Quality-based Artifact Modeling for Facial Deepfake Detection in Videos

Sara Concas Simone Maurizio la Cava Roberto Casula Giulia Orrù
Giovanni Puglisi Gian Luca Marcialis
University of Cagliari, Cagliari (Italy)
{sara.concas90c,simonem.lac,roberto.casula,giulia.orru,giovanni.puglisi,marcialis}@unica.it

Abstract

Facial deepfakes are becoming more and more realistic, to the point that it is often difficult for humans to distinguish between a fake and a real video. However, it is acknowledged that deepfakes contain artifacts at different levels; we hypothesize a connection between manipulations and visible or non-visible artifacts, especially where the subject's movements are difficult to reproduce in detail. Accordingly, our approach relies on different quality measures, No-Reference (NR) and Full-Reference (FR), over the detected faces in the video. The measurements allow us to adopt a frame-by-frame approach to build an effective matrix-based representation of a video sequence. We show that the results obtained by this basic feature set for a neural network architecture constitute the first step that encourages the empowerment of this representation, aimed to extend our investigation to further deepfake classes. The FaceForensics++ dataset is chosen for experiments, which allows the evaluation of the proposed approach over different deepfake generation algorithms.

1. Introduction

Deepfakes have emerged as one of the most noteworthy and potentially dangerous innovations. With the term ”deepfake”, we identify all the techniques related to alterations of digital contents affecting humans; such alterations are carried out by deep learning methods [36][1]. In deepfakes [23], we may have a face-swap, where the face of a targeted subject replaces the face of another one, or an expression-swap, also called reenactment, where the facial expressions are manipulated, till the manipulation of attributes (soft-biometrics) such as age, gender, makeup and so on. Finally, the face synthesis generates the whole appearance *ex novo*.

Besides the benign use of deepfakes, the phenomena of bullying among teenagers, fake news, and, in general, misinformation are fertile ground for their malicious use. It is a common opinion that the two most harmful types of deepfakes are based on face-swap and reenactment [36]: use cases include spreading fake news, fake pornography, and blackmailing. In such deepfakes, the salient information that can be exploited for detection is mainly concentrated in the face. For example, in [5], the authors drive the attention of their system to the targeted subject's mouth, nose, and ears.

Extending their intuition, in this paper we hypothesize that targeting movements of specific facial parts allows us to track artifacts induced by deep learning techniques. These ”temporal artifacts” change the variation dynamic of the pixel in those areas, impacting the quality of the view. Accordingly, we investigate the use of quality measures computed for each video frame (no-reference methods) and by comparing subsequent frames (full-reference methods). We focus on the mouth and the eyes, for example, where we hypothesize a possible quality degradation due to the impossibility of reproducing the facial movements in detail. We also include the same computation approach on the whole face image to keep a high-level quality dynamic. A handcrafted matrix-based representation is obtained and used to train a neural network classifier to determine whether the video is manipulated. Only facial deepfakes are considered.

We selected the FaceForensics++ [28] dataset for experiments, where different generation approaches of facial deepfakes are sampled. Reported results show that the selected quality measures are suitable for obtaining classifiers that can generalize over never-seen-before manipulation approaches. In our opinion, this is the first evidence of the possibility of detecting artifacts that are invariant with respect to the methods of manipulation, although in here we are limited by those adopted in the FaceForensics++ dataset.

The remainder of this paper is organized as follows. Section 2 reviews the current literature on deepfake detection. Section 3 describes the proposed approach. Section 4 reports the data and the experimental protocol employed to

[1] In the Merrian-Webster and Oxford dictionaries, deep learning approaches to obtain deepfakes are not mentioned and the definition simply refers to “an image or recording that has been convincingly altered and manipulated to misrepresent someone as doing or saying something that was not actually done or said”.

DOI: 10.1109/CVPRW63382.2024.00389

conduct our evaluation, while Section 5 reports the obtained results. Finally, conclusions are drawn in Section 6.

2. Related Work

The deepfake generation process tends to leave traces in the form of artifacts in both the spatial [6] and frequency domains [11], especially in specific regions of the face [31]. For this reason, several studies concentrate their analysis on face image portions. In [18], for example, the authors propose DFT-MF, a deepfake detection model based on mouth-driven features. Their purpose is to crop the mouth region and analyze the movement of the lips individually. Similarly, the high-level semantic irregularities of the lips in videos are exploited by the LipForensics approach [14], employing a pre-trained spatio-temporal network for a lip speech analysis task, fine-tuned to accomplish deepfake detection. Lip movement is also exploited in [37] where the authors use a self-supervised audio-visual transformer; the mouth motion representations are learned in such a way that the paired video and audio representations are close, while unpaired ones are not. The audio information is also exploited in [38], which proposes synchronizing visual and auditory modalities, allowing for better generalization of unknown deepfakes. Similarly, this information is exploited in RealForensics [13], trying to generalize to never-seen-before forgery methods. This is done by focusing on natural facial behavior and appearance in real videos to detect visual-only forged videos.

In 2021, Li et al. [20], starting from the observation that fake samples usually lack symmetry in corresponding face portions, extracted the features from symmetrical face patches and computed an angular distance to verify how similar the two portions are. In 2022, the authors of [2] tried to find a metric to accurately describe the structural similarity between real and fake images. They considered MSE (Mean Squared Error) and SSIM (Structural Similarity Index Measure) in combination with morphological tools. They found that combining SSIM with erosion and dilation yields the most satisfactory results. In [10], starting with the observation that during deepfake production most traces are left on the edges of faces, the authors separate the face edges from the video frames, extract the edge bands, and train an EfficientNet-B3 network. To exploit the information in specific patches of the images, Ju et al. [19] fuse the global information related to the whole image with the local information contained in multiple patches, picked by a dedicated selection module.

