A Novel Assessment Framework for Learning-based Deepfake Detectors in Realistic Conditions

Yuhang Lu and Touradj Ebrahimi

Multimedia Signal Processing Group (MMSPG), École Polytechnique Fédérale de Lausanne (EPFL), CH-1015 Lausanne, Switzerland

ABSTRACT

Detecting manipulations in facial images and video has become an increasingly popular topic in media forensics community. At the same time, deep convolutional neural networks have achieved exceptional results on deepfake detection tasks. Despite the remarkable progress, the performance of such detectors is often evaluated on benchmarks under constrained and non-realistic situations. In fact, current assessment and ranking approaches employed in related benchmarks or competitions are unreliable. The impact of conventional distortions and processing operations found in image and video processing workflows, such as compression, noise, and enhancement, is not sufficiently evaluated. This paper proposes a more rigorous framework to assess the performance of learning-based deepfake detectors in more realistic situations. This framework can serve as a broad benchmarking approach for both general model performance assessment and the ranking of proponents in a competition. In addition, a stochastic degradation-based data augmentation strategy driven by realistic processing operations is designed, which significantly improves the generalization ability of two deepfake detectors.

Keywords: Assessment Framework, Deepfake Detection, Data Augmentation

1. INTRODUCTION

In recent years, the development of deep convolutional neural networks (DCNNs) and free access to large-scale datasets have led to significant progress over the generation of realistic forgery content. Deepfakes refer to manipulated face contents by deep learning tools. The recent advancement of such techniques and wide availability of open-source software has simplified the creation of such face manipulations, posing serious public concerns. To counteract the misuse of these deepfake techniques and malicious attacks, detecting manipulations in facial images and video has become a popular topic in the media forensics community and receives increasing attention from both academia and businesses.

Nowadays, multiple grand challenges, competitions, and public benchmarks[1–3] are organized to assist the progress of deepfake detection. At the same time, with the advanced deep learning techniques and large-scale datasets, numerous detection methods[4–10] have been published and have reported promising results on different benchmarks. However, most of the recent detection methods are developed under constrained and less realistic conditions. Similarly, the conventional assessment approach, which exists in different benchmarks, often samples test data from the same distribution as training data and cannot reflect model performance in more complex situations.

In fact, it has long been shown that DCNN-based methods are vulnerable to real-world perturbations and processing operations.[11–13] In more realistic conditions, images can face unpredictable distortions from the extrinsic environment, such as noise and poor illumination conditions, or constantly undergo various processing operations to ease their distribution. In the context of this paper, a deployed deepfake detector could mistakenly block a pristine yet heavily compressed image. On the other hand, a malicious agent could also fool the detector by simply adding imperceptible noise to fake media contents. Moreover, current learning-based deepfake detectors often suffer from poor generalization ability facing new manipulation techniques or unseen human faces. Therefore, a more reliable and systematic approach is desired firsthand in order to assess the performance of a

Further author information: (Send correspondence to the authors)
E-mail: yuhang.lu@epfl.ch, touradj.ebrahimi@epfl.ch

Table 1: Deepfake Detection Challenge (DFDC)[1] top-5 prize winners and their corresponding results.

Team name	Overall log loss
Selim Seferbekov[21]	0.4279
WM[22]	0.4284
NTechLab[23]	0.4345
Eighteen Years Old[24]	0.4347
The Medics[25]	0.4371

deepfake detector in more realistic scenarios. At the same time, a generic approach to improve the robustness of the detectors is also desired.

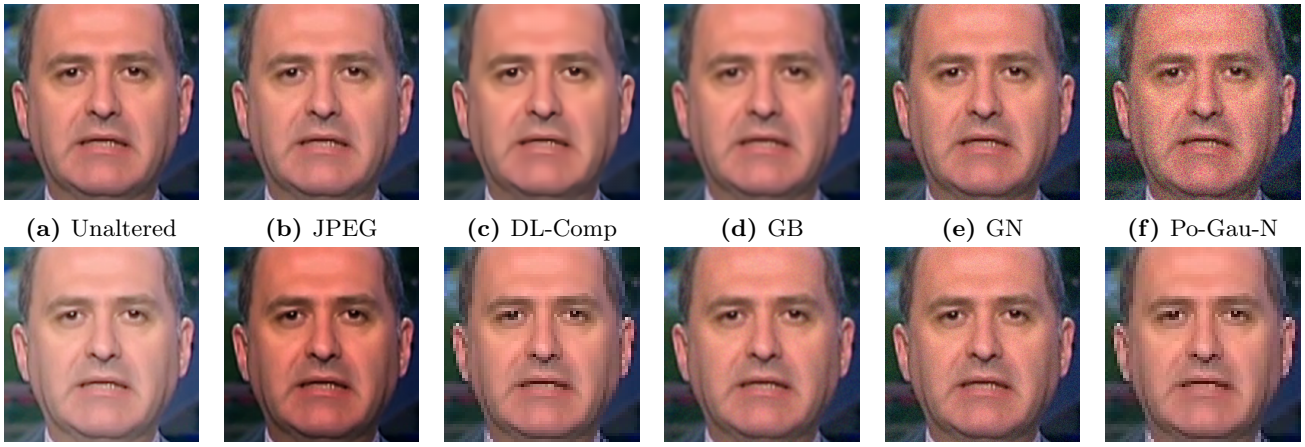
In this work, the following contributions have been made.

