Temporal surface frame anomalies for deepfake video detection

Andrea Ciamarra
University of Florence, Florence, Italy
Universitas Mercatorum, Rome, Italy
andrea.ciamarra@unifi.it

Roberto Caldelli
CNIT, Florence, Italy
Universitas Mercatorum, Rome, Italy
roberto.caldelli@cnit.it

Alberto Del Bimbo
University of Florence, Florence, Italy
alberto.delbimbo@unifi.it

Abstract

Looking at a video sequence where a foreground person is represented is not as time ago anymore. Deepfakes have revolutionized our way to watch at such contents and nowadays we are more often used to wonder if what we are seeing is real or is just a mystification. In this context of generalized disinformation, the need for reliable solutions to help common users, and not only, to make an assessment on this kind of video sequences is strongly upcoming. In this paper, a novel approach which leverages on temporal surface frame anomalies in order to reveal deepfake videos is introduced. The method searches for possible discrepancies, induced by deepfake manipulation, in the surfaces belonging to the captured scene and in their evolution along the temporal axis. These features are used as input of a pipeline based on deep neural networks to perform a binary assessment on the video itself. Experimental results witness that such a methodology can achieve significant performance in terms of detection accuracy.

1. Introduction

Throughout history, the proliferation of disinformation has posed a significant challenge in our communication society. However, with the advent of deep learning (DL), this issue has escalated beyond control. The term "Deepfake" (DF) has become widely known and it is frequently encountered on the internet, on social networks and news platforms. It refers to content generated through DL models like text-to-image and GAN-based tools. Anyway, at the beginning, Deepfake specifically referred to the fabrication of convincingly realistic video sequences, achieved by manipulating facial expressions or swapping faces to alter the original video's context and significance. This specific application remains highly detrimental, as it seeks to deceive audiences by presenting distorted audio-visual narratives. On the other side of the barricade, there are the deepfake detectors that try to reliably detect such fabricated content. Typically, detection techniques exploit the traces left behind during the fake content generation process. These methods aim to uncover alterations in order to identify falsified content. Various features have been explored in literature for this purpose, inspecting images at the pixel level, employing frequency domain analysis or investigating temporal inconsistencies in the specific case of videos.

In this work, we precisely focus on this last issue and we propose a method which, by resorting to novel features based on the entire environmental conditions existing at the time of video capturing, analyses the temporal anomalies present within the sequence to make an assessment on its integrity. In fact, it is reasonable to presume that Deepfake creation alters this intrinsic information, particularly in its evolution during time, providing a basis for manipulation detection. Hopefully, the manipulation process should not be good enough to coherently reproduce frame-by-frame all these small relations that are dependent on physical elements present during scene capture, such as variations of the illumination sources, changes in the correlations between the different objects/surfaces and the camera due to their respective spatial positions in time and so on.

In order to exploit such possible inconsistencies, we have explored novel features which provide a thorough pixel-level description of the diverse surfaces involved during sequence capture. These features are named surface frames (**SF**) and they have been introduced in [24] with the objective, through a learning-based approach, to estimate 2DoF (degrees of freedom) camera orientation from a single RGB image. Though designed for indoor scenes, they give a complete and detailed characterisation of the observed scene (see Sec. 3.1). In particular, these surface

