



# DeepFakes detection across generations: Analysis of facial regions, fusion, and performance evaluation

Ruben Tolosana <sup>a,\*</sup>, Sergio Romero-Tapiador <sup>a</sup>, Ruben Vera-Rodriguez <sup>a</sup>, Ester Gonzalez-Sosa <sup>b</sup>, Julian Fierrez <sup>a</sup>

<sup>a</sup> Biometrics and Data Pattern Analytics Lab, Universidad Autonoma de Madrid, Spain

<sup>b</sup> Nokia Bell Labs, Spain



## ARTICLE INFO

### Keywords:

Fake news  
DeepFakes  
Media forensics  
Face manipulation  
Fake detection  
Benchmark  
Databases

## ABSTRACT

Media forensics has attracted a tremendous attention in the last years in part due to the increasing concerns around DeepFakes. Since the release of the initial DeepFakes databases of the 1st generation such as UADFV and FaceForensics++ up to the latest databases of the 2nd generation such as Celeb-DF and DFDC, many visual improvements have been carried out, making fake videos almost indistinguishable to the human eye. This study provides an in-depth analysis of both 1st and 2nd DeepFakes generations in terms of fake detection performance. Two different methods are considered in our experimental framework: (i) the traditional one followed in the literature based on selecting the entire face as input to the fake detection system, and (ii) a novel approach based on the selection of specific facial regions as input to the fake detection system. Fusion techniques are applied both to the facial regions and also to three different state-of-the-art fake detection systems (Xception, Capsule Network, and DSP-FWA) in order to further increase the robustness of the detectors considered. Finally, experiments regarding intra- and inter-database scenarios are performed.

Among all the findings resulting from our experiments, we highlight: (i) the very good results achieved using facial regions and fusion techniques with fake detection results above 99% Area Under the Curve (AUC) for UADFV, FaceForensics++, and Celeb-DF v2 databases, and (ii) the necessity to put more efforts on the analysis of inter-database scenarios to improve the ability of the fake detectors against attacks unseen during learning.

## 1. Introduction

Fake images and videos including facial information generated by digital manipulations, in particular with DeepFakes methods (Tolosana et al., 2020; Verdoliva, 2020; Rathgeb et al., 2021), have become a great public concern recently (Citron, 2019; Cellan-Jones, 2019). The very popular term “DeepFakes” is referred to a deep learning based technique able to create fake videos by swapping the face of a person by the face of another person.<sup>1</sup> Open software and mobile applications such as ZAO<sup>2</sup> allow nowadays to automatically generate fake videos by anyone, without a prior knowledge of the task.

Digital manipulations based on face swapping are known in the literature as Identity Swap, and they are usually based on computer graphics and deep learning techniques (Tolosana et al., 2020). Since the initial publicly available fake databases, such as the UADFV database (Li et al., 2018), up to the recent Celeb-DF and Deepfake Detection Challenge (DFDC) databases (Li et al., 2020; Dolhansky et al.,

2019), many visual improvements have been carried out, increasing the realism of fake videos. As a result, Identity Swap databases can be divided into two different generations. Fig. 1 shows some examples of identity swap databases of the 1st and 2nd generations of DeepFakes.

In general, fake videos of the 1st generation are characterised by, among other factors: (i) low-quality synthesised faces, (ii) different colour contrast among the synthesised fake mask and the skin of the original face, (iii) visible boundaries of the fake mask, (iv) visible facial elements from the original video, (v) low pose variations, and (vi) strange artifacts among sequential frames. Also, they usually consider controlled scenarios in terms of camera position and light conditions. Many of these aspects have been successfully improved in databases of the 2nd generation. For example, the recent DFDC database considers different acquisition scenarios (i.e., indoors and outdoors), light conditions (i.e., day, night, etc.), distances from the person to the camera, pose variations, etc.

\* Corresponding author.

E-mail address: [ruben.tolosana@uam.es](mailto:ruben.tolosana@uam.es) (R. Tolosana).

<sup>1</sup> <https://www.youtube.com/watch?v=Ulv0EW7l5rs>.

<sup>2</sup> <https://apps.apple.com/cn/app/id1465199127>.



Fig. 1. Examples of the weaknesses present in identity swap databases of the 1st generation and the excellent naturalness achieved in the 2nd generation of DeepFakes.

The present study provides an exhaustive analysis of both 1st and 2nd DeepFakes generations using state-of-the-art fake detectors. Fig. 2 graphically summarises our proposed framework. Two different approaches are considered to detect fake videos: (i) the traditional one followed in the literature and based on selecting the entire face as input to the fake detection system, and (ii) a novel approach based on the selection of specific facial regions as input to the fake detection system. To the best of our knowledge, this is the first study that proposes the detection of DeepFakes videos using fusion approaches of selected fake detection systems and facial regions.

The main contributions of this study are as follow:

- A novel DeepFakes detection approach based on: (i) the selection of specific facial regions, and (ii) fusion approaches in order to develop more robust fake detectors. Two different fusion approaches are considered: fusion at system level considering three different state-of-the-art fake detection systems (Xception, Capsule Network, and DSP-FWA), and fusion at facial region level. All the DeepFakes detection models are publicly available in GitHub.<sup>3</sup>
- The fusion results achieved outperform the state of the art with fake detection results above 99% Area Under the Curve (AUC) for the UADFV, FaceForensics++, and Celeb-DF v2 databases.
- An in-depth comparison in terms of performance among Identity Swap databases of the 1st and 2nd generation. We provide an analysis of the discriminative power of the different facial regions between the 1st and 2nd generations, and also between fake detectors. In addition, we study the generalisation ability of state-of-the-art fake detectors to unseen databases of the 1st and 2nd generations, which is currently one of the key challenges in the field (Tolosana et al., 2020; Verdoliva, 2020; Rathgeb et al., 2021).
- The analysis and proposal carried out in this study will benefit the research community by providing insights for: (i) the proposal of more robust fake detectors, e.g., through the fusion of

different facial regions depending on the scenario: light conditions, pose variations, and distance from the camera; and (ii) the improvement of the next generation of DeepFakes, focusing on the artifacts existing in specific facial regions.

A preliminary version of this article was published in Tolosana et al. (2021). This article significantly improves (Tolosana et al., 2021) in the following aspects: (i) we perform in the new Section 2 an in-depth literature review of the DeepFakes detection techniques, (ii) new tables and figures have been added/improved to provide a better comprehension of the results and the field, (iii) we perform an in-depth evaluation of the generalisation ability of the state-of-the-art fake detectors to unseen databases (inter-database scenarios), simulating this way real attack scenarios, (iv) we provide a more extensive evaluation considering a new state-of-the-art DeepFakes detector named DSP-FWA, (v) we analyse and propose fusion techniques to develop more robust fake detectors, (vi) we have updated the results and conclusions, outperforming the initial fake detection results presented in Tolosana et al. (2021).

The remainder of the paper is organised as follows. Section 2 summarises key related studies in the literature. Section 3 describes the proposed evaluation framework. Section 4 summarises all the databases considered in the experimental framework of this study. Sections 5 and 6 describe the experimental protocol and results achieved, respectively. Finally, Section 7 draws the final conclusions and points out future research lines.

## 2. Related works

Different approaches have been proposed in the literature to detect DeepFakes videos. Table 1 shows a comparison of the most relevant approaches in the area. For each study we include information related to the method, classifiers, best performance, and databases for research. It is important to remark that in some cases, different evaluation metrics are considered, e.g., AUC and Equal Error Rate (EER), which complicate the comparison among studies. Finally, the results highlighted in *italics* indicate the generalisation capacity of the detectors against unseen databases, i.e., those databases were not considered for training the fake detectors. Most of these results are extracted from Tolosana et al. (2020), Li et al. (2020).

