✓ Xception on Forensics++ with the same idea on train-test split.
2nd generation fakes are better. NN is not a complete black box.
Techniques to explain the result

DeepFakes Evolution: Analysis of Facial Regions and Fake Detection Performance

Ruben Tolosana, Sergio Romero-Tapiador, Julian Fierrez and Ruben Vera-Rodriguez
Biometrics and Data Pattern Analytics - BiDA Lab, Universidad Autonoma de Madrid
{ruben.tolosana, julian.fierrez, ruben.vera}@uam.es, sergio.romerotapiador@gmail.com

***Abstract*—Media forensics has attracted a lot of attention in the last years in part due to the increasing concerns around DeepFakes. Since the initial DeepFake databases from 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 exhaustive analysis of both 1st and 2nd DeepFake generations in terms of facial regions and fake detection performance. Two different methods are considered in our experimental framework: *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.**

Among all the findings resulting from our experiments, we highlight the poor fake detection results achieved even by the strongest state-of-the-art fake detectors in the latest DeepFake databases of the 2nd generation, with Equal Error Rate results ranging from 15% to 30%. These results remark the necessity of further research to develop more sophisticated fake detectors.

***Index Terms*—Fake News, DeepFakes, Media Forensics, Face Manipulation, Fake Detection, Benchmark**

I. Introduction

Fake images and videos including facial information generated by digital manipulations, in particular with DeepFake methods [1], [2], have become a great public concern recently [3], [4]. The very popular term "DeepFake" 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. Open software and mobile applications such as ZAO[1] allow nowadays to automatically generate fake videos by anyone, without a prior knowledge of the task. But, how real are these fake videos compared with the authentic ones[2]?

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 [1]. Since the initial publicly available fake databases, such as the UADFV database [5], up to the recent Celeb-DF and Deepfake Detection Challenge (DFDC) databases [6], [7], 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.

In general, fake videos of the 1st generation are characterised by: *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, and pose variations, among others. So, the question is, how easy is for a machine to automatically detect these kind of fakes?

Different fake detectors have been proposed based on the visual features existed in the 1st generation of fake videos. Yang *et al.* performed in [8] a study based on the differences existed between head poses using a full set of facial landmarks (68 extracted from DLib [9]) and those in the central face regions to differentiate fake from real videos. Once these features were extracted, Support Vector Machines (SVM) were considered for the final classification, achieving an Area Under the Curve (AUC) of 89.0% for the UADFV database [5].

The same authors proposed in [10] another approach based on the detection of face warping artifacts. They proposed a detection system based on Convolutional Neural Networks (CNN) in order to detect the presence of such artifacts from the face and the surrounding areas. Their proposed detection approach was tested using the UADFV and DeepfakeTIMIT databases [5], [11], outperforming the state of the art with 97.4% and 99.9% AUCs, respectively.

Agarwal *et al.* proposed in [12] a detection technique based on facial expressions and head movements. Their proposed approach achieved a final AUC of 96.3% over their own database, being robust against new manipulation techniques.

Finally, Sabir *et al.* proposed in [13] to detect fake videos through the temporal discrepancies across frames. They considered a Recurrent Convolutional Network similar to [14], trained end-to-end instead of using a pre-trained model. Their proposed detection approach was tested using FaceForensics++ database [15], achieving AUC results of 96.9% and 96.3% for the DeepFake and FaceSwap methods, respectively.

Therefore, very good fake detection results are already achieved on databases of the 1st generation, being an almost solved problem. But, what is the performance achieved on current Identity Swap databases of the 2nd generation?

