Unit-Selection Based Facial Video Manipulation Detection

Thomas Nielsen[1], Ali Khodabakhsh[2], Christoph Busch[3]

Abstract: Advancements in video synthesis technology have caused major concerns over the authenticity of audio-visual content. A video manipulation method that is often overlooked is inter-frame forgery, in which segments (or units) of an original video are reordered and rejoined while cut-points are covered with transition effects. Subjective tests have shown the susceptibility of viewers in mistaking such content as authentic. In order to support research on the detection of such manipulations, we introduce a large-scale dataset of 1000 morph-cut videos that were generated by automation of the popular video editing software Adobe Premiere Pro. Furthermore, we propose a novel differential detection pipeline and achieve an outstanding frame-level detection accuracy of 95%.

Keywords: Morph-cut, Video Manipulation, Interframe Forgery, Dataset, Video Manipulation Detection, Video Authenticity.

1 Introduction

Following the evolution of artificial intelligence and the rapid increase in the computational capacity of computers in recent decades, many novel video manipulation techniques have been introduced and became feasible. Despite the original intention of the developers of these techniques, many of them have the potential of being misused by malicious actors to spread disinformation for political and financial aims. Following the significant media attention to this problem after the introduction of Deepfakes, many research groups attempt to address the vulnerability [Ve20]. However, among video manipulation techniques, vulnerability to unit-selection based methods have been overlooked. Unlike Deepfakes and similar generation methods for which synthesis still requires a significant amount of expert knowledge and computational capacity, unit-selection based video manipulation can be flexibly done by commercial software such as Adobe Premiere Pro through their easy to use graphical user interface. Furthermore, subjective tests have shown unit-selection based manipulations to be more difficult to detect for humans than intra-frame manipulations [KRB19]. The use of seamless cut-point transitions is commonplace in media for shortening and summarizing the highlights of videos and they go unnoticed more often than not[4].

Due to the less computational cost and the higher video-realism of unit-selection based generation methods, these methods have been explored for synthesis early-on for appli-

[1] Department of Applied Mathematics and Computer Science, Technical University of Denmark, Richard Petersens Plads, Building 324, Kgs. Lyngby, Denmark, s144458@win.dtu.dk

[2] Department of Information Security and Communication Technology, Norwegian University of Science and Technology, Teknologiveien 22, Gjøvik, Norway, ali.khodabakhsh@ntnu.no

[3] Department of Information Security and Communication Technology, Norwegian University of Science and Technology, Teknologiveien 22, Gjøvik, Norway, christoph.busch@ntnu.no

[4] https://metro.co.uk/2018/12/13/viewers-baffled-child-appears-teleport-tv-interview-8244024/

cations like audio-visual synthesis and video dubbing [MV15]. Even though concatenative generation methods require long videos with constrained recording conditions to be seamless, thanks to searchable public archives of videos, there exists enough footage from interviews on celebrities and political figures for these methods to be feasible. The first automatic technique for face-animation was proposed by Bregler et al. in 1997 [BCS97]. They create a database of visemes[5] from existing footage and, given an input text, they retrieve the visemes and concatenate them using morphing to synthesize a new sentence. More recently, Berthouzoz et al. [BLA12] introduced an editing tool to place visible cuts and seamless transitions in interview videos based on text transcript, which was further developed into the morph-cut transition in Adobe Premiere Pro[6] as a replacement for B-roll[7] and jump-cut transitions[8] for video summarization. Mattheyses and Verhelst [MV15] and Johnston and Elyan [JE19] provide an overview of existing unit-selection based manipulation methods. Among the existing datasets, the biggest that includes inter-frame forgery is VTD 2016 [ASAS16] which is comprised of 33 videos, 6 of which contain inter-frame forgery. Johnston and Elyan [JE19] provide a review of existing video tampering datasets.