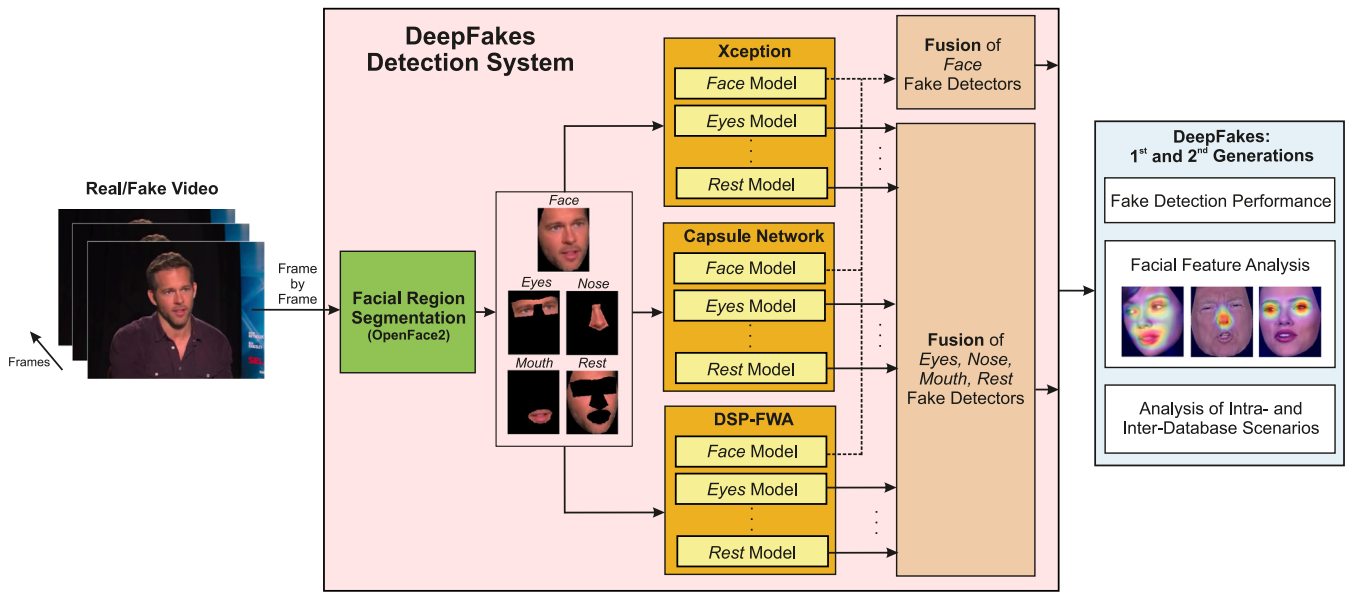
The first studies in the area focused on the visual artifacts present in the 1st generation of fake videos. Matern et al. proposed in Matern et al. (2019) fake detectors based on simple visual aspects such as eye colour, missing reflections, and missing details in the teeth areas, achieving a final 85.1% AUC.

Approaches based on the detection of the face warping artifacts have also been studied in the literature. Li et al. proposed in Li and Lyu (2019), Li et al. (2020) fake detectors based on Convolutional Neural Networks (CNN) in order to detect the presence of such artifacts from the face and the surrounding areas, achieving state-of-the-art results.

Undoubtedly, fake detectors based on pure deep learning features are the most popular ones: feeding the networks with as many real/fake videos as possible and letting the networks to automatically extract the discriminative features. In general, these fake detectors have achieved state-of-the-art results using popular network architectures such as Xception (Rössler et al., 2019; Dolhansky et al., 2019), novel ones such as Capsule Networks (Nguyen et al., 2019), and novel training techniques based on attention mechanisms (Dang et al., 2020).

Fake detectors based on the image and temporal discrepancies across frames have also been proposed in the literature. Sabir et al. proposed in Sabir et al. (2019) a Recurrent Convolutional Network, similar to Güera and Delp (2018), trained end-to-end instead of using a pre-trained model. Their proposed detection approach was tested using FaceForensics++ (Rössler et al., 2019), achieving AUC results above 96%. An interesting approach in order to improve the interpretability of the results has been recently presented in Trinh et al. (2021).

<sup>3</sup> [https://github.com/BiDALab/DeepFakes\\_FacialRegions](https://github.com/BiDALab/DeepFakes_FacialRegions).



**Fig. 2.** Architecture of our proposed framework to analyse: (i) fake detection performance, (ii) facial features, and (iii) intra- and inter-database scenarios in DeepFakes video databases of the 1st and 2nd generations. Two different approaches are studied: (i) selecting the entire face as input to the fake detection system (*Face*), and (ii) selecting specific facial regions (*Eyes*, *Nose*, *Mouth*, and *Rest*). Different fusion approaches are finally studied in order to develop more robust DeepFakes detectors.

**Table 1**

Comparison of different state-of-the-art fake detectors. Results in *italics* indicate the generalisation capacity of the detectors against unseen databases. FF++ = FaceForensics++, AUC = Area Under the Curve, Acc. = Accuracy, EER = Equal Error Rate.

Study	Method	Classifiers	Best Performance	Databases
<a href="#">Matern et al. (2019)</a>	Visual Features	Logistic RegressionMLP	AUC = 85.1% AUC = 78.0% AUC = 66.2% AUC = 55.1%	Own FF++/DFD DFDC Preview Celeb-DF
<a href="#">Li and Lyu (2019)</a> , <a href="#">Li et al. (2020)</a>	Face Warping Features	CNN	AUC = 97.7% AUC = 93.0% AUC = 75.5% AUC = 64.6%	UADFV FF++/DFD DFDC Preview Celeb-DF
<a href="#">Rössler et al. (2019)</a>	Mesoscopic Features Steganalysis Features Deep learning features	CNN	Acc. $\approx$ 94.0% Acc. $\approx$ 98.0% Acc. $\approx$ 100.0%  Acc. $\approx$ 93.0% Acc. $\approx$ 97.0% Acc. $\approx$ 99.0%	FF++ (DeepFakes, LQ) FF++ (DeepFakes, HQ) FF++ (DeepFakes, RAW)  FF++ (FaceSwap, LQ) FF++ (FaceSwap, HQ) FF++ (FaceSwap, RAW)
<a href="#">Nguyen et al. (2019)</a>	Deep learning features	Capsule Networks	AUC = 61.3% AUC = 96.6% AUC = 53.3% AUC = 57.5%	UADFV FF++/DFD DFDC Preview Celeb-DF
<a href="#">Dang et al. (2020)</a>	Deep learning features	CNN + Attention mechanism	AUC = 99.4% EER = 3.1%	DFFD
<a href="#">Dolhansky et al. (2019)</a>	Deep learning features	CNN	Precision = 93.0% Recall = 8.4%	DFDC Preview
<a href="#">Sabir et al. (2019)</a>	Image + Temporal features	CNN + RNN	AUC = 96.9% AUC = 96.3%	FF++ (DeepFakes, LQ) FF++ (FaceSwap, LQ)
<a href="#">Trinh et al. (2021)</a>	Image + Temporal features	Dynamic Prototype Network	AUC = 99.2% AUC = 71.8%	FF++ (FaceSwap, HQ) Celeb-DF
<a href="#">Li et al. (2018)</a>	Eye blinking features	LRCN	AUC = 99.0%	UADFV
<a href="#">Jung et al. (2020)</a>	Eye blinking features	Distance	Acc. = 87.5%	Own
<a href="#">Ciftci et al. (2020)</a>	Blood changes features	SVM/CNN	Acc. = 76.7%	FF++ (DeepFakes)
<a href="#">Qi et al. (2020)</a>	Blood changes features	CNN + Attention mechanism	Acc. = 100.0% Acc. = 100.0% Acc. = 64.1%	FF++ (FaceSwap) FF++ (DeepFakes) DFDC Preview
<a href="#">Hernandez-Ortega et al. (2021)</a>	Blood changes features	CNN + Attention mechanism	AUC = 98.2% AUC = 99.9%	DFDC Celeb-DF
<a href="#">Zhu et al. (2021)</a>	3D Decomposition	CNN + Attention mechanism	AUC = 99.5% AUC = 66.1%	FF++ (Identity swap) DFDC

The authors proposed Dynamic Prototype Network (DPNet), an interpretable and effective solution that utilises dynamic representations (i.e., prototypes) to explain DeepFakes temporal artifacts. Competitive performances were achieved on DeepFakes databases such as Celeb-DF, providing easy referential explanations of DeepFakes dynamics.