[1]https://apps.apple.com/cn/app/id1465199127
[2]https://www.youtube.com/watch?v=UlvoEW7l5rs

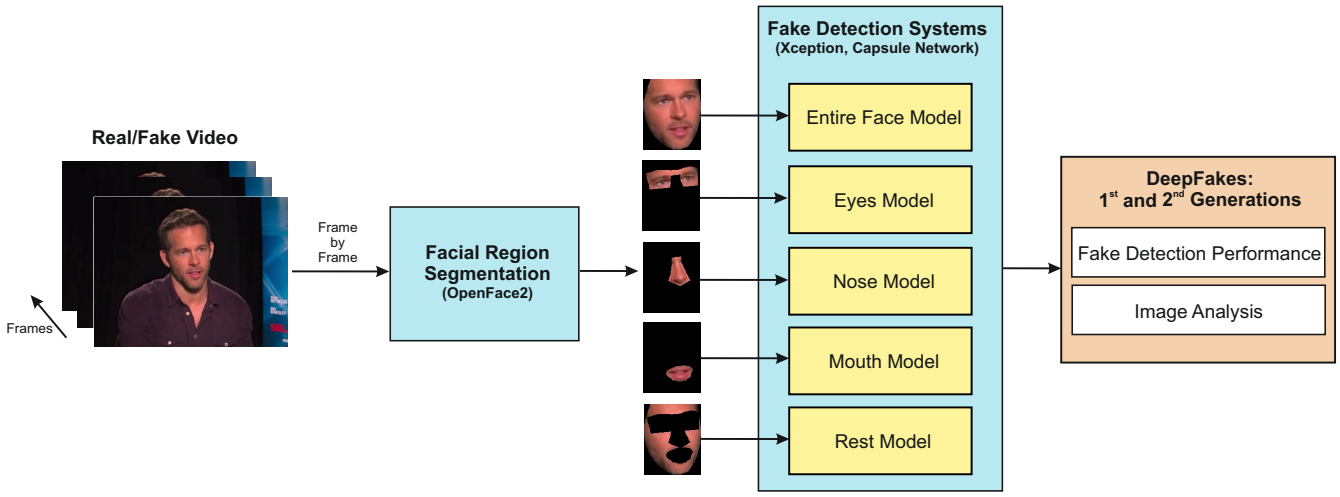

Fig. 1. Architecture of our evaluation framework to analyse both facial regions and fake detection performance in DeepFake video databases of the 1st and 2nd generations. Two different approaches are studied: *i)* selecting the entire face as input to the fake detection system, and *ii)* selecting specific facial regions.

The present study provides an exhaustive analysis of both 1st and 2nd DeepFake generations using state-of-the-art fake detectors. 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 [1], and *ii)* a novel approach based on the selection of specific facial regions as input to the fake detection system. The main contributions of this study are as follow:

- An in-depth comparison in terms of performance among Identity Swap databases of the 1st and 2nd generation. In particular, two different state-of-the-art fake detectors are considered: *i)* Xception, and *ii)* Capsule Network.
- An analysis of the discriminative power of the different facial regions between the 1st and 2nd generations, and also between fake detectors.

The analysis carried out in this study will benefit the research community for many different reasons: *i)* insights for 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.

The remainder of the paper is organised as follows. Sec. II describes our proposed evaluation framework. Sec. III summarises all databases considered in the experimental framework of this study. Sec. IV and V describe the experimental protocol and results achieved, respectively. Finally, Sec. VI draws the final conclusions and points out future research lines.

II. Proposed Evaluation Framework

Fig. 1 graphically summarises our evaluation framework. It comprises two main modules: *i)* facial region segmentation, described in Sec. II-A, and *ii)* fake detection systems, described in Sec. II-B.

A. 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 [16]. This toolbox extracts 68 total landmarks for each face. Fig. 2 shows an example of the 68 landmarks (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 [7]. 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. 2. Once each facial region is segmented, the remaining part of the face is discarded (black background as depicted in Fig. 1). 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.