In the context of facial video manipulation, a substantial amount of research is oriented towards intra-frame facial video manipulation detection [Ve20]. However, there exists a gap in knowledge with regards to detection of unit-selection based facial video manipulation, and to the best of our knowledge, there are no dataset and no proposed detection method that explicitly address this vulnerability. Nonetheless, Among the proposed methods for the detection of intra-frame manipulations, some utilize inter-frame information for detection to a limited extent. The authors in [GD18] and [Sa19] exploit the inter-frame dependencies to detect frame-by-frame manipulations via a convolutional long short-term memory (LSTM) network and a recurrent neural network respectively. Amerini et al. [Am19] use estimation of the optical flow field as input to a convolutional neural network (CNN) for the detection of inter-frame inconsistencies.

To reduce the visibility of concatenation points in inter-frame forgery, simple gradual transitions such as interpolation, warping, and morphing, as well as more advanced methods such as face-specific warping [Da11] and intermediate frame mining [BLA12] can be used. Examples of advanced transitions that are already available in video editing software are Adobe Premiere Pro Morph-cut (Figure 1) and Avid[9] Fluid Morph. Despite the core algorithms of these transitions being trade secrets, the name of these transitions implies the use of morphing in some form. Consequently, single-image face morphing detection algorithms that are developed in the context of biometric presentation attack detection become relevant for detection. Scherhag et al. [Sc19] provide a recent survey of existing morphing attack detection methods. Asaad and Jassim [AJ17] used the responses of uniform local binary pattern (LBP) extractors on the image to build a Vietoris-Rips complex for detection.

[5] Visemes denote the shape of the mouth when pronouncing specific phonemes. Visemes and phonemes do not share a one-to-one correspondence.

[6] https://www.adobe.com/products/premiere.html

[7] In B-roll transition, a supplemental footage is intercut with the main shot to cover the cuts.

[8] In jump-cut transition, the cut is kept as it is, causing an abrupt jump in the resulting footage.

[9] https://www.avid.com/

Wandzik et al. [WKG18] use high-level features of pretrained face recognition networks as input for a linear SVM classifier.

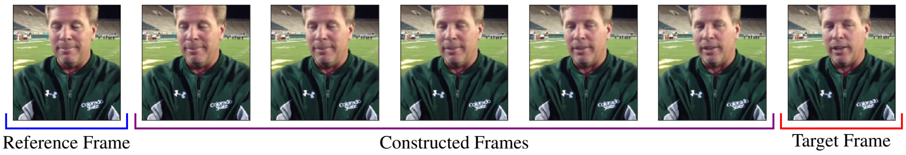

Fig. 1: An example of a morph-cut transition.

Another set of relevant detection methods can be adopted from general-purpose inter-frame forgery detection, namely frame-insertion and frame-deletion detection methods. Siatara and Mehtre [SM16] provide an overview of the existing inter-frame forgery detection methods. Notably, Chao et al. [CJS12] detect manipulated videos by using the consistency in the total optical flow values in the X and Y directions. More recently, Bakas and Naskar [BN18] used 3D convolutional neural networks with a special difference layer to detect out of place frames in the video sequence.

In this work, we introduce a large-scale dataset of videos containing morph-cut transitions based on videos collected from the wild.[10] To the best of our knowledge, the Morph Cut dataset is the first of its kind and enables the training of deep learning solutions for the detection task. Furthermore, we introduce a robust neural detection pipeline, capable of detecting the morph-cut position at the frame level in a video. The rest of this article is organized as follows: The dataset and the proposed detector are introduced in Section 2. The experiment setup is explained in Section 3 and the results are discussed in Section 4. Finally, the paper is concluded in Section 5.

2 Methodology

Due to the lack of datasets containing a sufficiently large number of unit-selection based manipulation in the literature, we decided to generate a dataset and provide it publicly to stimulate further research in inter-frame forgery detection. In this section, we summarize the construction process of the new Morph Cut dataset along with the description of our proposed method for detecting the inter-frame forgeries.