Finally, we pay special attention to the fake detectors based on physiological measures. Eye blinking has been studied to detect fake videos in Li et al. (2018), Jung et al. (2020). In Li et al. (2018), Li et al. proposed Long-Term Recurrent Convolutional Networks (LRCN) to capture the temporal dependencies in the human eye blinking. Their method was evaluated on the UADFV database, achieving a final 99.0% AUC. More recently, Jung et al. proposed a different approach named DeepVision. Their proposed approach achieved an accuracy of 87.5% over an in-house database.

Changes in the blood flow have also been studied in Ciftci et al. (2020), Qi et al. (2020), Hernandez-Ortega et al. (2021) using remote photoplethysmography (rPPG) techniques (Hernandez-Ortega et al., 2020). In Ciftci et al. (2020), the authors considered classifiers based on Support Vector Machines (SVM) and CNN, achieving a final 76.7% accuracy for the DeepFakes videos of FaceForensics++. Qi et al. (2020), developed a more sophisticated detector named DeepRhythm, based on two modules: (i) motion-magnified spatial-temporal representation, and (ii) dual-spatial-temporal attention. Accuracies of 100% were achieved on FaceForensics++ database. However, poor generalisation results were achieved on unseen databases (not considered during training) such as DFDC Preview (Acc. = 64.1%). A similar approach was also considered in Hernandez-Ortega et al. (2021), proposing DeepFakesON-Phys, a fake detector based on a Convolutional Attention Network (CAN) which extracts spatial and temporal information from video frames. AUC results above 98% were achieved on Celeb-DF and DFDC databases.

An interesting detection approach was recently presented in Zhu et al. (2021). The authors proposed to use decomposition techniques, which reversibly decompose an image into several constituent elements, being possible to highlight the hidden forgery details. In particular, the authors proposed methods to disentangle the face region into 3D shape, common texture, identity texture, ambient light, and direct light. Promising results were achieved on FaceForensics++ (99.6% AUC) and DFDC (67.8% AUC).

To the best of our knowledge, despite the large number of fake detectors proposed in the literature, there are no studies that provide an in-depth evaluation of both 1st and 2nd DeepFakes generations and both intra- and inter-database scenarios using state-of-the-art fake detectors. In addition, we benchmark our proposed DeepFakes detection approach based on: (i) the selection of specific facial regions, and (ii) fusion approaches in order to develop more robust fake detectors.

### 3. Proposed evaluation framework

Fig. 2 graphically summarises our evaluation framework. It comprises three main modules: (i) facial region segmentation, described in Section 3.1, (ii) fake detection systems, described in Section 3.2, and (iii) fusion techniques, described in Section 3.3.

#### 3.1. Facial region segmentation

Two different approaches are studied: (i) segmenting the entire face as input to the fake detection system, and (ii) segmenting only specific facial regions.

Regarding the second approach, 4 different facial regions are selected: *Eyes*, *Nose*, *Mouth*, and *Rest* (i.e., the part of the face obtained after removing the eyes, nose, and mouth from the entire face). For the segmentation of each region, we consider the open-source toolbox OpenFace2 (Baltrusaitis et al., 2018). This toolbox extracts 68 total landmarks for each face. Fig. 3 shows an example of the 68 landmarks

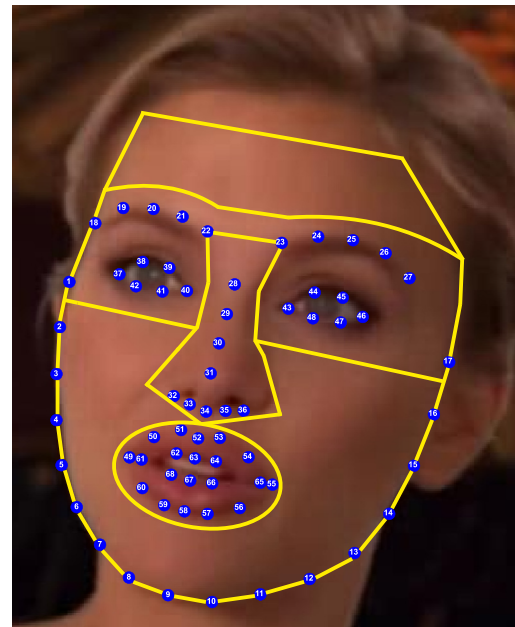


Fig. 3. Example of the different facial regions (i.e., *Eyes*, *Nose*, *Mouth*, and *Rest*) extracted using the 68 facial landmarks provided by OpenFace2 (Baltrusaitis et al., 2018).

(blue circles) extracted by OpenFace2 over a frame of the Celeb-DF database. It is important to highlight that OpenFace2 is robust against pose variations, distance from the camera, and light conditions, extracting reliable landmarks even for challenging databases such as the DFDC database (Dolhansky et al., 2019). The specific key landmarks considered to extract each facial region are as follow:

- *Eyes*: using landmark points from 18 to 27 (top of the mask), and using landmarks 1, 2, 16, and 17 (bottom of the mask).
- *Nose*: using landmark points 22, 23 (top of the mask), from 28 to 36 (line and bottom of the nose), and 40, 43 (width of the middle-part of the nose).
- *Mouth*: using landmark points 49, 51–53, 55, and 57–59 to build a circular/elliptical mask.
- *Rest*: extracted after removing eyes, nose, and mouth masks from the entire face.

Each facial region is highlighted by yellow lines in Fig. 3. Once each facial region is segmented, the remaining part of the face is discarded (black background as depicted in Fig. 2). Also, for each facial region, we keep the same image size and resolution as the original face image to perform a fair evaluation among facial regions and the entire face, avoiding therefore the influence of other pre-processing aspects such as interpolation.

#### 3.2. Fake detection systems

Three different state-of-the-art fake detection approaches are considered in our evaluation framework:

- *Xception* (Chollet, 2017): this network has achieved very good fake detection results in recent studies (Rössler et al., 2019; Dolhansky et al., 2019; Neves et al., 2020; Dang et al., 2020). Xception is a CNN architecture inspired by Inception (Szegedy et al., 2015), where Inception modules have been replaced with depthwise separable convolutions. In our evaluation framework, we follow the same training approach considered in Rössler et al. (2019): (i) we first consider the Xception model pre-trained with



ImageNet (Deng et al., 2009), (ii) we change the last fully-connected layer of the ImageNet model by a new one (two classes, real or fake), (iii) we fix all weights up to the final fully-connected layer and re-train the network for few epochs, and finally (iv) we train the whole network for 20 more epochs and choose the best performing model based on validation accuracy.

- **Capsule Network** (Nguyen et al., 2019): we consider the same detection approach proposed by Nguyen et al., which is publicly available in GitHub.<sup>4</sup> It is based on the combination of traditional CNN and recent Capsule Networks, which require fewer parameters to train compared with traditional CNN (G.E. Hinton and Frosst, 2018). In particular, the authors proposed to use part of the VGG19 model pre-trained with ImageNet database for the feature extractor (from the first layer to the third max-pooling layer). The output of this pre-trained part is concatenated with 10 primary capsules and finally 2 output capsules (real and fake). In our evaluation framework, we train only the capsules following the procedure described in Nguyen et al. (2019).
- **DSP-FWA** (Li et al., 2020): this detection system is an improved version of the original one presented in Li and Lyu (2019), based on the detection of the warping artifacts included in DeepFakes videos. The authors included in this new version Spatial Pyramid Pooling (He et al., 2015) to better handle the variations in the resolutions of the faces. We consider the same implementation details released by the authors in GitHub.<sup>5</sup>

Finally, as shown in Fig. 2, it is important to highlight that we train a specific fake detector per database and facial region.