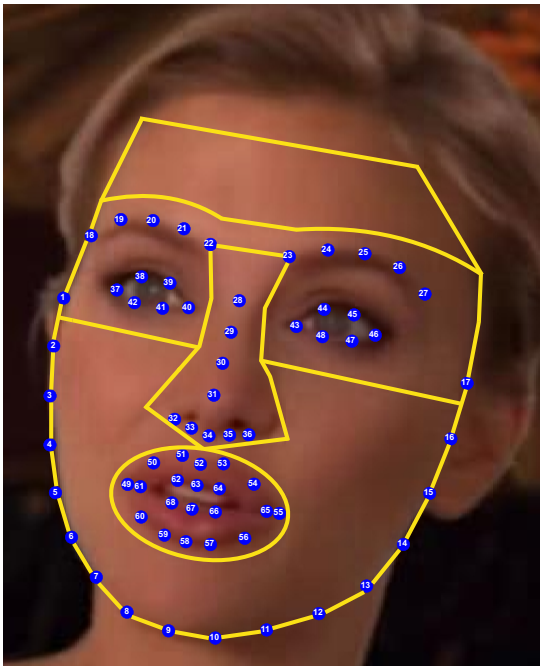

Fig. 2. Example of the different facial regions (i.e., *Eyes*, *Nose*, *Mouth*, and *Rest*) extracted using the 68 facial landmarks provided by OpenFace2 [16].

B. Fake Detection Systems

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

- *Xception* [17]: this network has achieved very good fake detection results in recent studies [7], [15], [18], [19]. Xception is a CNN architecture inspired by Inception [20], where Inception modules have been replaced with depthwise separable convolutions. In our evaluation framework, we follow the same training approach considered in [15]: *i)* we first consider the Xception model pre-trained with ImageNet [21], *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* [22]: we consider the same detection approach proposed by Nguyen *et al.*, which is publicly available in GitHub[3]. It is based on the combination of traditional CNN and recent Capsule Networks, which require fewer parameters to train compared with traditional CNN [23]. 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 [22].

Finally, as shown in Fig. 1, it is important to highlight that we train a specific fake detector per database and facial region.

[3]https://github.com/nii-yamagishilab/Capsule-Forensics-v2

TABLE I
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) [5]	49 (Youtube)	49 (FakeApp)
FaceForensics++ (2019) [15]	1,000 (Youtube)	1,000 (FaceSwap)

2nd Generation

Database	*Real Videos*	*Fake Videos*
Celeb-DF (2019) [6]	408 (Youtube)	795 (DeepFake)
DFDC Preview (2019) [7]	1,131 (Actors)	4,119 (Unknown)

III. 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 I summarises their main features.

A. UADFV

The UADFV database [5] comprises 49 real videos downloaded from Youtube, which were used to create 49 fake videos through the FakeApp mobile application[4], 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 294×500 pixels, and 11.14 seconds on average.

B. FaceForensics++

The FaceForensics++ database [15] was introduced in 2019 as an extension of the original FaceForensics [24], which was focused only on Expression Swap manipulations. FaceForensics++ contains 1,000 real videos extracted from Youtube. Fake videos were generated using both computer graphics and deep learning approaches (1,000 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[5]. This algorithm consists of face alignment, Gauss Newton optimization and image blending to swap the face of the source person to the target person.

C. Celeb-DF

The aim of the Celeb-DF database [6] was to generate fake videos of better visual quality compared with their original UADFV database. This database consists of 408 real videos extracted from Youtube, corresponding to interviews of 59

[4]https://fakeapp.softonic.com/

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

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 795 videos were created using DeepFake technology, swapping faces for each pair of the 59 subjects. The final videos are in MPEG4.0 format.

D. DFDC

The DFDC database [7] 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 1,131 real videos from 66 paid actors, ensuring realistic 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 4,119 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 others 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.

IV. Experimental Protocol

All databases have been divided into non-overlapping datasets, development ($\simeq$ 80% of the identities) and evaluation ($\simeq$ 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 [15], 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 [7] as the authors already considered this concern. Finally, for the Celeb-DF database, we consider real/fake videos of 40 and 19 different identities for the development and evaluation datasets, respectively.

V. Experimental Results

Two experiments are considered: *i)* Sec. V-A considers the traditional scenario of feeding the fake detectors with the entire face, and *ii)* Sec. V-B analyses the discriminative power of each facial region. Finally, we compare in Sec. V-C the results achieved in this study with the state of the art.

A. Entire Face Analysis

Table II shows the fake detection performance results achieved in terms of Equal Error Rate (EER) and AUC 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 database and fake detection approach, we remark in **bold** the best performance results achieved.