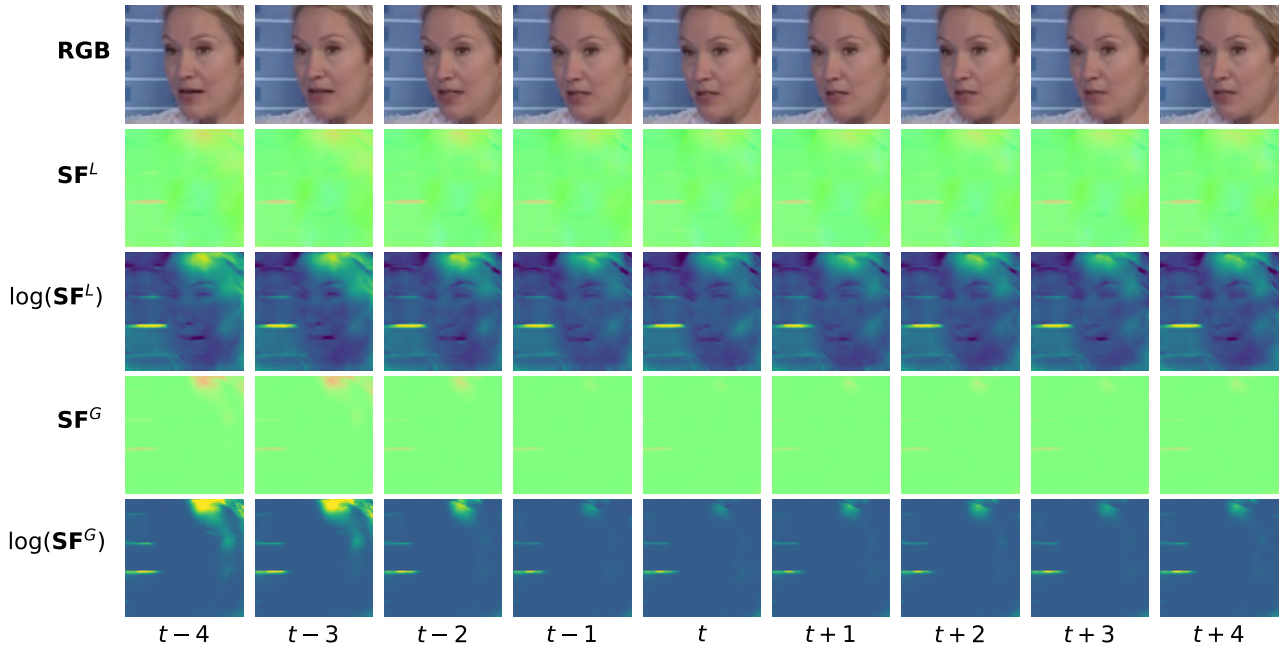

Figure 1. A visual example of the surface frames (local $\mathbf{SF}^L$ and global $\mathbf{SF}^G$) for a group of consecutive frames; we also calculate the logarithm of each surface frames, i.e. $log(\mathbf{SF}^L)$ and $log(\mathbf{SF}^G)$ for sake of visibility.

frames provide both a global ($\mathbf{SF}^G$) and a local ($\mathbf{SF}^L$) description of the captured scene. A sample visualisation of such surface frames (both global and local) for a group of consecutive frames is pictured in Fig. 1 where their logarithmic function is also depicted for sake of visibility. The considered video shows a fake case in which the facial expressions have been altered. We can observe that thoroughly temporal variations can be highlighted while processing consecutive images in terms of surface frames. For instance, the upper part of the forehead is changing through time for both the surface representations (see $log(\mathbf{SF}^L)$ and $log(\mathbf{SF}^G)$, even though the face positions are almost steady. Such scene characteristics can be noted using surface frame representations and this will happen for any other kind of video and content. However, our purpose is to exploit these temporal evolving surface details that could significantly change between real and fake videos, since the deepfake alterations are often applied frame-by-frame. We believe that such discrepancies can be temporally observed using recurrent architectures like LSTMs, able to retain spatio-temporal contexts such as sequential surface frames. In other words, in this paper the evolution in time of such surface frames has been investigated particularly to highlight possible anomalies in deepfake videos with respect to pristine sequences.

The main contributions of the present work are the following:

(i) we propose the temporal evolution of surface frames (**t-SF**) to provide a pixel-level description of the diverse surfaces involved during sequence acquisition.

(ii) we show how such time surface frames can be adopted to exploit inconsistencies and consequently to distinguish between pristine and fake video sequences.

(iii) we carry out an experimental analysis by taking into account of various deepfake forgeries and neural network architectures, achieving a significant detection accuracy up to 90% on average.

This is the paper layout: Sec. 2 gives a general overview of the related works in particular dealing with video deepfake detection, while Sec. 3 is dedicated to the proposed methodology with some specific details about surface frames provided in Sec. 3.1. Sec. 4 presents and discusses some of the main experimental results, and Sec. 5 draws conclusions giving some ideas for eventual future works.

2. Related Works

The capability to generate synthetic contents is one hot topic in the field of Deepfakes. The last decade has registered dramatic advancements in generative models, e.g. [1, 4, 9], that nowadays is making possible to recreate or manipulate the information, such as images and videos. It is notewor-

thy that altering human faces mines the privacy and the security of people. Several implementations [16] have been designed to change facial expressions or directly substituting faces between a source image and a target one, that can also be done e.g. via CNNs [12], conditional GANs [18] or based on facial landmark alignment [3]. To contrast this phenomenon, Deepfake detection methods [17] have arisen as a mean to prevent misinformation using deep learning. To do that, existing works look for inconsistencies that are inherently left into the content during the fake manipulation process. Such anomalies can be analysed either at image or video level. Most of the existing methods addresses the detection task as a binary classification problem. Several works [15] have posed the detection problem by looking for dissimilarities within single images, in terms of specialised features, e.g. depth map [14], even exploiting subtle local attention mechanism [28] or feature consistency across regions [29]. More recently, other approaches have proposed to employ the spatial and frequency domains via a graph learning [23], or to uncover common forgery features in a multitask learning strategy [26], or to consider differences between the explicit (source) and implicit (target) identity of face in swapping manipulations [10], or even to look for subtle geometrical aspects related to facial regions [5].

