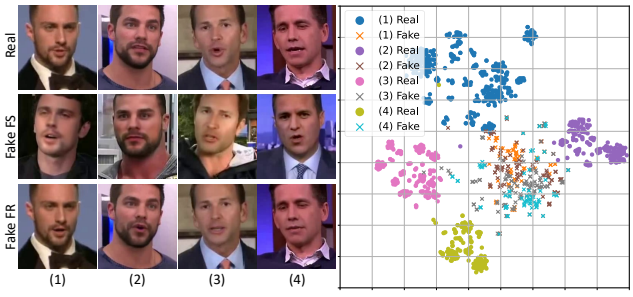

Figure 6. The t-SNE visualizations of features extracted from real and fake videos of four individuals. Each dot represents the feature relative to a video segment.

gle 3-second snippet is sufficient to reach perfect detection. Detecting facial reenactment is more challenging, as well known, with an AUC that starts at 70% with a single segment and grows slowly with the size of the video to reach an 80% plateau with about 20 seconds of video.

Influence of size and variety of the reference set. Results are shown in the right graph of Fig.5 and are averaged on all types of manipulations. It appears that a single POI video is not sufficient to achieve good performance, no matter how long the video is, with an AUC always at 84% or below. Instead, using multiple videos, the AUC improves rapidly beyond 90% to reach 92% when at least 40 segments are available. In summary, variety seems very important, more than sheer data size. In Fig.5, we observe that results on FR manipulations are worse than on FS ones, since the former better preserves the characteristics of the identity, and so they are more challenging for our approach.

t-SNE representation. In Fig.6, we show a scatter plot of the 2D projections, obtained by means of the t-SNE [15] dimensionality reduction technique, of some selected embedded vectors (concatenation of audio and video features). The vectors are extracted from real (circles) and fake (crosses) videos of the four identities shown on the left part of the figure. These were selected so as to present rather similar features in order not to induce systematic biases in the projections. Even so, embedded vectors relative to different individuals appear to from relatively compact clusters, well separated from one another. Moreover, with few exceptions, their manipulated counterparts turn out to be quite distant from the corresponding real videos.

4.3. Comparative Performance Analysis

State-of-the-art approaches. We consider approaches for which publicly available implementations are available. The methods used for our comparison are ensemble frame-based methods methods: Seferbekov [46], Real Forensics [21], temporal-based methods: FTCN (Fully Temporal Convolution Network) [56], LipForensics [22], audio-visual methods: MDS-based FD [7], Joint AV [57], and identity-based ones: ICT (Identity Consistency Transformer) [18], ID-Reveal [10]. A detailed description of these approaches can be found in the supplemental document.

Training. In order to ensure a fair comparison, Seferbekov, FTCN, LipForensics and Real Forensics are all trained on the FaceForensics++ dataset [44], which includes different types of facial manipulation (both FS and FR). MDS-based FD and Joint AV are trained on DFDC [16], which includes also audio manipulations. ICT is trained on a publicly available data-set of pristine faces of 1 million identities, MS-Celeb-1M [20]. ID-Reveal, instead, is trained on VoxCeleb2 [8] like our own method. In all cases, we ensure that training and test data do not overlap. Note that the reference set for ICT-Ref, ID-Reveal and the proposal are built from the same set of pristine videos.

Generalization analysis. Results are reported in Table 2 in terms of AUC and accuracy. The top part of the table refers to high-quality (HQ) data, with videos compressed using H.264 coding with factor 23 and original audio tracks. The bottom part is for low-quality (LQ) data, with videos compressed using H.264 coding with factor 40 and 25 fps, and audio tracks compressed using Advanced Audio Coding (AAC) with a sample-rate of 16K and a bit-rate of 48K. As a preliminary observation, note that all methods have much better figures in terms of AUC than Accuracy, confirming the difficulty in setting a global threshold and the need to properly calibrate each technique to make it work in realistic conditions.

On HQ videos, the proposed method outperforms by 2.3% (AUC) and 1.5% (Accuracy) on the average the best reference (Real Forensics and Seferbekov respectively) and more markedly the other methods. The supervised audio-visual approach obtains very good performance on pDFDC since it was trained on the DFDC dataset, but poorly generalize to other datasets. Most competitive methods are Seferbekov and Real Forensics. On LQ videos, however, the lead of POI-Forensics over all reference methods becomes very large, with an average gain of 11% (AUC) and 12% (Accuracy) over the second best. Indeed, in this challenging scenario, only identity-based approaches keep providing a good performance.