- A realistic assessment framework is proposed to evaluate and benchmark the performance of learning-based deepfake detection systems. To the best of our knowledge, this is the first framework that rigorously evaluates deepfake detection in realistic situations.
- Inspired by real-world data degradation process, a stochastic degradation-based augmentation (SDAug) method driven by typical image and video processing operations is designed for deepfake detection tasks. It brings remarkable improvement in the robustness of general detectors.
- A flexible Python toolbox is developed and the source code of the proposed assessment framework is released to facilitate relevant research activities.

2. RELATED WORK

Face manipulation detection. Deepfake detection is often treated as a binary classification problem in computer vision. Early on, solutions based on facial expressions,[14] head movements[15] and eye blinking[16] were proposed. Current studies leverage deep learning techniques to address such detection problems. Zhou et al.[17] proposed to detect the deepfakes with a two-stream neural network. Rössler et al.[4] retrained an XceptionNet[18] with manipulated face dataset which outperforms their proposed benchmark. Nguyen et al.[5] combined traditional CNN and Capsule networks,[19] which require fewer parameters. Attention mechanisms have also been applied to further improve the training process of the detection system. Dang et.al[20] proposed a detection system based on attention mechanism. Zhao et al.[6] designed multi-attention head to predict multiple spatial attention maps. Their proposed attention map can be easily implemented and inserted into existing backbone networks. Besides focusing on the spatial domain, recent works[7–10] attempt to resolve the problem in the frequency domain. These methods transform the image to the frequency domain via DCT transformation and separate information according to frequency band, which leads to better performance. In this paper, two widely used deepfake detectors[4,5] are adopted for experiments.

Deepfake detection competitions review. To assist in a faster progress and better advancement of deepfake detection tasks, numerous large-scale benchmarks, competitions, and challenges[1–4] have been organized, the results of which have been made publicly available. Meta partnered with some academic experts and industry leaders and created the Deepfake Detection Challenge (DFDC)[1] in 2019. The competition provided a large incentive, i.e. 1 million USD, for experts in computer vision and deepfake detection to dedicate time and computational resources to train models for benchmarking. More recently, the Trusted Media Challenge (TMC)[3] was organized by AI Singapore with a total prize pool of up to 500k USD in order to explore how artificial intelligence technologies could be leveraged to combat fake media. Nevertheless, after a thorough investigation of the benchmarking results, a new question emerges: *Can the assessment approach adopted by the competitions reflect their performance in realistic scenarios?* Although both challenges tried to simulate real-world conditions by preprocessing part of the testing data with some common video processing techniques, they do not really differentiate the detectors. As shown in Table 1, the final results of the top-5 prize winners from DFDC[1] are extremely close and the ranking seems to be easily affected by some random noise, for example simply taking out a few fake samples or adding slightly more severe blurry effect. The current ranking approach in these



(a) Unaltered **(b)** JPEG **(c)** DL-Comp **(d)** GB **(e)** GN **(f)** Po-Gau-N

(g) GammaCorr **(h)** GammaCorr **(i)** Low Res **(j)** GN+GB **(k)** JPEG+GN **(l)** JPEG+LR

Figure 1: Example of a typical frame in the test dataset after applying various operations. Some notations are explained as following. DL-Comp: learning-based compression. GB: Gaussian blur. GN: Gaussian noise. Po-Gau-N: Poissonian Gaussian noise. GammaCorr: Gamma correction. +: mixture.

competitions is not reliable. A more rigorous framework is introduced in this work, which is able to differentiate the detectors in multiple dimensions, i.e. general performance, general robustness in realistic conditions, and robustness to specific impacting factors.

Robustness benchmark. In recent years, research has been conducted to explore the robustness of CNN-based methods toward real-world image corruption. Dodge and Karam[11] measured the performance of image classification models with data disturbed by noise, blurring, and contrast changes. In,[26] Hendrycks et al. presented a corrupted version of ImageNet[27] to benchmark the robustness of image recognition models against common image manipulations.[28–30] focused on a safety-critical task, autonomous driving, and provided robustness benchmark for various relevant vision tasks, such as object detection and semantic segmentation. Similar work has been done for face recognition tasks,[12,13,31] analyzed the robustness of CNN-based face recognition models towards face variations caused by illumination change, occlusion, and standard image processing operations. In media forensics community, StirMark[32] tested the robustness of image watermarking algorithms. The ALASKA#2 dataset[33] was created following a careful evaluation of ISO parameters, JPEG compression and noise level on FlickR images, etc., in order to help researchers in designing way more general and robust steganographic and steganalysis methods. Similarly, two popular deepfake detection benchmarks, DFDC[1] and Deeperforensics-1.0[2] also adopted standard processing operations to part of the testing data. They randomly applied distortions to a small portion of test data and considered only one severity level for each processing operation. However, the way they evaluate a detector's robustness is not systematic enough. The assessment results cannot rigorously show to which extent the detector is affected by the distorted data, nor help identify which factors show more significant influence on the detector's performance. There is a lack of a fair and flexible methodology that systematically compares the performance of deepfake detectors in realistic situations. In this work, a new assessment framework is introduced to solve this problem.

