Lu and Ebrahimi
EURASIP Journal on Image and Video Processing (2024) 2024:6
https://doi.org/10.1186/s13640-024-00621-8

EURASIP Journal on Image and Video Processing

RESEARCH Open Access

Assessment framework for deepfake detection in real-world situations

Yuhang Lu[1*] and Touradj Ebrahimi[1]

*Correspondence: yuhang.lu@epfl.ch

[1] Multimedia Signal Processing Group (MMSPG), École Polytechnique Fédérale de Lausanne (EPFL), Lausanne, Switzerland

Abstract

Detecting digital face manipulation in images and video has attracted extensive attention due to the potential risk to public trust. To counteract the malicious usage of such techniques, deep learning-based deepfake detection methods have been employed and have exhibited remarkable performance. However, the performance of such detectors is often assessed on related benchmarks that hardly reflect real-world situations. For example, the impact of various image and video processing operations and typical workflow distortions on detection accuracy has not been systematically measured. In this paper, a more reliable assessment framework is proposed to evaluate the performance of learning-based deepfake detectors in more realistic settings. To the best of our acknowledgment, it is the first systematic assessment approach for deepfake detectors that not only reports the general performance under real-world conditions but also quantitatively measures their robustness toward different processing operations. To demonstrate the effectiveness and usage of the framework, extensive experiments and detailed analysis of four popular deepfake detection methods are further presented in this paper. In addition, a stochastic degradation-based data augmentation method driven by realistic processing operations is designed, which significantly improves the robustness of deepfake detectors.

Keywords: Assessment framework, Deepfake detection, Data augmentation

1 Introduction

In recent years, the rapid development of deep convolutional neural networks (DCNNs) and ease of access to large-scale datasets have led to significant progress on a broad range of computer vision tasks and meanwhile created a surge of new applications. For example, the recent advancement of generative adversarial networks (GANs) [1–3] has made it possible to generate realistic forged contents that are difficult for humans to distinguish from their authentic counterparts. In particular, current deep learning-based face manipulation techniques [4–7] are capable of changing the expression, attributes, and even identity of a human face image, the outcome of which refers to the popular term 'Deepfake'. The recent development of such technologies and the wide availability of open-source software has simplified the creation of deepfakes, increasingly damaging our trust in online media and raising serious public concerns. To counteract the misuse of these deepfake techniques and malicious

attacks, detecting manipulations in facial images and video has become a hot topic in the media forensics community and has received increasing attention from both academia and businesses.

Nowadays, multiple grand challenges, competitions, and public benchmarks [8–10] 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, 11–16] have been published and have reported promising results on different datasets. But some studies [17, 18] have shown that the detection performance significantly drops in the cross-dataset scenario, where the fake samples are forged by other unknown manipulation methods. Therefore, cross-dataset evaluation has become an important step in recent studies to better show the advantages of deepfake detection methods, encouraging researchers [19–21] to propose detection methods with better generalization ability to different types of manipulations.

Nevertheless, another scenario that commonly exists in the real world has received little attention from researchers. In fact, it has long been shown that DCNN-based methods are vulnerable to real-world perturbations and processing operations [22–24] in different vision tasks. In more realistic conditions, images and video 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 content. To the best of our acknowledgment, most of the current deep learning-based deepfake detection methods are developed based on constrained and less realistic face manipulation datasets, and therefore, they are not robust enough in real-world situations. Similarly, the conventional assessment approach, which exists in various benchmarks, often directly samples test data from the same distribution as training data and can hardly reflect model performance in more complex situations. In fact, most of the existing deepfake detection methods only report their performance on some well-known benchmarks in the community.

Therefore, a more reliable and systematic approach is desired firsthand to assess the performance of a deepfake detector in more realistic scenarios and further motivate researchers to develop robust detection methods. In this paper, a comprehensive assessment framework for deepfake detection in real-world conditions has been conceived for both image and video deepfakes. Notably, the realistic situations are simulated by applying common image and video processing operations to the test data. The performance of multiple deepfake detectors is measured under the impact of various real-world processing operations. At the same time, a generic approach to improve the robustness of the detectors has been proposed.

In summary, 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 systematically evaluates deepfake detectors in realistic situations.

- The performance of several popular deepfake detection methods has been evaluated and analyzed with the proposed performance evaluation framework. The extensive results demonstrate the necessity and effectiveness of the assessment approach.
- Inspired by the 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 different detectors.
- A flexible Python toolbox is developed and the source code of the proposed assessment framework is released to facilitate relevant research activities.