2.1 Morph Cut Dataset

The development of deep learning-based detectors requires large-scale datasets. Consequently, as the manual generation of datasets of such scale is impractical, the generation process needs to be automated. Adobe Premiere Pro is a well-known popular video editing application that features a seamless morph-cut transition for cut-point concatenation. Furthermore, Adobe Systems provide the scripting language named Extendscript which can

[10] The instructions on how to download the Morph Cut dataset are available at `http://ali.khodabakhsh.org/research/morphcut/`

be used for automation of repetitive tasks in video editing. As such, Adobe Premiere Pro morph-cut transition is the perfect candidate to be used for the generation of the dataset. To achieve a seamless transition, the frames before and after transition need to be similar with regards to the background as well as the general body posture.

To ensure the quality of the generated data, we relied on a much larger video dataset consisting of interview videos as the basis for video selection. Thereafter, based on the movements of face bounding-box after face detection in the videos and the structural similarity of the frames to one another, the videos were ranked and the most suitable videos were selected for the application of morph-cut. Subsequently, the transition is applied to the videos at random points during the interview and the resulting manipulated videos were manually investigated for videos with visible artifacts to be discarded.

2.2 Morph-cut Detection

The unit-selection based video synthesis requires smooth transitions at the cut-points to cover the abrupt changes between the frame before and after. As such, it is safe to assume the existence of frame interpolation during the transition in one form or another. During frame interpolation, the content of the new frame in-between is generated based on the information available in the frame before and after. In contrast, pristine frames contain a natural variability that is not completely explainable based on the information in the frame before and after. Let us consider the frame in the middle to be consisting of two factors, p for the redundant information that is inferable from the frame before and after, and u for the unpredictable natural variability. A good frame interpolation would be able to infer p accurately, however, inference of u is an ill-defined problem. If during the design and training of a frame interpolation method, no mechanism is considered for ignoring u, the objective function would force the interpolation method to generate an average u which minimizes the penalty, yet never occurs in the pristine data. This phenomenon often results in synthetic samples described as over-smooth.

Considering any two frame interpolation methods with the aforementioned characteristics, we hypothesize that the predicted intermediate frames would show more similarity to each other than to the pristine data. The rationale behind this is that the p factor would exist in both pristine and synthetic frames, yet the u factor would only properly occur in pristine data while the frame interpolation methods each would generate an over-smooth average u. Thus it is reasonable for the difference between the natural u and the average u to be greater than the difference between two average us generated by the two synthesis methods. To use this behavior for interpolation detection, for each frame, the interpolated parallel can be generated from the frame before and after with any other good interpolation method that fits the aforementioned description. Next, the prediction error can be measured as the difference between the interpolated frame and the observed one. Consequently, this difference can be used for distinguishing pristine frames from interpolated ones by using a distance measure. Alternatively, this prediction error *image* can be fed to a classifier which specializes in the detection of interpolated frames for better performance.

3 Experiment Setup

We provide the large-scale Morph Cut dataset for the task of unit-selection based facial video manipulation detection training and testing on which we empirically verify the detection hypothesis. Furthermore, in our benchmark we perform the detection task with four applicable detection methods from the literature. The details of the dataset along with the experiment setup is explained in the following.

3.1 Morph Cut Dataset Details

The VoxCeleb2 [NCZ17] dataset is used as a basis for video selection, which contains a collection of interview videos from celebrities hosted on the video-sharing platform YouTube. The videos are ranked based on the face bounding-box movements, and on the suitable videos, uniform random sampling is applied to select candidate points for morph-cut. Next, the candidates with high structural similarity index [WB09] are selected and two morph-cut transitions are automatically added to each video using Extendscript. The Morph Cut dataset contains $1,000$ videos with an average duration of 2.75 seconds. This dataset adds up to $\sim 83,000$ frames with $\sim 27,500$ morphed frames and a ratio of 33% morphed frames to pristine ones. The videos are split three sets corresponding to training, validation, and the test data according to numbers in Table 1. The video parameters are summarized in Table 2. The videos are accompanied by frame-level labels corresponding to whether each frame is morphed or pristine. All reported results are based on frame-level classification performance between the morphed frames and the pristine ones.

