A Case Study on How Beautification Filters Can Fool Deepfake Detectors

Alexandre Libourel, Sahar Husseini, Nelida Mirabet-Herranz, Jean-Luc Dugelay

A Case Study on How Beautification Filters Can Fool Deepfake Detectors

Alexandre Libourel*[1], Sahar Husseini*[1,2], Nelida Mirabet-Herranz*[1], Jean-Luc Dugelay[1]
[1]*Department of Digital Security EURECOM*
[2]*Docaposte Biometrics Lab*
Biot, France
{libourel, husseini, mirabet, dugelay}@eurecom.fr

***Abstract*—If telling the difference between a real video and a deepfake is difficult, with the proliferation of beautification filters on social networks it becomes nearly impossible to differentiate between a real video, a video enhanced by a filter, and a video with its original identity replaced. Therefore, is it possible to fool state-of-the-art (SotA) detectors by simply applying a beautification filter to the manipulated video? In this paper, we study the impact of beautification filters on Celeb-DF-B, a novel database created by applying popular social media beautification filters to a subset of real and fake videos from the Celeb-DF dataset. We assessed three SotA passive deepfake detectors, comparing their performance against that of human evaluators. The results indicate that filters significantly alter the behavior of the three detectors studied, resulting in a notable decrease in the video-level AUC when classifying beautified videos. In the context of human-level performance, the use of filters similarly influences human decision-making, affecting the accurate categorization of videos as either real or fake.**

***Index Terms*—Deepfake detection, Social media filters, Beautification, Subjective evaluation**

I. Introduction

Creating, sharing and visualizing videos has become a daily activity for mobile users in the past decade. Yet, determining the authenticity of those videos is becoming a challenge due to the popularity and availability of deepfake methods [16]. Generative Adversarial Networks (GAN) [10] and, more recently, Denoising Diffusion Probabilistic Models (DDPM) [13] have led to significant advances in the generation of synthetic media, namely video deepfakes. These fake videos aim to portray a person in a situation they have not experienced in a highly realistic way to deceive the human eye. Although various types of deepfake videos exist, the currently most prevalent involves a *face swap* between a target person and an individual in a video. While these fake images can be seen as entertainment for the film and advertising industries, they also raise issues of privacy and credibility with the massive sharing on social networks. On one hand, research efforts have been directed toward evaluating how effective deepfakes are at deceiving human perception. There is a prevailing assumption that deepfakes are highly realistic [17], however, the accuracy of this belief, particularly in the context of automatically generated deepfakes, remains to be thoroughly investigated. It is essential to determine whether such deepfakes pose a significant threat to the human ability to discern video authenticity and it is of interest to explore the relationship between human perceptual accuracy and the effectiveness of automated deepfake detection systems. On the other hand, automatic classification between genuine and fake videos has been the focus of research in past years. AI-based deepfake detectors are being developed and made publicly available for the average user. A distinction is made between passive and active deepfake detectors [25]. Active detectors rely on the modification of the original videos before they are used for deepfake generation like watermarking or adversarial attacks. Passive detectors are trained to learn and detect intrinsic features of manipulated content without interfering before the manipulation.

Social media platforms offer a diverse range of tools referred to as "filters" designed to automatically enhance a user's image, demanding minimal or no user proficiency [21]. Certain types of filters are designed to tweak different facial features such as skin, lips, eyes, and nose to enhance the beauty of the user. We will refer to those filters as *beautification filters*. Some common modifications are makeup addition, narrow noses, skin tanning and smoothening. *Beautification filters* have been demonstrated as a disturbance factor for AI facial processing tasks such as face recognition and gender classification [21]. Despite deepfake detection technology being challenged against several video processing operations [20], its robustness against social media beautification effects has not yet been tested.

In this paper, we study the behavior of 3 SotA passive deepfake detectors trained on the FaceForensics++ (FF++) dataset [24]. Our objective is to test the robustness of deepfake detectors against beautified videos and measure the impact of the *beautification filters* on the classification score. Moreover, we compare the performance of those detectors with the ability of an average user to classify real and fake videos when they are beautified. The pipeline is presented in Figure 1. The key contributions of this study include:

- We introduce a new benchmark dataset, the Celeb-DB-B database based on a subset of videos from the Celeb-DF dataset and composed of 928 videos balanced in terms of four categories Real, Real-Beautified, Fake, Fake-

*Equal contribution to this work.