Analysing the fake videos of the 1st generation, AUC values close to 100% are achieved, proving how easy it is for both 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 DeepFake databases of the 2nd generation, a high performance degradation is observed in both fake detectors when using Celeb-DF and DFDC databases. In particular, an average 23.05% EER is achieved for Xception whereas for Capsule Network, the average EER is 22.84%. As a result, an average absolute worsening of around 20% EER is produced for both fake detectors compared with the databases of the 1st generation. This degradation is specially substantial for the Celeb-DF database, with EER values of 28.55% and 24.29% for Xception and Capsule Network fake detectors, respectively. These results prove the higher realism achieved in the 2nd in comparison with the 1st DeepFake generation.

Finally, we would like to highlight the importance of selecting different identities (not only videos) for the development and final evaluation of the fake detectors, as we have done in our experimental framework. As an example of how relevant this aspect is, Table III shows the detection performance results achieved using Xception for the *Same* and *Different* identities between development and evaluation of Celeb-DF. As can be seen, much better results are obtained for the scenario of considering the *Same* identities, up to 5 times better compared with the *Different* identities scenario. The *Same* identities scenario generates a misleading result because the network is learning intrinsic features from the identities, not the key features to distinguish among real and fake videos. Therefore, poor results are expected to be achieved when testing with other identities. This is a key aspect not considered in the experimental protocol of many previous studies.

B. Facial Regions Analysis

Table II 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 addition, we also visualise in Fig. 3 which part of the image is more important for the final decision, for both real and fake examples. We consider the popular heatmap visualisation technique Grad-CAM [25]. Similar Grad-CAM results are obtained for both Xception and Capsule Network.

In general, as shown in Table II, the facial region *Eyes* provides the best results whereas the *Rest* (i.e., the remaining

TABLE II

FAKE DETECTION PERFORMANCE RESULTS IN TERMS OF EER (%) AND AUC (%) OVER THE FINAL EVALUATION DATASETS. TWO APPROACHES ARE CONSIDERED AS INPUT TO THE FAKE DETECTION SYSTEMS: *i)* SELECTING THE ENTIRE FACE (*Face*), AND *ii)* SELECTING SPECIFIC FACIAL REGIONS (*Eyes*, *Nose*, *Mouth*, *Rest*). 1ST GENERATION DATABASES: UADFV AND FACEFORENSIC++. 2ND GENERATION DATABASES: CELEB-DF AND DFDC. FOR EACH 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, RESPECTIVELY.

Xception	*Face*		*Eyes*		*Nose*		*Mouth*		*Rest*	
	EER (%)	AUC (%)	EER (%)	AUC (%)	EER (%)	AUC (%)	EER (%)	AUC (%)	EER (%)	AUC (%)
UADFV (2018) [5]	**1.00**	**100**	2.20	99.70	13.50	94.70	12.50	95.40	7.90	97.30
FaceForensics++ (2019) [15]	**3.31**	**99.40**	14.23	92.70	21.97	86.30	13.77	93.90	22.37	85.50
Celeb-DF (2019) [6]	**28.55**	**83.60**	29.40	77.30	38.46	64.90	39.37	65.10	43.55	60.10
DFDC Preview (2019) [7]	**17.55**	**91.17**	23.82	83.90	26.80	81.50	27.59	79.50	29.94	76.50

Capsule Network	*Face*		*Eyes*		*Nose*		*Mouth*		*Rest*	
	EER (%)	AUC (%)	EER (%)	AUC (%)	EER (%)	AUC (%)	EER (%)	AUC (%)	EER (%)	AUC (%)
UADFV (2018) [5]	2.00	99.90	**0.28**	**100**	3.92	99.30	3.20	99.56	12.30	94.83
FaceForensics++ (2019) [15]	**2.75**	**99.52**	10.29	95.32	17.51	90.09	9.66	96.18	21.58	86.61
Celeb-DF (2019) [6]	**24.29**	**82.46**	30.58	76.64	37.39	66.24	35.36	67.75	36.64	68.56
DFDC Preview (2019) [7]	**21.39**	**87.45**	25.06	83.12	26.53	81.50	27.92	78.14	32.56	72.42

TABLE III

FAKE DETECTION RESULTS IN TERMS OF EER (%) USING XCEPTION OVER THE FINAL EVALUATION DATASET OF CELEB-DF. TWO SCENARIOS ARE ANALYSED REGARDING WHETHER THE SAME IDENTITIES ARE USED FOR THE DEVELOPMENT AND FINAL EVALUATION OF THE DETECTORS OR NOT. IN BOTH SCENARIOS, DIFFERENT VIDEOS (REAL AND FAKE) ARE CONSIDERED IN EACH DATASET.