### 3.3. Fusion

Fusion techniques are finally considered to develop more robust fake detectors. Two different approaches are considered: (i) fusion of 3 complementary state-of-the-art fake detectors (Xception, Capsule Network, and DSP-FWA), and (ii) fusion of different facial regions (*Eyes*, *Nose*, *Mouth*, and *Rest*).

In all cases we consider fusion techniques at score level (Kittler et al., 1998; Fierrez et al., 2018). In particular, we consider a simple but very useful fusion technique based on a weighted sum of the matching scores, which has achieved very good results in many information fusion problems in biometrics (Singh et al., 2019). The final fusion score  $s_f$  is obtained as (Fierrez et al., 2018):

$$s_f = k_1 \cdot s_1 + k_2 \cdot s_2 + \dots + k_N \cdot s_N \quad (1)$$

where  $s_n$  with  $n = 1, \dots, N$  is the matching score of a specific fake detection system or facial region system  $n$ , and  $k_n$  is the fusion weighted coefficient associated to it. The higher the  $k_n$  value is, the more important the system  $n$  will be for the final decision. Also, it is important to highlight that the sum of all weighted coefficients  $k$  is set to 1. These values are heuristically determined as detailed in Section 6.

## 4. Databases

Four different public databases are considered in the experimental framework of this study. In particular, two databases of the 1st generation (UADFV and FaceForensics++) and two recent databases of the 2nd generation (Celeb-DF and DFDC). Table 2 summarises their main features.

**Table 2**

Identity swap publicly available databases of the 1st and 2nd generations considered in our experimental framework.

1st Generation		
Database	Real Videos	Fake Videos
UADFV (2018) (Li et al., 2018)	49 (Youtube)	49 (FakeApp)
FaceForensics++ (2019) (Rössler et al., 2019)	1000 (Youtube)	1000 (FaceSwap)
2nd Generation		
Database	Real Videos	Fake Videos
Celeb-DF v2 (2019) (Li et al., 2020)	890 (Youtube)	5639 (DeepFakes)
DFDC Preview (2019) (Dolhansky et al., 2019)	1131 (Actors)	4119 (Unknown)

### 4.1. UADFV

The UADFV database (Li et al., 2018) comprises 49 real videos downloaded from Youtube, which were used to create 49 fake videos through the FakeApp mobile application,<sup>6</sup> swapping in all of them the original face with the face of the actor Nicolas Cage. Therefore, only one identity is considered in all fake videos. Each video represents one individual, with a typical resolution of  $294 \times 500$  pixels, and 11.14 s on average.

### 4.2. FaceForensics++

The FaceForensics++ database (Rössler et al., 2019) was introduced in 2019 as an extension of the original FaceForensics (Rössler et al., 2018), which was focused only on Expression Swap manipulations. FaceForensics++ contains 1000 real videos extracted from Youtube. Fake videos were generated using both computer graphics and deep learning approaches (1000 fake videos for each approach). In this study we focus on the computer graphics approach where fake videos were created using the publicly available FaceSwap algorithm.<sup>7</sup> This algorithm consists of face alignment, Gauss Newton optimisation, and image blending to swap the face of the source person to the target person.

### 4.3. Celeb-DF v2

The Celeb-DF v2 database (Li et al., 2020) aims to provide fake videos of better visual qualities, similar to the popular videos that are shared on the Internet.<sup>8</sup> This database consists of 890 real videos extracted from Youtube, corresponding to interviews of 59 celebrities with a diverse distribution in terms of gender, age, and ethnic group. In addition, these videos exhibit a large range of variations in aspects such as the face sizes (in pixels), orientations, lighting conditions, and backgrounds. Regarding fake videos, a total of 5639 videos were created using a refined version of a public DeepFakes generation algorithm, improving aspects such as the low resolution of the synthesised faces and colour inconsistencies.

### 4.4. DFDC preview

The DFDC database (Dolhansky et al., 2019) is one of the latest public databases, released by Facebook in collaboration with other companies and academic institutions such as Microsoft, Amazon, and the MIT. In the present study we consider the DFDC preview dataset consisting of 1131 real videos from 66 paid actors, ensuring realistic

<sup>4</sup> <https://github.com/nii-yamagishilab/Capsule-Forensics-v2>.

<sup>5</sup> <https://github.com/danmohaha/DSP-FWA>.

<sup>6</sup> <https://fakeapp.softonic.com/>.

<sup>7</sup> <https://github.com/MarekKowalski/FaceSwap>.

<sup>8</sup> <https://www.youtube.com/c/CtrlShiftFace/videos>.

variability in gender, skin tone, and age. It is important to remark that no publicly available data or data from social media sites were used to create this dataset, unlike other popular databases. Regarding fake videos, a total of 4119 videos were created using two different unknown approaches for fakes generation. Fake videos were generated by swapping subjects with similar appearances, i.e., similar facial attributes such as skin tone, facial hair, glasses, etc. After a given pairwise model was trained on two identities, they swapped each identity onto the other's videos.

It is important to highlight that the DFDC database considers different acquisition scenarios (i.e., indoors and outdoors), light conditions (i.e., day, night, etc.), distances from the person to the camera, and pose variations, among others.

## 5. Experimental protocol

All databases have been divided into non-overlapping datasets, development ( $\approx 80\%$  of the identities) and evaluation ( $\approx 20\%$  of the identities). It is important to remark that each dataset comprises videos from different identities (both real and fake), unlike some previous studies. This aspect is very important in order to perform a fair evaluation and predict the generalisation ability of the fake detection systems against unseen identities. For example, for the UADFV database, all real and fake videos related to the identity of Donald Trump were considered only for the final evaluation of the models. For the FaceForensics++ database, we consider 860 development videos and 140 evaluation videos per class (real/fake) as proposed in Rössler et al. (2019), selecting different identities in each dataset (one fake video is provided for each identity). For the DFDC Preview database, we follow the same experimental protocol proposed in Dolhansky et al. (2019) as the authors already considered this concern. Finally, for the Celeb-DF v2 database, we consider real/fake videos of 40 and 19 different identities for the development and evaluation datasets, respectively.

In addition, it is important to highlight that the evaluation is carried out at frame level as in all related works (Tolosana et al., 2020), not video level, using popular metrics in the literature such as the Area Under the Curve (AUC) and Equal Error Rate (EER).

## 6. Experimental results

Five sets of experiments are performed:

- Section 6.1 considers the traditional scenario of feeding the fake detectors with the entire face.
- Section 6.2 analyses the discriminative power of each facial region.

The two previous sets of experiments are conducted intra-database, i.e., training and testing the fake detectors within the same database.

- Section 6.3 performs an inter-database analysis to study the generalisation ability of the fake detectors to unseen digital manipulations, simulating this way real scenarios.
- Section 6.4 analyses how fusion techniques can further enhance fake detectors on both intra- and inter-database scenarios.
- Section 6.5 compares the results achieved in this study with the state of the art.

### 6.1. Entire face analysis

Table 3 shows the fake detection performance results achieved in terms of AUC and EER over the final evaluation datasets of both 1st and 2nd generations of fake videos. The results achieved using the entire face are indicated as *Face*. For each approach and database, we remark in **bold** the best performance results achieved.

For the 1st generation, AUC values close to 100% are achieved for all three detection systems, proving how easy it is for the systems to

detect fake videos of the 1st generation. In terms of EER, higher fake detection errors are achieved when using the FaceForensics++ database (around 3% EER), proving to be more challenging than the UADFV database.