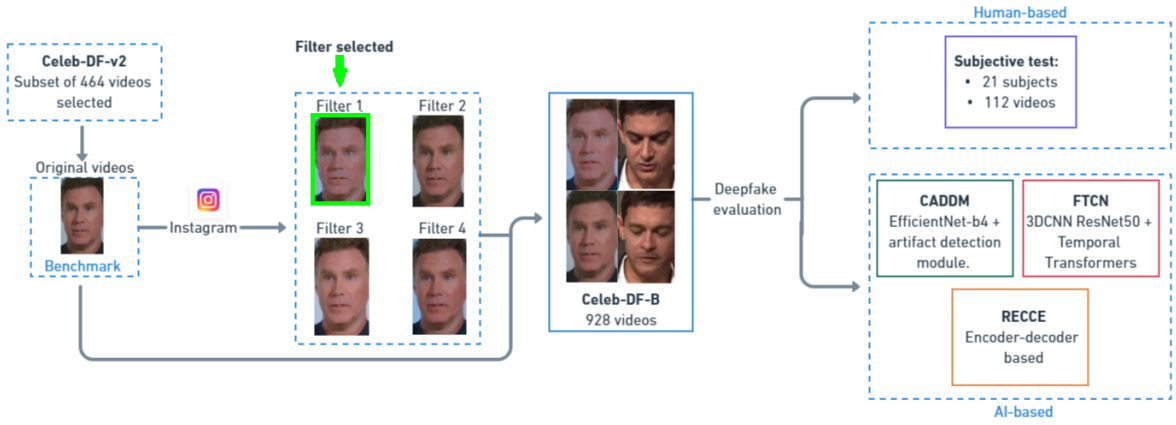

Fig. 1. Pipeline of the proposed study. A subset of 464 videos (50% Real and 50% Fake) are selected. Each video is uploaded to the social network Instagram, where one of the four different filters is randomly selected and applied to it. The four filters uniformly appear in the Celeb-DF-B database. The final database has a size of 928 videos and it is used to perform a human-based deepfake detection and to evaluate the robustness of three SotA AI-based detectors.

Beautified;

- We study the impact of those filters on three deepfake detectors finding a drop in performance for video-level AUC and revealing how social media beautification can be used to make fake videos look more authentic;
- Finally a subjective evaluation is conducted to investigate whether the utilization of beautification filters presents challenges for human observers when distinguishing between the authenticity of deepfake and real videos.

The paper is organized as follows. Section II detail related works and Section III the creation of the Celeb-DF-B dataset. In Sections IV and V, we present the performance of deepfake detectors and users on the Celeb-DF-B dataset. Finally, conclusions are summarized in Section VI

II. Related works

Assessment of deepfake detectors robustness: To ensure that models are suitable for detecting any types of deepfakes, deepfake databases have been created [6], [19], [24]. By having a great number of people with different facial attributes, expressions, and movements, deepfake detectors are more able to learn specific features related to identity manipulation. Despite a large number of different identities in the datasets, recent work highlighted how binary classifiers learn the identity representation of the people in the dataset, leading to a bad generalization of the classification performances against unseen datasets [7]. To assess the good generalization of the deepfake detectors, studies have been conducted in real-case scenarios, i.e. detecting deepfake uploaded online with video compression [16], [22], [29]. The higher the compression rate, the lower they can correctly classify. Indeed, the effect of compression can be seen in the classification AUC of the deepfake detectors with the low-quality videos of FaceForensics++ [3], [8], [28].