Although some of the literature works facing the detection problem by processing single images, video-level discrepancies have been also investigated. Video deepfake approaches exploit spatio-temporal contexts by means of recurrent architectures, e.g. RNNs and LSTMs, so to pay attention to visual inconsistency across consecutive frames. To this aim, Sabir et al. [21] propose to use a CNN-RNN approach, in order to catch temporal discrepancies, as manipulations are performed frame-by-frame. Also LSTMs have been used to detect fakes by exploiting consecutive frame-level features [7]. Rather than elaborating features from the entire face images, Li et al. [13] observed the eye blinking frequency over the time in order to catch possibly crucial anomalies in the eye area across consecutive faces. Observing that deepfakes are generated by splicing synthesized face region into the original image, Yang et al. [27] propose to reveal errors in landmark locations from head poses estimated, by using the 2D landmarks in real and fake parts of the face. Also connected with the motion artifacts over the video frames, Caldelli et al. [2] demonstrate that synthetic motion patterns can be found in fake sequences by looking into abnormal optical flows calculated over the faces. Differently, Sun et al. [22] propose to look for precise geometric features that can reveal subtle unnatural expressions or facial organ movements using sequences of facial landmarks, by exploiting temporal information from a two-stream RNN. Xu et al. [25] transform a video clip into a predefined layout in order to retain relevant spatio-temporal contexts in video frames. Gu et al. [6] integrate both spatial and temporal features in a unified 2D CNN framework. Zheng et al. [30] propose to restrict the network capability to classify by temporal-related artifacts rather than looking into spatial artifacts for detection to improve generalisation. Leveraging of semantically high-level irregularities through time has been even investigated on mouth movements [8]. Raza et al. [19] introduce a unified multimodal approach, by analysing both visual and audio streams simultaneously. Different than [10, 23, 26], we deal with deepfake video detection in a single modality approach, avoiding the need of processing additional data [19]. Considering that temporal inconsistency relates to video sequences, we propose to focus on anomalies that are connected with the surface frames estimated over the time. Instead of analysing 2D landmark positions [22] or limiting our approach to just look into discrepancy within single images, as done e.g. in [14, 28, 29], we introduce to investigate geometrical temporal anomalies in terms of surface frames. Such characteristics can exploit the per-pixel facial properties that are described according to orientations of unit vectors, namely normals, tangents and bitangents, with respect to certain points of view. In particular, we consider to analyse spatio-temporal contexts of consecutive surface frames extracted from an input video sequence, either in the local camera coordinates $\mathbf{t}\text{-}\mathbf{SF}^L$ or in the global up-right coordinate system $\mathbf{t}\text{-}\mathbf{SF}^G$ (see Fig. 2). These two diverse scene aspects that are acquired in pristine images can be inherently affected by manipulations, that could show with different variability through time with respect to the fake content. In Sec. 3 we will provide more details about our proposed approach and the surface frames.

3. Proposed method

When a deepfake manipulation is applied to a pristine video to alter the face and/or the expressions of the represented person, this has to be consistently performed for all the frames belonging to the sequence or, at least, for a consecutive group of pictures related to that intended modification. It is expected that such alteration should impact, in some way, on the original time flow of the different correlations existing among frames temporally close; such correlations are basically due to the way the acquisition has happened. In fact, at the time of video capturing diverse components contribute to the construction of the final grabbed video that is then represented in terms of RGB pixels. All of this is due to the sources of illumination, to the presence of different objects and surfaces in those specific positions and, above all, with respect to the camera that is recording the video. Furthermore, the shadows, the partial occlusions and the light reflections can generate an univocal intrinsic fingerprint that is embedded within the recorded video file. In addition to this and most importantly, it is consequential to think that this will have a particular evolution in time throughout the frames. It is plausible to deem that deepfake

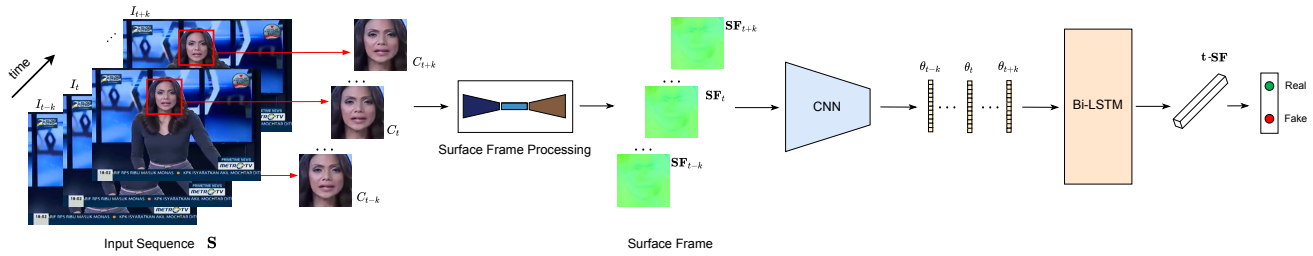

Figure 2. The proposed video deepfake detection pipeline. Given a sequence $\mathbf{S}$, a group of $2k+1$ frames, i.e. $\{I_{t-k}, \ldots, I_t, \ldots, I_{t+k}\}$ is considered. For each video frame I_t, we first extract the face crop C, then we calculate its surface frame $\mathbf{SF}$ through UpRightNet [24] and we scale the values in $[0, 255]$. Next, we extract the features θ_i by using a CNN backbone and we give them to a bidirectional LSTM in order to retain spatio-temporal contexts. We then append a linear layer to map the $\mathbf{t}$-$\mathbf{SF}$ feature into two classes. Finally, we end-to-end train the whole architecture to classify whether the content is real or fake.

manipulation should break all of this and generate some inconsistencies accordingly, in particular along the temporal axis.

