– Supplemental Document –
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

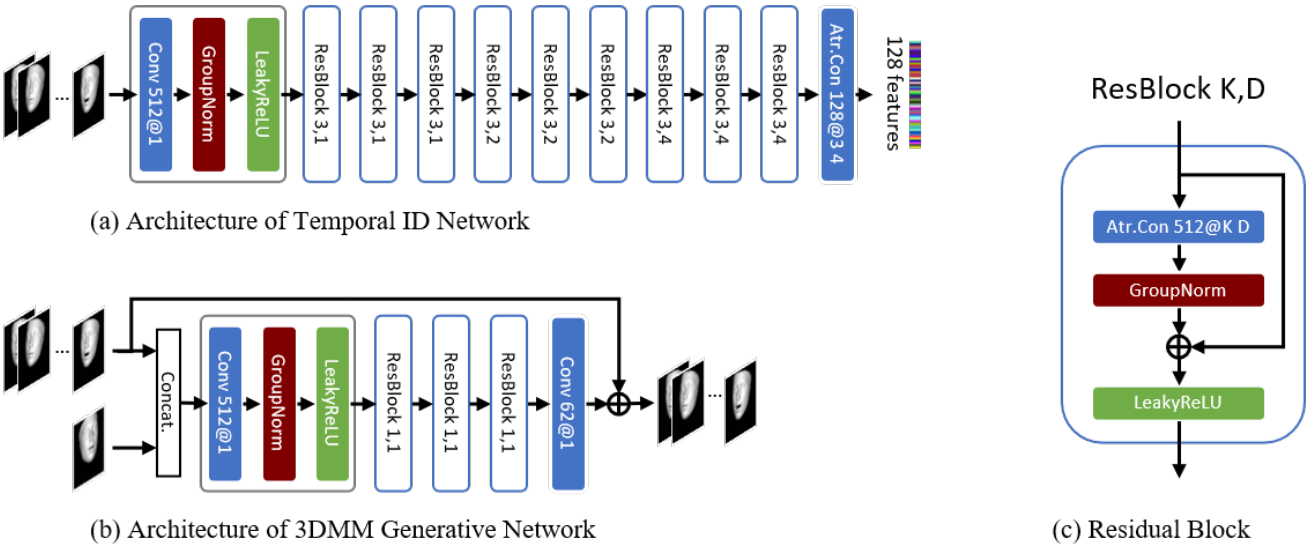

Figure 1: Architecture of our proposed Temporal ID Network and 3DMM Generative Network.

In this supplemental document, we report the details of our architectures used for the Temporal ID Network and the 3DMM Generative Network (Sec. 1). Moreover, we briefly describe the state of the art DeepFake methods we compare to, (see Sec. 2). In Sec. 3 and Sec. 4, we present additional results to prove the generalization capability of our method. In Sec. 5, we include scatter plots that show the separability of videos of different subjects in the embedding space. Finally, we analyze a real case on the web (see Sec. 6).

1. Architectures

Temporal ID Network We leverage a convolution neural network architecture that works along the temporal direction and is composed by eleven layers (see Fig.1 (a)). We use Group Normalization [14] and LeakyReLU nonlinearity for all layers except the last one. Moreover, we adopt à-trous convolutions (also called dilated convolution), instead of classic convolutions in order to increase the receptive fields without increasing the trainable parameters. The first layer increases the number of channels from 62 to 512, while the successive ones are inspired by the ResNet architecture [8] and include a residual block as shown in Fig.1 (c). The parameters K and D of the residual blocks are the dimension of the filter and the dilatation factor of the à-trous convolution, respectively. The last layer reduces the channels from 512 to 128. The receptive field of the whole network is equal to 51 frames which is around 2 seconds.