Subjective evaluation of deepfake: The subjective evaluation of deepfakes can be divided into two distinct categories. The first category involves assessing the quality of generated deepfake videos/images, considering criteria such as visual quality, realism, and coherence [14]. The second category, which is the focal point of our paper, focuses on assessing the capability of human evaluators to effectively differentiate between real and fake videos or images. In the study conducted by Bray et al. [2] 273 participants were involved, each assessing 20 images. After familiarizing themselves with 20 deepfake images and receiving descriptions of 10 common artifacts found in deepfakes, their task was to detect deepfakes while indicating their confidence in their decisions. The study revealed that the mean accuracy ranged from 60% to 64%, indicating a lack of human performance. In a similar approach, Korshunov et al. [17] conducted a subjective evaluation with 60 participants who viewed 120 (60 deepfakes and 60 real) videos. Participants were tasked with determining whether the video was fake or real. The study compared human evaluations with machine-based deepfake detectors, revealing limitations in both. Interestingly, algorithms struggled to identify deepfakes that were easily discernible by humans, underscoring the disparities in their detection capabilities. While the existing literature has contributed valuable insights into human-based deepfake detection, our study prominently highlights the specific influence of beautification filters on human accuracy in detecting deepfakes.

Impact of social media filters on AI-based facial processing tasks: Social media filters alter face images in different manners spanning from basic color transformations to the incorporation of virtual elements into a scene. *Beautification filters* subtly alter facial features, making it challenging to discern the changes without a reference image. In the literature, some researchers are dedicated to crafting facial filters, such as imperceptible skin smoothing techniques [27], that are not easy to detect to the naked eye. In opposition, some studies are examining the impact of filters on facial processing tasks. Inside those, an area of study is focused on evaluating the potential harm that the addition of artificial elements, such

TABLE I
CHARACTERISTICS OF THE SELECTED INSTAGRAM FILTERS. TRAITS MODIFICATIONS WERE ASSESSED BY VISUAL INSPECTION OF PIXEL DIFFERENCES BETWEEN ORIGINAL AND FILTERED IMAGES.

Filter	Color	Skin	Makeup	Eyes	Nose	Lips
BROWN	x	x	x		x	
California dreamin	x	x			x	x
Relax! You Pretty!	x	x			x	x
Hawaii Grain	x	x	x	x	x	x

as flower crowns or puppy ears, to face images via social media filters might pose to face recognition systems [1], [12]. More in line with our investigation, other works assessed the impact that *beautification filters* have on society [23] and in the estimation of different biometrics traits namely face, gender and weight [21].

III. DATASET

In this section, we introduce the protocol employed in the creation of the Celeb-DF-B database, its composition, and the specific social media filters chosen for face beautification.

The Celeb-DF [18] dataset consists of 590 real videos and 5639 DeepFake videos. The average duration of all videos is approximately 13 seconds, with a standard frame rate of 30 frames per second. The real videos are sourced from publicly accessible YouTube content corresponding to interviews featuring 59 celebrities. Among these, for the creation of the Celeb-DF-B database, we chose a subset consisting of 232 real and 232 fake videos. The selection of videos followed three criteria: 1) an equal sampling from each identity in the real videos; 2) pairing each real video with a fake counterpart created through FaceSwap; and 3) maintaining a balance between the source and driving identities of the selected fake videos.

Once the data was sampled, *beautification filters* were applied to the videos as depicted in Figure 1. Instagram was selected as the filter provider due to its large selection of available *beautification filters* which users regularly apply to enhance their multimedia content. In Table I we present the selected *beautification filters* along with the facial traits modified by them. Each of the 464 non-beautified videos is beautified with one of the four selected filters resulting in the creation of 928 videos that constitute the Celeb-DF-B dataset. Example frames of the videos belonging to the Celeb-DF-B database are displayed in Figure 2.

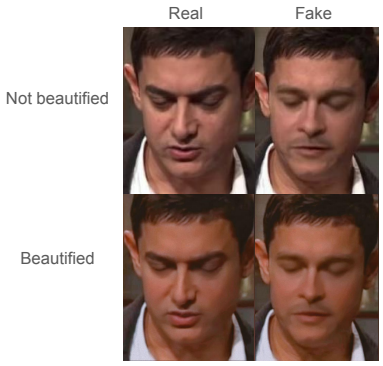

Fig. 2. Frames extracted from four distinct videos within the Celeb-DF-B database are depicted here.

IV. EXPERIMENTAL SETUP AND RESULTS

A. Deepfake Detectors

We selected 3 different passive deepfake detectors for this experiment: CADDM [7], RECCE [3], and FTCN [30].