Set	Count
Train	700
Dev	150
Test	150

Tab. 1: The number of videos in each set of the constructed Morph Cut dataset.

Video parameters
MPEG-4 (Base Media / Version 2)
480p (854 × 480)
30 FPS (Frames-Per-Second)
AVC (NTSC)

Tab. 2: The parameters used to create each video in the constructed Morph Cut dataset.

3.2 Proposed Detector

For the detector's reference frame-interpolation method, the pre-trained CyclicGen [Li19] convolutional neural network is used. For a given pair of frames, this network produces a high-quality intermediate interpolated frame. Using this network, for each frame in a video, a corresponding interpolated frame is synthesized based on the frame before and after, and the prediction error is calculated in terms of a difference image. The resulting prediction error *images* on cropped face regions are then converted to gray-scale and fed to a simple convolutional neural network for frame-level classification. The input to the network is augmented with the *context* prediction error images of two frames before and after, resulting in an input shape of $64 \times 64 \times 5$. The training and evaluation pipeline is visualized in Figure 2 and the classifier network architecture is summarized in Table 3.

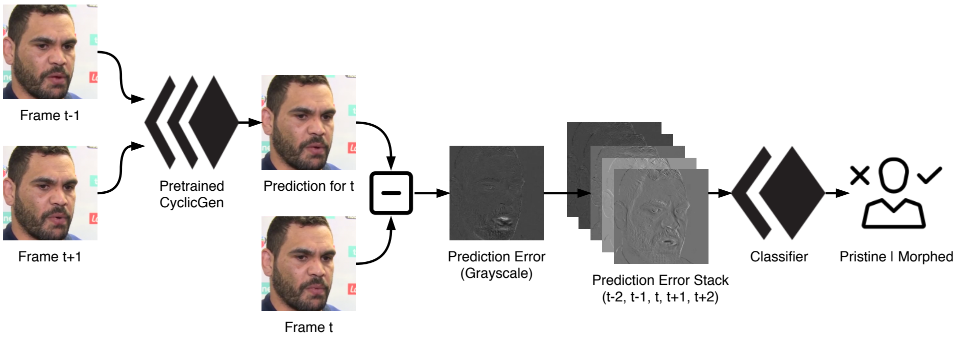

Fig. 2: The training and evaluation pipeline in the proposed method.

Layer	Output Shape	Parameters
Conv2D	(62, 62, 128)	Kernel=(3,3)
MaxPooling2D	(31, 31, 128)	Pool=(2,2)
Conv2D	(29, 29, 128)	Kernel=(3,3)
MaxPooling2D	(14, 14, 128)	Pool=(2,2)
Conv2D	(12, 12, 256)	Kernel=(3,3)
MaxPooling2D	(6, 6, 256)	Pool=(2,2)
Conv2D	(4, 4, 512)	Kernel=(3,3)
MaxPooling2D	(2, 2, 512)	Pool=(2,2)
Flatten	(2048)	
Dense	(512)	
DropOut	(512)	
Dense	(2)	

Tab. 3: The network architecture of the classifier. The network contains 1.6M trainable parameters.

3.3 Baseline Methods

For baseline methods to be used in our benchmark, we relied on recently published and reproducible detection methods for face-morph detection [AJ17], time-aware Deepfake detection [GD18], inter-frame forgery detection [BN18], and general purpose image classification [Ch17]. Among the four methods, [GD18] and [BN18] utilize temporal information while [AJ17] and [Ch17] rely only on static face images. All methods provide frame-level decision.