3DMM Generative Network As described in the main paper, the 3DMM Generative Network is fed by two 3DMM feature vectors. The two feature vectors are concatenated which results in a single input vector of 124 channels. The

network is formed by five layers: a layer to increase the channels from 124 to 512, three residual blocks, and a last layer to decrease the channels from 512 to 62. The output is summed to the input 3DMM feature vector to obtain the generated 3DMM feature vector (see Fig.1 (b)). All the convolutions have a dimension of filter equal to one in order to work frame-by-frame.

2. Comparison methods

In the main paper, we compare our approach with several state of the art DeepFake detection methods, that are described in following:

Frame-based methods

(i) MesoNet [1]: is one of the first CNN methods proposed for DeepFake detection which uses dilated convolutions with inception modules.

(ii) Xception [4]: is a relatively deep neural network that is achieving a very good performance compared to other CNNs for video DeepFake detection [9].

(iii) FFD (Facial Forgery Detection) [5]: is a variant of Xception, including an attention-based layer, in order to focus on high-frequency details.

(iv) Efficient-B7 [12]: has been proposed by Tan et al. and is pre-trained on ImageNet using the strategy described in [15], where the network is trained with injected noise (such as dropout, stochastic depth, and data augmentation) on both labeled and unlabeled images.

Ensemble methods

(v) ISPL (Image and Sound Processing Lab) [3]: employs an ensemble of four variants of Efficienet-B4. The networks are trained using different strategies, such as self-attention mechanism and triplet siamese strategy. Data augmentation is performed by applying several operations, like downscaling, noise addition and JPEG compression.

(vi) Seferbekov [10]: is the algorithm proposed by the winner of the Kaggle competition (Deepfake Detection Challenge) organized by Facebook [6]. It uses an ensemble of seven Efficientnet-B7 that work frame-by-frame. The networks are pre-trained using the strategy described in [15]. The training leverages data augmentation, that, beyond some standard operations, includes a cut-out that drops specific parts of the face.

Temporal-based methods

(vii) ResNet + LSTM: is a method based on Long Short Term Memory (LSTM) [7]. In detail, a ResNet50 is used to extract frame-level features from 20 frames uniformly extracted from the video. These features are provided to a LSTM that classifies the whole video.

		High Quality (HQ)		Low Quality (LQ)	
Acc(%) / AUC		DFD FR	DFD FS	DFD FR	DFD FS
MesoNet	Mean	57.0 / 0.65	54.0 / 0.57	58.1 / 0.61	52.7 / 0.53
	Max	54.5 / 0.55	52.6 / 0.47	54.2 / 0.55	51.6 / 0.48
Xception	Mean	51.9 / 0.74	78.5 / 0.93	49.8 / 0.48	58.5 / 0.63
	Max	58.9 / 0.71	80.4 / 0.92	46.1 / 0.44	51.6 / 0.59
Effic.-B7	Mean	53.1 / 0.75	88.2 / 0.97	50.2 / 0.48	58.5 / 0.64
	Max	62.5 / 0.73	79.4 / 0.96	45.2 / 0.44	55.9 / 0.66
FFD	Mean	53.6 / 0.57	75.3 / 0.83	53.9 / 0.55	64.9 / 0.72
	Max	52.5 / 0.56	60.2 / 0.78	50.9 / 0.50	51.7 / 0.64
AVG	Mean	53.9 / 0.68	74.0 / 0.83	53.0 / 0.53	58.7 / 0.63
	Max	57.1 / 0.64	68.2 / 0.78	49.1 / 0.48	52.7 / 0.59

Table 1: Video-level detection accuracy and AUC of frame-based methods. We compare two strategies: averaging the score over 32 frames in a video and taking the maximum score. Results are obtained on the DFD dataset on HQ videos and LQ ones, split in facial reenactment (FR) and face swapping (FS) manipulations.

(viii) Eff.B1 + LSTM: This is a variant of the approach described above, where the ResNet architecture is replaced by EfficientNet-B1.

Identity-based methods

(ix) A&B (Appearance and Behavior) [2]: is an identity-based approach that includes a face recognition network and a network that is based on head movements. The behavior recognition system encodes the information about the identity through a network that works on a sequence of attributes related to the movement [13].