CADDM detects traces of forgery on the frame level. First, the image is passed through an EfficientNet-b4 [26] backbone to extract useful features for the classification task. Then, it detects forgery locations on different scales through an artifact detection module trained with a custom Multi-scale Face Swap algorithm to generate forgery location ground truth. The average of the scores between the individual frames becomes the classification score of the video. With this architecture, the model focuses more on local forgeries instead of learning face distribution to perform better while detecting fakes of unseen faces.

RECCE is an encoder-decoder-based model. The encoder is based on Xception [5]. The reconstruction network has been trained in an unsupervised manner to learn the representation of real faces. The face frames are passed through the encoder-decoder architecture. Then, encoder and decoder features are agglomerated together with the residual images (i.e. the difference between the reconstructed and the original frame) to classify each frame as fake or genuine. The video's classification score is computed as the average score between each frame.

FTCN is a model trained to detect temporal inconsistencies in videos. Because deepfakes are generated frame by frame, they are likely to present temporal incoherences. FTCN network has a Resnet50 3DCNN [11] backbone to extract temporal features and a Temporal Transformers [9] as a classifier. Therefore, FTCN does not look for manipulations on each image independently but on a sequence of frames.

B. Experimental setup

Metrics: To evaluate the three selected deepfake detectors, we compute the video-level Area under the Curve AUC of the Receiver Operating Characteristic (ROC) curve and the False Negative Rate (FNR), i.e., the proportion of fake videos recognized as genuine, which, in a real-case scenario, is desirable to minimize. Additionally, we analyze the histogram of the classification scores before and after beautification to gain a better understanding of the behavior of deepfake detectors on beautified videos.

Evaluation protocol: We follow the evaluation process defined in [7]. We extract 32 frames at equal intervals to obtain 32 classification scores. Each evaluation score represents a real number between 0 and 1 for real and fake videos, respectively. The video score is then computed as the average of all the individual scores. FTCN, on the other hand, extracts a sequence of N consecutive frames from the video. To maintain consistency with the evaluation of CADDM and RECCE, we

TABLE II
TYPE OF DATA FROM FF++ SEEN BY EACH DEEPFAKE DETECTORS

Model	Compression	Seen Face Manipulation	Seen fake samples
CADDM [7]	Raw	DF, F2F, FSh, FS, NT	Yes
RECCE [4]	c23	DF, F2F, FSh, FS, NT	No
FTCN [30]	c23	DF, F2F, FS, NT	Yes

set $N = 32$. In our study, we define the positive class as 'fake videos' and the negative class as 'genuine videos'.

Implementation details: Our implementations of CADDM[1] RECCE[2] and FTCN[3] are based on publicly available GitHub projects. All three deepfake detector models are trained on FaceForensics++ [24] and use a backbone trained on ImageNet to extract features from images. FF++ contains respectively 5000 and 1000 fake and real videos divided into three subsets: train, val, and test. Five manipulation techniques were used to generate the fake videos. They are either face reenactment (Face2Face: F2F, NeuralTexture: NT) or FaceSwap (Deepfake: DF, FaceSwap: FS, FaceShifter: FSh) based methods. All the 6000 videos exist in 3 versions: *raw*, *High-Quality (c23)*, and *Low Quality (c40)*. Table II gives a summary of the specific data seen by each model during their training on FF++. For more information about the training of the three models, please refer to their corresponding publications.

C. Experimental results

In Table III and Figure 3, we present various results of our experiments. From Table III, we can observe that all detectors suffer a drop of approximately 15% in AUC when tested with beautification filters. In Figure 3 (a), we see the impact of the beautification process on the FNR. Specifically, beautified videos reduce the FNR for CADDM and FTCN. However, the False Negative Rate for RECCE is higher for beautified videos. This presents a significant issue, as fake videos may appear authentic due to the simple application of a beautification filter. In contrast to CADDM and FTCN, RECCE did not encounter any fake videos during its training as presented in Table II. Thus, RECCE did not learn any specific features associated with face manipulation. Even if beautification introduces minor artifacts, it removes some of the manipulation introduced by deepfakes. However, supervised trained models such as CADDM and FTCN can detect these minor artifacts.