The idea of the here proposed method (see Fig. 2 for the whole procedure pipeline) is to try to exploit such anomalies in order to perform video deepfake detection.
Let us consider a frame I_t at a certain time t of a video sequence $\mathbf{S}$. Moreover, let us take, around this frame, a group of pictures so that $I_{t\pm k}$ with k which defines the semi-length of the group $\{I_{t-k}, \ldots, I_t, \ldots, I_{t+k}\}$ made of $2k+1$ frames. For each frame, firstly the person's face is detected by using *dlib* [11] and then cropped at a predefined resolution of 224×224. After that, the bunch of $2k+1$ cropped faces are passed to a pretrained model *UpRightNet* [24], which extracts the surface frames of every face crop, indicated as $\mathbf{SF}$ in Fig. 2 and presented in detail within Sec. 3.1. This packet of pictures is then used as input of a convolutional neural network (CNN) that plays the role of a backbone to generate feature vectors θ_i. Such vectors are then passed to a bi-directional LSTM (Bi-LSTM) which will learn the spatio-temporal contexts determined by the subsequent surface frames and we retain the last output named $\mathbf{t}$-$\mathbf{SF}$. In order to finally provide an assessment on the authenticity of the central frame of the group (I_t), we append at the end of the pipeline a linear layer that maps $\mathbf{t}$-$\mathbf{SF}$ into a 2D feature, which is used to classify the content as real or fake. This is done iteratively for the other frames of the sequence. At training time, we first pre-train the CNN backbone on a single frame instance without Bi-LSTM and then the whole architecture is end-to-end trained. Different kinds of CNN backbones have been investigated.

3.1. The Surface Frames

When a scene is captured, being a single image or a video, inherent information about objects' surfaces and, for instance, specific regions of a human face are grabbed; all these physical structures intrinsically contribute to define the diverse shapes, lights and colors finally represented by the image pixels. In fact, all the pixel values are determined by various combined factors directly related to both the pixel position (i, j) in the image plane and also to the corresponding camera location; moreover, further external factors, such as illumination (magnitude and direction), shadows and reflections, provide an additional contribution that is recorded when using a digital camera. Presumably, it can be assumed that an altered video (image), though visually looking realistic and consistent, probably does not contain anymore the same innate characteristics as obtained during the camera acquisition process and could evidence some inconsistencies. According to this, we have decided to model such anomalies by resorting to surface frames ($\mathbf{SF}$) that can be obtained as output of a pretrained surface estimator named *UpRightNet* [24]. Such estimator is basically a supervised encoder-decoder network that, given a single RGB picture, can compute the two degrees of freedom (DoF, that is roll and pitch) of the camera; to achieve this, UpRightNet generates two intermediate representations, called surface frames, one calculated from a global (scene) up-right coordinate system $\mathbf{SF}^G$ and one from the local camera reference system $\mathbf{SF}^L$. Such surface frames can be considered as effective descriptors of the acquired scene and possibly useful to the task of deepfake detection. Every surface frame $\mathbf{SF}$ is basically constituted, at each pixel location (i, j), by a 3×3 matrix which describes the components of three mutually orthogonal vectors, i.e. normal, tangent and bitangent respectively $(\mathbf{n}(i,j), \mathbf{t}(i,j), \mathbf{b}(i,j) \in \mathbb{R}^3)$. Given that such surface frames convey important and diverse information of the scene characteristics, we have decided to verify if a possible combination of them could provide effective features in particular for the task of video deepfake detection. A preliminary view of this has been shown by just using the z-component of the global surface frame in [5]. In this work, we have tried to examine both the surface frames ($\mathbf{SF}^G$ and $\mathbf{SF}^L$) and, in addition to this, their temporal evolution in order to enhance the capabilities to reveal some discrepancies in a fake video sequence with respect to a pristine one.

Surface frames contain different valuable information concerning the recorded scene. There are several possible combinations of these surface representations that could be designed to address our task. We left this depth investigation for future works. However, by getting inspiration from [5], it has been decided to just use the z-component of the three vectors (n_z, t_z and b_z) whose values are rescaled to $[0, 255]$ so obtaining a three channel image ($\mathbb{R}^{H\times W\times 3}$ where H and W are height and width of the cropped face respectively). In particular, we consider to analyse these specific surface features in both the local and global representations, by temporally parsing them over time in a video sequence. In the experiments, we calculate the surface frames from 224×224 face patches; to do that, the UpRightNet pretrained weights are frozen, as in [5], and then we follow [24] by rescaling the input image to 288×384 before running the model on it. Then the generated output is successively downscaled from 288×384 to the original dimension of 224×224.

4. Experimental results

Different experimental tests have been carried out in order to verify if the proposed approach and, above all, the features based on the temporal surface frames could grant a significant distinctiveness for deepfake video detection purposes. The following Sec. 4.1 describes the adopted set-up while Sec. 4.2 presents the main experimental results.

Table 1. Accuracy for the global and local surface frames (with or without temporal) with respect to the five forgeries for ResNet50.