We want to add to the contributions above a further step into understanding the role of local and global artifacts generated in the video by the deep learning process. Analyzing many deepfake videos by visual inspection points out that similar artifacts are common to many deepfake generation algorithms and could be roughly described in a similar way. This is also recalled in the previously cited works. The acknowledged fact that artifacts alter the normal variation flow of a certain facial feature (the mouth, for example) during its movements also suggests a degradation of the visual image quality. Moreover, a large plethora of quality measurements have been proposed in the literature to assess image quality with an ”objective” value or set of values. Accordingly, we chose a selection of quality measurements as the best mean to verify our hypothesis about deepfake-made artifacts. Our idea is supported by their successful use in presentation attack detection concerning many biometric traits since 2013 [12] and 2016 [26]. In both these works, the authors employed Full-Reference and No-Reference quality assessment measures: when dealing with Full-Reference metrics, the authors of [12] filtered the considered image with a low-pass Gaussian kernel. Then, they computed the quality between the original image and the filtered one.

Section 3 details our approach and lists the selected quality measurements.

3. Proposed Approach

As already reported in previous sections, this paper's main aim is to develop a deepfake detection method with a stronger relationship with the alterations of visual image quality than classical end-to-end deep learning-based approaches, keeping satisfactory detection accuracy also on unseen manipulation (i.e., generalizability). We have carefully selected manually engineered features to achieve this goal for a comprehensive analysis strategy that combines spatial and temporal information extracted from a video sequence. Specifically, the proposed methodology is based on the hypothesis that deepfake videos exhibit unique artifacts, particularly in regions of the face subject to motion, like mouth, nose, and eyes, and that these artifacts can be detected through a combination of quality measures.

Our method consists of different stages described in Fig. 1. Since temporal inconsistencies can be detected from consecutive frames, we defined an appropriate time window to analyze the set of frames. We extract several patches from each frame focused on areas of the face most prone to manipulation artifacts, such as the eyes, nose, and mouth. From these patches, we extract different quality measures capable of describing various artifacts.

Feature extraction is repeated twice: the first time on the original frame and the second time after applying a high-pass filter to the frame (Fig. 2). This second step is motivated by the literature [1], which points out that deepfake samples show different behaviors at high frequencies compared to real samples. The extracted quality features serve as input to two separate Convolutional Neural Networks (CNN). This architecture is chosen due to the faster and more efficient calculation compared to more complex solutions and the possibility of more easily interpreting the

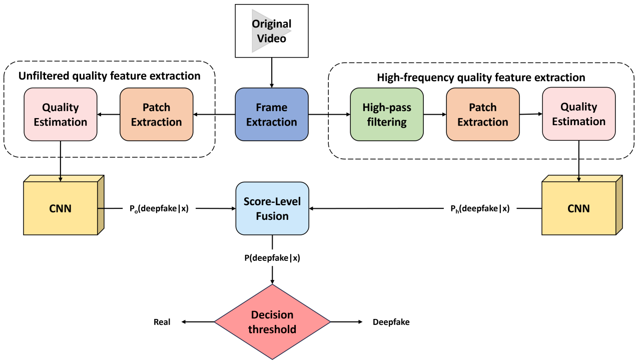
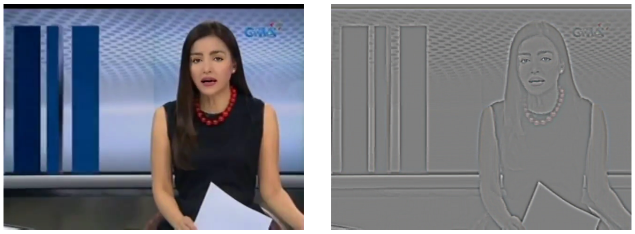


Figure 1. General scheme of the proposed approach: frames are extracted from each video sequence, and the related quality measures are calculated from them for each facial patch. The resulting matrix is input to a Convolutional Neural Network.



(a) Original Frame

(b) High-pass filtered frame.

Figure 2. Example from the FaceForensics++ dataset [28]. The original frame extracted from a video in the dataset (a) and its high-pass filtered version (b).

network's decision based on the particular input. We finally perform a score-level fusion to get the system's final decision.

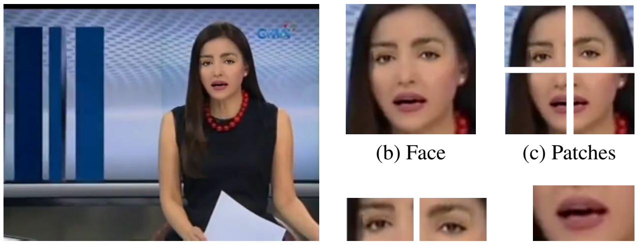
In the following sections, the processing steps are described in detail.

3.1. Frames and Patches Extraction

After the extraction of a sequence of frames according to a pre-defined time window, we extracted several patches: the entire face of the person, the left eye, the right eye, the mouth, and four quadrants representing four facial regions, as shown in Fig. 3. After this stage, we proceeded to compute the quality measures, including Full-Reference (FR) methods (i.e., methods that compare the considered image to a reference image) and No-Reference (NR) methods (that try to assess the quality of an image without comparison to other images).

3.2. Quality Assessment Measures

An overview of the quality assessment measures can be seen in Fig. 4. We selected five NR techniques, where a quality score is assigned to the single patch, and five FR methods, based on comparisons among patches.



(b) Face

(c) Patches

(a) Original frame

(d) Eyes

(e) Mouth

Figure 3. Example from the FaceForensics++ dataset [28]. The patches are extracted from a single frame: original frame (a), face (b), face patches (c), right and left eyes (d), and mouth (e).

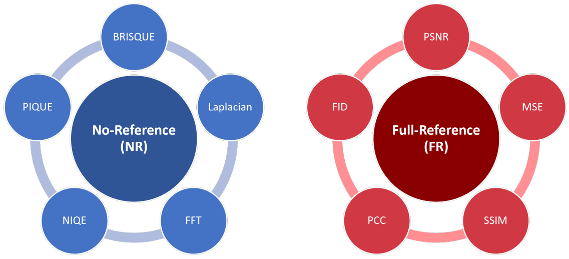


Figure 4. Summary scheme of the different quality measures applied to the extracted patches.

