# ID-Reveal: Identity-aware DeepFake Video Detection

Davide Cozzolino[1] Andreas Rössler[2] Justus Thies[2,3] Matthias Nießner[2] Luisa Verdoliva[1]

[1]University Federico II of Naples [2]Technical University of Munich
[3]Max Planck Institute for Intelligent Systems, Tübingen

## Abstract

*A major challenge in DeepFake forgery detection is that state-of-the-art algorithms are mostly trained to detect a specific fake method. As a result, these approaches show poor generalization across different types of facial manipulations, e.g., from face swapping to facial reenactment. To this end, we introduce ID-Reveal, a new approach that learns temporal facial features, specific of how a person moves while talking, by means of metric learning coupled with an adversarial training strategy. The advantage is that we do not need any training data of fakes, but only train on real videos. Moreover, we utilize high-level semantic features, which enables robustness to widespread and disruptive forms of post-processing. We perform a thorough experimental analysis on several publicly available benchmarks. Compared to state of the art, our method improves generalization and is more robust to low-quality videos, that are usually spread over social networks. In particular, we obtain an average improvement of more than 15% in terms of accuracy for facial reenactment on high compressed videos.*

## 1. Introduction

Recent advancements in synthetic media generation allow us to automatically manipulate images and videos with a high level of realism. To counteract the misuse of these image synthesis and manipulation methods, the digital media forensics field got a lot of attention [41, 40]. For instance, during the past two years, there has been intense research on DeepFake detection, that has been strongly stimulated by the introduction of large datasets of videos with manipulated faces [35, 33, 15, 31, 25, 28, 19].

However, despite excellent detection performance, the major challenge is how to generalize to previously unseen methods. For instance, a detector trained on face swapping will drastically drop in performance when tested on a facial reenactment method. This unfortunately limits practicality as we see new types of forgeries appear almost on a daily basis. As a result, supervised detection, which requires extensive training data of a specific forgery method, cannot immediately detect a newly-seen forgery type.

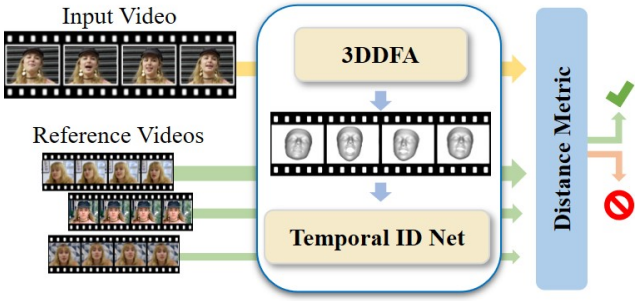


Figure 1: ID-Reveal is an identity-aware DeepFake video detection. Based on reference videos of a person, we estimate a temporal embedding which is used as a distance metric to detect fake videos.

This mismatch and generalization issue has been addressed in the literature using different strategies, ranging from applying domain adaptation [12, 5] or active learning [17] to strongly increasing augmentation during training [43, 15] or by means of ensemble procedures [15, 7]. A different line of research is relying only on pristine videos at training time and detecting possible anomalies with respect to forged ones [24, 11, 13]. This can help to increase the generalization ability with respect to new unknown manipulations but does not solve the problem of videos characterized by a different digital history. This is quite common whenever a video is spread over social networks and posted multiple times by different users. In fact, most of the platforms often reduce the quality and/or the video resolution.

Note also that current literature has mostly focused on face-swapping, a manipulation that replaces the facial identity of a subject with another one, however, a very effective modification is facial reenactment [39], where only the expression or the lips movements of a person are modified (Fig. 2). Recently, the MIT Center for Advanced Virtual-

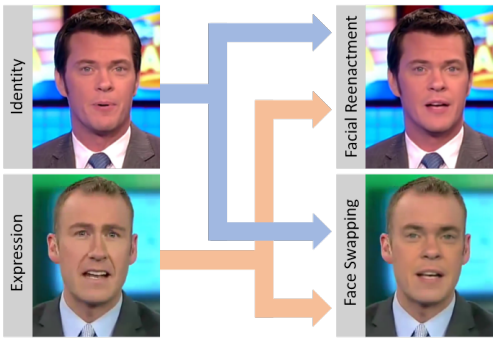


Figure 2: Automatic face manipulations can be split in two main categories: facial reenactment and face-swapping. The first one alters the facial expression preserving the identity. The second one modifies the identity of a person preserving the facial expression.

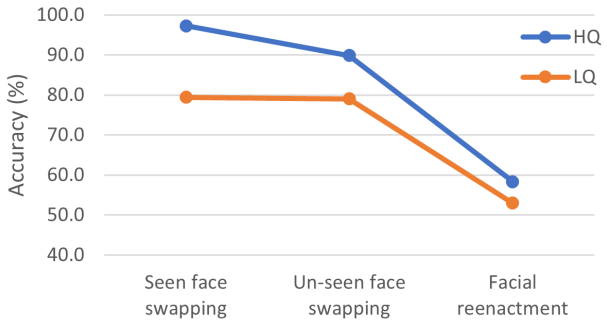


Figure 3: Accuracy results (binary classification task) of the winner of the Deepfake Detection Challenge [36] trained on DFDC dataset [15] and tested on different datasets: the preview DFDC [16] (seen face swapping) and FaceForensics++ [35] both on face swapping and facial-reenactment. Results are presented on high quality (HQ) and low quality (LQ) videos.

ity created a DeepFake video of president Richard Nixon [1]. The synthetic video shows Nixon giving a speech he never intended to deliver, by modifying only the lips movement and the speech of the old pristine video. The final result is impressive and shows the importance to develop forgery detection approaches that can generalize on different types of facial manipulations.

To better highlight this problem, we carried out an experiment considering the winning solution of the recent DeepFake Detection Challenge organized by Facebook on Kaggle platform. The performers had the possibility to train their models using a huge dataset of videos (around 100k fake videos and 20k pristine ones with hundreds of different identities). In Fig. 3, we show the results of our experiment. The model was first tested on a dataset of real and deepfake videos including similar face-swapping manipulations, then we considered unseen face-swapping manipulations and finally videos manipulated using facial reenactment. One can clearly observe the significant drop in performance in this last situation. Furthermore, the test on low quality compressed videos shows an additional loss and the final value for the accuracy is no more than a random guess.