Accuracy (↑)	**Surface Frames**			
FF++ forgeries	**SF^G** [5]	**t-SF^G**	**SF^L**	**t-SF^L**
F2F	0.734	**0.761**	0.895	**0.929**
NT	0.667	**0.696**	0.829	**0.929**
DF	0.761	**0.785**	0.893	**0.900**
FSH	0.728	**0.735**	0.798	**0.874**
FS	0.693	**0.727**	**0.851**	**0.851**
Average	0.717	**0.741**	0.853	**0.897**

Table 2. Accuracy for the global and local surface frames (with or without temporal) with respect to the five forgeries for EfficientNet-B0.

Accuracy (↑)	**Surface Frames**			
FF++ forgeries	**SF^G** [5]	**t-SF^G**	**SF^L**	**t-SF^L**
F2F	0.772	**0.791**	0.944	**0.962**
NT	**0.701**	0.628	0.885	**0.869**
DF	**0.790**	0.751	**0.929**	0.883
FSH	**0.762**	0.719	0.857	**0.896**
FS	**0.756**	**0.756**	**0.905**	0.809
Average	**0.756**	0.729	**0.904**	0.884

Table 3. Accuracy for the global and local surface frames (with or without temporal) with respect to the five forgeries for Xception.

Accuracy (↑)	**Surface Frames**			
FF++ forgeries	**SF^G** [5]	**t-SF^G**	**SF^L**	**t-SF^L**
F2F	0.772	**0.826**	0.916	**0.962**
NT	0.713	**0.742**	0.855	**0.890**
DF	0.783	**0.812**	0.902	**0.939**
FSH	0.753	**0.769**	0.818	**0.905**
FS	0.739	**0.796**	0.848	**0.936**
Average	0.752	**0.789**	0.868	**0.926**

4.1. The experimental set-up

The experimental tests have been performed on FaceForensics++ (FF++) [20] dataset. Such dataset is well-known and mostly used in deepfake detection experiments. It is composed by 1000 original videos and by 5000 fake ones, derived from those real ones by applying 5 different deepfake forgeries (1000 for each forgery respectively): two reenactment methods (identity does not change), Face2Face (F2F) and NeuralTextures (NT), and three swapping methods, DeepFakes (DF), FaceShifter (FSH) and FaceSwap (FS).

The video sequences have variable frame size in the interval 272×480 and 1920×1080 and three types of compression levels, we have taken into account the intermediate case named c23 (high quality - HQ). Each set of 1000 videos is subdivided into training/validation/testing respectively with this ratio 72/14/14.

Frames are sampled by the video sequence and face images of size 224×224 are cropped following the procedure described in [20]; such face crops are passed as input of the pipeline in Fig. 2. In the experiments, we have evaluated three different backbone architectures: ResNet50, EfficientNet-B0 and Xception, all of them pretrained on ImageNet. Due to the fact that Xception accepts inputs of size 299×299, the face crops are upscaled to the required resolution. The experimental tests have been done in Pytorch with an NVIDIA TITAN RTX; the loss used during training is the cross entropy with two classes. Training is run for 30 epochs with a batch size of 32 and SGD is the optimizer with momentum 0.9, weight decay 0.0001 and learning rate 0.001.

4.2. Quantitative analysis of performances

In this section, different experimental results in order to investigate the capacity of temporal surface frames to distinguish between pristine and fake videos are discussed. Results are presented in terms of accuracy at frame-level ($Acc = \frac{TP+TN}{TP+TN+FP+FN}$) and of AUC (Area Under Curve) by analysing the behaviours of the diverse kinds of features: global and local without the temporal component

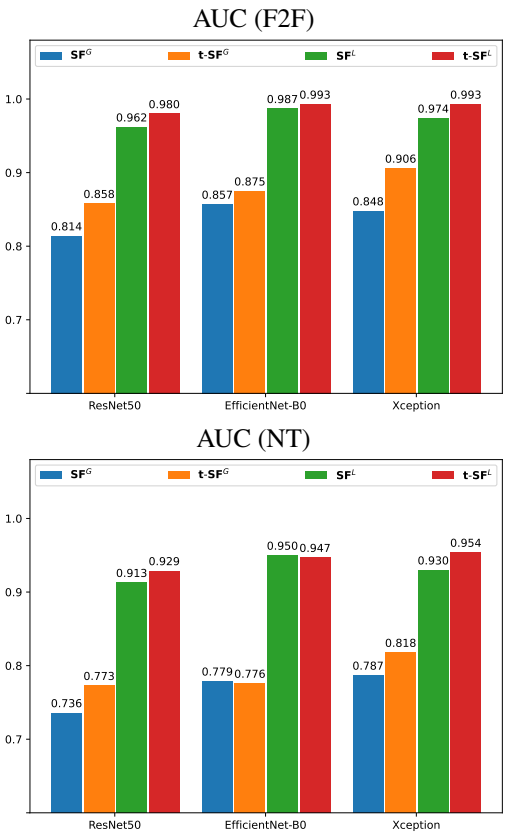

Figure 3. AUC for the two reenactment FF++ forgeries (F2F and NT) for the global and local surface frames (with or without temporal) with respect to the three backbone architectures.