3. PROPOSED ASSESSMENT FRAMEWORK

In this section, the common realistic influencing factors that affect the performance of deepfake detectors are first introduced. Then, the proposed assessment framework is described in order to provide a fair comparison for deepfake detectors under more realistic situations.

3.1 Realistic Influencing Factors

In a real-world scenario, the media is often processed by various digital image processing operations. In more adverse cases, malicious deepfakes can be slightly corrupted to fool the detector while maintaining good perceptual quality. In general, our framework contains six categories of processing operations or corruptions with more

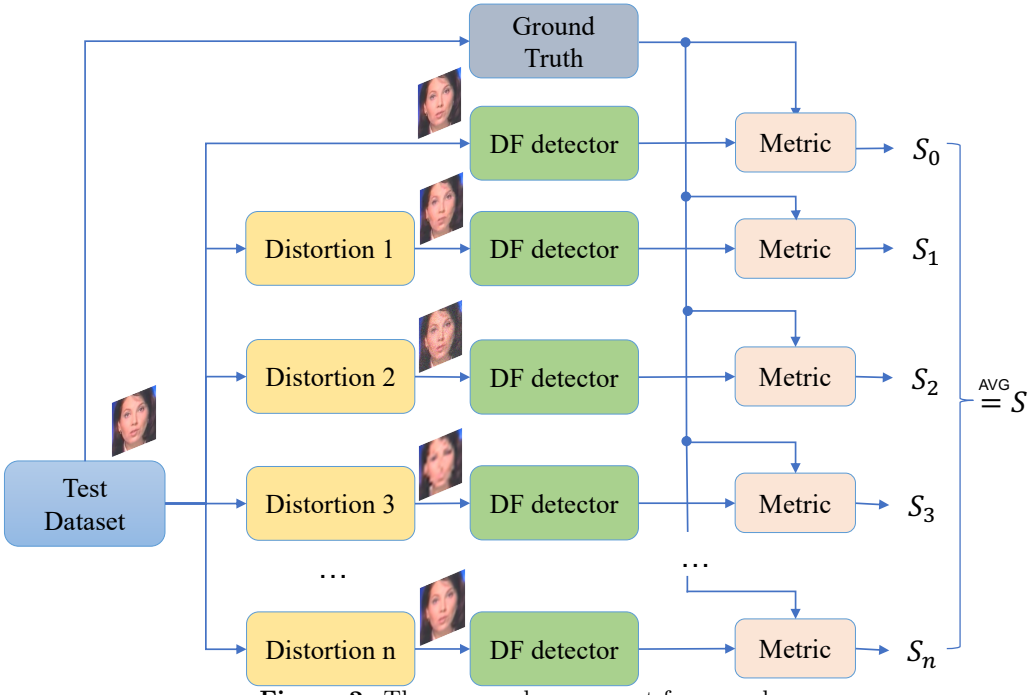


Figure 2: The proposed assessment framework.

than ten minor types. Each type consists of over five different severity levels. The details of all operations used in evaluations are described below with the illustration of a typical example in Fig. 1. Specifically, the following factors are considered in our assessment framework.

Noise: Noise is a typical distortion especially when images are captured in a low illumination condition. To simulate the noise, an Additive White Gaussian Noise (AWGN) is applied to the data and the pixel values are clipped to [0, 255]. In this paper, the variance value σ is selected in a range from 5 to 50. In addition, Poissonian-Gaussian noise[34] is also included to better reflect the realistic noise levels, whose parameters are learned from a group of real noisy pictures.

Resizing: Resizing is one of the most commonly used image and video processing operations. It refers to changing the dimensions of the media content to fit the display or other purposes. On the other hand, low-resolution data can significantly reduce the performance of modern deep learning-based detectors[35,36] due to a lack of discriminative information. This is often the case for those earlier image or video contents that are of poor quality. In this framework, the impact of resizing operation on low-resolution image is simulated by first downscaling the images and then upscaling back using bicubic interpolation.

Compression: Lossy compression refers to the class of data encoding methods that remove unnecessary or less important information and only use partial data to represent the content. These techniques are used to reduce data size for efficient storage and transmission content and are widely applied to image and video processing. In this framework, the JPEG compression artifacts are applied and the impact of different quality factors, i.e. from 10 to 95, to deepfake detection system is evaluated. As deep learning-based compression techniques are becoming increasingly popular in this community, an AI-based image compression technique[37] is also considered in this framework with 3 compression qualities to choose from.

Denoising: A typical way to reduce noise is by smoothing, which is a low-pass filtering applied to the image. The denoising operation is often applied to image and video contents after being acquired by the camera but at the same time it tends to blur the media content and results in a reduction of details, which is harmful to the detection system. To measure the impact of denoising operation, the blurry effect is simulated in our framework by applying Gaussian filters with kernel size σ ranging from 3 to 11. Meanwhile, learning-based

denoising techniques are gradually deployed in practice. They recover a noisy image with higher quality but often bring unpredictable artifacts. The impact of applying DnCNN technique[38] is assessed in our framework.

Enhancement: In realistic conditions, the image data captured in the wild can suffer from poor illumination. Image enhancement is frequently used to adjust the media content for better display. In this assessment framework, the contrast and brightness of the test data is modified by both linear and nonlinear adjustments. The former simply adds or reduces a constant pixel value while the latter applies gamma correction.