To better understand the behavior of deepfake detectors on beautified videos, we analyzed the histogram of the classification scores before and after beautification. In Figure 4, we illustrate the difference in the distribution of the classification scores of the deepfake detectors on Celeb-DF-B for beautified and non-beautified videos. A score of 0 is the lowest probability that a video is fake according to a deepfake detector while a score of 1 represents the highest probability. For CADDM and FTCN, we can observe higher confidence scores for the beautified videos, indicating they are more likely to be

[1]https://github.com/megvii-research/CADDM
[2]https://github.com/VISION-SJTU/RECCE
[3]https://github.com/yinglinzheng/FTCN

TABLE III
AUC SCORE OF EACH DETECTOR W/O AND W/ BEAUTIFICATION ON CELEB-DF-B

Model	AUC (↑)	
	w/o beautification	w/ beautification
CADDM [7]	0.91	0.76 (↓ 0.15)
RECCE [4]	0.81	0.66 (↓ 0.15)
FTCN [30]	0.80	0.64 (↓ 0.16)

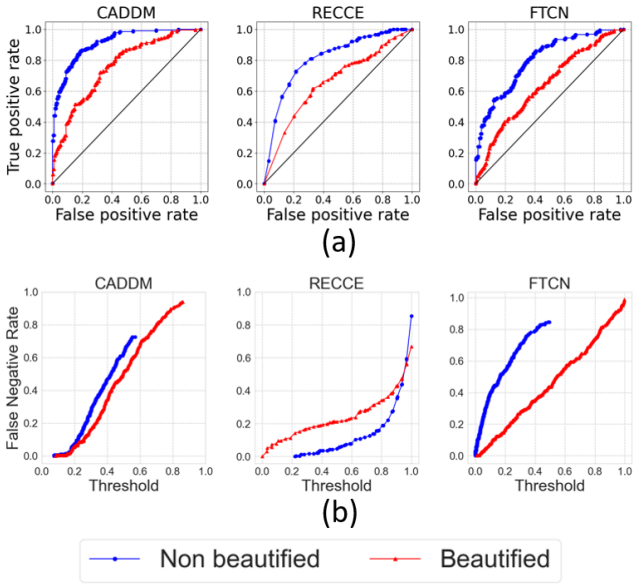

Fig. 3. Result of the evaluation on Celeb-DF-B with the the 3 detectors. a) The video-level AUC of the ROC curve and b) The False Negative Rate for different classification score thresholds

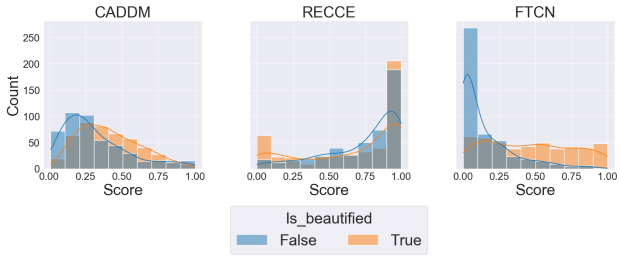

Fig. 4. The histogram of the classification score. CADDM and FTCN tend to see beautification as additional face manipulation whereas RECCE finds fake videos more realistic after the beautification process.

detected as fake. On average, all the confidence scores of the videos are shifted by +0.1 and +0.3, respectively, for CADDM and FTCN after beautification. This behavior was expected since beautification may present manipulation clues. However, the behavior is slightly different for RECCE. On average, beautified videos appear more authentic than the original ones, with an average score shift of -0.07. In summary, for RECCE, beautified videos tend to appear more real than their non-beautified counterparts.

V. SUBJECTIVE EVALUATION

In this section, we present the subjective evaluation conducted to assess the impact of applying a beautification filter

to both deepfake and real videos. We performed a subjective evaluation of deepfake videos, using a web-based framework for crowdsourcing experiments. The primary objective of this subjective test was to investigate whether the utilization of such filters presents challenges for human observers when distinguishing between the authenticity of deepfake videos and real videos. To achieve this goal, we selected a total of 112 videos (56 real and 56 deepfakes) from the Celeb-DF-B database. The selection process involved the following steps:

Fake Video Selection: We randomly picked 7 videos for each type of beautification filter from the fake video category, resulting in 28 videos. These same videos were included without the filter in the subjective test dataset.