It is also worth noting that current approaches are often used as black-box models and it is very difficult to predict the result because in a realistic scenario it is impossible to have a clue about the type of manipulation that occurred. The lack of reliability of current supervised deep learning methods pushed us to take a completely different perspective, avoiding to answer to a binary question (real or fake?) and instead focusing on wondering if the face under test preserves all the biometric traits of the involved subject.

Following this direction, our proposed method turns out to be able to generalize to different manipulation methods and also shows robustness w.r.t. low-quality data. It can reveal the identity of a subject by highlighting inconsistencies of facial features such as temporal consistent motion. The underlying CNN architecture comprises three main components: a facial feature extractor, a temporal network to detect biometric anomalies (temporal ID network) and a generative adversarial network that tries to predict person-specific motion based on the expressions of a different subject. The networks are trained only on real videos containing many different subjects [9]. During test time, in addition to the test video, we assume to have a set of pristine videos of the target person. Based on these pristine examples, we compute a distance metric to the test video using the embedding of the temporal ID network (Fig. 1). Overall, our main contributions are the following:

- We propose an example-based forgery detection approach that detects videos of facial manipulations based on the identity of the subject, especially the person-specific face motion.
- An extensive evaluation that demonstrates the generalization to different types of manipulations even on low-quality videos, with a significant average improvement of more than 15% w.r.t. state of the art.

## 2. Related Work

Digital media forensics, especially, in the context of DeepFakes, is a very active research field. The majority of the approaches rely on the availability of large-scale datasets of both pristine and fake videos for supervised learning. A few approaches detect manipulations as anomalies w.r.t. features learned only on pristine videos. Some of these approaches verify if the behavior of a person in a video is consistent with a given set of example videos of this person. Our approach ID-Reveal is such an example-based forgery detection approach. In the following, we discuss the most related detection approaches.

[1] https://moondisaster.org

**Learned features** Afchar et al. [1] presented one of the first approaches for DeepFake video detection based on supervised learning. It focuses on mesoscopic features to analyze the video frames by using a network with a low number of layers. Rössler et al. [35] investigated the performance of several CNNs architectures for DeepFake video detection and showed that very deep networks are more effective for this task, especially, on low-quality videos. To train the networks, the authors also published a large-scale dataset. The best performing architecture XceptionNet [8] was applied frame-by-frame and has been further improved by follow-up works. In [14] an attention mechanism is included, that can also be used to localize the manipulated regions, while in Kumar et al. [27] a triplet loss has been applied to improve performance on highly compressed videos.

Orthogonally, by exploiting artifacts that arise along the temporal direction it is possible to further boost performance. To this end, Guera et al. [20] propose using a convolutional Long Short Term Memory (LSTM) network. Masi et al. [32] propose to extract features by means of a two-branch network that are then fed into the LSTM: one branch takes the original information, while the other one works on the residual image. Differently in [46] a 3D CNN structure is proposed together with an attention mechanism at different abstraction levels of the feature maps.

Most of these methods achieve very good performance when the training set comprises the same type of facial manipulations, but performance dramatically impairs on unseen tampering methods. Indeed, generalization represents the Achilles' heel in media forensics. Augmentation can be of benefit to generalize to different manipulations as shown in [43]. In particular, augmentation has been extensively used by the best performing approaches during the DeepFake detection challenge [15]. Beyond the classic augmentation operations, some of them were particularly useful, e.g., by including cut-off based strategies on some specific parts of the face. In addition to augmentation, ensembling different CNNs have been also used to improve performance during this challenge [7, 17]. Another possible way to face generalization is to learn only on pristine videos and interpret a manipulation as an anomaly. This can improve the detection results on various types of face manipulations, even if the network never saw such forgeries in training. In [11] the authors extract the camera fingerprint information gathered from multiple frames and use those for detection. Other approaches focus on specific operations used in current DeepFake techniques. For example, in [28] the aim is to detect the blending operation that characterizes the face boundaries for most current synthetic face generation approaches.

A different perspective to improve generalization is presented in [12, 5], where few-shot learning strategies are applied. Thus, these methods rely on the knowledge of a few labeled examples of a new approach and guide the training process such that new embeddings can be properly separated from previous seen manipulation methods and pristine samples in a short retraining process.

**Features based on physiological signals** Other approaches look at specific artifacts of the generated videos that are related to physiological signals. In [29] it is proposed a method that detects eye blinking, which is characterized by a specific frequency and duration in videos of real people. Similarly, one can also use inconsistencies on head pose [45] or face warping artifacts [30] as identifiers for tampered content. Recent works are also using heart beat [18, 34] and other biological signals [10] to find inconsistencies both in spatial and along the temporal direction.

**Identity-based features** The idea of identity-based approaches is to characterize each individual by extracting some specific biometric traits that can be hardly reproduced by a generator [4, 3, 2]. The work by Agarwal et al. [4] is the first approach that exploits the distinct patterns of facial and head movements of an individual to detect fake videos. In [3] inconsistencies between the mouth shape dynamics and a spoken phoneme are exploited. Another related work is proposed in [2] to detect face-swap manipulations. The technique uses both static biometric based on facial identity and temporal ones based on facial expressions and head movements. The method includes standard techniques from face recognition and a learned behavioral embedding using a CNN powered by a metric-learning objective function. In contrast, our proposed method extracts facial features based on a 3D morphable model and focuses on temporal behavior through an adversarial learning strategy. This helps to improve the detection of facial reenactment manipulations while still consistently be able to spot face swapping ones.