Combinations: It is even more common that the media content suffers from multiple types of distortions and processing operations. Therefore, the mixture of two or three operations above is also considered, such as combining JPEG compression and Gaussian noise, making the test data better reflect more complex real-world scenarios.

3.2 Assessment Methodology

Current deep learning-based deepfake detectors are data-driven and rely heavily on the distribution of training set. Traditionally, the performance of a deepfake detector is simply evaluated with test dataset, which is in the same distribution of training set. Some benchmarks, such as,[1,2] randomly add perturbation to partial test data and mix up with the others. But there is not a standard for the proportion and strength of the manually applied perturbations, making the benchmarking results less reliable and insightful. The proposed assessment methodology thoroughly measures the impact of multiple influencing factors with different severity levels on the performance of deepfake detectors.

In this section, the principle and usage of our assessment framework is introduced in detail. First, the deepfake detector should be trained on its original target datasets, such as FaceForensics++.[4] The processing operations and corruptions in the framework are not applied on training data. Then, as illustrated in Figure. 2, multiple copies of the test set are created and each type of distortion with one severity level is applied to the copies independently. The standard test data together with different distorted data are fed to the to-be-evaluated deepfake detector respectively. Finally, the detector generates real or fake predictions and calculates performance metrics for each processed dataset. An overall evaluation score is obtained by averaging the scores from each distortion style and strength level.

In addition, in order to relieve the burden on storage caused by the multiple copies of test set, a Python toolbox is developed to address this problem in an online manner, which hard-codes the digital processing operations and makes the strength level a parameter. It operates in the same format as the famous Transforms module in TorchVison toolbox, and can be easily integrated into the evaluation process.

4. STOCHASTIC DEGRADATION-BASED AUGMENTATION

To reduce the negative impact of realistic distortions and post-processing operations on detection performance, an effective data augmentation approach is proposed which leads to a robustness improvement. Standard data augmentation methods enrich training data by introducing different transformations, such as translation and rotation. Although it has been shown to increase model generalization ability in many tasks, it brings limited performance improvement to detectors under realistic conditions. The proposed stochastic degradation-based augmentation (SDAug) method is motivated by a typical data acquisition and transmission pipeline in real world. The main novelty of the proposed augmentation technique resides in the fact that it is driven by the typical operations that images and video are subject to in realistic conditions. Based on the observation of data degradation process, a carefully designed augmentation chain is conceived, which produces augmented training data that are much closer to real-world conditions.

In general, the brightness and contrast of input image x are first modified by image enhancement operator $\boldsymbol{enh}$. Afterward, the image is convoluted with an image blurring kernel $\boldsymbol{f}$, followed by additive Gaussian noise $\boldsymbol{n}$. At the end, $\boldsymbol{JPEG}$ compression is applied to obtain the augmented training data x_{aug}. The augmentation chain is described by the following formula.

$$x_{\text{aug}} = \boldsymbol{JPEG}[(\boldsymbol{enh}(x) \circledast \boldsymbol{f}) + \boldsymbol{n}] \tag{1}$$

In addition, unlike common data augmentation techniques, the SDAug method is implemented in a stochastic manner. The term 'stochastic' can be interpreted in the following two aspects. Firstly, each aforementioned augmentation operation will occur with a certain probability in the augmentation chain. Secondly, each operation will use random severity level for every image. The realistic scenario is rather complex and not necessarily consists of multiple types of distortions and processing operations. A random mixture of several distortions and severity levels can create more diversity in the augmented training data. Moreover, the stochastic augmentation helps preserve more information from the original training data and therefore prevents from accuracy loss on the high-quality data. In detail, the augmentation operations are explained in sequence as follows.

Enhancement: The augmentation chain begins with an image enhancement operation. A probability of 50% is adopted to apply either a brightness or a contrast operation on the training data which will be then non-linearly modified by a factor randomly selected from $[0.5, 1.5]$.

Smoothing: Image blurring operation is then applied with a selected probability of 50%. Either Gaussian blur or Average blur filter is used with a kernel size varying in the range $[3, 15]$.

Additive Gaussian Noise: For each batch of training data, a probability of 30% is adopted to add a Gaussian noise. The standard deviation of the Gaussian noise varies randomly in the interval $[0, 50]$.

JPEG Compression : Finally, JPEG compression is applied with a selected probability of 70%. The quality factor corresponding to the compression is randomly chosen in the range $[10, 95]$.

5. EXPERIMENTS AND RESULTS

5.1 Implementation Details

5.1.1 Datasets

Two widely used face manipulation datasets are selected for extensive experimentation to demonstrate the effectiveness of the proposed augmentation technique.

FaceForensics++,[4] denoted by FFpp, contains 1000 pristine and 4000 manipulated video generated by four different deepfake creation algorithms. Additionally, raw video contents are compressed with two quality parameters using the AVC/H.264 codec, denoted as C23 and C40. In the experiments, the training set is denoted as *FFpp-Raw* or *FFpp-C23* when the model is trained on single-quality-level data, while it is denoted as *FFpp-Full* when data of all three quality levels are involved for training. On the contrary, only uncompressed data are used for the final assessment.