Real Video Selection: Subsequently, we chose 7 videos for each filter type from the real video dataset, and once again, we incorporated these same videos without filters into the subjective test dataset.

Test protocol: To establish a consistent benchmark for comparison with typical deepfake detection algorithms, we presented human subjects with cropped face regions. Furthermore, we extended the boundary by an additional 100 pixels into the background to assess the algorithms' ability to handle background information.

Before the evaluation, participants received comprehensive explanations of the test procedures and completed practice tests to ensure their understanding. To optimize efficiency and prevent fatigue during the evaluation, we divided the test dataset randomly into three batches. This approach allowed participants to complete each test in separate sessions, with breaks in between. On average, each test batch lasted approximately 15 minutes, consistent with the standard recommendations [15]. The evaluation involved 21 participants with diverse backgrounds. Each video was shown to the participants three times consecutively. After viewing each video, following a procedure similar to that of Korshunov et al. [17], participants were asked, "Is the person's face in the video real or fake?". Then, they were then asked to identify the specific features or characteristics that influenced their judgment regarding the video's authenticity. The available feature options included: 1. Face contour, 2. Shadow inconsistency, 3. Inconsistency between eyes, 4. Eye blinking, 5. Mouth, 6. Teeth, 7. Lip motion, 8. Head motion, 9. Face/body mismatch, 10. Contextual mismatch, 11. Skin texture, and 12. Video quality.

A. Subjective evaluation results

Table IV displays the results of the subjective assessment outlined in Section V. The data within the table offers valuable insights into human performance in discerning deepfake videos from authentic ones, explaining the influence of beautification filters on human accuracy. The results suggest that human accuracy for non-beautified videos is higher (69%) than for beautified videos (66%), implying that human judgments are more effective at distinguishing between real and fake videos when no beautification is applied.

Furthermore, our analysis uncovers a significant contrast in recall rates between beautified (76%) and non-beautified (70%) videos. Recall, also known as sensitivity or true positive rate, measures the ability of a classifier to correctly identify positive instances among all actual positive instances. In the context of deepfake detection, a higher recall implies that the deepfake detection model or human evaluators are better at spotting deepfakes when the videos are beautified.

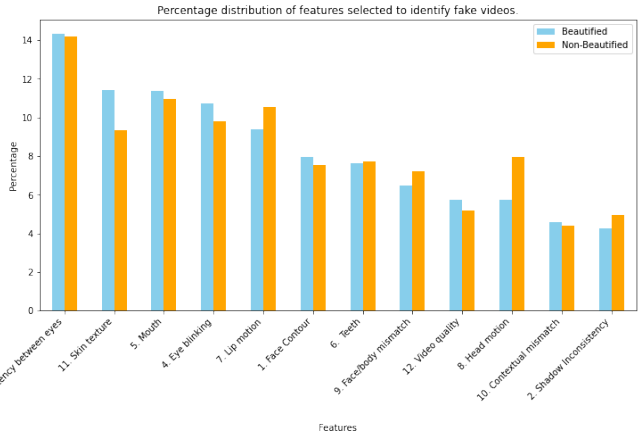

Fig. 5. Identifying prominent features influencing authenticity assessment in subjective evaluation.

The increased recall rate in our study implies that evaluators excel at identifying deepfakes when beautification filters are present. However, it's crucial to consider this alongside the accuracy findings. The observed accuracy rates suggest that while human subjects improve in detecting deepfakes with applied filters, they also tend to misclassify more genuine videos as fake in this scenario. This underscores the impact of beautification filters on human detection capabilities: they not only aid in recognizing deepfakes but may also lead to a higher rate of false positives, where non-deepfake videos are mistakenly identified as deepfakes.

In the subjective evaluation, participants were also tasked with identifying the specific features or characteristics that played a role in shaping their judgment regarding the video's authenticity. Among the provided feature options, the inconsistency between eyes stood out as the most frequently noted feature in both beautified and non-beautified videos. An interesting finding is that many participants highlighted alterations in skin texture as a factor influencing their categorization of videos as fake, with a notably higher percentage observed in beautified videos, as depicted in Figure 5.