3.2.1 No-Reference Measures

This section summarizes the NR quality measures adopted. We refer to the original papers for further details.

- BRISQUE (Blind/Referenceless Image Spatial Quality Evaluator) [24] generates a quality score by quantifying potential losses of "naturalness" in the picture as a result of distortions using scene statistics of locally normalized luminance coefficients. The statistical features of the image are analyzed and compared to a model created starting from pictures of known quality.
- Laplacian Operator [4] is a differential operator given by the divergence of the gradient of a function in the Euclidean space. It highlights the areas of rapid intensity change, thus performing edge detection. By computing the variance, it is possible to evaluate the edges contained in the image and, therefore, the level of blurriness.
- Fast Fourier Transform (FFT) [9]; when the blur increases, the number of high-frequency components decreases. We exploited this property by computing the FFT, removing the low frequencies, and finally computing the mean value of the magnitude representation: the smaller the mean, the blurrier the image.
- NIQE (Natural Image Quality Evaluator) [25] and PIQUE (Perception-based Image QUality Evaluator) [34] measurements. In particular, NIQE exploits the measurable deviations from statistical regularities typical of natural and undistorted images, while PIQUE extracts local fea-

tures by considering only perceptually significant spatial regions, as a human observer would do.

3.2.2 Full-Reference Measures

FR quality assessment measures are employed to compare an image under analysis with a reference one. We proceeded accordingly, using two adjacent frames in the observed time window.

- Signal-To-Noise Ratio (PSNR), which represents the expression for the ratio between a signal's maximum power and the noise's power applied to the same signal. Several studies analyze its effectiveness in evaluating the quality of a compressed image with respect to the original uncompressed one [17].
- Mean Squared Error (MSE) computes the error between the distorted image and the original one at the pixel level [3].
- Structural Similarity Index Measure (SSIM) [35] focuses on the similarity between two images, trying to predict their perceived quality.
- Pearson correlation coefficient (PCC) [7] measures the linear correlation between the images.
- The ”Fréchet Inception Distance” (FID) method proposed in [16], originally used to measure the similarity of produced pictures to genuine ones, to assess the effectiveness of GANs (Generative Adversarial Networks) in image generation.

3.3. Creation of the Feature Matrices

Starting from the sequence of N_f frames related to a predefined time window of the video under analysis, the matrices of features M_{or} and M_{hf} corresponding to original and high-pass filtered images are computed as follows. From each frame F_t, a set of N_p patches are extracted ($P_t = \{p_{1,t}, p_{2,t}, \ldots, p_{N_p,t}\}$) selecting the regions most prone to manipulation artifacts such as face, eyes, mouth, etc. (see Fig. 3). The features related to the frame under analysis F_t are then computed by exploiting the quality measures defined in the previous sections. Specifically, considering the set of patches P_t and a set of N_q^{NR} No-Reference (NR) quality measures $Q^{NR} = \{Q_1^{NR}, Q_2^{NR}, \ldots, Q_{N_q^{NR}}^{NR}\}$ the following matrix of features M^{NR} is computed:

$$M_{r,t}^{NR} = Q_k^{NR}(p_{i,t}) \tag{1}$$

with $p_{i,t} \in P_t$, $Q_k^{NR} \in Q^{NR}$, $r = N_p(k-1)+i$, $i = 1, \ldots, N_p$ and $k = 1, \ldots, N_q^{NR}$.

Later, taking into account pairs of corresponding patches related to consecutive frames, a set of N_q^{FR} Full-Reference (FR) quality measures $Q^{FR} = \{Q_1^{FR}, Q_2^{FR}, \ldots, Q_{N_q^{FR}}^{FR}\}$ can be applied to compute a matrix of features M^{FR} as follows:

$$M_{r,t}^{FR} = Q_k^{FR}(p_{i,t}, p_{i,t+1}) \tag{2}$$

with $p_{i,t} \in P_t, p_{i,t+1} \in P_{t+1}, Q_k^{FR} \in Q^{FR}$, $r = N_p(k-1)+i$, $i = 1, \ldots, N_p$ and $k = 1, \ldots, N_q^{NR}$.

FR quality measures can also be applied to compare patches of the same frame F_t. Specifically, in the proposed solution, a further feature matrix M^{FRSF} has been built comparing patches related to the left eye and the right one.

$$M_{r,t}^{FRSF} = Q_k^{FR}(p_{i,t}, p_{j,t}) \tag{3}$$

where $p_{i,t}, p_{j,t}$ are the left and right eye patches belonging to P_t, $r = k$, $k = 1, \ldots, N_q^{FR}$.

Finally, the vertical concatenation of the computed matrices $M^{NR} \in \mathbb{R}^{(N_p * N_q^{NR}) \times N_f}$, $M^{FR} \in \mathbb{R}^{(N_p * N_q^{FR}) \times N_f}$, $M^{FRSF} \in \mathbb{R}^{N_q^{FR} \times N_f}$ allow us to generate the feature matrix $M_{or} \in \mathbb{R}^{(N_p * N_q^{NR} + (N_p+1) * N_q^{FR}) \times N_t}$. Note that starting from the high-pass filtered version of the same video, applying (1), (2), (3), the M_{hr} feature matrix can also be derived.

Considering eight patches for each frame ($N_p = 8$) as depicted in Fig. 3, five No-Reference and Full-Reference quality measures ($N_q^{NR} = 5$, $N_q^{FR} = 5$) as described in Section 3.2, two 85 × N_f matrices of features (M_{or} and M_{hr}) are computed.

3.4. Neural Network Architecture

The feature matrices M_{or} and M_{hf} are the input of two CNNs with the same architecture. Note that our goal is not to propose a new neural network architecture but a method that could be applied to several, even more complex, models. Fig. 5 shows the architecture, which is made up of two 2D convolution layers followed by max-pooling layers and a dropout to avoid the overfitting problem. A dense layer for classification that ends with 2 neurons completes the network.