This article is an extended version of our recent publication [25]. The additional contents of this paper are summarized as follows.

- More recent deepfake detection methods have been summarized and introduced in the related work section.
- The proposed assessment framework has been extended to support the evaluation of video deepfake detectors.
- The performance of two current state-of-the-art deepfake detection methods has been additionally evaluated using the assessment framework.
- More substantial experimental results have been presented to better demonstrate the necessity and usage of the assessment framework. The performance and characteristics of four popular deepfake detection methods are analyzed in depth based on the assessment results.
- The impact of different image compression operations on the performance of deepfake detectors is additionally studied in detail.
- More experiments, comparisons, and cross-manipulation evaluations have been conducted for the proposed stochastic degradation-based augmentation method. Its effectiveness and limitations are further analyzed.

2 Related work

2.1 Deepfake detection

Deepfake detection is often treated as a binary classification problem in computer vision. Early on, solutions based on facial expressions [26], head movements [27] and eye blinking [28] were proposed to address such detection problems. In recent years, the primary solution to this problem is by leveraging advanced neural network architectures. Zhou et al. [29] proposed to detect deepfakes with a two-stream neural network. Rössler et al. [4] retrained an XceptionNet [30] with manipulated face dataset which outperforms their proposed benchmark. Nguyen et al. [11] combined traditional CNN and Capsule networks [31], which require fewer parameters. Some video deepfake detectors [32–34] leveraged recurrent convolutional neural networks to track forgery clues from the temporal sequences. Other creative attempts in network architectures include, but are not limited to, multi-task autoencoders [35, 36], efficient networks [21, 37] and vision transformers [38, 39]. In addition, the attention mechanism, a well-known technique to highlight the informative regions, has also been applied to further improve the training process of the detection system.

Dang et.al [40] proposed a detection system based on an attention mechanism. Zhao et al. [12] designed multi-attention heads 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 work [13–16, 41] attempts to resolve the problem in the frequency domain. The theory behind them is based on the fact that current popular GAN-based image manipulation methods often introduce low-frequency clues due to the built-in up-sampling operation. These methods transform the image to the frequency domain via DCT transformation and separate information according to different frequency bands. As a result, the forgery traces are more effectively captured.

To tackle the generalization problem, one important branch of work directly trains models with fully synthetic data, which forces the models to learn more generic representations for deepfake detection. For example, Xray [42] and SBIs [21] methods manually generate blended faces during the training process as fake samples, which reproduce the blending artifacts existing in real-world GAN-synthesized deepfakes. Both methods have achieved remarkable performance and notable generalization ability to certain types of manipulation methods. But as explained by the authors, these methods are susceptible to many common perturbations, such as low-resolution and heavy compression. In this paper, four different types of deepfake detectors [4, 11, 21, 39] are adopted for experiments.

2.2 Deepfake detection competitions review

To assist in faster progress and better advancement of deepfake detection tasks, numerous large-scale benchmarks, competitions, and challenges [4, 8–10] 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) [8] 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) [10] was organized by AI Singapore with a total prize pool of up to 500k USD 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 [8] 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 blurriness effect.

The current ranking approach in these 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.

Table 1 Deepfake Detection Challenge (DFDC) [8] top-5 prize winners and their corresponding results

Team name	Overall log loss
Selim Seferbekov [43]	0.4279
WM [44]	0.4284
NTechLab [45]	0.4345
Eighteen Years Old [46]	0.4347
The Medics [47]	0.4371

2.3 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 [22] measured the performance of image classification models with data disturbed by noise, blurring, and contrast changes. In [48], Hendrycks et al. presented a corrupted version of ImageNet [49] to benchmark the robustness of image recognition models against common image manipulations. [50–52] focused on a safety-critical task, autonomous driving, and provided a robustness benchmark for various relevant vision tasks, such as object detection and semantic segmentation. Similar work has been done for face recognition tasks, [23, 24, 53, 54] analyzed the robustness of CNN-based face recognition models toward face variations caused by illumination change, occlusion, and standard image processing operations. In the media forensics community, StirMark [55] tested the robustness of image watermarking algorithms. The ALASKA#2 dataset [56] was created following a careful evaluation of ISO parameters, JPEG compression, and noise level on FlickR images, etc., to help researchers in designing way more general and robust steganographic and steganalysis methods. It is worth noting that two popular deepfake detection benchmarks, DFDC [8] and Deeperforensics−1.0 [9] 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

Nowadays, deepfakes are distributed on the internet in both image and video formats. Some of the detection methods are targeted for both cases, while others are specially designed for one type of deepfakes. The proposed assessment framework is designed in a way that the performance of a deepfake detector can be evaluated under either image or video scenarios.