TABLE IV
SUBJECTIVE EVALUATION RESULTS ON CELEB-DF-B DATASET FOR BEAUTIFIED AND NON-BEAUTIFIED VIDEOS

Metric	Non-beautified	Beautified
Accuracy	0.69	0.66
Recall	0.70	0.76

VI. CONCLUSION AND DISCUSSION

In this paper, we investigate the ongoing trend of digital face beautification through social media filters and its implications

for deepfake detection. The application of filters to facial multimedia is a user-friendly practice, as it does not demand any prior expertise, unlike other image editing techniques. This accessibility makes filters highly approachable for the average social media user. This study extends beyond AI-based detection, assessing three state-of-the-art deepfake detectors and the impact that the use of filters on deepfake videos has on human detection via a subjective evaluation. Experiments are conducted in our proposed Celeb-DF-B database showing how the application of filters significantly shifts the scores of the assessed deepfake detectors and changes the perceived information for human subjects.

Our findings reveal that, depending on the classifier used, even easy-to-use social media filters can significantly increase the likelihood of a deepfake video being wrongly classified as authentic. This not only challenges the robustness of current deepfake detection methods but also raises important questions about the reliability of these systems in real-world scenarios, where such filters are commonly used. We highlighted that deepfake detection is not just a matter of identifying sophisticated manipulations but also understanding how common alterations can impact these systems. Future challenges include mitigating the effects of beautification filters. In scenarios requiring access to secure locations or sensitive information, such as government facilities, financial institutions, or military installations, it becomes imperative to minimize the risk of an impersonation attack. Retraining deep learning-based detectors with beautified data might not guarantee a solution, as filters are being created daily, making generalization difficult. Given the substantial impact of beautification filters, the use of a dedicated filter detection method is strongly advisable.

ACKNOWLEDGMENTS

The contributions of Nelida Mirabet-Herranz, Sahar Husseini, and Alexandre Libourel are respectively supported by the following entities: the European CHIST-ERA program, facilitated by the French National Research Agency (ANR) under the XAIface project (grant agreement CHIST-ERA-19-XAI-011), Docaposte biometrics lab, the french ASTRID program, facilitated by the Defence Innovation Agency and the ANR under the DeTOX project. We express our sincere appreciation for the invaluable support extended by these organizations.

REFERENCES

[1] C. Botezatu, M. Ibsen, C. Rathgeb, and C. Busch, "Fun selfie filters in face recognition: Impact assessment and removal," *IEEE Transactions on Biometrics, Behavior, and Identity Science*, vol. 5, no. 1, 2022.

[2] S. D. Bray, S. D. Johnson, and B. Kleinberg, "Testing human ability to detect 'deepfake'images of human faces," *Journal of Cybersecurity*, vol. 9, no. 1, p. tyad011, 2023.

[3] J. Cao, C. Ma, T. Yao, S. Chen, S. Ding, and X. Yang, "End-to-end reconstruction-classification learning for face forgery detection," in *2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2022, pp. 4103–4112.

[4] ——, "End-to-end reconstruction-classification learning for face forgery detection," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2022, pp. 4113–4122.

[5] F. Chollet, "Xception: Deep learning with depthwise separable convolutions," 2017.

[6] B. Dolhansky, J. Bitton, B. Pflaum, J. Lu, R. Howes, M. Wang, and C. C. Ferrer, "The deepfake detection challenge (dfdc) dataset," 2020.

[7] S. Dong, J. Wang, R. Ji, J. Liang, H. Fan, and Z. Ge, "Implicit identity leakage: The stumbling block to improving deepfake detection generalization," 2023.

[8] S. Dong, J. Wang, J. Liang, H. Fan, and R. Ji, "Explaining deepfake detection by analysing image matching," in *European Conference on Computer Vision*. Springer, 2022, pp. 18–35.

[9] A. Dosovitskiy, L. Beyer, A. Kolesnikov, D. Weissenborn, X. Zhai, T. Unterthiner, M. Dehghani, M. Minderer, G. Heigold, S. Gelly, J. Uszkoreit, and N. Houlsby, "An image is worth 16x16 words: Transformers for image recognition at scale," 2021.