Regarding the fake videos of the 2nd generation, different results are achieved depending on the database. For the Celeb-DF v2 database, good results are obtained by all fake detection approaches with an average AUC of 97.90%. As a result, and despite the good visual quality of the fake videos included in Celeb-DF, the three fake detectors considered are able to extract discriminative features between real and fake videos. We believe one of the main reasons that explains the good results achieved is the large number of fake videos included in the second version of Celeb-DF (5639). In fact, these results are highly degraded when training the systems using only the first version of Celeb-DF (795 fake videos), with an average AUC of around 85% (Tolosana et al., 2021). It is important to remark that different identities are considered in the development and evaluation datasets of our experimental protocol in order to avoid biased results.

The last database of the 2nd generation is DFDC. In this case, worse results are achieved compared with the other fake video databases. The three fake detectors obtain an average AUC of 88.85%, an absolute worsening of 9.05% AUC compared with the Celeb-DF v2 database. These results are even worse when looking at the EERs, proving to be one of the most challenging databases. This can be produced due to fake videos were generated using two different approaches, being more difficult for the fake detectors to extract robust features for both fake approaches at the same time.

Finally, it is also interesting to analyse the performance achieved by each specific fake detection approach. In general, for the fake videos of the 1st generation, the three detection approaches provide very similar results close to 100% AUC. However, different trends are observed when analysing fake videos of the 2nd generation. For the Celeb-DF v2 database, DSP-FWA slightly outperforms the other fake detectors whereas for the DFDC database, Xception clearly achieves the best results.

### 6.2. Facial regions analysis

Table 3 also includes the results achieved for each specific facial region: *Eyes*, *Nose*, *Mouth*, and *Rest*. For each database and fake detection approach, we remark in **blue** and **orange** the facial regions that provide the **best** and **worst** results, respectively. It is important to remark that a separate fake detection model is trained for each facial region and database.

In general, the facial region *Eyes* provides the best results whereas the *Rest* (i.e., the remaining part of the face after removing eyes, nose, and mouth) provides the worst results.

For the UADFV database, the *Eyes* region provides AUC values close to 100%, similar to the results achieved using the entire *Face*. This aspect was initially remarked by Matern et al. (2019), proposing features based on the missing reflection details of the eyes, achieving good results.

Regarding the FaceForensics++ database, the *Mouth* is the facial region that achieves the best results with an average 95.02% AUC. This is produced due to the lack of details in the teeth (blurred) and also the lip inconsistencies among the original face and the synthesised. Similar results are obtained when using the *Eyes* with an average AUC of 93.36%. Finally, fake detection systems based on the *Rest* of the face provide the worst results, with an average 86.24% AUC. This may happen due to both colour contrast and visible boundaries were further improved in FaceForensics++ compared with the UADFV database.

For the Celeb-DF v2 database, the best results are achieved when using the *Eyes* region, with an average 93.37% AUC, close to the results achieved using the entire *Face*, an average 97.90% AUC.

Regarding the DFDC database, worse detection results are obtained compared with the Celeb-DF v2 database. In particular, the facial

**Table 3**

Intra-database results in terms of AUC (top) and EER (bottom) over the final evaluation datasets. Two approaches are considered as input to the fake detectors: (i) selecting the entire face (*Face*), and (ii) selecting specific facial regions (*Eyes*, *Nose*, *Mouth*, *Rest*). For each approach and database, we remark in **bold** the best fake detection results, and in **blue** and **orange** the facial regions that provide the **best** and **worst** results. Fusion results are remarked in green colour.

AUC (%)	1st Generation						2nd Generation					
	UADFV			FaceForensics++			Celeb-DF v2			DFDC Preview		
	Xception	Capsule	DSP-FWA	Xception	Capsule	DSP-FWA	Xception	Capsule	DSP-FWA	Xception	Capsule	DSP-FWA
<i>Face</i>	<b>100</b>	99.90	99.74	99.40	99.52	99.48	98.30	96.80	98.59	<b>91.17</b>	87.45	87.94
<i>Fusion</i>	<b>100</b>				<b>99.79</b>		<b>99.04</b>			<b>90.61</b>		
<i>Eyes</i>	99.70	<b>100</b>	98.76	92.70	95.32	92.07	95.20	91.10	93.80	83.90	83.12	79.68
<i>Nose</i>	94.70	99.30	96.92	86.30	90.09	85.21	78.00	80.20	78.00	81.50	81.50	74.91
<i>Mouth</i>	95.40	99.56	90.88	93.90	96.18	94.99	84.80	82.10	83.80	79.50	78.14	77.96
<i>Rest</i>	97.30	94.83	91.99	85.50	86.61	86.81	84.40	78.20	84.00	76.50	72.42	65.04
<i>Fusion</i>	99.59	<b>100</b>	97.43	95.58	<b>98.00</b>	95.63	<b>95.50</b>	92.15	94.24	<b>84.43</b>	83.68	81.93

EER (%)	1st Generation						2nd Generation					
	UADFV			FaceForensics++			Celeb-DF v2			DFDC Preview		
	Xception	Capsule	DSP-FWA	Xception	Capsule	DSP-FWA	Xception	Capsule	DSP-FWA	Xception	Capsule	DSP-FWA
<i>Face</i>	1.00	2.00	1.00	3.31	2.75	3.35	6.41	7.98	5.98	<b>17.55</b>	21.39	22.37
<i>Fusion</i>	<b>0.40</b>				<b>1.96</b>		<b>4.74</b>			<b>20.44</b>		
<i>Eyes</i>	2.20	<b>0.28</b>	6.07	14.23	10.29	15.02	11.11	15.26	12.75	23.82	25.06	26.60
<i>Nose</i>	13.50	3.92	11.50	21.97	17.51	22.49	29.05	26.82	28.95	26.80	26.53	32.33
<i>Mouth</i>	12.50	3.20	11.84	13.77	9.66	13.25	23.49	24.53	23.78	27.59	27.92	28.97
<i>Rest</i>	7.90	12.30	16.20	22.37	21.58	21.24	23.48	26.93	23.74	29.94	32.56	38.21
<i>Fusion</i>	1.90	<b>0.20</b>	7.30	10.45	<b>9.47</b>	13.56	<b>10.27</b>	14.93	12.28	<b>23.41</b>	24.09	26.40

region *Eyes* provides an average 82.23% AUC, an absolute worsening of 11.14% AUC compared with the *Eyes* facial region for the Celeb-DF v2 database. These results prove the significant quality improvement of the DeepFakes databases of the 2nd generation for most facial regions.

Finally, it is also interesting to analyse the discriminative power of each facial region regarding the specific fake detector. Fig. 4 shows correct and wrong decisions of the fake detectors (based on the complete *Face*) with their corresponding Grad-CAM heatmaps (Selvaraju et al., 2017), representing the most useful areas inside the face image for each fake detector.

Analysing the correct decisions of the fake detectors, in most cases the facial region around the eyes seems to be the most discriminative to distinguish between real and fake videos. Also, other central regions of the face such as the teeth and the nostrils are used sometimes to make a better decision.

It is also interesting to analyse the wrong decisions made by the fake detectors. For the 1st generation of fake videos, most errors of the Xception and Capsule Network are produced due to the outer part of the face is selected to make the decision. For the DSP-FWA, the errors are produced due to the features extracted from the eyes and mouth are not discriminative enough. Regarding the 2nd generation, most errors are produced due to the lack of robust features as: (i) the eyes are (almost) closed, and (ii) the mouth is (almost) closed. Thus, a good approximation in order to develop better fake detectors could be to incorporate a new module to detect the eye/mouth opening before predicting whether the image is real or fake.

### 6.3. Inter-database analysis

This experiment evaluates the generalisation ability of the fake detectors to unseen digital manipulations. Table 4 shows the fake detection performance results in terms of AUC over the final evaluation datasets. We consider the approach of feeding the fake detectors with the entire *Face*. Training databases are indicated in rows whereas evaluation databases are indicated in columns. The results included in the diagonal of the tables represent the intra-database scenario studied in previous experiments.

In general, we can see poor generalisation results in all fake detection approaches compared with the intra-database scenario. Analysing the intra-database scenario for all detection approaches together, an average 96.52% AUC is achieved. However, this result is significantly degraded on inter-database scenarios with an average 65.81% AUC.