Robustness to adversarial attacks. Adding adversarial noise to an image can easily fool a classifier, and this has been shown to apply also to deepfake detectors [24, 41]. In [33] the fast gradient sign method [19] has been used to simulate malicious attacks. An EfficientNet-B4 network [48] with two fully connected layers has been trained as a deepfake detector and then attacked by means of adversarial examples. Our method and the selected reference ones were all tested on this dataset, obtaining the results reported in

	AUC/ACC	pDFDC	DF-TIMIT	FakeAVCel.	KoDF	AVG
High quality	Seferbekov	85.5 / 72.0	95.7 / 87.6	98.6 / 95.0	88.4 / 79.2	92.0 / 83.5
	FTCN	72.3 / 63.9	100.0 / 87.4	84.0 / 64.9	76.5 / 63.0	83.2 / 69.8
	LipForensics	68.7 / 60.0	98.8 / 78.0	97.6 / 83.3	92.9 / 56.1	89.5 / 69.3
	Real Forensics	85.2 / 70.4	100.0 / 99.5	88.3 / 76.2	95.6 / 63.1	92.3 / 77.3
	MDS-based FD	98.0 / 93.4	56.6 / 55.8	64.7 / 61.7	65.3 / 63.1	71.3 / 68.5
	Joint AV	53.2 / 52.8	50.0 / 50.0	55.1 / 48.6	49.4 / 49.2	51.9 / 50.1
	ICT	77.1 / 70.7	87.8 / 77.1	68.2 / 63.9	62.5 / 58.9	73.9 / 67.7
	ICT-Ref	87.6 / 79.8	100.0 / 94.7	71.9 / 64.5	79.4 / 60.3	84.7 / 74.8
	ID-Reveal	91.3 / 80.4	99.0 / 92.8	70.2 / 60.3	87.6 / 63.7	87.0 / 74.3
	POI-Forensics (ours)	95.2 / 86.7	99.2 / 85.7	94.1 / 86.6	89.9 / 81.1	**94.6/ 85.0**
Low quality	Seferbekov	62.6 / 54.0	81.8 / 71.4	61.7 / 58.5	79.6 / 55.9	71.4 / 59.9
	FTCN	51.3 / 50.0	78.3 / 58.9	37.6 / 42.1	68.4 / 58.6	58.9 / 52.4
	LipForensics	46.5 / 45.8	70.9 / 62.1	58.3 / 53.8	85.7 / 52.3	65.3 / 53.5
	Real Forensics	55.8 / 57.1	81.1 / 73.6	52.9 / 49.2	88.0 / 56.3	69.4 / 59.0
	MDS-based FD	95.6 / 90.0	52.6 / 52.2	61.1 / 58.6	64.6 / 62.4	68.7 / 65.8
	Joint AV	53.4 / 52.5	51.2 / 50.2	55.2 / 48.4	50.4 / 49.6	52.6 / 50.2
	ICT	72.0 / 66.4	84.4 / 74.1	66.9 / 61.5	61.1 / 59.3	71.1 / 65.3
	ICT-Ref	82.4 / 73.8	100.0 / 96.5	71.2 / 59.6	78.1 / 62.3	82.9 / 73.1
	ID-Reveal	86.5 / 73.9	93.6 / 75.5	70.8 / 58.9	85.0 / 63.8	84.0 / 68.0
	POI-Forensics (ours)	93.9 / 82.0	98.2 / 76.3	94.4 / 88.7	89.0 / 81.5	**93.9/ 82.1**

Table 2. Comparison with state-of-the-art on pDFDC, DeepfakeTIMIT, FakeAVCelebV2, and KoDF. Results are shown on high-quality (HQ) videos and low-quality (LQ) ones.

	Sef.	FTCN	Lip For.	Real For.	MDS-FD	Joint AV	ICT	ICT-ref	ID-Ref.	POI-For.
AUC	61.5	58.3	54.8	55.5	55.4	47.1	61.5	78.0	73.4	**80.5**
Pd@10%	23.7	10.3	16.2	3.7	13.1	10.7	19.1	48.2	27.6	**49.6**
ACC	56.7	52.6	48.0	47.1	53.4	49.8	53.7	65.3	67.1	**73.0**

Table 3. Results on the attacked KoDF subset.