[10] I. J. Goodfellow, J. Pouget-Abadie, M. Mirza, B. Xu, D. Warde-Farley, S. Ozair, A. Courville, and Y. Bengio, "Generative adversarial networks," 2014.

[11] K. Hara, H. Kataoka, and Y. Satoh, "Learning spatio-temporal features with 3d residual networks for action recognition," 2017.

[12] P. Hedman, V. Skepetzis, K. Hernandez-Diaz, J. Bigun, and F. Alonso-Fernandez, "On the effect of selfie beautification filters on face detection and recognition," *Pattern Recognition Letters*, vol. 163, pp. 104–111, 2022.

[13] J. Ho, A. Jain, and P. Abbeel, "Denoising diffusion probabilistic models," 2020.

[14] S. Husseini and J.-L. Dugelay, "A comprehensive framework for evaluating deepfake generators: Dataset, metrics performance, and comparative analysis," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2023, pp. 372–381.

[15] C. Keimel, J. Habigt, C. Horch, and K. Diepold, "Qualitycrowd—a framework for crowd-based quality evaluation," in *2012 Picture coding symposium*. IEEE, 2012, pp. 245–248.

[16] M. Kombrink and Z. Geradts, "The influence of compression on the detection of deepfake videos," *Artificial Intelligence (AI) in Forensic Sciences*, p. 174, 2023.

[17] P. Korshunov and S. Marcel, "Subjective and objective evaluation of deepfake videos," in *ICASSP 2021-2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2021, pp. 2510–2514.

[18] Y. Li, X. Yang, P. Sun, H. Qi, and S. Lyu, "Celeb-df (v2): a new dataset for deepfake forensics [j]," *arXiv preprint arXiv*, 2019.

[19] ——, "Celeb-df: A large-scale challenging dataset for deepfake forensics," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2020, pp. 3207–3216.

[20] Y. Lu and T. Ebrahimi, "Impact of video processing operations in deepfake detection," *arXiv preprint arXiv:2303.17247*, 2023.

[21] N. Mirabet-Herranz, C. Galdi, and J.-L. Dugelay, "Impact of digital face beautification in biometrics," in *2022 10th European Workshop on Visual Information Processing (EUVIP)*. IEEE, 2022, pp. 1–6.

[22] A. Mitra, S. P. Mohanty, P. Corcoran, and E. Kougianos, "A machine learning based approach for deepfake detection in social media through key video frame extraction," *SN Computer Science*, vol. 2, pp. 1–18, 2021.

[23] P. Riccio, B. Psomas, F. Galati, F. Escolano, T. Hofmann, and N. M. Oliver, "Openfilter: A framework to democratize research access to social media ar filters," in *Thirty-sixth Conference on Neural Information Processing Systems Datasets and Benchmarks Track*, 2022.

[24] A. Rössler, D. Cozzolino, L. Verdoliva, C. Riess, J. Thies, and M. Nießner, "FaceForensics++: Learning to detect manipulated facial images," in *International Conference on Computer Vision (ICCV)*, 2019.

[25] N. A. Shelke and S. S. Kasana, "A comprehensive survey on passive techniques for digital video forgery detection," *Multimedia Tools and Applications*, vol. 80, pp. 6247–6310, 2021.

[26] M. Tan and Q. V. Le, "Efficientnet: Rethinking model scaling for convolutional neural networks," 2020.

[27] S. Velusamy, R. Parihar, R. Kini, and A. Rege, "Fabsoften: Face beautification via dynamic skin smoothing, guided feathering, and texture restoration," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops*, 2020, pp. 530–531.

[28] Z. Yan, P. Sun, Y. Lang, S. Du, S. Zhang, W. Wang, and L. Liu, "Multimodal graph learning for deepfake detection," 2023.

[29] P. Yu, Z. Xia, J. Fei, and Y. Lu, "A survey on deepfake video detection," *Iet Biometrics*, vol. 10, no. 6, pp. 607–624, 2021.

[30] Y. Zheng, J. Bao, D. Chen, M. Zeng, and F. Wen, "Exploring temporal coherence for more general video face forgery detection," 2021.