Note that all the techniques are compared at video level. Hence, if a method works frame-by-frame, we average the probabilities obtained from 32 frames uniformly extracted from the video. Furthermore, to validate this choice, we compare averaging with the maximum strategy. Results are reported in Tab. 1 using the same experimental setting of Tab. 2 of the main paper. The results prove the advantage to use the averaging operation with respect to the maximum value: the increase in terms of AUC is around 0.04, while the accuracy increases (on average) of about 3%.

3. Additional results

To show the ability of our method to be agnostic to the type of manipulation, we test our proposal on additional datasets, that are not included in the main paper. In Tab. 2 we report the analysis on the dataset FaceForensics++ (FF++) [9]. Results are split for facial reenactment (FR) and face swapping (FS) manipulations. It is important to underline that this dataset does not provide information about multiple videos of the same subject, therefore, for identity-based approaches, the first 6 seconds of each pristine video

	High Quality (HQ)		Low Quality (LQ)	
Acc(%) / AUC	FF++ FR	FF++ FS	FF++ FR	FF++ FS
MesoNet	55.4 / 0.58	57.1 / 0.61	55.4 / 0.57	57.3 / 0.62
Xception	55.6 / 0.58	79.0 / 0.89	51.9 / 0.57	69.2 / 0.79
Efficient-B7	54.9 / 0.59	85.4 / 0.93	50.6 / 0.54	65.6 / 0.80
FFD	54.4 / 0.56	69.2 / 0.75	53.5 / 0.56	63.3 / 0.70
ISPL	56.6 / 0.59	74.2 / 0.83	53.3 / 0.55	68.8 / 0.76
Seferbekov	58.3 / 0.62	89.9 / 0.97	53.0 / 0.55	79.4 / 0.87
ResNet + LSTM	55.0 / 0.58	59.0 / 0.63	56.2 / 0.58	61.9 / 0.66
Eff.B1 + LSTM	57.2 / 0.62	81.8 / 0.90	54.1 / 0.58	69.0 / 0.78
A&B	72.2 / 0.78	89.0 / 0.97	51.5 / 0.53	51.9 / 0.65
ID-Reveal (Ours)	78.3 / 0.87	93.6 / 0.99	74.8 / 0.83	81.9 / 0.97

Table 2: Video-level detection accuracy and AUC of our approach compared to state-of-the-art methods. Results are obtained on the FF++ 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.

are used as reference dataset, while the last 6 seconds are used to evaluate the performance (we only consider videos of at least 14 seconds duration, thus, obtaining 360 videos for each manipulation method). For the FF++ dataset, our method obtains always better performance in both the cases of high-quality videos and low-quality ones.

As a further analysis, we test our method on a recent method of face reenactment, called FOMM (First-Order Motion Model) [11]. Using the official code of FOMM, we created 160 fake videos using the pristine videos of DFD, some examples are in Fig. 2. Our approach on these videos achieves an accuracy of 85.6%, and an AUC of 0.94 which further underlines the generalization of our method with respect to a new type of manipulation.

4. Robustness to different contexts

We made additional experiments to understand that for our method it is not necessary that the reference videos are similar to the manipulated ones in terms of environment, lighting, or distance from the subject. To this end, we show results in Fig. 3 obtained for the DFD FR and DFD FS datasets, where information about the video context (kitchen, podium, outside, talking, meeting, etc.) is available. While the reference videos and the under-test videos differ, our method shows robust performance. Results seem only affected by the variety of poses and expressions present in the reference videos (the last reference video in the table contains the most variety in motion, thus yielding better results).

Figure 2: Aligned examples of created FOMM videos. From top to bottom: source videos, target sequences, and manipulations created using First-Order Motion Model [11].