Table 3. As expected, all performance metrics reduce dramatically for all methods. Only our method, and to a lesser extent ICT-ref and ID-Reveal, keep providing a reasonable performance in this scenario. Note that for the sake of fairness, the proposed method does not use the audio information in this experiment, as it is not subject to the attack.

5. Conclusions

We introduced POI-Forensics, an audio-visual deepfake detection method that is trained on real videos to learn a POI representation useful to reveal if the identity of the video under test is the person that is claimed to be. Our experiments show that by leveraging audio-visual information, we can handle a wide variety of manipulations. Most important, since our detector learns a real-world data model, it does not depend on patterns introduced by specific deepfake generators. In addition, the multimodal analysis could be further enriched by including more information from other domains such as textual data.

Acknowledgment. We gratefully acknowledge the support of this research by a TUM-IAS Hans Fischer Senior Fellowship, a TUM-IAS Rudolf Mößbauer Fellowship and a Google Faculty Research Award. This material is also based on research sponsored by the Defense Advanced Research Projects Agency (DARPA) and the Air Force Research Laboratory (AFRL) under agreement number FA8750-20-2-1004. 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 or the U.S. Government. In addition, this work has received funding by the European Union under the Horizon Europe vera.ai project, Grant Agreement number 101070093, and the ERC Starting Grant Scan2CAD (804724). It is also supported by the PREMIER project, funded by the Italian Ministry of Education, University, and Research within the PRIN 2017 program.

References

[1] S. Agarwal, H. Farid, T. El-Gaaly, and S. Lim. Detecting deep-fake videos from appearance and behavior. In *IEEE international workshop on information forensics and security (WIFS)*, 2020. 2, 3

[2] S. Agarwal, H. Farid, O. Fried, and M. Agrawala. Detecting deep-fake videos from phoneme-viseme mismatches. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR) Workshops*, 2020. 2, 3

[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 (CVPR) Workshops*, 2019. 2, 3

[4] S. Agarwal, L. Hu, E. Ng, T. Darrell, H. Li, and A. Rohrbach. Watch those words: Video falsification detection using word-conditioned facial motion. In *IEEE Winter conference on Applications of Computer Vision (WACV)*, 2023. 2, 3

[5] M. Boháček and H. Farid. Protecting world leaders against deep fakes using facial, gestural, and vocal mannerisms. In *Proceedings of the National Academy of Sciences*, 2022. 2, 3

[6] N. Bonettini, E.D. Cannas, S. Mandelli, L. Bondi, P. Bestagini, and S. Tubaro. Video Face Manipulation Detection Through Ensemble of CNNs. In *IEEE International Conference on Pattern Recognition (ICPR)*, 2020. 3

[7] K. Chugh, P. Gupta, A. Dhall, and R. Subramanian. Not made for each other- audio-visual dissonance-based deepfake detection and localization. In *ACM International Conference on Multimedia*, 2020. 3, 7

[8] J.S. Chung, A. Nagrani, and A. Zisserman. VoxCeleb2: Deep speaker recognition. In *Interspeech*, 2018. 4, 7

[9] D. Cozzolino, G. Poggi, and L. Verdoliva. Extracting camera-based fingerprints for video forensics. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR) Workshops*, 2019. 2, 3

[10] D. Cozzolino, A. Rössler, J. Thies, M. Nießner, and L. Verdoliva. ID-Reveal: Identity-aware DeepFake Video Detection. In *IEEE International Conference on Computer Vision (ICCV)*, 2021. 2, 3, 4, 7

[11] D. Cozzolino, J. Thies, A. Rössler, M. Nießner, and L. Verdoliva. SpoC: Spoofing camera fingerprints. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR) Workshops*, 2021. 2

[12] D. Cozzolino, J. Thies, A. Rössler, C. Riess, M. Nießner, and L. Verdoliva. ForensicTransfer: Weakly-supervised domain adaptation for forgery detection. *arXiv preprint arXiv:1812.02510*, 2018. 3

[13] H. Dang, F. Liu, J. Stehouwer, X. Liu, and A.K. Jain. On the detection of digital face manipulation. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2020. 3

[14] S. Das, S. Seferbekov, A. Datta, Md. S. Islam, and Md. R. Amin. Towards solving the deepfake problem: An analysis on improving deepfake detection using dynamic face augmentation. In *IEEE International Conference on Computer Vision (ICCV) Workshops*, 2021. 3