## 3. Proposed Method

ID-Reveal is an approach for DeepFake detection that uses prior biometric characteristics of a depicted identity, to detect facial manipulations in video content of the person. Any manipulated video content based on facial replacement results in a disconnect between visual identity as well as biometrical characteristics. While facial reenactment preserves the visual identity, biometrical characteristics such as the motion are still wrong. Using pristine video material of a target identity we can extract these biometrical features and compare them to the characteristics computed on a test video that is potentially manipulated. In order to be able to generalize to a variety of manipulation methods, we avoid training on a specific manipulation method, instead, we solely train on non-tampered videos. Additionally, this

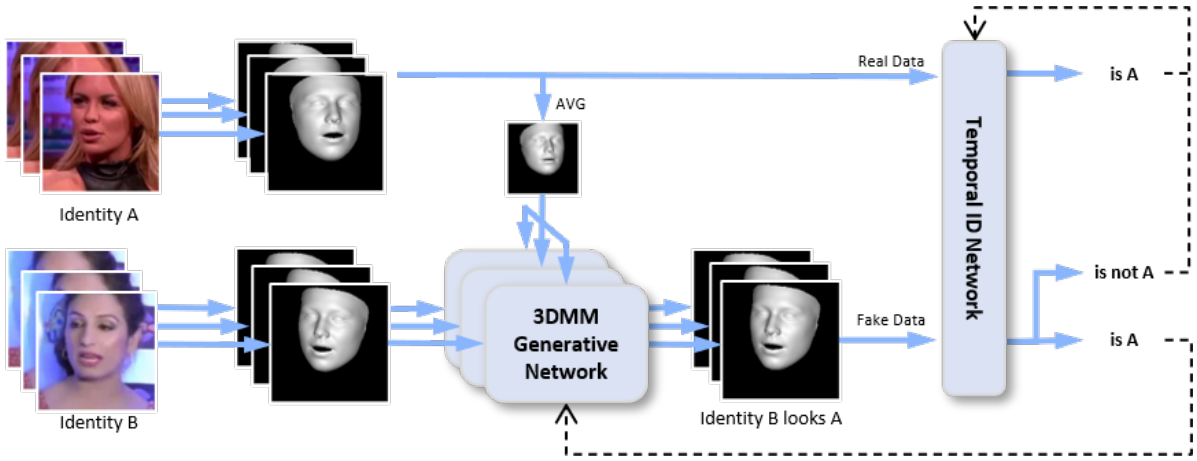


Figure 4: ID-Reveal is based on two neural networks, the Temporal ID Network as well as the 3DMM Generative Network, which interact with each other in an adversarial fashion. Using a three-dimensional morphable model (3DMM), we process videos of different identities and train the Temporal ID Network to embed the extracted features such that they can be separated in the resulting embedding space based on their containing identity. In order to incentivize this network to focus on temporal aspects rather than visual cues, we jointly train the 3DMM Generative Network to transform extracted features to fool its discriminative counterpart.

allows us to leverage a much larger training corpus in comparison to the facial manipulation datasets [35, 15].

Our proposed method consists of three major components (see Fig. 4). Given a video as input, we extract a compact representation of each frame using a 3D Morphable model (3DMM) [6]. These extracted features are input to the Temporal ID Network which computes an embedded vector. During test time, a metric in the embedding space is used to compare the test video to the previously recorded biometrics of a specific person. However, in order to ensure that the Temporal ID Network is also based on behavioral instead of only visual information, we utilize a second network, called 3DMM Generative Network, which is trained jointly in an adversarial fashion (using the Temporal ID Network as discriminator). In the following, we will detail the specific components and training procedure.

**Feature Extraction** Our employed networks are based on per-frame extracted facial features. Specifically, we utilize a low dimensional representation of a face based on a 3D morphable model [6]. The morphable model represents a 3D face by a linear combination of principle components for shape, expression, and appearance. These components are computed via a principle component analysis of aligned 3D scans of human faces. A new face can be represented by this morphable model, by providing the corresponding coefficients for shape, expression, and appearance. To retrieve these parameters from video frames, one can use optimization-based analysis-by-synthesis approaches [39] or learned regression. In our method, we rely on the regression framework of Guo et al. [21] which predicts a vector of 62 coefficients for each frame. Note that the 62 parameters, contain 40 coefficients for the shape, 10 for the expression, and additional 12 parameters for the rigid pose of the face (represented as a $3 \times 4$ matrix). In the following, we denote the extracted 3DMM features of video $i$ of the individual $c$ at frame $t$ by $x_{c,i}(t) \in \mathbb{R}^{62}$.

**Temporal ID Network** The Temporal ID Network $\mathcal{N}_T$ processes the temporal sequence of 3DMM features through convolution layers that work along the temporal direction in order to extract the embedded vector $y_{c,i}(t) = \mathcal{N}_T\left[x_{c,i}(t)\right]$. To evaluate the distance between embedded vectors, we adopt the squared Euclidean distance, computing the following similarity:

$$S_{c,i,k,j}(t) = -\frac{1}{\tau} \min_{t'} \left\| y_{c,i}(t) - y_{k,j}(t') \right\|^2 \qquad (1)$$

As a metric learning loss, similar to the Distance-Based Logistic Loss [42], we adopt a log-loss on a suitably defined probability [13]. Specifically, for each embedded vector $y_{c,i}(t)$, we build the probability through softmax processing as:

$$p_{c,i}(t) = \frac{\sum_{j \neq i} e^{S_{c,i,c,j}(t)}}{\sum_{j \neq i} e^{S_{c,i,c,j}(t)} + \sum_{k \neq c} \sum_{j} e^{S_{c,i,k,j}(t)}}, \qquad (2)$$

Thus, we are considering all the similarities with respect to the pivot vector $y_{c,i}(t)$ in our probability definition $p_{c,i}(t)$. Note that to obtain a high probability value it is only necessary that at least one similarity with the same individual is much larger than similarities with other individuals. Indeed, the loss proposed here is a less restrictive loss compared to the current literature, where the aim is to achieve a high similarity for all the coherent pairs [22, 23, 44]. The adopted

metric learning loss is then obtained from the probabilities through the log-loss function:

$$\mathcal{L}_{rec} = \sum_{c,i,t} -log\left(p_{c,i}(t)\right). \tag{3}$$

In order to tune hyper-parameters during training, we also measure the accuracy of correctly identifying a subject. It is computed by counting the number of times where at least one similarity with the same individual is larger than all the similarities with other individuals. The Temporal ID Network is first trained alone using the previously described loss, and afterward it is fine-tuned together with the 3DMM Generative Network, which we describe in the following paragraph.