	Face	*Eyes*	*Nose*	*Mouth*	*Rest*
Same identities	**5.66**	12.06	23.44	17.81	21.58
Different identities	**28.55**	29.40	38.46	39.37	43.55

part of the face after removing eyes, nose, and mouth) provides the worst results.

For the UADFV database, the *Eyes* provides EER values close to the results achieved using the entire *Face*. It is important to highlight the results achieved by the Capsule Network as in this case the fake detector based only on the *Eyes* has outperformed the case of feeding the detector with the entire *Face* (2.00% vs. 0.28% EER). The discriminative power of the *Eyes* facial region was preliminary studied by Matern *et al.* in [26], proposing features based on the missing reflection details of the eyes. Also, in this particular database, Xception achieves good results using the *Rest* of the face, 7.90% EER. This is produced due to the different colour contrast among the synthesised fake mask and real skin, and also to the visible boundaries of the fake mask. These aspects can be noticed in the examples included in Fig. 3.

Regarding the FaceForensics++ database, the *Mouth* is the facial region that achieves the best result for both Xception and Capsule Network with EER values of 13.77% and 9.66%. 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*. It is interesting to see in Fig. 3 how the decision of the fake detection systems is mostly based on a single eye (the same happens in other databases such as UADFV). Finally, the fake detection system based on the *Rest* of the face provides the worst result, EER values of 22.37% and 21.58% for Xception and Capsule Network, respectively. This may happen because both colour contrast and visible boundaries were further improved in FaceForensics++ compared with the UADFV database.

It is also interesting to analyse the ability of each approach for the detection of fake videos of the 1st generation. In general, much better results are obtained using Capsule Networks compared with Xception. For example, regarding the UADFV database, EER absolute improvements of 1.92%, 9.58%, and 9.30% are obtained for the *Eyes*, *Nose*, and *Mouth*, respectively.

Analysing the Celeb-DF database of the 2nd generation, the best results for local regions are achieved when using the *Eyes* of the face, with EER values around 30%, similar to using the entire *Face* for Xception. It is important to remark that this EER is over 13 times higher than the original 2.20% and 0.28% EERs achieved by Xception and Capsule Network on the UADFV. Similar poor detection results, around 40% EER, are obtained when using other facial regions, being one of the most challenging databases nowadays. Fig. 3 depicts some fake examples of Celeb-DF, showing very realistic features such as the colour contrast, boundaries of the mask, quality