[15] L. Van der Maaten and G. Hinton. Visualizing data using t-SNE. *Journal of machine learning research*, 9(11), 2008. 7

[16] B. Dolhansky, J. Bitton, B. Pflaum, J. Lu, R. Howes, M. Wang, and C. Canton Ferrer. The deepfake detection challenge dataset. *arXiv preprint arXiv:2006.07397*, 2020. 7

[17] B. Dolhansky, R. Howes, B. Pflaum, N. Baram, and C. Canton Ferrer. The deepfake detection challenge (DFDC) preview dataset. *arXiv preprint arXiv:1910.08854*, 2019. 6

[18] 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 *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2022. 2, 3, 7

[19] I. Goodfellow, J. Shlens, and C. Szegedy. Explaining and harnessing adversarial examples. In *International Conference on Learning Representations (ICLR)*, 2015. 7

[20] Y. Guo, L. Zhang, Y. Hu, X. He, and J. Gao. MS-Celeb-1M: A Dataset and Benchmark for Large-Scale Face Recognition. In *European Conference on Computer Vision (ECCV)*, 2016. 7

[21] A. Haliassos, R. Mira, S. Petridis, and M. Pantic. Leveraging real talking faces via self-supervision for robust forgery detection. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2022. 3, 7

[22] A. Haliassos, K. Vougioukas, S. Petridis, and M. Pantic. Lips don't lie: A generalisable and robust approach to face forgery detection. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2021. 3, 7

[23] K. Hasam, T. Shahroz, K. Minha, and S.S. Woo. FakeAVCeleb: A novel audio-video multimodal deepfake dataset. In *Thirty-fifth Conference on Neural Information Processing Systems Datasets and Benchmarks Track (Round 2)*, 2021. 5, 6

[24] S. Hussain, P. Neekhara, M. Jere, F. Koushanfar, and J. McAuley. Adversarial deepfakes: Evaluating vulnerability of deepfake detectors to adversarial examples. In *IEEE Winter conference on Applications of Computer Vision (WACV)*, 2021. 2, 3, 7

[25] H. Jeon, Y. Bang, J. Kim, and S. Woo. T-GD: Transferable GAN-generated Images Detection Framework. In *International Conference on Machine Learning (ICML)*, 2020. 3

[26] Y. Jia, Y. Zhang, R. Weiss, Q. Wang, J. Shen, F. Ren, P. Nguyen, R. Pang, I. Lopez Moreno, Y. Wu, et al. Transfer learning from speaker verification to multispeaker text-to-speech synthesis. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2018. 6

[27] S.A. Khan and H. Dai. Video transformer for deepfake detection with incremental learning. In *ACM International Conference on Multimedia*, 2021. 3

[28] P. Khosla, P. Teterwak, C. Wang, A. Sarna, Y. Tian, P. Isola, A. Maschinot, C. Liu, and D. Krishnan. Supervised contrastive learning. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2020. 4

[29] P. Korshunov, M. Halstead, D. Castan, M. Graciarena, M. McLaren, B. Burns, A. Lawson, and S. Marcel. Tampered speaker inconsistency detection with phonetically aware

audio-visual features. In *International Conference on Machine Learning (ICML) Workshops*, 2019. 3

[30] P. Korshunov and S. Marcel. Deepfakes: a new threat to face recognition? assessment and detection. *arXiv preprint arXiv:1812.08685*, 2018. 6

[31] P. Korshunov and S. Marcel. Speaker inconsistency detection in tampered video. In *European Signal Processing Conference (EUSIPCO)*, 2018. 3

[32] I. Korshunova, W. Shi, J. Dambre, and L. Theis. Fast face-swap using convolutional neural networks. In *IEEE International Conference on Computer Vision (ICCV)*, 2017. 1, 2

[33] P. Kwon, J. You, G. Nam, S. Park, and G. Chae. KoDF: A large-scale korean deepfake detection dataset. In *IEEE International Conference on Computer Vision (ICCV)*, 2021. 6, 7

[34] J. Li, H. Xie, J. Li, Z. Wang, and Y. Zhang. Frequency-aware discriminative feature learning supervised by single-center loss for face forgery detection. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2021. 3

[35] L. Li, J. Bao, T. Zhang, H. Yang, D. Chen, F. Wen, and B. Guo. Face X-ray for more general face forgery detection. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2020. 3