Celeb-DFv2[39] is another high-quality dataset, with 590 pristine celebrity video and 5639 fake video. The test data is selected as recommended by[39] while the training and validation set was split in 80% and 20% accordingly.

For both datasets, 100 frames are randomly sampled from each video for training purposes and 32 frames are extracted for validation and testing. Extracted frames were pre-processed and cropped around the face regions using the dlib toolbox.[40] The face regions are finally resized into 300x300 pixels before feeding to the network.

5.1.2 Detection Methods

Experiments have been conducted with two learning-based deepfake detectors, both of which have reported excellent performance on popular benchmarks.

Capsule-Forensics[5] achieves high detection accuracy and meanwhile maintains a rather small amount of parameters by combining conventional CNN and Capsule network.[19]

XceptionNet[18] is a popular CNN architecture in many computer vision tasks. It achieved excellent performance in the FaceForensics++ benchmark on both compressed and uncompressed contents.

5.1.3 Training Details

Both detectors were trained with Adam optimizer with $\beta_1 = 0.9$, $\beta_2 = 0.999$. The Capsule-Forensics model is trained from scratch for 25 epochs with a learning rate of 5×10^{-4}, and the XceptionNet model is trained for 10 epochs with a learning rate of 1×10^{-3}.

Table 2: AUC (%) scores of two detectors tested on unaltered and distorted variants of FFpp and Celeb-DF test set respectively. Capsule-Forensics detector is shortened as *Capsule*. *Raw*, *C23* and *Full* refer to different quality settings of FFpp. The suffix *+SDAug* denotes the proposed stochastic degradation-based augmentation technique.

Methods	TrainSet	Unaltered	JPEG		DL-Comp			Gau Noise			Pois-Gau Noise	Gau Blur			Gamma Corr			Resize		
			95	60	30	High	Med	Low	5	30	50		3	7	11	0.1	0.75	1.3	2.5	x4	x8	x16
Capsule	FFpp-Raw	99.20	97.91	76.48	59.60	55.24	54.50	50.92	61.80	51.26	50.84	55.63	67.19	58.22	52.26	50.50	98.86	99.17	96.12	55.42	52.18	53.10
	FFpp-C23	96.32	95.09	95.76	74.91	56.96	57.42	81.57	84.51	58.63	50.51	70.59	85.21	53.94	52.04	52.08	95.06	96.72	92.91	79.33	64.62	50.33
	FFpp-Full	94.52	94.95	93.97	84.50	99.01	96.77	88.95	89.03	57.95	51.11	64.87	85.72	58.83	56.05	56.02	93.86	93.87	85.44	87.05	69.93	54.15
	Celeb-DF	99.76	99.80	99.33	96.51	99.01	96.77	88.95	97.35	63.30	55.32	-	99.15	96.54	90.58	48.33	99.71	99.71	93.44	95.67	75.16	68.35
XceptionNet	FFpp-Raw	99.56	76.77	56.00	54.20	50.16	50.37	50.10	50.12	50.36	50.70	51.02	68.76	55.61	50.70	54.66	98.66	99.57	70.45	68.60	55.80	50.45
	Celeb-DF	98.06	98.20	97.63	94.98	96.23	90.23	75.46	95.92	63.19	55.93	-	97.32	87.22	78.05	53.25	97.63	98.34	89.02	85.47	59.40	49.21
Capsule	FFpp-Raw+SDAug	98.16	97.97	96.36	94.08	93.81	71.41	59.74	97.05	83.51	75.09	90.04	96.86	90.32	80.31	60.17	97.68	98.18	96.91	93.54	79.22	58.05
XceptionNet	FFpp-Raw+SDAug	98.44	98.25	97.36	96.12	98.03	87.76	82.74	97.37	91.71	88.70	94.57	98.31	97.35	94.51	80.48	98.25	98.44	97.75	97.30	86.26	67.14

5.1.4 Performance Metrics

During the evaluation, the Accuracy (ACC), the Area Under Receiver Operating Characteristic Curve (AUC) were used as metrics in all experiments.