**3DMM Generative Network** The 3DMM Generative Network $\mathcal{N}_G$ is trained to generate 3DMM features similar to the features that we may extract from a manipulated video. Specifically, the generative network has the goal to output features that are coherent to the identity of an individual, but with the expressions of another subject. The generative network $\mathcal{N}_G$ works frame-by-frame and generates a 3DMM feature vector by combining two input feature vectors. Let $x_c$ and $x_k$ are the 3DMM feature vectors respectively of the individuals $c$ and $k$, then, $\mathcal{N}_G\left[x_k, x_c\right]$ is the generated feature vector with appearance of the individual $c$ and expressions of individual $k$. During training, we use batches that contain $N \times M$ videos of $N$ different individuals each with $M$ videos. In our experiments, we chose $M = N = 8$. To train the generative network $\mathcal{N}_G$, we apply it to pairs of videos of these $N$ identities. Specifically, for each identity $c$, we compute an averaged 3DMM feature vector $\overline{x}_c$. Based on this averaged input feature $\overline{x}_c$ and a frame feature $x_i(t)$ of a video of person $i$ (which serves as expression conditioning), we generate synthetic 3DMM features using the generator $\mathcal{N}_G$:

$$x^*_{c,i}(t) = \mathcal{N}_G\left[x_i(t), \overline{x}_c\right]. \tag{4}$$

The 3DMM Generative Network is trained based on the following loss:

$$\mathcal{L}_{\mathcal{N}_G} = \mathcal{L}_{adv} + \lambda_{cycle}\,\mathcal{L}_{cycle} \tag{5}$$

Where $\mathcal{L}_{cycle}$ is a cycle consistency used in order to preserve the expression. Specifically, the 3DMM Generative Network is applied twice, firstly to transform a 3DMM feature vector of the individual $i$ to identity $c$ and then to transform the generated 3DMM feature vector to identity $i$ again, we should obtain the original 3DMM feature vector. The $\mathcal{L}_{cycle}$ is defined as:

$$\mathcal{L}_{cycle} = \sum_{c,i,t} \left\| x_i(t) - \mathcal{N}_G\left[x^*_{c,i}(t), \overline{x}_i\right] \right\|^2. \tag{6}$$

The adversarial loss $\mathcal{L}_{adv}$ is based on the Temporal ID Network, i.e., it tries to fool the Temporal ID Network by generating features that are coherent for a specific identity. Since the generator works frame-by-frame, it can deceive the Temporal ID Network by only altering the appearance of the individual and not the temporal patterns. The adversarial loss $\mathcal{L}_{adv}$ is computed as:

$$\mathcal{L}_{adv} = \sum_{c,i,t} -log\left(p^*_{c,i}(t)\right), \tag{7}$$

where the probabilities $p^*_{c,i}(t)$ are computed using the equation 2, but considering the similarities evaluated between generated features and real ones:

$$S^*_{c,i,k,j}(t) = -\frac{1}{\tau} \min_{t'} \left\| \mathcal{N}_T\left[x^*_{c,i}(t)\right] - y_{k,j}(t') \right\|^2. \tag{8}$$

Indeed, the generator aims to increase the similarity between the generated features for a given individual and the real features of that individual. During training, the Temporal ID Network is trained to hinder the generator, through a loss obtained as:

$$\mathcal{L}_{\mathcal{N}_T} = \mathcal{L}_{rec} + \lambda_{inv}\mathcal{L}_{inv}. \tag{9}$$

where the loss $\mathcal{L}_{inv}$, contrary to $\mathcal{L}_{abv}$, is used to minimized the probabilities $p^*_{c,i}(t)$. Therefore, it is defined as:

$$\mathcal{L}_{inv} = \sum_{c,i,t} -log\left(1 - p^*_{c,i}(t)\right). \tag{10}$$

Overall, the final objective of the adversarial game is to increase the ability of the Temporal ID Network to distinguish real identities from fake ones.

**Identification** Given a test sequence depicting a single identity as well as a reference set of pristine sequences of the same person, we apply the following procedure: we first embed both the test as well as our reference videos using the Temporal ID Network pipeline. We then compute the minimum pairwise Euclidean distance of each reference video and our test sequence. Finally, we compare this distance to a fixed threshold $\tau_{id}$ to decide whether the behavioral properties of our testing sequence coincide with its identity, thus, evaluating the authenticity of our test video. The source code and the trained network of our proposal are publicly available[2].

## 4. Results

To analyze the performance of our proposed method, we conducted a series of experiments. Specifically, we discuss our design choices w.r.t. our employed loss functions and

[2] https://github.com/grip-unina/id-reveal


Figure 5: Aligned example images of DFD FS (Face-Swapping) as well as the newly created DFD FR (Facial Reenactment) datasets. From left to right: source videos, target sequences, DeepFakes and manipulations created using Neural Textures [38].

the adversarial training strategy based on an ablation study applied on a set of different manipulation types and different video qualities. In comparison to state-of-the-art DeepFake video detection methods, we show that our approach surpasses these in terms of generalizability and robustness.

### 4.1. Experimental Setup

Our approach is trained using the VoxCeleb2 development dataset [9] consisting of multiple video clips of several identities. Specifically, we use 5120 subjects for the training-set and 512 subjects for the validation-set. During training, each batch contains 64 sequences of 96 frames. The 64 sequences are formed by $M = 8$ sequences for each individual, with a total of $N = 8$ different individuals extracted at random from the training-set. Training is performed using the ADAM optimizer [26], with a learning rate of $10^{-4}$ and $10^{-5}$ for the Temporal ID Network and the 3DMM Generative Network, respectively. The parameters $\lambda_{cycle}$, $\lambda_{inv}$ and $\tau$ for our loss formulation are set to 1.0, 0.001 and 0.08 respectively. We first train the Temporal ID Network for 300 epochs (with an epoch size of 2500 iterations) and choose the best performing model based on the validation accuracy. Using this trained network, we enable our 3DMM Generative Network and continue training for a fixed 100 epochs. For details on our architectures, we refer to the supplemental document. For all experiments, we use a fixed threshold of $\tau_{id} = \sqrt{1.1}$ to determine whether the behavioral properties of a test video coincide with those of our reference videos. This threshold is set experimentally based on a one-time evaluation on 4 real and 4 fake videos from the original DFD [33] using the averaged squared euclidean distance of real and manipulated videos.