According to the cross-entropy cost function, their outcomes are interpreted as the probability of each sample belonging to the deepfake class. In the remainder of the paper, these values are referred to as scores and indicated with the terms s_{or} and s_{hf}, respectively.

3.5. Fusion of the Scores

It is well known that techniques like ensembling and multimodal or uni-modal fusion approaches are able to enhance the generalization capabilities of the systems and deal with both intra-class variations and inter-class similarities [27]. It is possible to apply fusion at different levels within a classification system: sensor, feature, score, and decision level. Also, in the field of deepfake detection, the use of such techniques is becoming increasingly popular, with the goal of attenuating the lack of generalization capabilities common to those systems [5, 31]. As in [8], we try to exploit the complementarity of the two systems by applying several score-level rules. In our case, the hypothesis is that

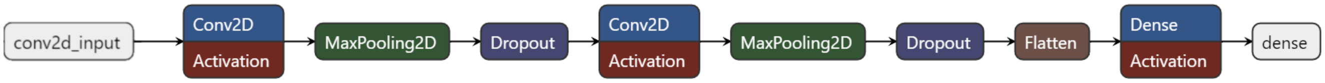


Figure 5. CNN architecture adopted for the experiments. Two 2D convolutional layers, followed by as many max-pooling layers, precede the dense layer and the two final neurons used for classification.

the quality measures extracted from the original image and the high-pass filtered one are influenced by different factors: it is likely that in the case of the filtered image, the measures are more influenced by details rather than its global appearance. To confirm the presence of such complementarity, we investigated some of the most promising rules described in [8]. In particular, we selected one non-parametric rule, namely the simple average. Let s_{or} and s_{hf} be the outcomes of the individual classifiers, namely, the one trained over the quality measures extracted from the original image and the filtered image, respectively.

The fusion of the such scores is obtained as follows:

$$s_{avg} = \frac{s_{or} + s_{hf}}{2} \tag{4}$$

We also considered a parametric rule, namely weighted average:

$$s_{Aavg} = \frac{A_{or} \cdot s_{or} + A_{hf} \cdot s_{hf}}{A_{or} + A_{hf}} \tag{5}$$

Values A_{or} and A_{hf} are the accuracy of the single model on a validation set. To calculate such accuracies, a threshold equal to 0.5 was used (if the sample obtained a score greater or equal to this value, it was classified as fake, real otherwise).

Finally, we used a logit-based perceptron as a stacked fusion rule, whose inputs are the outcomes of the individual deepfake detectors. Perceptron was trained by minimizing a cross-entropy cost function over a validation set. Its outcome is the final score.

4. Experimental Protocol

4.1. Dataset

In order to verify the validity of the proposed method, the chosen dataset was FaceForensics++ [28]. The peculiarity of this dataset is that it contains 1000 videos downloaded from YouTube and manipulated with 5 different methods. Therefore, it allows us to test our system on the basis of different manipulations, starting from the same original video. Although the applied techniques are different, the dataset contains two main types of manipulation: reenactment (two methods) and face-swap (three methods). In particular, among face-swap techniques, we find Faceswap[2], a graphics-based approach, Deepfakes[3], a learning-based approach built on an autoencoder architecture, and FaceShifter [21], based on two different networks for the face replacement and handling of the face occlusions. The last two reenactment methods are a learning-based and a graphics-based approach named NeuralTextures [29] and Face2Face [30], respectively.

Regarding the subdivision into training, validation, and test set, we used the official one proposed by the dataset authors in [28]: 720 videos for training, 140 for validation, and 140 for testing.

4.2. Training and Testing Protocol

To evaluate the proposed approach, we extracted $N_f = 300$ subsequent frames from each video of the considered dataset, corresponding to about ten seconds. Thus, we have one sample per video. An experimental investigation for $N_f < 300$, which we do not report in this paper for the sake of space, suggested a positive correlation between the system's performance and N_f; consequently, we opted for the maximum possible value, being $N_f = 300$ the size of the smallest available video.

After computing the feature matrices related to both the original video frames and the high-pass filtered ones, we trained the two CNNs[4] described in Section 3.4. For the convolution layers, a set of 32 filters with a kernel size equal to 5 has been set, using the ”same” padding technique. After each max-pooling layer (with a size of 3), a dropout with a fraction of 0.5 was applied. The training was conducted with a batch size of 64 and setting an early stopping based on a patient of 10 on the validation loss. For both networks, we chose *Adam* as an optimizer with a default learning rate of 0.001, *binary focal cross-entropy* as the loss function because of the disparity in the number of real and deepfake samples.

According to this paper's claims and the general approach in the literature, we subdivided our experimental setting into two protocols:

- Intra-manipulation: manipulation methods in the test set have also been employed in the training and validation sets, given that there is no sample overlap among such sets. In particular, two experiments have been conducted:

[2] https://github.com/MarekKowalski/FaceSwap

[3] https://github.com/deepfakes/faceswap

[4] https://github.com/fchollet/keras

Table 1. Intra-manipulation results in terms of False Negative Rate (FNR), False Positive Rate (FPR), Accuracy (Acc.), and Area Under the ROC Curve (AUC) on the FF++ dataset: the system has been trained on all the available methods and then tested on all the methods together (All), and on each method individually. The methods are Deepfakes (DF), Face2Face (F2F), FaceSwap (FS), NeuralTextures (NT), and FaceShifter (FSh). The best accuracy results are highlighted in bold.