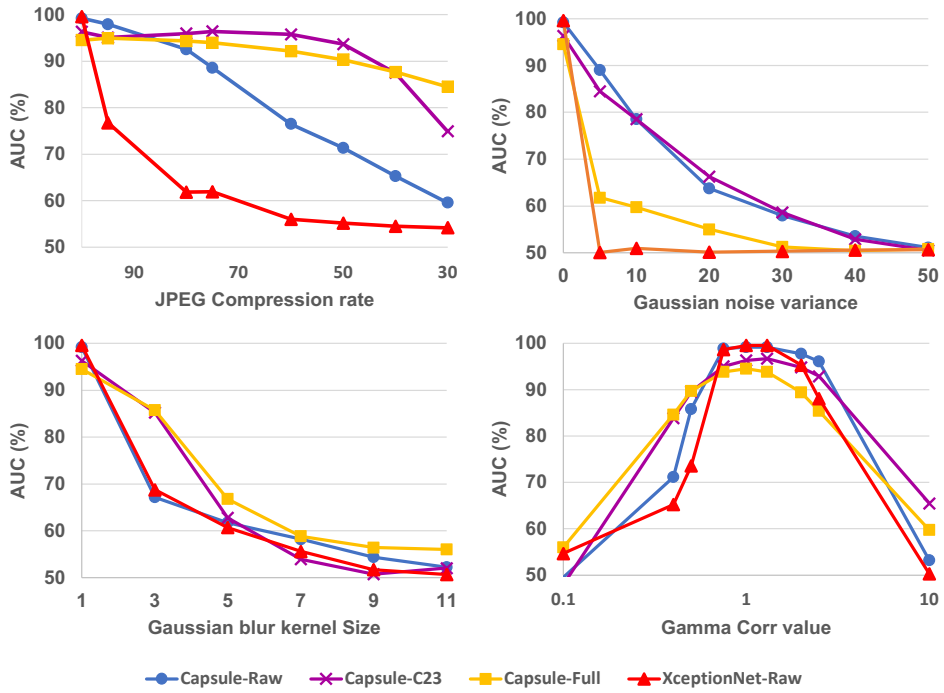


Figure 3: Assessment results of two models trained on FFpp dataset. The suffixes of legends refer to the qualities of the training data. *Full* means using all available quality data for training.

5.2 Results of Assessment Framework

The two deepfake detectors were trained on the original unaltered training sets of both FFpp and Celeb-DF. Table 2 shows the evaluation results using our assessment framework. Due to the page limit, only AUC scores and a subset of operations and severity levels are presented in this section.

In general, our findings draw the following conclusions. First, even mild real-world processing operations can have an obvious negative impact on detection accuracy. The two detectors present exceptional performance on unaltered FFpp testing data as expected, but then show severe performance deterioration on all kinds of modified data from the assessment framework, which indicates a lack of robustness.

Secondly, the two detectors are prone to be affected by different types of perturbation. When trained on the same dataset, CapsuleNet is generally more robust towards JPEG compression and synthetic noise, while XceptionNet at times presents slightly better results that could be of statistical nature. The results from our

assessment framework provide valuable guidance towards improving a specific deepfake detector. Moreover, among the considered influencing factors, noise and blurry effects are the most prominent for deepfake detectors. The performance of both detectors deteriorate rapidly after increasing the severity levels of the two distortions.

Finally, the impact of quality variants of training data on learning-based detectors has been analyzed. The Capsule-Forensics model trained only with very high-quality data (*FFpp-Raw*) will be extremely sensitive to nearly all kinds of realistic processing operations. On the contrary, mixing relatively low-quality data during the training process (*FFpp-Full*) slightly improves the robustness towards low-intensity processing and distortions, particularly for JPEG compression.

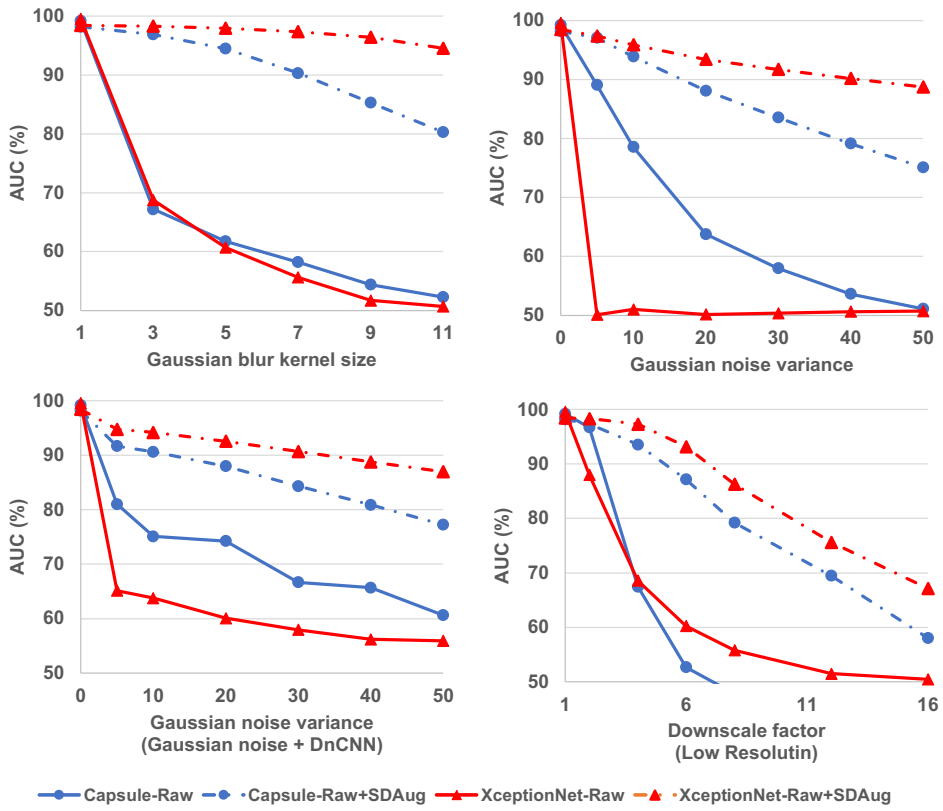


Figure 4: Performance comparison between model trained on FFpp-Raw only and trained with the proposed augmentation scheme.

5.3 Results with Augmentation

The last two rows of the Table 2 show the evaluation results of the two detectors trained on FFpp-Raw dataset together with the proposed augmentation strategy. The information regarding the models trained with the proposed stochastic degradation augmentation methods is denoted as **+SDAug**.