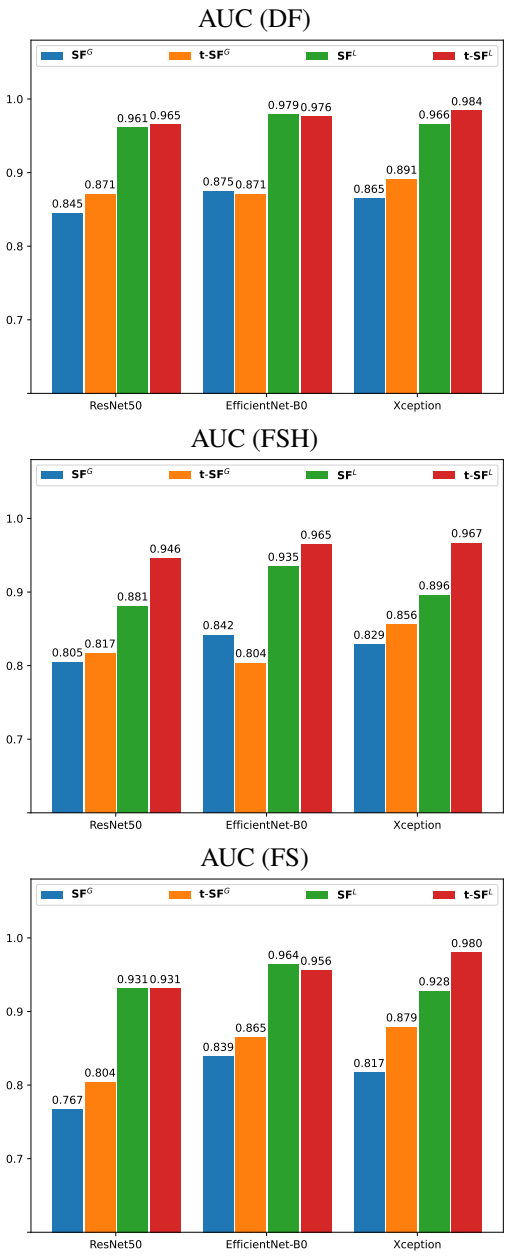

Figure 4. AUC for the three identity swapping FF++ forgeries (DF, FSH and FS) for the global and local surface frames (with or without temporal) with respect to the three backbone architectures.

($\mathbf{SF}^G$ and $\mathbf{SF}^L$ respectively) and with the temporal component ($\mathbf{t\text{-}SF}^G$ and $\mathbf{t\text{-}SF}^L$ respectively). The case of global surface frames (i.e. $\mathbf{SF}^G$) corresponds to what has been introduced in [5]. It can also be pointed out that, in practice, the case of the temporal surface frames reduces to the non-temporal one when $k = 0$ (both for global and local). In the experimental results presented hereafter, the value of $k = 2$ has been chosen so the group of pictures has a length of 5 (being $2k + 1$) that has been considered as a sufficient temporal interval to perceive an evolution in time. Therefore, we assess the central frame I_t of the group on the basis of $\{I_{t-2}, I_{t-1}, I_t, I_{t+1}, I_{t+2}\}$; groups of frames are sampled without overlapping. For a complete analysis, the performances related to the three different backbones and to each of the five possible forgeries are provided.

The accuracy at frame-level for the three different adopted backbones is reported in Tab. 1, Tab. 2 and Tab. 3 for ResNet50, EfficinetNet-B0 and Xception respectively. As it can be appreciated, all the different surface frames, with and without the temporal component, generally provide a significant distinctiveness demonstrating that they contain important information that allows to distinguish between real and fake videos. Going into details, it is possible to highlight two main aspects: first, the local surface frames $\mathbf{SF}^L$ ($\mathbf{t\text{-}SF}^L$) achieve a superior level of accuracy with respect to global ones $\mathbf{SF}^G$ ($\mathbf{t\text{-}SF}^G$) and second, the use of the temporal component gives an increment, though not always so evident, with respect to the non-temporal case. For instance, the situation when $\mathbf{t\text{-}SF}^L$ is used, very promising values of accuracy, often higher than 90%, are obtained:

such values are in general comparable with the state-of-the-art methods. If we look globally from Tab. 1 to Tab. 3, we can individuate a quite uniform trend confirmed for all the three diverse backbones, though in the case of Tab. 2 which refers to EfficientNet-B0, this is not always so well defined.