The first method is based on topological data analysis for image tampering detection described in the paper of the same name [AJ17]. This method was originally created to detect morphing attacks on face images by extracting features from the texture of the image itself, making the method sensitive to image tampering through the degradation of the image. For this method, we first extract the cropped faces from each frame in the dataset and construct a 1-skeleton of the full rips simplicial complex for each face image, which is then fed into an SVM classifier to attempt and recognize the morphed faces against the pristine ones.

The second method relies on recurrent neural networks for Deepfake detection [GD18]. The cropped face images are used as input to the network and all parameters are kept the same as described in the paper except we are training with fewer epochs. The third method relies on 3D convolutional neural networks for the detection of inter-frame forgery as described in [BN18]. Finally, due to the outstanding performance of the Xception-Net [Ch17] for Deepfake detection task, the pre-trained network is fine-tuned on the task of morph-cut detection on individual images.

4 Results and Discussion

Table 4 summarizes the detection accuracy of the proposed method in comparison to the baseline methods. The proposed method achieves the highest detection accuracy of 95.1% on the test set, followed surprisingly by the fine-tuned XceptionNet at 77.0%. The other three baseline methods show limited success in the detection of morph-cut frames. The detection-error-tradeoff (DET) curve for the top 3 best-performing methods is shown in Figure 3. In this figure, APCER stands for attack presentation classification error rate and BPCER stand for bona fide presentation classification error rate, which correspond to the missed detection and the false alarm rate of a biometric presentation attack detection system respectively following the ISO/IEC 30107 standard terminology[11]. The proposed method achieves an acceptable detection equal-error-rate (EER) of 4.95%.

Method	**Test Accuracy**
Topological Data Analysis [AJ17]	50.2%
Deepfake Video Detection [GD18]	59.0%
Inter-Frame Forgery C3D [BN18]	67.4%
Fine-tuned XceptionNet [Ch17]	77.0%
Proposed Method	95.1%

Tab. 4: The detection accuracy of the proposed method in comparison to the baseline methods. The results show the frame-level performance.

Examples of the prediction errors which are used as input to the classifier in the proposed method are visualized in Figure 4. Natural variations are clearly visible in prediction errors in pristine frames, while these variations are not observed in the morphed (interpolated) ones. Figure 5 shows the probability density distribution of average prediction error per frame over pristine and morphed frames. The morphed frame average prediction error distribution is shifted towards zero compared to the pristine distribution, confirming the hypothesis proposed in Section 2.2. The clear distinction between the pristine and morphed frame prediction errors visualized in Figure 4 and 5 show the effectiveness of prediction error *images* in isolating useful features for morphed face detection.

[11] `https://www.iso.org/obp/ui/#iso:std:iso-iec:30107:-3:ed-1:v1:en`

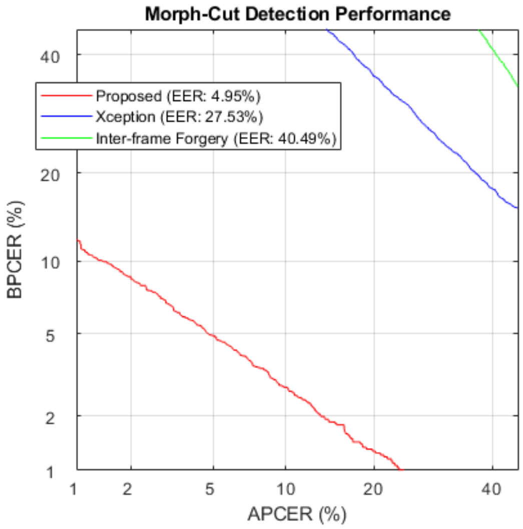

Fig. 3: The DET curve for the frame-level detection performance of the proposed method, the fine-tuned Xception-Net[Ch17], and the inter-frame forgery detection method[BN18]. The equal-error-rate (EER) value for the aforementioned methods is shown in the figure legend.