Test Method	**Original Image**				**HF Image**				**Fusion**											
									Simple Average				**Accuracy-based**				**Perceptron**			
	FNR	**FPR**	**Acc.**	**AUC**	**FNR**	**FPR**	**Acc.**	**AUC**	**FNR**	**FPR**	**Acc.**	**AUC**	**FNR**	**FPR**	**Acc.**	**AUC**	**FNR**	**FPR**	**Acc.**	**AUC**
All	0.54	3.53	98.94	99.95	0.37	5.49	98.74	99.94	0.33	2.94	99.22	99.98	0.33	2.94	99.22	99.98	0.33	1.37	**99.49**	99.97
DF	0.40	6.08	96.72	99.90	1.21	4.31	97.21	99.89	0.40	1.57	**99.00**	99.97	0.40	1.57	**99.00**	99.97	0.40	2.16	98.71	99.98
F2F	0.78	2.55	98.33	99.95	0.98	7.25	95.88	99.81	0.98	1.57	**98.73**	99.94	0.98	1.57	**98.73**	99.94	0.98	1.57	**98.73**	99.94
FS	0.00	3.53	98.11	99.96	0.00	4.90	97.37	99.98	0.00	1.37	**99.26**	99.99	0.00	1.37	**99.26**	99.99	0.00	1.57	99.16	99.99
NT	0.22	6.86	96.29	99.94	0.00	5.29	97.22	99.98	0.00	2.55	98.66	100.00	0.00	2.55	98.66	100.00	0.00	1.96	**98.97**	100.0
FSh	0.00	3.53	98.25	99.99	0.00	5.10	97.48	99.96	0.00	1.57	99.22	100.00	0.00	1.57	99.22	100.00	0.00	1.37	**99.32**	99.99

Table 2. Results in terms of False Negative Rate (FNR), False Positive Rate (FPR), Accuracy (Acc.), and Area Under the ROC Curve (AUC) on the FF++ dataset considering one deepfake category at a time: the system has been trained on one specific category between face-swap or reenactment and tested on the same one.

Test Manipulation	**Original Image**				**HF Image**				**Fusion**											
									Simple Average				**Accuracy-based**				**Perceptron**			
	FNR	**FPR**	**Acc.**	**AUC**	**FNR**	**FPR**	**Acc.**	**AUC**	**FNR**	**FPR**	**Acc.**	**AUC**	**FNR**	**FPR**	**Acc.**	**AUC**	**FNR**	**FPR**	**Acc.**	**AUC**
Face Swap	0.89	3.73	98.37	99.90	0.69	7.45	97.56	99.91	1.03	1.18	**98.93**	99.97	1.03	1.18	**98.93**	99.97	1.31	0.98	98.78	99.96
Expression Swap	1.75	0.78	98.58	99.89	1.24	2.94	98.18	99.84	1.03	0.59	99.12	99.93	0.72	0.59	**99.32**	99.93	1.24	0.59	98.99	99.92

(1) using the sets as described above, and (2) grouping the samples per deepfake generation category, namely, face swap and expression swap, and investigating the performance when we hypothesize being attacked by only one known category.

- Cross-manipulation: there is no overlap between the manipulation methods employed to generate training and test videos. Validation set is from methods characterizing the training set. The system is trained on four of five manipulations and tested on the remaining one. For example, we trained the system on Deepfakes, Face2Face, FaceSwap, and FaceShifter and tested it on NeuralTextures. As done in the previous scenario, we also investigated the performance by grouping the methods per deepfake category: face and expression swap, training on the former and testing on the latter, and *vice versa*.

5. Experimental Results

In this section, we present the results obtained from the experimental setup outlined earlier. Specifically, we show the performance of the intra-manipulation protocol in Subsection 5.1, followed by the results obtained in the cross-manipulation protocol in Subsection 5.2.

In all cases, reported values are related to a decision threshold over the final scores aimed at maximizing the accuracy estimated on the related validation set. We computed the percentage of misclassified deepfakes (false negative rate, FNR) and real samples (false positive rate, FPR) and the percentage of correctly classified samples (accuracy, Acc).

The parameter aimed at giving the overall view of the system's performance is the Area Under the ROC curve (AUC).

5.1. Intra-Manipulation Analysis

Table 1 reports that the accuracy is high (99.50% for the most-performing fusion method), and the Area Under the ROC Curve (AUC) values suggest slight differences if the decision threshold would vary around the chosen one. This confirms that quality measures can effectively describe artifacts in the observed deepfake generation methods. The difference in performance between the model based on the original image (”Original Image”) and the model based on the high-pass filtered image (”HF Image”) may suggest that even detected artifacts are different. For example, in the case of the DF method (Table 1, third row), the correct detection rate is 96.70% on original images and 97.20% on high-pass filtered ones. The fusion rules further improve the result. Even the simple average leads to an overall accuracy of 99%. This is the first confirmation of the complementarity of the two models and the diversity we may suppose on the artifacts detected.

In the second experiment, only one deepfake category was used at a time for training and testing. Table 2 makes it noticeable that, once again, the fusion of the two single models brings remarkable benefits. It is worth noting that individual classifiers exhibit a high accuracy, but there is a difference between OR and HF modules. The complementarity is still exploited by fusion.

Worth remarking, the fusion greatly reduces the false positive rate, while keeping the false negative rate around 0.3-1%. This means that, besides the property of keeping very low the probability of misclassifying a manipulated sample, detected artifacts, which can be misled in OR and HF images, are so complementary that the individual scores s_{or} and s_{hf} are strongly uncorrelated.

Table 3. Cross-manipulation results in terms of False Negative Rate (FNR), False Positive Rate (FPR), Accuracy (Acc.), and Area Under the ROC Curve (AUC) on the FF++ dataset: the system has been trained on 4 of the 5 available methods, and tested on the remaining one. The methods are Deepfakes (DF), Face2Face (F2F), FaceSwap (FS), NeuralTextures (NT), and FaceShifter (FSh). The best results are highlighted in bold.