In order to get another point of analysis of the achieved performances, we have also computed the values of AUC but, in this case, grouped with respect to the five different deepfake manipulations. Such results are shown in Fig. 3, where forgeries produce a deepfake reenactment without changing the identity of the represented person and in Fig. 4, in which forgeries determine an identity swapping. For each forgery the behaviours of the tested features with respect to the implemented backbone are depicted. Also in this case, the overall trends evidenced by the results within the previous tables are generally confirmed; by looking at the different kinds of forgery a uniform behaviour is registered with AUC values averagely lower for $\mathbf{SF}^G$ and $\mathbf{t}$-$\mathbf{SF}^G$, and higher for $\mathbf{SF}^L$ and $\mathbf{t}$-$\mathbf{SF}^L$. It is worthy noting that, for instance, for the case of the F2F forgery, the AUCs, for all the three different backbones, achieve remarkable values around 0.99 for $\mathbf{t}$-$\mathbf{SF}^L$ (red columns). It is interesting to verify again that using the temporal component is globally beneficial; through a visual inspection, it is immediate to appreciate that the columns in orange ($\mathbf{t}$-$\mathbf{SF}^G$) generally exceed in height the corresponding blue columns ($\mathbf{SF}^G$) and that the same happens between red and green columns which represent $\mathbf{SF}^L$ and $\mathbf{t}$-$\mathbf{SF}^L$ respectively. A slight exception to this can be noticed, as expected from the results in Tab. 2, for the case of the EfficientNet-B0 which is reported in the central part of the Fig. 3 and Fig. 4 where, sometimes, the blue/green columns ($\mathbf{SF}^G$/$\mathbf{SF}^L$) are a bit higher than their corresponding orange/red ones ($\mathbf{t}$-$\mathbf{SF}^G$/$\mathbf{t}$-$\mathbf{SF}^L$) respectively. This issue related, to the use of EfficientNet-B0, will be investigated in future works. It is worth to point out that, using Xception as backbone, we obtain best accuracy and highest AUC values, which is consistent for all the forgeries. Since Xception processes inputs at a size of 299×299, we had to upscale our 224×224 face crops. This is rather interesting if we considering that, in general, the operation of input rescaling affects the classification capability of a neural network. The performance increase is much more pronounced when the temporal aspects are analysed, both for the global and the local surface frames. In particular, we found that $\mathbf{t}$-$\mathbf{SF}^L$ achieves best performance, by gaining on average $+5.87\%$ over the $\mathbf{SF}^L$ (see Tab. 3). As an overall evaluation, it can be assessed that surface frames (with or without temporal component) provide an interesting characterisation of the possible anomalies induced by the deepfake video crafting that can be promisingly used to perform a distinction between real and fake video contents.

5. Conclusions

Distinguishing pristine contents from falsified, but realistic, ones is even more crucial, particularly in the case of video sequences representing persons in foreground: reliable instruments to perform this task are strictly required. In this work, we have introduced temporal surface frames, which are able to provide a thorough and punctual description of the captured scene, that can be used to highlights some anomalies injected by the deepfake generation process. According to the obtained experimental results, such features, in particular for the case when the temporal component is used, can globally grant a significant accuracy achieving AUC values of 0.99 as best cases. Being surface frames still quite new, it is necessary to go ahead with further investigations and this paves the way for various future works. Specifically, we deem that would be strategic to study the possible combination of these kinds of features with spatial frames and, moreover, to verify the actual impact of the number of pictures (related to the parameter k) for the LSTM learning phase.

Acknowledgment

This work was partially supported by the following projects: SERICS (PE00000014) under the NRRP MUR program funded by the EU - NGEU, FOSTERER funded by the Italian MUR PRIN 2022 program, AI4Debunk (GA n. 101135757) funded by the EU Horizon Europe Programme, AI4Media (H2020, GA n. 951911) and 04-FIN/RIC (Fin. Comp. 2023 UM).

References

[1] Rameen Abdal, Yipeng Qin, and Peter Wonka. Image2stylegan: How to embed images into the stylegan latent space? In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 4432–4441, 2019. 2

[2] Roberto Caldelli, Leonardo Galteri, Irene Amerini, and Alberto Del Bimbo. Optical flow based CNN for detection of unlearnt deepfake manipulations. *Pattern Recognition Letters*, 146:31–37, 2021. 3

[3] Dongyue Chen, Qiusheng Chen, Jianjun Wu, Xiaosheng Yu, and Jia Tong. Face swapping: realistic image synthesis based on facial landmarks alignment. *Mathematical Problems in Engineering*, 2019. 3

[4] Yunjey Choi, Minje Choi, Munyoung Kim, Jung-Woo Ha, Sunghun Kim, and Jaegul Choo. Stargan: Unified generative adversarial networks for multi-domain image-to-image translation. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 8789–8797, 2018. 2

[5] Andrea Ciamarra, Roberto Caldelli, Federico Becattini, Lorenzo Seidenari, and Alberto Del Bimbo. Deepfake detection by exploiting surface anomalies: the surfake approach. In *Proceedings of the IEEE/CVF Winter Conference on Ap-*

plications of Computer Vision, pages 1024–1033, 2024. 3, 4, 5, 6

[6] Zhihao Gu, Yang Chen, Taiping Yao, Shouhong Ding, Jilin Li, Feiyue Huang, and Lizhuang Ma. Spatiotemporal inconsistency learning for deepfake video detection. In *Proceedings of the 29th ACM international conference on multimedia*, pages 3473–3481, 2021. 3

[7] David Güera and Edward J Delp. Deepfake video detection using recurrent neural networks. In *15th IEEE international conference on advanced video and signal based surveillance (AVSS)*, pages 1–6, 2018. 3

[8] Alexandros Haliassos, Konstantinos Vougioukas, Stavros Petridis, and Maja Pantic. Lips don't lie: A generalisable and robust approach to face forgery detection. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 5039–5049, 2021. 3