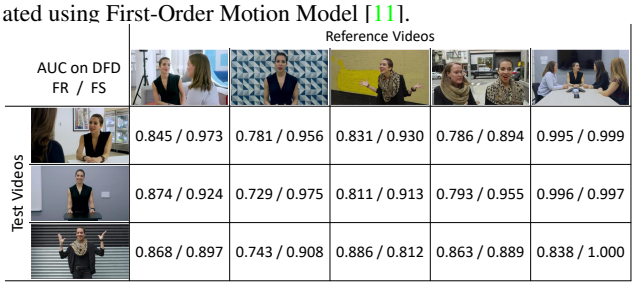

Figure 3: Average performance in terms of AUC evaluated on 28 actors of DFD FR and DFD FS datasets when test videos are in different contexts with respect to reference videos. Test videos: kitchen, podium-speech, outside laughing talking. Reference videos: angry talking, talking against wall, outside happy hugging, outside surprised, serious meeting.

5. Visualization of the embedded vectors

In this section, we include scatter plots that show the 2D orthogonal projection of the extracted temporal patterns. In particular, in Fig. 5 we show the scatter plots of embedded vectors extracted from 4 seconds long video snippets relative to two actors for the DFD dataset by using Linear Discriminant Analysis (LDA) and selecting the 2-D orthogonal projection that maximize the separations between the real videos of two actors and between real videos and fake ones. We can observe that in the embedding space the real videos relative to different actors are perfectly separated. Moreover, also the manipulated videos relative to an actor are well separated from the real videos of the same actor.

6. A real case on the web

We applied ID-Reveal to videos of Nicolas Cage downloaded from YouTube. We tested on three real videos, four DeepFakes videos, one imitator (a comic interpreting Nicolas Cage) and a DeepFake applied on the imitator. We eval-