Regardless of the fake detection approach, the following trends are observed: (i) the best generalisation results are usually obtained

**Table 4**

Inter-database results in terms of AUC (%) over the final evaluation datasets. For each training database and detection system, we remark in **bold** the best fake detection results, and in **blue** and **orange** the databases that provide the **best** and **worst** results..

		Evaluation			
	Xception	UADFV	FaceForensics++	Celeb-DF	DFDC
Training	UADFV	<b>100</b>	52.00	66.58	68.41
	FaceForensics++	79.52	<b>99.40</b>	53.62	45.05
	Celeb-DF	94.75	50.00	<b>98.30</b>	72.02
	DFDC	78.21	41.02	81.04	<b>91.17</b>
		Evaluation			
	Capsule	UADFV	FaceForensics++	Celeb-DF	DFDC
Training	UADFV	<b>99.90</b>	61.82	61.55	63.27
	FaceForensics++	60.25	<b>99.52</b>	54.76	36.60
	Celeb-DF	96.73	59.92	<b>96.80</b>	70.89
	DFDC	86.32	40.49	77.56	<b>87.45</b>
		Evaluation			
	DSP-FWA	UADFV	FaceForensics++	Celeb-DF	DFDC
Training	UADFV	<b>99.74</b>	54.58	59.08	67.12
	FaceForensics++	89.06	<b>99.48</b>	58.87	55.84
	Celeb-DF	98.21	46.28	<b>98.59</b>	77.65
	DFDC	85.81	45.84	78.50	<b>87.94</b>
		Evaluation			
	Fusion	UADFV	FaceForensics++	Celeb-DF	DFDC
Training	UADFV	<b>100</b>	53.53	64.33	67.42
	FaceForensics++	88.48	<b>99.79</b>	56.59	46.79
	Celeb-DF	97.46	59.45	<b>99.04</b>	74.31
	DFDC	85.73	42.16	81.31	<b>90.61</b>
		Evaluation			
	Fusion	UADFV	FaceForensics++	Celeb-DF	DFDC
Training	UADFV	<b>100</b>	53.53	64.33	67.42
	FaceForensics++	88.48	<b>99.79</b>	56.59	46.79
	Celeb-DF	97.46	59.45	<b>99.04</b>	74.31
	DFDC	85.73	42.16	81.31	<b>90.61</b>
		Evaluation			
	Fusion	UADFV	FaceForensics++	Celeb-DF	DFDC
Training	UADFV	<b>100</b>	53.53	64.33	67.42
	FaceForensics++	88.48	<b>99.79</b>	56.59	46.79
	Celeb-DF	97.46	59.45	<b>99.04</b>	74.31
	DFDC	85.73	42.16	81.31	<b>90.61</b>
		Evaluation			
	Fusion	UADFV	FaceForensics++	Celeb-DF	DFDC
Training	UADFV	<b>100</b>	53.53	64.33	67.42
	FaceForensics++	88.48	<b>99.79</b>	56.59	46.79
	Celeb-DF	97.46	59.45	<b>99.04</b>	74.31
	DFDC	85.73	42.16	81.31	<b>90.61</b>
		Evaluation			
	Fusion	UADFV	FaceForensics++	Celeb-DF	DFDC
Training	UADFV	<b>100</b>	53.53	64.33	67.42
	FaceForensics++	88.48	<b>99.79</b>	56.59	46.79
	Celeb-DF	97.46	59.45	<b>99.04</b>	74.31
	DFDC	85.73	42.16	81.31	<b>90.61</b>
		Evaluation			
	Fusion	UADFV	FaceForensics++	Celeb-DF	DFDC
Training	UADFV	<b>100</b>	53.53	64.33	67.42
	FaceForensics++	88.48	<b>99.79</b>	56.59	46.79
	Celeb-DF	97.46	59.45	<b>99.04</b>	74.31
	DFDC	85.73	42.16	81.31	<b>90.61</b>

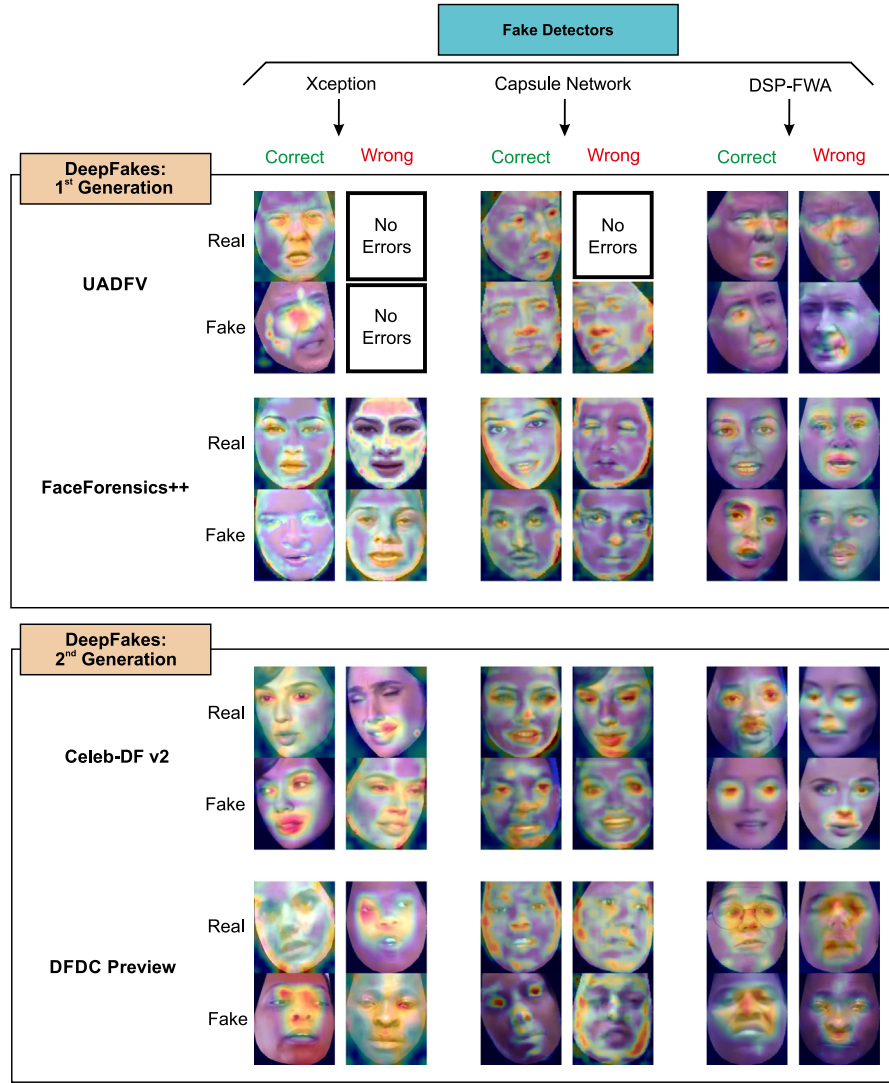


Fig. 4. Correct and wrong decisions of the fake detectors (based on the complete *Face*) with their corresponding Grad-CAM heatmaps, representing the most useful areas inside the face image for each fake detector.

approaches. For example, when FaceForensics++ is used for training, and UADFV for evaluation, AUC values of 79.52%, 60.25%, and 89.06% are achieved for Xception, Capsule, and DSP-FWA detection approaches, respectively.

#### 6.4. Fusion analysis

We finally analyse how fusion techniques can further improve the robustness of the fake detectors. Tables 3 and 4 show the **fusion** results of both intra- and inter-database scenarios in **green** colour.