Test Method	Original Image				HF Image				Fusion											
									Simple Average				Accuracy-based				Perceptron			
	FNR	FPR	Acc.	AUC	FNR	FPR	Acc.	AUC	FNR	FPR	Acc.	AUC	FNR	FPR	Acc.	AUC	FNR	FPR	Acc.	AUC
DF	0.81	3.73	97.71	99.91	1.21	3.92	97.41	99.83	0.81	1.57	98.81	99.95	0.81	1.57	98.81	99.95	0.81	1.37	**98.91**	99.95
F2F	0.39	7.65	95.98	99.88	2.55	2.16	97.65	99.82	1.57	1.96	98.24	99.93	1.57	1.96	98.24	99.93	1.18	0.98	**98.92**	99.91
FS	0.68	5.10	96.95	99.95	0.00	5.88	96.84	99.98	0.23	1.18	**99.26**	100.00	0.23	1.18	**99.26**	100.00	0.00	1.37	**99.26**	100.00
NT	0.00	6.86	96.39	99.99	0.22	4.71	97.42	99.98	0.00	1.37	99.28	100.00	0.00	1.37	99.28	100.00	0.00	0.78	**99.59**	100.00
FSh	0.00	5.10	97.48	99.97	5.96	5.49	94.27	98.13	2.88	0.59	**98.25**	99.93	2.88	0.59	**98.25**	99.93	2.31	2.16	97.77	99.88

Table 4. Results in terms of False Negative Rate (FNR), False Positive Rate (FPR), Accuracy (Acc.) and Area Under the ROC Curve (AUC) on the FF++ dataset, training on one deepfake category (face-swap or expression-swap) and testing on the other.

Test Manipulation	Original Image				HF Image				Fusion											
									Simple Average				Accuracy-based				Perceptron			
	FNR	FPR	Acc.	AUC	FNR	FPR	Acc.	AUC	FNR	FPR	Acc.	AUC	FNR	FPR	Acc.	AUC	FNR	FPR	Acc.	AUC
Face Swap	1.58	1.37	98.47	99.90	5.77	4.31	94.61	98.23	2.34	1.18	97.96	99.94	2.34	1.18	97.96	99.94	1.44	0.59	**98.78**	99.96
Expression Swap	1.13	2.94	98.24	99.91	1.24	3.73	97.91	99.80	0.93	0.98	**99.05**	99.94	0.93	0.98	**99.05**	99.94	0.93	1.57	98.85	99.94

5.2. Cross-Manipulation Analysis

This is the most realistic scenario, where a performance decrease is usually observed.

A first set of experiments was carried out by training each system on 4 out of 5 techniques and testing the remaining one. Results are reported in Table 3.

First of all, looking at the performance of the individual classifiers, it is very noticeable how the performance is still high. The performance over never-seen-before manipulations adds that individual classifiers are able to detect artifacts common to the investigated deepfake generation methods.

We also noticed that fixing the decision threshold means staying around the so-called $0\% - 1\%$FNR operational working point. This means that the observed performance is related to a case where detecting correctly deepfake samples with a low probability of failure leaves the possibility that some real samples are misclassified with higher frequency. However, the fusion strongly reduces this occurrence without a substantial variation to the reference operational point above: we stay around the $0\% - 1\%$FNR operational point, with a significant decrease in the related FPR. Moreover, the overall accuracy ensures the performance is way better than that of individual classifiers.

Among other deepfake generation methods, FaceShifter tests turn out to be the most particular. The last row of Table 3 points out that the model trained on HF images performs lower than others, reaching only 94.3%. We may motivate this by hypothesizing that artifacts of FaceShifter are less evident in HF images. Moreover, the FaceShifter method has recently been included in the FF++ dataset, constituting a more sophisticated deepfake creation technique. This apparent correlation between the novelty of FaceShifter and the reduced performance in HF images points out the rapidity with which deepfake generation methods are refining their characteristics over time. However, the fusion of the models leads to an accuracy of even 98.2%, thus overcoming this issue.

In the second experiment, one deepfake category was used at a time for training, while the other was used for testing. Table 4 confirms the advantages of the fusion between the two single models, still highlighting the lower performance of the HF module than the OR one.

5.3. Comparison with the State of the Art

Table 5 reports the comparison in terms of the AUC of the proposed approach (”Ours” rows) with the Xception network, which is considered as a sort baseline [14]. We also selected four reference methods that exhibited the best performance over the literature and adopted the same experimental protocol. For completeness, we added some characteristics of each method: the input to the network, if a preliminary feature extraction step is performed, the number of network parameters (if available), the architecture typology, and the use of auxiliary datasets.

The first thing to notice is that our architecture is much lighter than the reported ones, including Xception net. Second, a certain complexity in the architecture typology characterizes the most competitive approaches. Third, they are often coupled with an auxiliary dataset, probably due to the huge number of parameters to be set. Fourth, no feature extraction step is performed: the cropped face or part of it is simply the input to the selected architecture, whose size is motivated by the large number of features to process.

Still, the proposed method is based on the modeling of artifacts by a set of quality measurements for video-based deepfake detection; it includes a light architecture and no additional data for training (for which a careful selection is often necessary); nevertheless, it showed to be in the same

Table 5. Comparison of the proposed approach with other state-of-the-art methods. The columns DF-NT report the AUC value obtained by training on 3 out of 4 categories between Deepfakes (DF), Face2Face (F2F), FaceSwap (FS), and NeuralTextures (NT) and tested on the remaining one. FaceShifter (FSh) is kept out of training, as done in the papers used for comparison, in this case, the system is trained on the other 4 manipulations. The last column reports the average AUC of all five methods. About backbone columns, MS-TCN corresponds to Multi-scale Temporal Convolutional Network [22], RN18 corresponds to ResNet-18 [15], R(2+1)D-18 corresponds to ResNet-18 with (2+1)D convolutions [33], and CSN corresponds to Channel-Separated Convolutional Network [32].