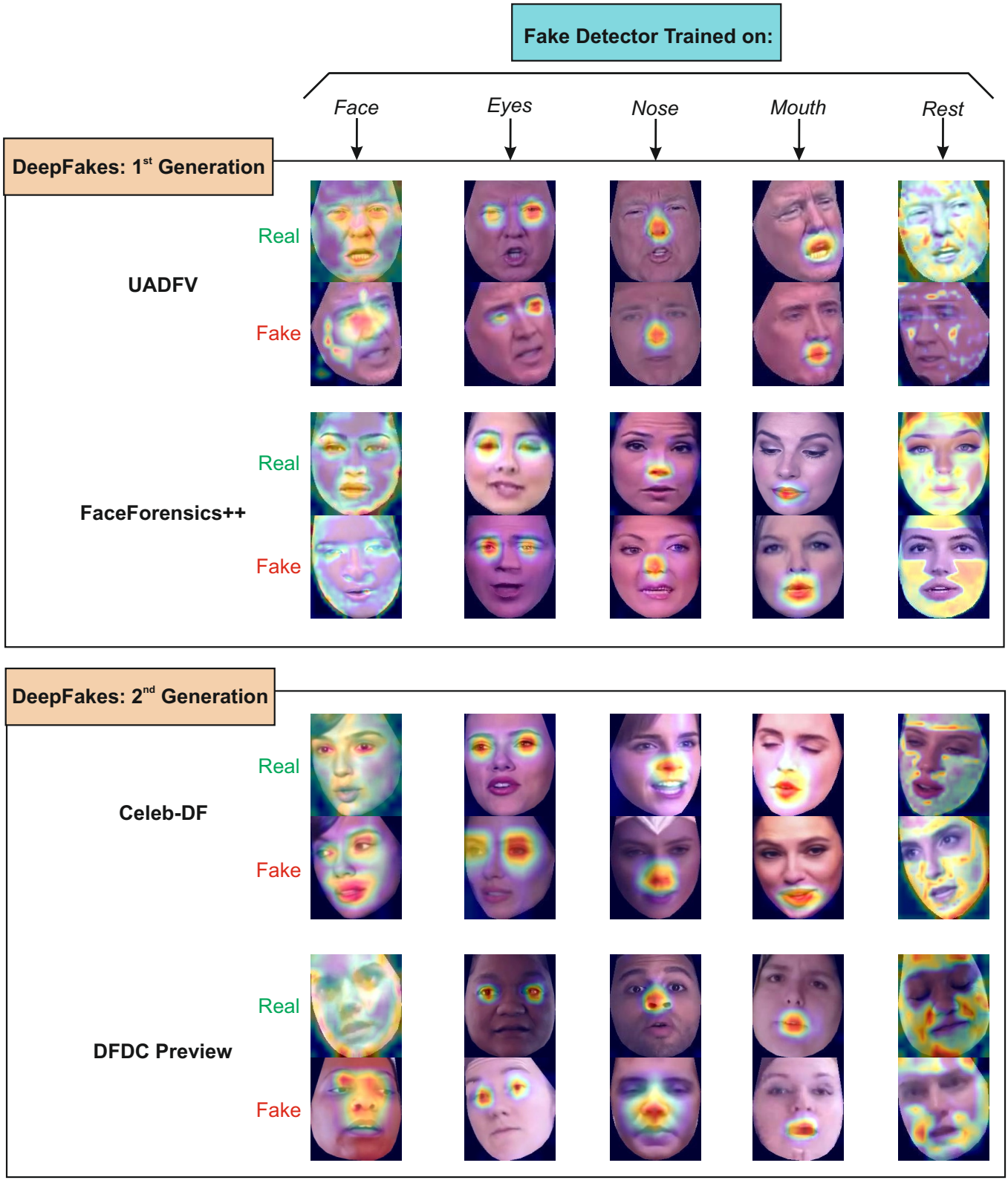

Fig. 3. Real and fake image examples of the DeepFake video databases evaluated in the present paper with their corresponding Grad-CAM heatmaps, representing the facial features most useful for each fake detector (i.e., *Face*, *Eyes*, *Nose*, *Mouth*, and *Rest*).

TABLE IV

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 [6].

Study	Method	Classifiers	AUC Results (%)			
			UADFV [5]	FF++ [15]	Celeb-DF [6]	DFDC [7]
Yang *et al.* [8]	Head Pose Features	SVM	89.0	*47.3*	*54.6*	*55.9*
Li *et al.* [6]	Face Warping Features	CNN	97.7	*93.0*	*64.6*	*75.5*
Afchar *et al.* [27]	Mesoscopic Features	CNN	*84.3*	84.7	*54.8*	*75.3*
Sabir *et al.* [13]	Image + Temporal Features	CNN + RNN	-	96.3	-	-
Dang *et al.* [19]	Deep Learning Features	CNN + Attention Mechanism	98.4	-	*71.2*	-
Present Study	Deep Learning Features	Xception [17]	**100**	99.4	**83.6**	**91.1**
		Capsule Network [22]	**100**	**99.5**	82.4	87.4

of the eyes, teeth, nose, etc.

Regarding the DFDC database, better detection results are obtained compared with the Celeb-DF database. In particular, the facial region *Eyes* also provides the best results with EER values of 23.82% and 25.06%, an absolute improvement of 5.58% and 5.52% EER compared with the *Eyes* facial region of Celeb-DF. Despite this performance improvement, the EER is still much worse compared with the databases of the 1st generation.