We first analyse in Table 3 the fusion results achieved using the entire *Face*. In this case, we perform fusion at system level, summing the scores provided by the three fake detectors (Xception, Capsule Network, and DSP-FWA). As all approaches achieve very similar fake detection results, the weighted parameter  $k$  of each approach is equal for all systems. In general, the fusion approach outperforms individual detection approaches, achieving very good results above 99% AUC in UADFV, FaceForensics++, and Celeb-DF v2 databases. For the DFDC database, the fusion achieves a bit worse results compared with Xception.

Fusion can also be considered at facial regions. In particular, we include in Table 3 the results achieved when fusing the facial regions *Eyes* and *Mouth*. In this case, and due to the facial region *Eyes* has proven to be more discriminative than the *Mouth*, we set the weighted

parameters  $k$  to 0.75 and 0.25 for the *Eyes* and *Mouth*, respectively. In all databases, the final fusion of the facial regions achieves better results compared with the individual facial regions, and close to the fusion results achieved using the entire *Face*, especially for the 1st generation of fake videos.

Finally, we also show in Table 4 how fusion techniques can improve the detection of fake videos over inter-database scenarios. The average performances of each individual fake detector on this scenario are 65.19%, 64.18%, and 68.05% AUC for the Xception, Capsule, and DSP-FWA detectors. The fusion approach slightly outperforms the individual approaches, especially for the Xception and Capsule, with an average 68.13% AUC.

#### 6.5. Comparison with the state of the art

Table 5 compares the AUC results achieved in the present study with the state of the art. Different detection methods are considered: head pose variations (Yang et al., 2019), face warping artifacts (Li et al., 2020), mesoscopic features (Afchar et al., 2018), image and temporal features (Sabir et al., 2019; Trinh et al., 2021), eye blinking features (Li et al., 2018), pure deep learning features in combination with attention mechanisms (Dang et al., 2020), and 3D decomposition features (Zhu et al., 2021). The best results achieved for each database are remarked



**Table 5**

Comparison in terms of AUC (%) of different state-of-the-art fake detectors with the present study. The best results achieved for each database are remarked in **bold**. Results in *italics* indicate that the evaluated database was not used for training (Li et al., 2020).

Study	Method	Classifiers	AUC Results (%)			
			UADFV (Li et al., 2018)	FF++ (Rössler et al., 2019)	Celeb-DF v2 (Li et al., 2020)	DFDC-Preview (Dolhansky et al., 2019)
Yang et al. (2019)	Head Pose Features	SVM	89.0	47.3	54.6	55.9
Li et al. (2020)	Face Warping Features	CNN	97.7	93.0	64.6	75.5
Afchar et al. (2018)	Mesoscopic Features	CNN	84.3	84.7	54.8	75.3
Sabir et al. (2019)	Image + Temporal features	CNN + RNN	–	96.3	–	–
Trinh et al. (2021)	Image+ Temporal features	Dynamic Prototype Network	–	99.2	71.8	–
Li et al. (2018)	Eye blinking features	LRCN	99.0	–	–	–
Dang et al. (2020)	Deep learning features	CNN + Attention mechanism	98.4	–	71.2	–
Zhu et al. (2021)	3D Decomposition Features	CNN + Attention mechanism	–	99.5	–	66.1
Present Study	Deep learning features	Fusion of CNN	<b>100</b>	<b>99.8</b>	<b>99.1</b>	<b>90.6</b>

in **bold**. Results in *italics* indicate that the evaluated database was not used for training. These results are extracted from Li et al. (2020).

Note that the comparison in Table 5 is not always under the same datasets and protocols, therefore it must be interpreted carefully. Despite of that, it is patent that our fusion approach achieves state-of-the-art results in all databases. In particular, this improvement is significantly higher for the Celeb-DF v2 and DFDC-Preview databases.

## 7. Conclusions

This study has performed an analysis of the DeepFakes evolution in terms of fake detection performance. Popular databases such as UADFV and FaceForensics++ of the 1st generation, as well as the latest databases of the 2nd generation such as Celeb-DF and DFDC, are considered in the analysis.

Two different approaches have been followed in our evaluation framework to detect fake videos: (i) selecting the entire face as input to the fake detection system, and (ii) selecting specific facial regions such as the eyes or nose, among others, as input to the fake detectors. Fusion techniques have also been studied to further increase the robustness of the fake detectors.

Regarding the fake detection performance, we highlight the very good results achieved using fusion techniques with AUC values above 99% in UADFV, FaceForensics++, and Celeb-DF v2 databases. Also, fusion of multiple detectors has proven to be very useful under inter-database scenarios. Nevertheless, different approaches could be further studied to increase the robustness of the fake detectors against unseen attacks (Singh et al., 2020; Cozzolino et al., 2018; Baweja et al., 2020). Finally, we remark the significant improvements achieved at image level in some facial regions such as the nose, mouth, and outer part of the face in the databases of the 2nd generation, with AUC results much lower than in the 1st generation.

The analysis carried out in this study provides useful insights about the improvements achieved between 1st and 2nd DeepFakes generations, not only at the performance level of fake detectors but also in the naturalness of the DeepFakes. In addition, the insights extracted from this work could be very valuable for: (i) the development of better fake detectors, not only based on DeepFakes, but to other emerging face manipulations (Raja et al., 2020; Kim et al., 2018; Zhou et al., 2019; Fried et al., 2019), and (ii) the proposal of the next generation of face manipulations, more realistic than the current ones.

## CRedit authorship contribution statement

**Ruben Tolosana:** Conceptualization, Investigation, Writing – original draft, Writing – review & editing, Visualization, Funding acquisition. **Sergio Romero-Tapiador:** Conceptualization, Investigation, Writing – review & editing, Visualization. **Ruben Vera-Rodriguez:** Conceptualization, Writing – review & editing, Visualization, Funding acquisition. **Ester Gonzalez-Sosa:** Conceptualization, Writing – review & editing, Visualization. **Julian Fierrez:** Conceptualization, Writing – review & editing, Visualization, Funding acquisition.

## Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

## Acknowledgements

This work has been supported by projects: PRIMA (H2020-MSCA-ITN-2019-860315), TRESPASS-ETN (H2020-MSCA-ITN-2019-860813), BIBECA, Spain (MINECO/FEDER RTI2018-101248-B-I00).

## References

- Afchar, D., Nozick, V., Yamagishi, J., Echizen, I., 2018. MesoNet: A compact facial video forgery detection network. In: Proc. International Workshop On Information Forensics And Security.
- Baltrusaitis, T., Zadeh, A., Lim, Y., Morency, L., 2018. OpenFace 2.0: FAcial behavior analysis toolkit. In: Proc. IEEE International Conference On Automatic Face & Gesture Recognition.
- Baweja, Y., Oza, P., Perera, P., Patel, V.M., 2020. Anomaly detection-based unknown face presentation attack detection. arXiv preprint arXiv:2007.05856.
- Cellan-Jones, R., 2019. Deepfake videos double in nine months. <https://www.bbc.com/news/technology-49961089>.
- Chollet, F., 2017. Xception: Deep learning with depthwise separable convolutions. In: Proc. IEEE/CVF Conference On Computer Vision And Pattern Recognition.
- Ciftci, U.A., Demir, I., Yin, L., 2020. FakeCatcher: DEtection of synthetic portrait videos using biological signals. IEEE Trans. Pattern Anal. Mach. Intell.
- Citron, D., 2019. How DeepFake undermine truth and threaten democracy. <https://www.ted.com>.
- Cozzolino, D., Thies, J., Rössler, A., Riess, C., Nießner, M., Verdoliva, L., 2018. ForensicTransfer: WEakly-supervised domain adaptation for forgery detection. arXiv preprint arXiv:1812.02510.
- Dang, H., Liu, F., Stehouwer, J., Liu, X., Jain, A., 2020. On the detection of digital face manipulation. In: Proc. IEEE/CVF Conference On Computer Vision And Pattern Recognition.
- Deng, J., Dong, W., Socher, R., Li, L., Li, K., Fei-Fei, L., 2009. ImageNet: A Large-scale hierarchical image database. In: Proc. IEEE/CVF Conference On Computer Vision And Pattern Recognition.
- Dolhansky, B., Howes, R., Pflaum, B., Baram, N., Ferrer, C., 2019. The deepfake detection challenge (DFDC) preview dataset. arXiv:1910.08854.
- Fierrez, J., Morales, A., Vera-Rodriguez, R., Camacho, D., 2018. Multiple classifiers in biometrics. Part 1: Fundamentals and review. Inf. Fusion 44, 57–64.
- Fried, O., Tewari, A., Zollhöfer, M., Finkelstein, A., Shechtman, E., Goldman, D.B., Genova, K., Jin, Z., Theobalt, C., Agrawala, M., 2019. Text-based editing of talking-head video. ACM Trans. Graph. 38 (4), 1–14.
- G.E. Hinton, S.S., Frosst, N., 2018. Matrix capsules with EM routing. In: Proc. International Conference On Learning Representations Workshop.
- Güera, D., Delp, E., 2018. Deepfake video detection using recurrent neural networks. In: Proc. IEEE Int. Conference On Advanced Video And Signal Based Surveillance.
- He, K., Zhang, X., Ren, S., Sun, J., 2015. Spatial pyramid pooling in deep convolutional networks for visual recognition. IEEE Trans. Pattern Anal. Mach. Intell. 37 (9), 1904–1916.
- Hernandez-Ortega, J., Fierrez, J., Morales, A., Diaz, D., 2020. A comparative evaluation of heart rate estimation methods using face videos. In: Proc. IEEE International Workshop On Medical Computing.
- Hernandez-Ortega, J., Tolosana, R., Fierrez, J., Morales, A., 2021. DeepFakesON-Phys: DeepFakes detection based on heart rate estimation. In: Proc. 35th AAAI Conference On Artificial Intelligence Workshops.