Model	**Input**	**Feature Extraction**	**Backbone**	**Auxiliary Dataset**	**Parameters**	**Test Method - AUC (%)**					
						DF	**F2F**	**FS**	**NT**	**FSh**	**AVG**
Xception [28]	Face Crop	No	XceptionNet	No	22.8M	93.90	86.80	51.20	79.70	72.00	76.72
LipForensics [14]	Mouth Crop	No	3D Convolution + RN18 + MS-TCN	Yes	24.8M	99.70	99.70	90.10	99.10	97.10	97.14
AV DFD [38]	Face Crop + Audio	No	R(2+1)D-18	No	-	99.99	99.79	90.48	98.32	-	97.15
Zhao et al. [37]	Mouth Crop	No	3D Convolution + RN18 + Transformer	Yes	36M	98.50	98.30	91.90	96.40	97.80	96.58
RealForensics [13]	Face Crop	No	CSN + RN18	Yes	21.4M	**100.00**	99.70	97.10	99.20	99.70	99.14
Ours - Original	85x300 Feature Matrix	Yes	5 Layers CNN	No	45.5K	99.84	99.90	99.96	99.97	**99.88**	99.91
Ours - HF					45.5K	99.69	99.69	99.94	99.96	98.87	99.63
Ours - Simple Average					90K	99.94	**99.91**	**99.99**	**99.99**	99.87	**99.94**

performance rank as the best reported one.

6. Conclusions

In this paper, we described a novel approach to model artifacts in digitally manipulated videos where facial deepfakes are taken into account.

We showed that using a battery of quality measurements applied to the detected facial region in each frame or subsequent frames allows to point out the presence of artifacts generated by deepfakes methods. Although these artifacts are different depending on the spatial frequency range under exploration (as we showed on original and high-pass filtered frames), they are common to different manipulation methods. We confirmed this claim by computing a feature matrix of quality measurements on the face detected in a video sequence. Specifically, we focused on the facial regions where the movements have shown to be difficult to reproduce in detail; consequently, the manipulation process leads to (un)perceptible incoherences.

We tested the proposed approach on a well-known benchmark dataset, including different manipulation approaches that represented two main categories: face-swap and expression-swap. The light architecture adopted and the noticeable help of fusion rules confirmed our proposal's effectiveness, especially compared to other state-of-the-art solutions.

In our opinion, the proposed approach is an informative way to model artifact detection in digitally manipulated videos. Further works will rely on the extension of the experiments and an in-depth study of the proposed feature matrix, even including data with different video characteristics, such as in terms of compression, also to make it possible to evaluate the proposed approach in a cross-dataset scenario. Similarly, an explainability analysis should be included to highlight the actual relevance of the single pairs of patches and quality measures. The goal is to use it as the input to an intrinsically explainable-to-human architecture, able to point out what artifacts are detected when verifying the presence of facial deepfakes in a video sequence.

Acknowledgment

This work is supported by SERICS (PE00000014) under the Italian Ministry of University and Research (MUR) National Recovery and Resilience Plan funded by the European Union - NextGenerationEU and within the PRIN 2022 PNRR - BullyBuster 2 – the ongoing fight against bullying and cyberbullying with the help of artificial intelligence for the human wellbeing (CUP: P2022K39K8).

References

[1] M.A. Amin, Y. Hu, H. She, J. Li, Y. Guan, and M.Z. Amin. Exposing deepfake frames through spectral analysis of color channels in frequency domain. In *11th Int. Work. on Biometrics and Forensics (IWBF)*, pages 1–6, 2023. 2

[2] R. Arun and S.S. Verma. Image forgery detection using structural similarity and morphological tools. In *2022 10th Int. Conf. on Reliability, Infocom Technologies and Optimization (Trends and Future Directions) (ICRITO)*, pages 1–5, Oct 2022. 2

[3] I. Avcibas, B. Sankur, and K. Sayood. Statistical evaluation of quality measures. j electron imaging. *J. Electronic Imaging*, 11:206–223, 04 2002. 4

[4] R. Bansal, G. Raj, and T. Choudhury. Blur image detection using laplacian operator and open-cv. In *2016 Int. Conf. System Modeling Advancement in Research Trends (SMART)*, pages 63–67, 2016. 3

[5] N. Bonettini, E. Cannas, S. Mandelli, L. Bondi, P. Bestagini, and S. Tubaro. Video face manipulation detection through ensemble of cnns. In *2020 25th Int. Conf. on Pattern Recognition (ICPR)*, pages 5012–5019, Jan 2021. 1, 4

[6] L. Chai, D. Bau, S.-N. Lim, and P. Isola. What makes fake images detectable? understanding properties that generalize. In H. Bischof A. Vedaldi, T. Brox, and J.-M. Frahm, editors, *Computer Vision – ECCV 2020*, pages 103–120, Cham, 2020. Springer International Publishing. 2

[7] I. Cohen, Y. Huang, J. Chen, and J. Benesty. Pearson correlation coefficient. *Noise reduction in speech processing*, pages 1–4, 2009. 4

[8] S. Concas, S.M. La Cava, G. Orrù, C. Cuccu, J. Gao, X. Feng, G.L. Marcialis, and F. Roli. Analysis of score-level fusion rules for deepfake detection. *Applied Sciences*, 12(15), 2022. 4, 5

[9] K. De and V. Masilamani. Image sharpness measure for blurred images in frequency domain. *Procedia Engineering*, 64:149–158, 2013. International Conference on Design and Manufacturing (IConDM2013). 3

[10] Z. Deng, B. Zhang, S. He, and Y. Wang. Deepfake detection method based on face edge bands. In *9th Int. Conf. on Digital Home (ICDH)*, pages 251–256, Oct 2022. 2

[11] R. Durall, M. Keuper, and J. Keuper. Watch your up-convolution: Cnn based generative deep neural networks are failing to reproduce spectral distributions. In *IEEE/CVF Conf. on Computer Vision and Pattern Recognition (CVPR)*, pages 7887–7896, 2020. 2

[12] J. Galbally, S. Marcel, and J. Fierrez. Image quality assessment for fake biometric detection: Application to iris, fingerprint, and face recognition. *IEEE Trans. on Image Processing*, 23(2):710–724, 2014. 2

[13] A. Haliassos, R. Mira, S. Petridis, and M. Pantic. Leveraging real talking faces via self-supervision for robust forgery detection. In *IEEE/CVF Conf. on Computer Vision and Pattern Recognition (CVPR)*, pages 14930–14942, 2022. 2, 8