[9] Jonathan Ho, Ajay Jain, and Pieter Abbeel. Denoising diffusion probabilistic models. *Advances in neural information processing systems*, 33:6840–6851, 2020. 2

[10] Baojin Huang, Zhongyuan Wang, Jifan Yang, Jiaxin Ai, Qin Zou, Qian Wang, and Dengpan Ye. Implicit identity driven deepfake face swapping detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 4490–4499, 2023. 3

[11] Davis E King. Dlib-ml: A machine learning toolkit. *The Journal of Machine Learning Research*, 10:1755–1758, 2009. 4

[12] Iryna Korshunova, Wenzhe Shi, Joni Dambre, and Lucas Theis. Fast face-swap using convolutional neural networks. In *Proceedings of the IEEE international conference on computer vision*, pages 3677–3685, 2017. 3

[13] Yuezun Li, Ming-Ching Chang, and Siwei Lyu. In ictu oculi: Exposing ai created fake videos by detecting eye blinking. In *IEEE Internat. workshop on information forensics and security*, pages 1–7, 2018. 3

[14] Luca Maiano, Lorenzo Papa, Ketbjano Vocaj, and Irene Amerini. Depthfake: a depth-based strategy for detecting deepfake videos. *arXiv preprint arXiv:2208.11074*, 2022. 3

[15] Asad Malik, Minoru Kuribayashi, Sani M Abdullahi, and Ahmad Neyaz Khan. Deepfake detection for human face images and videos: A survey. *Ieee Access*, 10:18757–18775, 2022. 3

[16] Momina Masood, Mariam Nawaz, Khalid Mahmood Malik, Ali Javed, Aun Irtaza, and Hafiz Malik. Deepfakes generation and detection: State-of-the-art, open challenges, countermeasures, and way forward. *Applied intelligence*, 53(4): 3974–4026, 2023. 3

[17] Thanh Thi Nguyen, Quoc Viet Hung Nguyen, Dung Tien Nguyen, Duc Thanh Nguyen, Thien Huynh-The, Saeid Nahavandi, Thanh Tam Nguyen, Quoc-Viet Pham, and Cuong M Nguyen. Deep learning for deepfakes creation and detection: A survey. *Computer Vision and Image Understanding*, 223:103525, 2022. 3

[18] Kyle Olszewski, Zimo Li, Chao Yang, Yi Zhou, Ronald Yu, Zeng Huang, Sitao Xiang, Shunsuke Saito, Pushmeet Kohli, and Hao Li. Realistic dynamic facial textures from a single image using gans. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 5429–5438, 2017. 3

[19] Muhammad Anas Raza and Khalid Mahmood Malik. Multimodaltrace: Deepfake detection using audiovisual representation learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 993–1000, 2023. 3

[20] Andreas Rossler, Davide Cozzolino, Luisa Verdoliva, Christian Riess, Justus Thies, and Matthias Nießner. Faceforensics++: Learning to detect manipulated facial images. In *Proc. of the IEEE/CVF international conference on computer vision*, pages 1–11, 2019. 5

[21] Ekraam Sabir, Jiaxin Cheng, Ayush Jaiswal, Wael AbdAlmageed, Iacopo Masi, and Prem Natarajan. Recurrent convolutional strategies for face manipulation detection in videos. *Interfaces (GUI)*, pages 80–87, 2019. 3

[22] Zekun Sun, Yujie Han, Zeyu Hua, Na Ruan, and Weijia Jia. Improving the efficiency and robustness of deepfakes detection through precise geometric features. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 3609–3618, 2021. 3

[23] Yuan Wang, Kun Yu, Chen Chen, Xiyuan Hu, and Silong Peng. Dynamic graph learning with content-guided spatial-frequency relation reasoning for deepfake detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 7278–7287, 2023. 3

[24] Wenqi Xian, Zhengqi Li, Matthew Fisher, Jonathan Eisenmann, Eli Shechtman, and Noah Snavely. Uprightnet: geometry-aware camera orientation estimation from single images. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 9974–9983, 2019. 1, 4, 5

[25] Yuting Xu, Jian Liang, Gengyun Jia, Ziming Yang, Yanhao Zhang, and Ran He. Tall: Thumbnail layout for deepfake video detection. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 22658–22668, 2023. 3

[26] Zhiyuan Yan, Yong Zhang, Yanbo Fan, and Baoyuan Wu. Ucf: Uncovering common features for generalizable deepfake detection. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 22412–22423, 2023. 3

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

[28] Hanqing Zhao, Wenbo Zhou, Dongdong Chen, Tianyi Wei, Weiming Zhang, and Nenghai Yu. Multi-attentional deepfake detection. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 2185–2194, 2021. 3

[29] Tianchen Zhao, Xiang Xu, Mingze Xu, Hui Ding, Yuanjun Xiong, and Wei Xia. Learning self-consistency for deepfake detection. In *Proc. of the IEEE/CVF international conference on computer vision*, pages 15023–15033, 2021. 3

[30] Yinglin Zheng, Jianmin Bao, Dong Chen, Ming Zeng, and Fang Wen. Exploring temporal coherence for more general video face forgery detection. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 15044–15054, 2021. 3