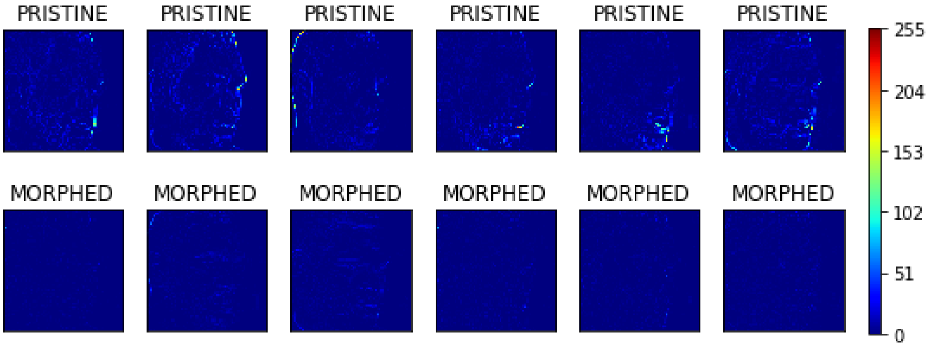

Fig. 4: Example of prediction error *images* of cropped faces in a six-frame sequence of pristine frames (top) and morph-cut frames (bottom) in a video. The images visualize the absolute gray value difference per pixel between the interpolation output and the actual frame.

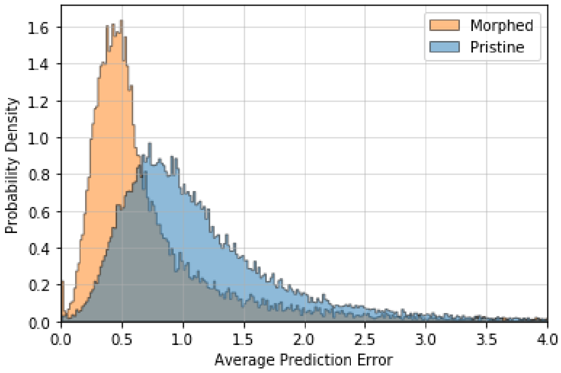

Fig. 5: The probability density distribution of average prediction error per frame for pristine and morphed frames across the dataset.

5 Conclusion

In this article, we addressed the problem of unit-selection based facial video manipulation by providing the first large-scale dataset of videos manipulated by popular video-editing software. Furthermore, we proposed a detection method that relies on frame-interpolation prediction-errors as discriminative features for the detection of morphed frames. The proposed method outperforms the baseline methods by a wide margin. The high frame-level performance of the proposed method shows its capacity in reliably detecting unit-selection based video manipulation and confirms the detection hypothesis that synthetic frames demonstrate higher similarity to each other than to pristine ones.

References

[AJ17] Asaad, Aras; Jassim, Sabah: Topological data analysis for image tampering detection. In: International Workshop on Digital Watermarking. Springer, pp. 136–146, 2017.

[Am19] Amerini, Irene; Galteri, Leonardo; Caldelli, Roberto; Del Bimbo, Alberto: Deepfake Video Detection through Optical Flow Based CNN. In: The IEEE International Conference on Computer Vision (ICCV) Workshops. Oct 2019.

[ASAS16] Al-Sanjary, Omar Ismael; Ahmed, Ahmed Abdullah; Sulong, Ghazali: Development of a video tampering dataset for forensic investigation. Forensic Science International, 266:565 – 572, 2016.

[BCS97] Bregler, Christoph; Covell, Michele; Slaney, Malcolm: Video Rewrite: Driving Visual Speech with Audio. In: SIGGRAPH. SIGGRAPH '97, ACM Press/Addison-Wesley Publishing Co., USA, p. 353–360, 1997.

[BLA12] Berthouzoz, Floraine; Li, Wilmot; Agrawala, Maneesh: Tools for Placing Cuts and Transitions in Interview Video. ACM Trans. Graph., 31(4), July 2012.