[36] H. Liu, X. Li, W. Zhou, Y. Chen, Y. He, H. Xue, W. Zhang, and N. Yu. Spatial-phase shallow learning: Rethinking face forgery detection in frequency domain. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2021. 3

[37] I. Loshchilov and F. Hutter. Decoupled weight decay regularization. *arXiv preprint arXiv:1711.05101*, 2017. 5

[38] Y. Luo, Y. Zhang, J. Yan, and W. Liu. Generalizing face forgery detection with high-frequency features. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2021. 3

[39] F. Marra, C. Saltori, G. Boato, and L. Verdoliva. Incremental learning for GAN-generated image detection. In *IEEE international workshop on information forensics and security (WIFS)*, 2019. 3

[40] T. Mittal, U. Bhattacharya, R. Chandra, A. Bera, and D. Manocha. Emotions don't lie: An audio-visual deepfake detection method using affective cues. In *ACM International Conference on Multimedia*, 2018. 3

[41] P. Neekhara, B. Dolhansky, J. Bitton, and Cristian Canton Ferrer. Adversarial threats to deepfake detection: A practical perspective. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2021. 2, 3, 7

[42] Y. Nirkin, Y. Keller, and T. Hassner. FSGAN: Subject agnostic face swapping and reenactment. In *IEEE International Conference on Computer Vision (ICCV)*, 2019. 1, 2

[43] K.R. Prajwal, R. Mukhopadhyay, V.P. Namboodiri, and C.V. Jawahar. A lip sync expert is all you need for speech to lip generation in the wild. In *ACM International Conference on Multimedia*, 2020. 1, 2, 3

[44] A. Rössler, D. Cozzolino, L. Verdoliva, C. Riess, J. Thies, and M. Nießner. Faceforensics++: Learning to detect manipulated facial images. In *IEEE International Conference on Computer Vision (ICCV)*, 2019. 7

[45] C. Sanderson and B.C. Lovell. Multi-region probabilistic histograms for robust and scalable identity inference. In *International Conference on Biometrics (ICB)*, 2009. 6

[46] S. Seferbekov. *DeepFake Detection (DFDC) Team Sefer.* https://github.com/selimsef/dfdc_deepfake_challenge. 7

[47] K. Sohn. Improved deep metric learning with multi-class n-pair loss objective. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2016. 4

[48] M. Tan and Q. Le. Efficientnet: Rethinking model scaling for convolutional neural networks. In *International Conference on Machine Learning (ICML)*, 2019. 7

[49] R. Tolosana, R. Vera-Rodriguez, J. Fierrez, A. Morales, and J. Ortega-Garcia. Deepfakes and beyond: A survey of face manipulation and fake detection. *Information Fusion*, pages 131–148, 2020. 2

[50] L. Verdoliva. Media forensics and deepfakes: an overview. *IEEE Journal of Selected Topics in Signal Processing*, 14(5):910–932, 2020. 2

[51] C. Wang and W. Deng. Representative forgery mining for fake face detection. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2021. 3

[52] J. Wang, Z. Wu, X. Ouyang, W.and Han, J. Chen, Y.-G. Jiang, and S.-N. Li. M2tr: Multi-modal multi-scale transformers for deepfake detection. In *International Conference on Multimedia Retrieval*, 2022. 3

[53] Y. Wu and K. He. Group normalization. In *European Conference on Computer Vision (ECCV)*, 2018. 4

[54] H. Zhao, W. Zhou, D. Chen, T. Wei, W. Zhang, and N. Yu. Multi-attentional deepfake detection. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2021. 3

[55] T. Zhao, X. Xu, M. Xu, H. Ding, Y. Xiong, and W. Xia. Learning self-consistency for deepfake detection. In *IEEE International Conference on Computer Vision (ICCV)*, 2021. 3

[56] Y. Zheng, J. Bao, D. Chen, M. Zeng, and F. Wen. Exploring temporal coherence for more general video face forgery detection. In *IEEE International Conference on Computer Vision (ICCV)*, 2021. 3, 7

[57] Y. Zhou and S.-N. Lim. Joint audio-visual deepfake detection. In *IEEE International Conference on Computer Vision (ICCV)*, 2021. 3, 7

[58] X. Zhu, H. Wang, H. Fei, Z. Lei, and S.Z. Li. Face forgery detection by 3d decomposition. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2021. 3