[14] A. Haliassos, K. Vougioukas, S. Petridis, and M. Pantic. Lips don't lie: A generalisable and robust approach to face forgery detection. In *IEEE/CVF Conf. on Computer Vision and Pattern Recognition (CVPR)*, pages 5037–5047, 2021. 2, 7, 8

[15] K. He, X. Zhang, S. Ren, and J. Sun. Deep residual learning for image recognition. In *IEEE Conf. on Computer Vision and Pattern Recognition*, pages 770–778, 2016. 8

[16] M. Heusel, H. Ramsauer, T. Unterthiner, B. Nessler, and S. Hochreiter. Gans trained by a two time-scale update rule converge to a local nash equilibrium. In *Proceedings of the 31st International Conference on Neural Information Processing Systems*, NIPS'17, page 6629–6640, Red Hook, NY, USA, 2017. Curran Associates Inc. 4

[17] Q. Huynh-Thu and M. Ghanbari. Scope of validity of psnr in image/video quality assessment. *Electronics Letters*, 44:800–801(1), June 2008. 4

[18] M.T. Jafar, M. Ababneh, M. Al-Zoube, and A. Elhassan. Forensics and analysis of deepfake videos. In *2020 11th International Conference on Information and Communication Systems (ICICS)*, pages 053–058, April 2020. 2

[19] Y. Ju, S. Jia, L. Ke, H. Xue, K. Nagano, and S. Lyu. Fusing global and local features for generalized ai-synthesized image detection. In *IEEE Int. Conf. on Image Processing (ICIP)*, pages 3465–3469, Oct 2022. 2

[20] G. Li, Y. Cao, and X. Zhao. Exploiting facial symmetry to expose deepfakes. In *2021 IEEE International Conference on Image Processing (ICIP)*, pages 3587–3591, 2021. 2

[21] L. Li, J. Bao, H. Yang, D. Chen, and F. Wen. Advancing high fidelity identity swapping for forgery detection. In *IEEE/CVF Conf. on Computer Vision and Pattern Recognition (CVPR)*, pages 5073–5082, 2020. 5

[22] B. Martinez, P. Ma, S. Petridis, and M. Pantic. Lipreading using temporal convolutional networks. In *IEEE Int. Conf. on Acoustics, Speech and Signal Processing (ICASSP)*, pages 6319–6323. IEEE, 2020. 8

[23] Y. Mirsky and W. Lee. The creation and detection of deepfakes: A survey. *ACM Computing Surveys*, 54:1–41, 01 2021. 1

[24] A. Mittal, A.K. Moorthy, and A.C. Bovik. No-reference image quality assessment in the spatial domain. *IEEE Trans. on Image Processing*, 21(12):4695–4708, Dec 2012. 3

[25] A. Mittal, R. Soundararajan, and A.C. Bovik. Making a "completely blind" image quality analyzer. *IEEE Signal Processing Letters*, 20(3):209–212, 2013. 3

[26] P. Pravallika and K. Satya Prasad. Svm classification for fake biometric detection using image quality assessment: Application to iris, face and palm print. In *2016 Int. Conf. on Inventive Computation Technologies (ICICT)*, volume 1, pages 1–6, Aug 2016. 2

[27] A. Ross and K. Nandakumar. *Fusion, Score-Level*, pages 611–616. Springer US, Boston, MA, 2009. 4

[28] A. Rössler, D. Cozzolino, L. Verdoliva, C. Riess, J. Thies, and M. Niessner. Faceforensics++: Learning to detect manipulated facial images. In *IEEE/CVF Int. Conf. on Computer Vision (ICCV)*, pages 1–11, Oct 2019. 1, 3, 5, 8

[29] J. Thies, M. Zollhöfer, and M. Nießner. Deferred neural rendering: image synthesis using neural textures. *ACM Trans. Graph.*, 38(4), jul 2019. 5

[30] J. Thies, M. Zollhöfer, M. Stamminger, C. Theobalt, and M. Nießner. Face2face: real-time face capture and reenactment of rgb videos. *Commun. ACM*, 62(1):96–104, dec 2018. 5

[31] R. Tolosana, S. Romero-Tapiador, R. Vera-Rodriguez, E. Gonzalez-Sosa, and J. Fierrez. Deepfakes detection across generations: Analysis of facial regions, fusion, and performance evaluation. *Engineering Applications of Artificial Intelligence*, 110:104673, 2022. 2, 4

[32] D. Tran, H. Wang, L. Torresani, and M. Feiszli. Video classification with channel-separated convolutional networks. In *IEEE/CVF Int. Conf. on Computer Vision*, pages 5552–5561, 2019. 8

[33] D. Tran, H. Wang, L. Torresani, J. Ray, Y. LeCun, and M. Paluri. A closer look at spatiotemporal convolutions for action recognition. In *IEEE Conf. on Computer Vision and Pattern Recognition*, pages 6450–6459, 2018. 8

[34] N. Venkatanath, D. Praneeth, B. Maruthi Chandrasekhar, S.S. Channappayya, and S.S. Medasani. Blind image quality evaluation using perception based features. In *2015 21th Nat. Conf. on Communications (NCC)*, pages 1–6, 2015. 3

[35] Z. Wang, A.C. Bovik, H.R. Sheikh, and E.P. Simoncelli. Image quality assessment: from error visibility to structural similarity. *IEEE Trans. on Image Processing*, 13(4):600–612, April 2004. 4

[36] T. Zhang. Deepfake generation and detection, a survey. *Multimedia Tools and Applications*, 81(5):6259–6276, February 2022. 1

[37] H. Zhao, W. Zhou, D. Chen, W. Zhang, and N. Yu. Self-supervised transformer for deepfake detection. *arXiv preprint*, 2022. 2, 8

[38] Y. Zhou and S.-N. Lim. Joint audio-visual deepfake detection. In *IEEE/CVF Int. Conf. on Computer Vision (ICCV)*, pages 14780–14789, 2021. 2, 8