To summarise this section, we have observed significant improvements in the realism of DeepFakes of the 2nd in comparison with the 1st generation for some specific facial regions. In particular, for the *Nose*, *Mouth*, and the edge of the face (*Rest*). This realism provokes a lot of fake detection errors even for the advanced detectors explored in the present paper, which result in EER values between 24% and 44% depending on the database. The quality of the *Eyes* has also been improved, but it is still the facial region most useful to detect fake images, as depicted in Fig. 3.

C. Comparison with the State of the Art

Finally, we compare in Table IV the AUC results achieved in the present study with the state of the art. Different methods are considered to detect fake videos: head pose variations [8], face warping artifacts [6], mesoscopic features [27], image and temporal features [13], and pure deep learning features in combination with attention mechanisms [19]. The best results achieved for each database are remarked in **bold**. Results in *italics* indicate that the evaluated database was not used for training. These results are extracted from [6].

Note that the comparison in Table IV is not always under the same datasets and protocols, therefore it must be interpreted with care. Despite of that, it is patent that both Xception and Capsule Network fake detectors have achieved state-of-the-art results in all databases. In particular, for Celeb-DF and DFDC, Xception obtains the best results whereas for FaceForensics++, Capsule Network is the best. The same good results are obtained by both detectors on UADFV.

VI. CONCLUSIONS

In this study we have performed an exhaustive analysis of the DeepFakes evolution, focusing on facial regions and fake detection performance. Popular databases such as UADFV and FaceForensics++ of the 1st generation, as well as the latest databases such as Celeb-DF and DFDC of the 2nd generation, 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 detection system.

Regarding the fake detection performance, we highlight the very poor results achieved in the latest DeepFake video databases of the 2nd generation with EER values around 20-30%, compared with the EER values of the 1st generation ranging from 1% to 3%. In addition, we remark the significant improvements in the realism achieved at image level in some facial regions such as the nose, mouth, and edge of the face in DeepFakes of the 2nd generation, resulting in fake detection results between 24% and 44% EERs.

The analysis carried out in this study provides useful insights for the research community, e.g.: *i)* for 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.

ACKNOWLEDGMENTS

This work has been supported by projects: PRIMA (H2020-MSCA-ITN-2019-860315), TRESPASS-ETN (H2020-MSCA-ITN-2019-860813), BIBECA (MINECO/FEDER RTI2018-101248-B-I00), and Accenture. Ruben Tolosana is supported by Consejería de Educación, Juventud y Deporte de la Comunidad de Madrid y Fondo Social Europeo.

REFERENCES

[1] R. Tolosana, R. Vera-Rodriguez, J. Fierrez, A. Morales, and J. Ortega-Garcia, "DeepFakes and Beyond: A Survey of Face Manipulation and Fake Detection," *Information Fusion*, 2020.

[2] L. Verdoliva, "Media Forensics and DeepFakes: an Overview," *arXiv preprint arXiv:2001.06564*, 2020.

[3] D. Citron, "How DeepFake Undermine Truth and Threaten Democracy," 2019. [Online]. Available: https://www.ted.com

[4] R. Cellan-Jones, "Deepfake Videos Double in Nine Months," 2019. [Online]. Available: https://www.bbc.com/news/technology-49961089

[5] Y. Li, M. Chang, and S. Lyu, "In Ictu Oculi: Exposing AI Generated Fake Face Videos by Detecting Eye Blinking," in *Proc. IEEE International Workshop on Information Forensics and Security*, 2018.

[6] Y. Li, X. Yang, P. Sun, H. Qi, and S. Lyu, "Celeb-DF: A Large-Scale Challenging Dataset for DeepFake Forensics," in *Proc. IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2020.

[7] B. Dolhansky, R. Howes, B. Pflaum, N. Baram, and C. C. Ferrer, "The Deepfake Detection Challenge (DFDC) Preview Dataset," *arXiv preprint arXiv:1910.08854*, 2019.