| | Acc(%) / AUC | MSL | Triplet | ours |
|---|---|---|---|---|
| w.o. adversarial | DFD FR HQ | **73.8** / 0.83 | 73.6 / 0.85 | **73.8** / **0.86** |
| | LQ | 66.6 / 0.77 | 73.3 / 0.81 | **79.1** / **0.87** |
| | DFD FS HQ | **87.8** / 0.94 | 83.5 / 0.95 | 82.6 / **0.96** |
| | LQ | 74.0 / 0.92 | **77.0** / 0.93 | 73.0 / **0.94** |
| w. adversarial | DFD FR HQ | 68.9 / 0.81 | 71.6 / 0.85 | **75.6** / **0.89** |
| | LQ | 69.1 / 0.77 | 73.9 / 0.81 | **81.8** / **0.90** |
| | DFD FS HQ | **87.8** / 0.94 | 84.7 / 0.94 | 84.8 / **0.96** |
| | LQ | 78.0 / 0.92 | **78.9** / 0.91 | 78.1 / **0.94** |

Table 1: Accuracy and AUC for variants of our approach. We compare three different losses: the multi-similarity loss (MSL), the triplet loss and our proposed loss (eq.3). In addition, we present the results obtained with and without the adversarial learning strategy on high quality (HQ) and low quality (LQ) videos manipulated using Facial Reenactment (FR) and Face Swapping (FS).

### 4.2. Ablation Study

In this section, we show the efficacy of the proposed loss and the adversarial training strategy. For performance evaluation of our approach, we need to know the involved identity (the source identity for face-swapping manipulations and the target identity for facial reenactment ones). Based on this knowledge, we can set up the pristine reference videos used to compute the final distance metric. To this end, we chose a controlled dataset that includes several videos of the same identity, i.e., the recently created dataset of the Google AI lab, called DeepFake Dataset (DFD) [33]. The videos contain 28 paid actors in 16 different contexts, furthermore, for each subject there are pristine videos provided (varying from 9 to 16). In total, there are 363 real and 3068 DeepFakes videos. Since the dataset only contains face-swapping manipulations, we generated 320 additional videos that include 160 Face2Face [39] and 160 Neural Textures [38] videos. Some examples are shown in Fig. 5.

Performance is evaluated at video level using a leave-one-out strategy for the reference-dataset. In detail, for each video under test, the reference dataset only contains pristine videos with a different context from the one under test. The evaluation is done both on high quality (HQ) compressed videos (constant rate quantization parameter equal to 23) using H.264 and low quality (LQ) compressed videos (quantization parameter equal to 40). This scenario helps us to consider a realistic situation, where videos are uploaded to the web, but also to simulate an attacker who further compresses the video to hide manipulation traces.

We compare the proposed loss to the triplet loss [23] and the multi-similarity loss (MSL) [44]. For these two losses, we adopt the cosine distance instead of the Euclidean one as proposed by the authors. Moreover, hyper-parameters are chosen to maximize the accuracy to correctly identify a

subject in the validation set. Results for facial reenactment (FR) and face swapping (FS) in terms of Area Under Curve (AUC) and accuracy both for HQ and LQ videos are shown in Tab. 1. One can observe that our proposed loss gives a consistent improvement over the multi-similarity loss (5.5% on average) and the triplet loss (2.8% on average) in terms of AUC. In addition, coupled with the adversarial training strategy performance, it gets better for the most challenging scenario of FR videos with a further improvement of around 3% for AUC and of 6% (on average) in terms of accuracy.

### 4.3. Comparisons to State of the Art

We compare our approach to several state-of-the-art DeepFake video detection methods. All the techniques are compared using the accuracy at video level. Hence, if a method works frame-by-frame, we average the probabilities obtained from 32 frames uniformly extracted from the video, as it is also done in [7, 1].

**State of the art approaches** The methods used for our comparison are frame-based methods: MesoNet [1], Xception [8], FFD (Facial Forgery Detection) [14], Efficient-B7 [37]; ensemble methods: ISPL (Image and Sound Processing Lab) [7], Seferbekov [36]; temporal-based methods: Eff.B1 + LSTM, ResNet + LSTM [20] and an Identity-based method: A&B (Appearance and Behavior) [2]. A detailed description of these approaches can be found in the supplemental document. In order to ensure a fair comparison, all supervised approaches (frame-based, ensemble and temporal-based methods) are trained on the same dataset of real and fake videos, while the identity-based ones (A&B and our proposal) are trained instead on VoxCeleb2 [9].

**Generalization and robustness analysis** To analyze the ability to generalize to different manipulation methods, training and test come from different datasets. Note that we will focus especially on generalizing from face swapping to facial reenactment.

In a first experiment we test all the methods on the DFD Google dataset that contains both face swapping and facial reenactment manipulations, as described in Section 4.2. In this case all supervised approaches are trained on DFDC [15] with around 100k fake and 20k real videos. This is the largest DeepFake dataset publicly available and includes five different types of manipulations[3]. Experiments on (HQ) videos, with a compression factor of 23, and on low-quality (LQ) videos, where the factor is 40 are presented in terms of accuracy and AUC in Tab. 2. Most methods suffer from a huge performance drop when going from face-swapping to facial-reenactment, with an accuracy that often borders 50%, equivalent to coin tossing. The likely reason is that the DFDC training set includes mostly face-swapping videos, and methods with insufficient generalization ability are unable to deal with different manipulations. This does not hold for ID-Reveal and A&B, which are trained only on real data and, hence, have an almost identical performance with both types of forgeries. For facial reenactment videos, this represents a huge improvement with respect to all competitors. In this situation it is possible to observe a sharp performance degradation of most methods in the presence of strong compression (LQ videos). This is especially apparent with face-swapping, where some methods are very reliable on HQ videos but become almost useless on LQ videos. On the contrary, ID-Reveal suffers only a very small loss of accuracy on LQ videos, and outperforms all competitors, including A&B, by a large margin.