In comparison, it is evident that training with the stochastic degradation-based augmentation technique on the same dataset remarkably improves the performance on nearly all kinds of processed data even with intense severity. Previous experiments show that the detectors are more vulnerable to synthetic noises and blurry effects. The first two sub-figures in Fig. 4 further illustrate the impact of increasing the severity of the two distortions. The data augmentation scheme significantly improves the robustness and meanwhile still maintains high performance on original unaltered data.

It is worth noting that the performance improves not only on the four types of processing operations that appear during data augmentation but also on other different kinds of distortions. As shown in the Table 2 and

Table 3: Cross-dataset evaluation on Celeb-DFv2 (AUC(%)) after training on FFpp dataset. The suffixes *+DAug* denotes that the models is trained with the our proposed augmentation chain but without the stochastic manner. The suffixes *+SDAug* denotes that the model is trained with the stochastic degradation-based augmentation technique.

Deepfake Detector	Augmentation Method	FFpp	Celeb-DFv2
Capsule	No Aug	99.20	54.39
	DAug	93.51	68.39
	SDAug	97.82	**71.86**
XceptionNet	No Aug	99.56	50.00
	DAug	78.64	62.81
	SDAug	98.44	**73.88**

the last two sub-figures in Fig. 4, both detectors are much more robust towards learning-based compression, low-resolution effects, and other mixed distortions.

Finally, a cross-dataset assessment has been conducted to evaluate the generalization ability of the models trained with the proposed augmentation scheme on unseen datasets. The results are shown in Table 3. The selected detectors are trained on FFpp dataset but tested on Celeb-DFv2 test set for frame-level AUC scores. The two methods both obtain very low scores on the new dataset. On the contrary, the proposed augmentation scheme brings a significant performance improvement for both detectors on Celeb-DFv2, showing its capability to improve the generalization ability of deepfake detectors on unseen forensic face contents. Moreover, the results in Table 3 demonstrate the effectiveness of the stochastic mechanism. Although the model trained with degradation-based augmentation (DAug) improves the performance on Celeb-DFv2 dataset, the AUC scores on original FFpp test set degrades heavily. The SDAug shows the most significant improvement on generalization ability and meanwhile maintains high performance on original high-quality data.

6. CONCLUSION

Many detectors are designed to be as high performing as possible on specific benchmarks. But this often results in sacrificing generalization ability to more realistic situations. The proposed assessment framework is capable of assessing detectors in more realistic conditions and provides valuable insights on designing more robust techniques. A carefully conceived augmentation chain based on a natural data degradation process is proposed and significantly improves the model's robustness against various distortions. In the future, a statistical hypothesis test will be conducted to better validate the effectiveness of the proposed augmentation technique.

ACKNOWLEDGMENTS

The authors acknowledge support from CHIST-ERA project XAIface (CHIST-ERA-19-XAI-011) with funding from the Swiss National Science Foundation (SNSF) under grant number 20CH21 195532.

REFERENCES

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

[2] Jiang, L., Li, R., Wu, W., Qian, C., and Loy, C. C., "Deeperforensics-1.0: A large-scale dataset for real-world face forgery detection," (2020).

[3] Chen, W., Chua, B., and Winkler, S., "Ai singapore trusted media challenge dataset," *arXiv preprint arXiv:2201.04788* (2022).

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

[5] Nguyen, H. H., Yamagishi, J., and Echizen, I., "Use of a capsule network to detect fake images and videos," *ArXiv* **abs/1910.12467** (2019).

[6] Zhao, H., Zhou, W., Chen, D., Wei, T., Zhang, W., and Yu, N., "Multi-attentional deepfake detection," *2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)* , 2185–2194 (2021).

[7] Liu, H., Li, X., Zhou, W., Chen, Y., He, Y., Xue, H., Zhang, W., and Yu, N., "Spatial-phase shallow learning: rethinking face forgery detection in frequency domain," in [*Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*], 772–781 (2021).

[8] Qian, Y., Yin, G., Sheng, L., Chen, Z., and Shao, J., "Thinking in frequency: Face forgery detection by mining frequency-aware clues," in [*European conference on computer vision*], 86–103, Springer (2020).

[9] Li, J., Xie, H., Li, J., Wang, Z., and Zhang, Y., "Frequency-aware discriminative feature learning supervised by single-center loss for face forgery detection," in [*Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*], 6458–6467 (2021).

[10] Luo, Y., Zhang, Y., Yan, J., and Liu, W., "Generalizing face forgery detection with high-frequency features," in [*Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*], 16317–16326 (2021).

[11] Dodge, S. F. and Karam, L., "Understanding how image quality affects deep neural networks," *2016 Eighth International Conference on Quality of Multimedia Experience (QoMEX)* , 1–6 (2016).

[12] Mehdipour Ghazi, M. and Kemal Ekenel, H., "A comprehensive analysis of deep learning based representation for face recognition," in [*Proceedings of the IEEE conference on computer vision and pattern recognition workshops*], 34–41 (2016).

[13] Grm, K., Štruc, V., Artiges, A., Caron, M., and Ekenel, H. K., "Strengths and weaknesses of deep learning models for face recognition against image degradations," *Iet Biometrics* **7**(1), 81–89 (2018).