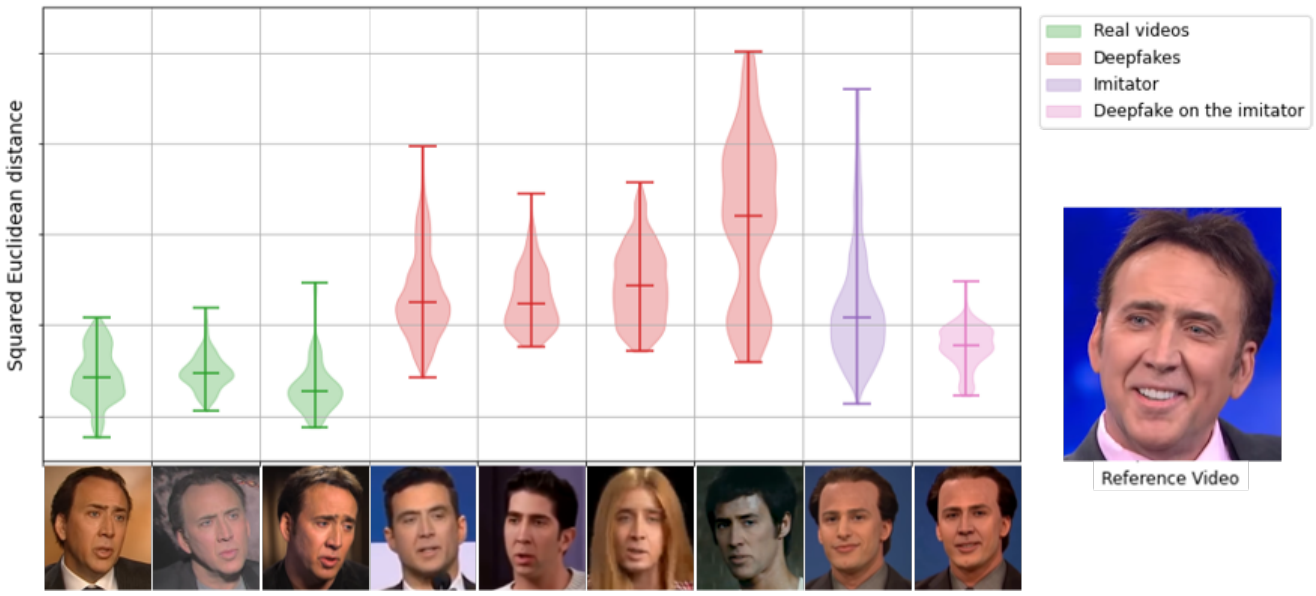

Figure 4: Distributions of squared Euclidean distances of 9 videos downloaded from YouTube with respect to a real reference video of Nicolas Cage. From left to right: 3 real videos, 4 DeepFakes, a video from an imitator and a DeepFake driven by the imitator.

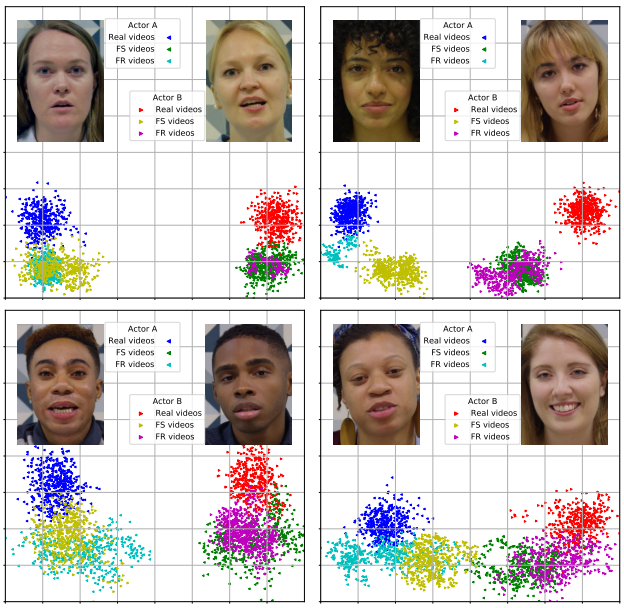

Figure 5: Scatter plots of embedded vectors extracted from 4 seconds long video snippets relative to couple of actors. We included both face swapping (FS) and facial reenactment (FR).

uate the distributions of distance metrics that are computed as the minimum pairwise squared Euclidean distance in the embedding space of 4 seconds long video snippets extracted from the pristine reference video and the video under test. In Fig. 4, we report these distributions using a violin plot.

We can observe that the lowest distances are relative to real videos (green). For the DeepFakes (red) all distances are higher and, thus, can be detected as fakes. An interesting case is the video related to the imitator (purple), that presents a much lower distance since he is imitating Nicolas Cage. A DeepFake driven by the imitator strongly reduces the distance (pink), but is still detected by our method.

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. 2

[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. 2

[3] 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. 2

[4] François Chollet. Xception: Deep learning with depthwise separable convolutions. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 1251–1258, 2017. 2

[5] 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. 2

[6] 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. 2

[7] 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. 2

[8] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 770–778, 2016. 1

[9] 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. 2

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

[11] Aliaksandr Siarohin, Stéphane Lathuilière, Sergey Tulyakov, Elisa Ricci, and Nicu Sebe. First order motion model for image animation. In *Advances in Neural Information Processing Systems (NeurIPS)*, volume 32, 2019. 3

[12] Mingxing Tan and Quoc Le. Efficientnet: Rethinking model scaling for convolutional neural networks. In *International Conference on Machine Learning*, pages 6105–6114, 2019. 2

[13] Olivia Wiles, A. Sophia Koepke, and Andrew Zisserman. Self-supervised learning of a facial attribute embedding from video. In *British Machine Vision Conference*, 2018. 2

[14] Yuxin Wu and Kaiming He. Group normalization. In *Proceedings of the European conference on computer vision (ECCV)*, pages 3–19, 2018. 1

[15] Qizhe Xie, Minh-Thang Luong, Eduard Hovy, and Quoc V Le. Self-training with noisy student improves imagenet classification. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 10687–10698, 2020. 2