| | High Quality (HQ) | | Low Quality (LQ) | |
|---|---|---|---|---|
| Acc(%) / AUC | DFD FR | DFD FS | DFD FR | DFD FS |
| MesoNet | 57.0 / 0.65 | 54.0 / 0.57 | 58.1 / 0.61 | 52.7 / 0.53 |
| Xception | 51.9 / 0.74 | 78.5 / 0.93 | 49.8 / 0.48 | 58.5 / 0.63 |
| Efficient-B7 | 53.1 / 0.75 | 88.2 / 0.97 | 50.2 / 0.48 | 58.5 / 0.64 |
| FFD | 53.6 / 0.57 | 75.3 / 0.83 | 51.3 / 0.53 | 63.9 / 0.69 |
| ISPL | 61.4 / 0.71 | 85.2 / 0.93 | 53.9 / 0.55 | 64.9 / 0.72 |
| Seferbekov | 55.8 / 0.77 | **91.8 / 0.98** | 49.4 / 0.47 | 61.9 / 0.67 |
| ResNet + LSTM | 52.2 / 0.56 | 60.0 / 0.65 | 56.1 / 0.62 | 58.7 / 0.64 |
| Eff.B1 + LSTM | 53.6 / 0.72 | 86.6 / 0.95 | 50.9 / 0.57 | 61.6 / 0.76 |
| A&B | 74.1 / 0.78 | 75.6 / 0.77 | 59.5 / 0.60 | 63.2 / 0.61 |
| ID-Reveal | **75.6 / 0.87** | 84.8 / 0.96 | **81.8 / 0.90** | **78.1 / 0.94** |

Table 2: Video-level detection accuracy and AUC of our approach compared to state-of-the-art methods. Results are obtained on the DFD dataset on HQ videos and LQ ones, split in facial reenactment (FR) and face swapping (FS) manipulations. Training for supervised methods is carried out on DFDC, while for identity-based methods on VoxCeleb2.

In another experiment, we use FaceForensics++ [35] (HQ) for training the supervised methods, while the identity based methods are always trained on the VoxCeleb2 dataset [9]. For testing, we use the preview DFDC Facebook dataset [16] and CelebDF [31]. The preview DFDC dataset [16] is composed only of face-swapping manipulations of 68 individuals. For each subject there are 3 to 39 pristine videos with 3 videos for each context. We consider 44 individuals which have at least 12 videos (4 contexts); obtaining a total of 920 real videos and 2925 fake videos. CelebDF [31] contains 890 real videos and 5639 face-swapping manipulated videos. The videos are related to 59 individuals except for 300 real videos that do not have any information about the individual, hence, they cannot be included in our analysis. Results in terms of accuracy and

[3] https://www.kaggle.com/c/deepfake-detection-challenge

| | High Quality (HQ) | | Low Quality (LQ) | |
|---|---|---|---|---|
| Acc(%) / AUC | DFDCp | CelebDF | DFDCp | CelebDF |
| MesoNet | 53.6 / 0.74 | 50.1 / 0.75 | 51.9 / 0.67 | 50.1 / 0.67 |
| Xception | 72.0 / 0.79 | **77.2 / 0.88** | 59.9 / 0.61 | 55.0 / 0.58 |
| Efficient-B7 | 71.8 / 0.78 | 71.4 / 0.80 | 57.3 / 0.62 | 51.3 / 0.56 |
| FFD | 63.1 / 0.69 | 69.2 / 0.76 | 51.6 / 0.55 | 56.4 / 0.59 |
| ISPL | 69.6 / 0.78 | 71.2 / 0.83 | 52.0 / 0.71 | 50.8 / 0.61 |
| Seferbekov | 72.0 / 0.85 | 75.3 / 0.86 | 54.0 / 0.63 | 54.8 / 0.62 |
| ResNet + LSTM | 61.2 / 0.67 | 58.2 / 0.72 | 56.3 / 0.59 | 57.0 / 0.60 |
| Eff.B1 + LSTM | 67.2 / 0.75 | 75.3 / 0.84 | 51.0 / 0.54 | 55.3 / 0.58 |
| A&B | 65.2 / 0.60 | 54.0 / 0.56 | 61.7 / 0.59 | 52.6 / 0.55 |
| ID-Reveal | **80.4 / 0.91** | 71.6 / 0.84 | **73.9 / 0.86** | **64.4 / 0.80** |

Table 3: Video-level detection accuracy and AUC of our approach compared to state-of-the-art methods. Results are obtained on DFDCp and CelebDF on HQ videos and LQ ones. Training for supervised methods is carried out on FF++, while for identity-based methods on VoxCeleb2.

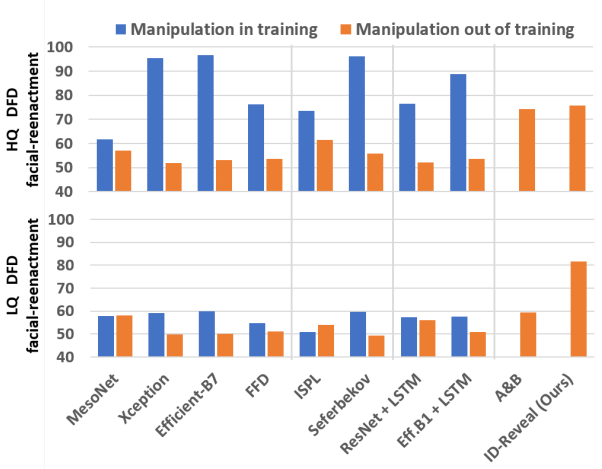


Figure 6: Binary detection accuracy of our approach compared to state-of-the-art methods. Results are obtained on the facial-reenactment DFD dataset both on HQ and LQ videos. We consider two different training scenarios for all the approaches that need forged videos in training: manipulation in training (blue bars), where the training set includes same type of manipulations present in test set (neural textures and face2face for facial reenactment), and manipulation out of training (orange bars), where we adopt DFDC that includes only face swapping.