[BN18] Bakas, Jamimamul; Naskar, Ruchira: A Digital Forensic Technique for Inter–Frame Video Forgery Detection Based on 3D CNN. In: International Conference on Information Systems Security. Springer, pp. 304–317, 2018.

[Ch17] Chollet, F.: Xception: Deep Learning with Depthwise Separable Convolutions. In: 2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR). pp. 1800–1807, 2017.

[CJS12] Chao, Juan; Jiang, Xinghao; Sun, Tanfeng: A Novel Video Inter-Frame Forgery Model Detection Scheme Based on Optical Flow Consistency. In: IWDW. IWDW'12, Springer-Verlag, Berlin, Heidelberg, p. 267–281, 2012.

[Da11] Dale, Kevin; Sunkavalli, Kalyan; Johnson, Micah K.; Vlasic, Daniel; Matusik, Wojciech; Pfister, Hanspeter: Video Face Replacement. In: SIGGRAPH Asia. SA '11, Association for Computing Machinery, New York, NY, USA, 2011.

[GD18] Güera, D.; Delp, E. J.: Deepfake Video Detection Using Recurrent Neural Networks. In: 2018 15th IEEE International Conference on Advanced Video and Signal Based Surveillance (AVSS). pp. 1–6, 2018.

[JE19] Johnston, Pamela; Elyan, Eyad: A review of digital video tampering: From simple editing to full synthesis. Digital Investigation, 29:67 – 81, 2019.

[KRB19] Khodabakhsh, A.; Ramachandra, R.; Busch, C.: Subjective Evaluation of Media Consumer Vulnerability to Fake Audiovisual Content. In: 2019 Eleventh International Conference on Quality of Multimedia Experience (QoMEX). pp. 1–6, 2019.

[Li19] Liu, Yu-Lun; Liao, Yi-Tung; Lin, Yen-Yu; Chuang, Yung-Yu: Deep Video Frame Interpolation using Cyclic Frame Generation. In: Proceedings of the 33rd Conference on Artificial Intelligence (AAAI). 2019.

[MV15] Mattheyses, Wesley; Verhelst, Werner: Audiovisual speech synthesis: An overview of the state-of-the-art. Speech Communication, 66:182 – 217, 2015.

[NCZ17] Nagraniy, Arsha; Chungy, Joon Son; Zisserman, Andrew: VoxCeleb: A large-scale speaker identification dataset. INTERSPEECH, 2017-August:2616–2620, 2017.

[Sa19] Sabir, Ekraam; Cheng, Jiaxin; Jaiswal, Ayush; AbdAlmageed, Wael; Masi, Iacopo; Natarajan, Prem: Recurrent Convolutional Strategies for Face Manipulation Detection in Videos. In: CVPR Workshops. June 2019.

[Sc19] Scherhag, U.; Rathgeb, C.; Merkle, J.; Breithaupt, R.; Busch, C.: Face Recognition Systems Under Morphing Attacks: A Survey. IEEE Access, 7:23012–23026, 2019.

[SM16] Sitara, K.; Mehtre, B.M.: Digital video tampering detection: An overview of passive techniques. Digital Investigation, 18:8 – 22, 2016.

[Ve20] Verdoliva, Luisa: Media forensics and deepfakes: an overview. arXiv preprint arXiv:2001.06564, 2020.

[WB09] Wang, Zhou; Bovik, Alan C.: Mean squared error: Lot it or leave it? A new look at signal fidelity measures. IEEE Signal Processing Magazine, 26(1):98–117, 2009.

[WKG18] Wandzik, L.; Kaeding, G.; Garcia, R. V.: Morphing Detection Using a General- Purpose Face Recognition System. In: 2018 26th European Signal Processing Conference (EUSIPCO). pp. 1012–1016, 2018.