- Jung, T., Kim, S., Kim, K., 2020. DeepVision: DEepfakes detection using human eye blinking pattern. *IEEE Access* 8, 83144–83154.
- Kim, H., Garrido, P., Tewari, A., Xu, W., Thies, J., Niessner, M., Pérez, P., Richardt, C., Zollhöfer, M., Theobalt, C., 2018. Deep video portraits. *ACM Trans. Graph.* 37 (4), 1–14.
- Kittler, J., Hatef, M., Duin, R.P., Matas, J., 1998. On combining classifiers. *IEEE Trans. Pattern Anal. Mach. Intell.* 20 (3), 226–239.
- Li, Y., Chang, M., Lyu, S., 2018. In icu oculi: Exposing AI generated fake face videos by detecting eye blinking. In: *Proc. IEEE International Workshop On Information Forensics And Security*.
- Li, Y., Lyu, S., 2019. Exposing DeepFake videos by detecting face warping artifacts. In: *Proc. IEEE/CVF Conf. On Computer Vision And Pattern Recognition Workshops*.
- Li, Y., Yang, X., Sun, P., Qi, H., Lyu, S., 2020. Celeb-DF: A large-scale challenging dataset for DeepFake forensics. In: *Proc. IEEE/CVF Conference On Computer Vision And Pattern Recognition*.
- Matern, F., Riess, C., Stamminger, M., 2019. Exploiting visual artifacts to expose DeepFakes and face manipulations. In: *Proc. IEEE Winter Applications Of Computer Vision Workshops*.
- Neves, J.C., Tolosana, R., Vera-Rodríguez, R., Lopes, V., Proença, H., Fierrez, J., 2020. GANprintR: Improved fakes and evaluation of the state-of-the-art in face manipulation detection. *IEEE J. Sel. Top. Signal Process.* 14, 1038–1048.
- Nguyen, H., Yamagishi, J., Echizen, I., 2019. Use of a capsule network to detect fake images and videos. *arXiv:1910.12467*.
- Qi, H., Guo, Q., Juefei-Xu, F., Xie, X., Ma, L., Feng, W., Liu, Y., Zhao, J., 2020. DeepRhythm: EXposing DeepFakes with attentional visual heartbeat rhythms. In: *Proc. ACM Multimedia Conference*.
- Raja, K., Ferrara, M., Franco, A., Spreeuwers, L., Batskos, I., de Wit, F., Gomez-Barrero, M., Scherhag, U., Fischer, D., Venkatesh, S., Singh, J.M., Li, G., Bergeron, L., Isadskiy, S., Ramachandra, R., Rathgeb, C., Frings, D., Seidel, U., Knopjes, F., Veldhuis, R., Maltoni, D., Busch, C., 2020. Morphing Attack Detection - Database, Evaluation Platform and Benchmarking. *IEEE Trans. Inf. Forensics Secur.*
- Rathgeb, C., Tolosana, R., Vera-Rodríguez, R., Busch, C., 2021. *Handbook Of Digital Face Manipulation And Detection: From DeepFakes to Morphing Attacks*. Springer.
- Rössler, A., Cozzolino, D., Verdoliva, L., Riess, C., Thies, J., Nießner, M., 2018. FaceForensics: A Large-scale video dataset for forgery detection in human faces. *arXiv:1803.09179*.
- Rössler, A., Cozzolino, D., Verdoliva, L., Riess, C., Thies, J., Nießner, M., 2019. FaceForensics++: LEarning to detect manipulated facial images. In: *Proc. IEEE/CVF International Conference On Computer Vision*.
- Sabir, E., Cheng, J., Jaiswal, A., AbdAlmageed, W., Masi, I., Natarajan, P., 2019. Recurrent convolutional strategies for face manipulation detection in videos. In: *Proc. IEEE/CVF Conference On Computer Vision And Pattern Recognition Workshops*.
- Selvaraju, R., Cogswell, M., Das, A., Vedantam, R., Parikh, D., Batra, D., 2017. Grad-CAM: Visual explanations from deep networks via gradient-based localization. In: *Proc. IEEE/CVF International Conference On Computer Vision*.
- Singh, M., Singh, R., Ross, A., 2019. A comprehensive overview of biometric fusion. *Inf. Fusion* 52, 187–205.
- Singh, R., Vatsa, M., Patel, V.M., Ratha, N., 2020. *Domain Adaptation For Visual Understanding*. Springer.
- Szegedy, C., Liu, W., Jia, Y., Sermanet, P., Reed, S., Anguelov, D., Erhan, D., Vanhoucke, V., Rabinovich, A., 2015. Going deeper with convolutions. In: *Proc. IEEE/CVF Conference On Computer Vision And Pattern Recognition*.
- Tolosana, R., Romero-Tapiador, S., Fierrez, J., Vera-Rodríguez, R., 2021. DeepFakes Evolution: Analysis of facial regions and fake detection performance. In: *Proc. International Conference on Pattern Recognition Workshops*.
- Tolosana, R., Vera-Rodríguez, R., Fierrez, J., Morales, A., Ortega-García, J., 2020. DeepFakes And beyond: A survey of face manipulation and fake detection. *Inf. Fusion* 64, 131–148.
- Trinh, L., Tsang, M., Rambhatla, S., Liu, Y., 2021. Interpretable and trustworthy deepfake detection via dynamic prototypes. In: *Proc. IEEE/CVF Winter Conference On Applications Of Computer Vision*.
- Verdoliva, L., 2020. Media forensics and DeepFakes: An overview. *IEEE J. Sel. Top. Signal Process.* 14, 910–932.
- Yang, X., Li, Y., Lyu, S., 2019. Exposing deep fakes using inconsistent head poses. In: *Proc. IEEE Int. Conference On Acoustics, Speech And Signal Processing*.
- Zhou, H., Liu, Y., Liu, Z., Luo, P., Wang, X., 2019. Talking face generation by adversarially disentangled audio-visual representation. In: *Proc. AAAI Conference On Artificial Intelligence*.
- Zhu, X., Wang, H., Fei, H., Lei, Z., Li, S.Z., 2021. Face forgery detection by 3D decomposition. In: *Proc. IEEE/CVF Conference On Computer Vision And Pattern Recognition*.