AUC at video-level are shown in Tab. 3. One can observe that also in this scenario our method achieves very good results for all the datasets, with an average improvement with respect to the best supervised approach of about 16% on LQ videos. Even the improvement with respect to the identity based approach A&B [2] is significant, around 14% on HQ videos and 13% on LQ ones. Again the performance of supervised approaches worsens in unseen conditions of low-quality videos, while our method preserves its good performance.

To gain better insights on both generalization and robustness, we want to highlight the very different behavior of supervised methods when we change the fake videos in training. Specifically, for HQ videos if the manipulation (in this case neural textures and face2face) is included in training and test, then performance are very high for all the methods, but they suddenly decrease if we exclude those manipulations from the training, see Fig. 6. The situation is even worse for LQ videos. Identity-based methods do not modify their performance since they do not depend at all on which manipulation is included in training.

## 5. Conclusion

We have introduced ID-Reveal, an identity-aware detection approach leveraging a set of reference videos of a target person and trained in an adversarial fashion. A key aspect of our method is the usage of a low-dimensional 3DMM representation to analyze the motion of a person. While this compressed representation of faces contains less information than the original 2D images, the gained type of robustness is a very important feature that makes our method generalize across different forgery methods. Specifically, the 3DMM representation is not affected by different environments or lighting situations, and is robust to disruptive forms of post-processing, e.g., compression. We conducted a comprehensive analysis of our method and in comparison to state of the art, we are able to improve detection qualities by a significant margin, especially, on low-quality content. At the same time, our method improves generalization capabilities by adopting a training strategy that solely focuses on non-manipulated content.

## 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. In addition, this material is 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 and AFRL or the U.S. Government. This work is also supported by the PREMIER project, funded by the Italian Ministry of Education, University, and Research within the PRIN 2017 program.

## References


[1] Darius Afchar, Vincent Nozick, Junichi Yamagishi, and Isao Echizen. Mesonet: a compact facial video forgery detection network. In *IEEE International Workshop on Information Forensics and Security*, pages 1–7, 2018. 3, 7

[2] Shruti Agarwal, Hany Farid, Tarek El-Gaaly, and Ser-Nam Lim. Detecting deep-fake videos from appearance and behavior. In *IEEE International Workshop on Information Forensics and Security (WIFS)*, pages 1–6, 2020. 3, 7, 8

[3] Shruti Agarwal, Hany Farid, Ohad Fried, and Maneesh Agrawala. Detecting deep-fake videos from phoneme-viseme mismatches. In *IEEE CVPR Workshops*, 2020. 3

[4] Shruti Agarwal, Hany Farid, Yuming Gu, Mingming He, Koki Nagano, and Hao Li. Protecting world leaders against deep fakes. In *IEEE CVPR Workshops*, June 2019. 3

[5] Shivangi Aneja and Matthias Nießner. Generalized zero and few-shot transfer for facial forgery detection. *arXiv preprint arXiv:2006.11863*, 2020. 1, 3

[6] Volker Blanz and Thomas Vetter. A morphable model for the synthesis of 3D faces. In *ACM Transactions on Graphics (Proc. of SIGGRAPH)*, pages 187–194, 1999. 4

[7] Nicolò Bonettini, Edoardo Daniele Cannas, Sara Mandelli, Luca Bondi, Paolo Bestagini, and Stefano Tubaro. Video Face Manipulation Detection Through Ensemble of CNNs. In *IEEE International Conference on Pattern Recognition (ICPR)*, 2020. https://github.com/polimi-ispl/icpr2020dfdc. 1, 3, 7

[8] François Chollet. Xception: Deep learning with depthwise separable convolutions. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 1251–1258, 2017. 3, 7

[9] Joon Son Chung, Arsha Nagrani, and Andrew Zisserman. Voxceleb2: Deep speaker recognition. In *Interspeech*, 2018. 2, 6, 7

[10] Umur Aybars Ciftci, Ilke Demir, and Lijun Yin. Fakecatcher: Detection of synthetic portrait videos using biological signals. *IEEE Transactions on Pattern Analysis and Machine Intelligence, in press*, 2020. 3

[11] Davide Cozzolino, Giovanni Poggi, and Luisa Verdoliva. Extracting camera-based fingerprints for video forensics. In *IEEE CVPR Workshops*, pages 130–137, 2019. 1, 3

[12] Davide Cozzolino, Justus Thies, Andreas Rössler, Christian Riess, Matthias Nießner, and Luisa Verdoliva. ForensicTransfer: Weakly-supervised domain adaptation for forgery detection. *arXiv preprint arXiv:1812.02510*, 2018. 1, 3

[13] Davide Cozzolino and Luisa Verdoliva. Noiseprint: A CNN-Based Camera Model Fingerprint. *IEEE Transactions on Information Forensics and Security*, 15:144–159, 2020. 1, 4

[14] Hao Dang, Feng Liu, Joel Stehouwer, Xiaoming Liu, and Anil K Jain. On the detection of digital face manipulation. In *IEEE Conference on Computer Vision and Pattern Recognition*, pages 5781–5790, 2020. http://cvlab.cse.msu.edu/project-ffd.html. 3, 7

[15] Brian Dolhansky, Joanna Bitton, Ben Pflaum, Jikuo Lu, Russ Howes, Menglin Wang, and Cristian Canton Ferrer. The deepfake detection challenge dataset. *arXiv preprint arXiv:2006.07397*, 2020. 1, 2, 3, 4, 7

[16] Brian Dolhansky, Russ Howes, Ben Pflaum, Nicole Baram, and Cristian Canton Ferrer. The deepfake detection challenge (DFDC) preview dataset. *arXiv preprint arXiv:1910.08854*, 2019. 2, 7

[17] Mengnan Du, Shiva K. Pentyala, Yuening Li, and Xia Hu. Towards generalizable deepfake detection with locality-aware autoencoder. In *ACM International Conference on Information and Knowledge Management*, pages 325–334, 2020. 1, 3

[18] Steven Fernandes, Sunny Raj, Eddy Ortiz, Iustina Vintila, Margaret Salter, Gordana Urosevic, and Sumit Jha. Predicting Heart Rate Variations of Deepfake Videos using Neural ODE. In *ICCV Workshops*, 2019. 3