[8] X. Yang, Y. Li, and S. Lyu, "Exposing Deep Fakes Using Inconsistent Head Poses," in *Proc. International Conference on Acoustics, Speech and Signal Processing*, 2019.

[9] D. King, "DLib-ML: A Machine Learning Toolkit," *Journal of Machine Learning Research*, vol. 10, pp. 1755–1758, 2009.

[10] Y. Li and S. Lyu, "Exposing DeepFake Videos By Detecting Face Warping Artifacts," in *Proc. IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops*, 2019.

[11] P. Korshunov and S. Marcel, "Deepfakes: a New Threat to Face Recognition? Assessment and Detection," *arXiv preprint arXiv:1812.08685*, 2018.

[12] S. Agarwal and H. Farid, "Protecting World Leaders Against Deep Fakes," in *Proc. IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops*, 2019.

[13] E. Sabir, J. Cheng, A. Jaiswal, W. AbdAlmageed, I. Masi, and P. Natarajan, "Recurrent Convolutional Strategies for Face Manipulation Detection in Videos," in *Proc. IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops*, 2019.

[14] D. Güera and E. Delp, "Deepfake Video Detection Using Recurrent Neural Networks," in *Proc. International Conference on Advanced Video and Signal Based Surveillance*, 2018.

[15] A. Rössler, D. Cozzolino, L. Verdoliva, C. Riess, J. Thies, and M. Nießner, "FaceForensics++: Learning to Detect Manipulated Facial Images," in *Proc. IEEE/CVF International Conference on Computer Vision*, 2019.

[16] T. Baltrusaitis, A. Zadeh, Y. Lim, and L. Morency, "OpenFace 2.0: Facial Behavior Analysis Toolkit," in *Proc. International Conference on Automatic Face & Gesture Recognition*, 2018.

[17] F. Chollet, "Xception: Deep Learning with Depthwise Separable Convolutions," in *Proc. IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2017.

[18] J. Neves, R. Tolosana, R. Vera-Rodriguez, V. Lopes, H. Proença, and J. Fierrez, "GANprintR: Improved Fakes and Evaluation of the State of the Art in Face Manipulation Detection," *IEEE Journal of Selected Topics in Signal Processing*, 2020.

[19] H. Dang, F. Liu, J. Stehouwer, X. Liu, and A. Jain, "On the Detection of Digital Face Manipulation," in *Proc. IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2020.

[20] C. Szegedy, W. Liu, Y. Jia, P. Sermanet, S. Reed, D. Anguelov, D. Erhan, V. Vanhoucke, and A. Rabinovich, "Going Deeper with Convolutions," in *Proc. IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2015.

[21] J. Deng, W. Dong, R. Socher, L. Li, K. Li, and L. Fei-Fei, "ImageNet: A Large-Scale Hierarchical Image Database," in *Proc. IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2009.

[22] H.H. Nguyen, J. Yamagishi and I. Echizen, "Use of a Capsule Network to Detect Fake Images and Videos," *arXiv preprint arXiv:1910.12467*, 2019.

[23] G.E. Hinton, S. Sabour and N. Frosst, "Matrix Capsules with EM Routing," in *Proc. International Conference on Learning Representations Workshop*, 2018.

[24] A. Rössler, D. Cozzolino, L. Verdoliva, C. Riess, J. Thies, and M. Nießner, "FaceForensics: A Large-Scale Video Dataset for Forgery Detection in Human Faces," *arXiv preprint arXiv:1803.09179*, 2018.

[25] R. Selvaraju, M. Cogswell, A. Das, R. Vedantam, D. Parikh, and D. Batra, "Grad-CAM: Visual Explanations from Deep Networks via Gradient-based Localization," in *Proc. IEEE International Conference on Computer Vision*, 2017.

[26] F. Matern, C. Riess, and M. Stamminger, "Exploiting Visual Artifacts to Expose DeepFakes and Face Manipulations," in *Proc. IEEE Winter Applications of Computer Vision Workshops*, 2019.

[27] D. Afchar, V. Nozick, J. Yamagishi, and I. Echizen, "MesoNet: a Compact Facial Video Forgery Detection Network," in *Proc. IEEE International Workshop on Information Forensics and Security*, 2018.