[14] Agarwal, S., Farid, H., Gu, Y., He, M., Nagano, K., and Li, H., "Protecting world leaders against deep fakes," in [*Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) Workshops*], (June 2019).

[15] Yang, X., Li, Y., and Lyu, S., "Exposing deep fakes using inconsistent head poses," *ICASSP 2019 - 2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)* , 8261–8265 (2019).

[16] Jung, T., Kim, S., and Kim, K., "Deepvision: Deepfakes detection using human eye blinking pattern," *IEEE Access* **8**, 83144–83154 (2020).

[17] Zhou, P., Han, X., Morariu, V. I., and Davis, L. S., "Two-Stream Neural Networks for Tampered Face Detection," in [*2017 IEEE Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*], 1831–1839 (July 2017). ISSN: 2160-7516.

[18] Chollet, F., "Xception: Deep learning with depthwise separable convolutions," *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)* , 1800–1807 (2017).

[19] Sabour, S., Frosst, N., and Hinton, G. E., "Dynamic routing between capsules," *Advances in neural information processing systems* **30** (2017).

[20] Dang, H., Liu, F., Stehouwer, J., Liu, X., and Jain, A. K., "On the detection of digital face manipulation," in [*Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern recognition*], 5781–5790 (2020).

[21] Seferbekov, S. https://github.com/selimsef/dfdc_deepfake_challenge.

[22] Hanqing, Z., Hao, C., and Wenbo, Z. https://github.com/cuihaoleo/kaggle-dfdc.

[23] Davletshin, A. https://github.com/NTech-Lab/deepfake-detection-challenge.

[24] Jing, S., Huafeng, S., Zhenfei, Y., Zheng, F., Guojun, Y., Siyu, C., Ning, N., and Yu, L. https://github.com/Siyu-C/RobustForensics.

[25] James, H. and Ian, P. https://github.com/jphdotam/DFDC/.

[26] Hendrycks, D. and Dietterich, T., "Benchmarking neural network robustness to common corruptions and perturbations," *Proceedings of the International Conference on Learning Representations* (2019).

[27] Deng, J., Dong, W., Socher, R., Li, L.-J., Li, K., and Fei-Fei, L., "Imagenet: A large-scale hierarchical image database," in [*2009 IEEE Conference on Computer Vision and Pattern Recognition*], 248–255 (2009).

[28] Michaelis, C., Mitzkus, B., Geirhos, R., Rusak, E., Bringmann, O., Ecker, A. S., Bethge, M., and Brendel, W., "Benchmarking robustness in object detection: Autonomous driving when winter is coming," *arXiv preprint arXiv:1907.07484* (2019).

[29] Kamann, C. and Rother, C., "Benchmarking the robustness of semantic segmentation models," *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)* (Jun 2020).

[30] Sakaridis, C., Dai, D., and Van Gool, L., "Acdc: The adverse conditions dataset with correspondences for semantic driving scene understanding," in [*Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*], 10765–10775 (October 2021).

[31] Karahan, S., Yildirum, M. K., Kirtac, K., Rende, F. S., Butun, G., and Ekenel, H. K., "How image degradations affect deep cnn-based face recognition?," in [*2016 international conference of the biometrics special interest group (BIOSIG)*], 1–5, IEEE (2016).

[32] Petitcolas, F. A., Anderson, R. J., and Kuhn, M. G., "Attacks on copyright marking systems," in [*International workshop on information hiding*], 218–238, Springer (1998).

[33] Cogranne, R., Giboulot, Q., and Bas, P., "Alaska# 2: Challenging academic research on steganalysis with realistic images," in [*2020 IEEE International Workshop on Information Forensics and Security (WIFS)*], 1–5, IEEE (2020).

[34] Foi, A., Trimeche, M., Katkovnik, V., and Egiazarian, K., "Practical poissonian-gaussian noise modeling and fitting for single-image raw-data," *IEEE Transactions on Image Processing* **17**(10), 1737–1754 (2008).

[35] Marciniak, T., Chmielewska, A., Weychan, R., Parzych, M., and Dabrowski, A., "Influence of low resolution of images on reliability of face detection and recognition," *Multimedia Tools and Applications* **74** (06 2013).

[36] Li, P., Prieto, L., Mery, D., and Flynn, P. J., "On low-resolution face recognition in the wild: Comparisons and new techniques," *IEEE Transactions on Information Forensics and Security* **14**, 2000–2012 (2019).

[37] Ballé, J., Minnen, D., Singh, S., Hwang, S. J., and Johnston, N., "Variational image compression with a scale hyperprior," in [*International Conference on Learning Representations*], (2018).

[38] Zhang, K., Zuo, W., Chen, Y., Meng, D., and Zhang, L., "Beyond a gaussian denoiser: Residual learning of deep cnn for image denoising," *IEEE Transactions on Image Processing* **26**(7), 3142–3155 (2017).

[39] Li, Y., Yang, X., Sun, P., Qi, H., and Lyu, S., "Celeb-df: A large-scale challenging dataset for deepfake forensics," in [*IEEE Conference on Computer Vision and Patten Recognition (CVPR)*], (2020).

[40] King, D. E., "Dlib-ml: A machine learning toolkit," *Journal of Machine Learning Research* **10**, 1755–1758 (2009).