[19] Gereon Fox, Wentao Liu, Hyeongwoo Kim, Hans-Peter Seidel, Mohamed Elgharib, and Christian Theobalt. Videoforensicshq: Detecting high-quality manipulated face videos. In *IEEE International Conference on Multimedia and Expo (ICME)*, pages 1–6, 2021. 1

[20] David Güera and Edward J Delp. Deepfake video detection using recurrent neural networks. In *IEEE International Conference on Advanced Video and Signal Based Surveillance*, 2018. 3, 7

[21] Jianzhu Guo, Xiangyu Zhu, Yang Yang, Fan Yang, Zhen Lei, and Stan Z Li. Towards fast, accurate and stable 3d dense face alignment. In *European Conference on Computer Vision (ECCV)*, 2020. 4

[22] Raia Hadsell, Sumit Chopra, and Yann LeCun. Dimensionality reduction by learning an invariant mapping. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, volume 2, pages 1735–1742, 2006. 4

[23] Elad Hoffer and Nir Ailon. Deep metric learning using triplet network. In *International Workshop on Similarity-Based Pattern Recognition*, pages 84–92. Springer, 2015. 4, 6

[24] Minyoung Huh, Andrew Liu, Andrew Owens, and Alexei A Efros. Fighting fake news: Image splice detection via learned self-consistency. In *European Conference on Computer Vision (ECCV)*, pages 101–117, 2018. 1

[25] Liming Jiang, Ren Li, Wayne Wu, Chen Qian, and Chen Change Loy. Deeperforensics-1.0: A large-scale dataset for real-world face forgery detection. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 2886–2895, 2020. 1

[26] Diederik P. Kingma and Jimmy Ba. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*, 2014. 6

[27] Akash Kumar, Arnav Bhavsar, and Rajesh Verma. Detecting deepfakes with metric learning. In *International Workshop on Biometrics and Forensics (IWBF)*, pages 1–6, 2020. 3

[28] Lingzhi Li, Jianmin Bao, Ting Zhang, Hao Yang, Dong Chen, Fang Wen, and Baining Guo. Face x-ray for more general face forgery detection. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 5001–5010, 2020. 1, 3

[29] Yuezun Li, Ming-Ching Chang, and Siwei Lyu. In Ictu Oculi: Exposing AI created fake videos by detecting eye blinking. In *IEEE international Workshop on Information Forensics and Security (WIFS)*, pages 1–7, 2018. 3

[30] Yuezun Li and Siwei Lyu. Exposing deepfake videos by detecting face warping artifacts. In *IEEE CVPR Workshops*, 2019. 3

[31] Yuezun Li, Pu Sun, Honggang Qi, and Siwei Lyu. Celeb-DF: A Large-scale Challenging Dataset for DeepFake Forensics. In *IEEE Conference on Computer Vision and Patten Recognition (CVPR)*, 2020. 1, 7

[32] Iacopo Masi, Aditya Killekar, Royston Marian Mascarenhas, Shenoy Pratik Gurudatt, and Wael AbdAlmageed. Two-branch recurrent networkfor isolating deepfakes in videos. In *European Conference on Computer Vision (ECCV)*, 2020. 3

[33] N. Dufour, A. Gully, P. Karlsson, A.V. Vorbyov, T. Leung, J. Childs and C. Bregler. DeepFakes detection Dataset, 2019. https://ai.googleblog.com/2019/09/contributing-data-to-deepfake-detection.html. 1, 6

[34] Hua Qi, Qing Guo, Felix Juefei-Xu, Xiaofei Xie, Lei Ma, Wei Feng, Yang Liu, and Jianjun Zhao. DeepRhythm: Exposing DeepFakes with Attentional Visual Heartbeat Rhythms. In *ACM International Conference on Multimedia*, pages 4318–4327, 2020. 3

[35] Andreas Rössler, Davide Cozzolino, Luisa Verdoliva, Christian Riess, Justus Thies, and Matthias Nießner. Faceforensics++: Learning to detect manipulated facial images. In *IEEE International Conference on Computer Vision (ICCV)*, pages 1–11, 2019. 1, 2, 3, 4, 7

[36] Selim Seferbekov. *DeepFake Detection (DFDC) Team Sefer.* https://github.com/selimsef/dfdc_deepfake_challenge. 2, 7

[37] Mingxing Tan and Quoc Le. Efficientnet: Rethinking model scaling for convolutional neural networks. In *International Conference on Machine Learning*, pages 6105–6114, 2019. 7

[38] Justus Thies, Michael Zollhöfer, and Matthias Nießner. Deferred neural rendering: Image synthesis using neural textures. *ACM Transactions on Graphics (Proc. of SIGGRAPH)*, 2019. 6

[39] Justus Thies, Michael Zollhöfer, Marc Stamminger, Christian Theobalt, and Matthias Nießner. Face2face: Real-time face capture and reenactment of rgb videos. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 2387–2395, 2016. 1, 4, 6

[40] 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. 1

[41] Luisa Verdoliva. Media forensics and deepfakes: an overview. *IEEE Journal of Selected Topics in Signal Processing*, 14(5):910–932, 2020. 1

[42] Nam N Vo and James Hays. Localizing and orienting street views using overhead imagery. In *European Conference on Computer Vision (ECCV)*, pages 494–509. Springer, 2016. 4

[43] Sheng-Yu Wang, Oliver Wang, Richard Zhang, Andrew Owens, and Alexei A Efros. CNN-generated images are surprisingly easy to spot... for now. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2020. 1, 3

[44] Xun Wang, Xintong Han, Weilin Huang, Dengke Dong, and Matthew R Scott. Multi-similarity loss with general pair weighting for deep metric learning. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 5022–5030, 2019. 4, 6

[45] Xin Yang, Yuezun Li, and Siwei Lyu. Exposing deep fakes using inconsistent head poses. In *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 8261–8265, 2019. 3

[46] Bojia Zi, Minghao Chang, Jingjing Chen, Xingjun Ma, and Yu-Gang Jiang. WildDeepfake: A Challenging Real-World Dataset for Deepfake Detection. In *ACM International Conference on Multimedia*, pages 2382–2390, 2020. 3