In the context of this paper, the main difference between the two scenarios resides in the real-world processing operations applied to the test data. In specific, in image

scenarios, we first extract frames from video and treat them as image deepfakes. The image processing operations are then applied to the forgery images. In video scenarios, we treat them as video deepfakes and directly apply video processing operations to the fake video.

In this section, the common realistic influencing factors and processing operations for image and video deepfakes are first introduced respectively. Then, the proposed assessment framework is described to provide a fair comparison for deepfake detectors under more realistic situations.

3.1 Realistic influencing factors for image deepfakes

In a real-world situation, the images are often processed by various digital image processing operations before being distributed. In more adverse cases, malicious deepfakes can be slightly corrupted to fool the detector while maintaining good perceptual quality. It is still unknown to which extent the popular deepfake detectors are able to make correct predictions. In this context, the most prominent factors have been considered in the assessment framework.

In general, the framework contains six categories of image processing operations or corruptions with more than ten minor types. Each type consists of multiple severity levels. The details of all operations used in evaluations are described below with the

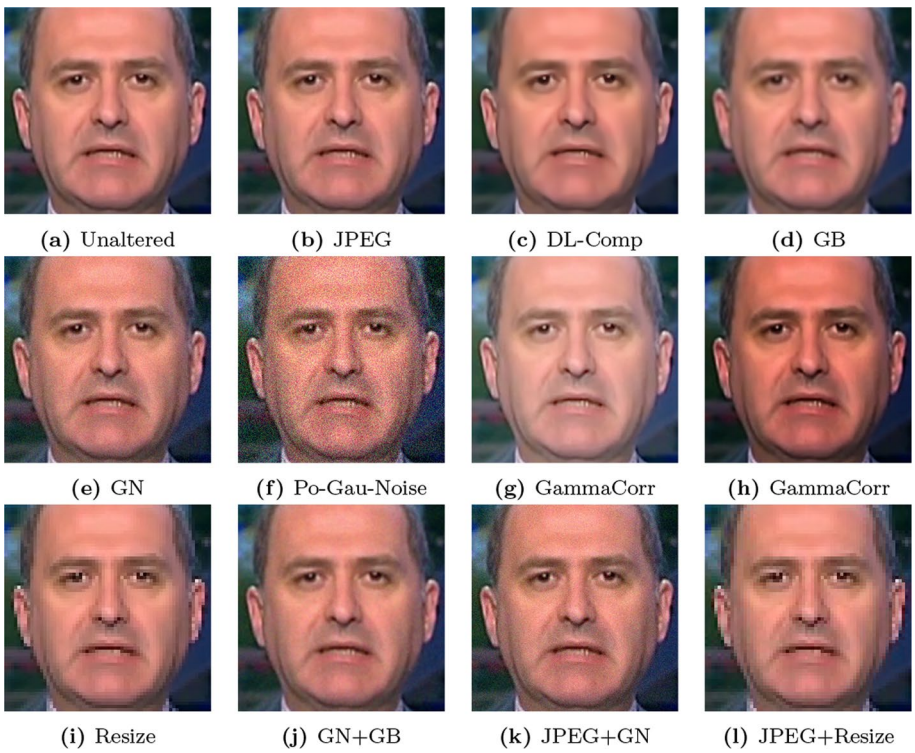

Fig. 1 Example of a typical image in the FFpp test set after applying various image processing operations. Some notations are explained as follows. DL-Comp: Deep learning-based compression. GB: Gaussian blur. GN: Gaussian noise. Po-Gau-Noise: Poissonian-Gaussian noise. GammaCorr: Gamma correction. Resize: Reduce resolution. +: Combination of two operations

illustration of a typical example in Fig. 1. In specific, the following factors are considered in the 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 [57] 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 processing operations. It refers to changing the dimensions of the media content to fit the display or other purposes. On the other hand, the resizing operation, more specifically the down-sampling operation, can significantly reduce the performance of modern deep learning-based detectors [58, 59] due to a loss of discriminative information. This is often the case for those earlier image contents that are of poor quality. In this framework, the impact of resizing operation is simulated by first downscaling the images and then upscaling back using bicubic interpolation.

Image 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 of content and are widely applied to image processing. In this framework, the JPEG compression artifacts are applied and the impact of different quality factors, i.e. from 30 to 95, on the deepfake detection system is evaluated. As deep learning-based compression techniques are becoming increasingly popular in this community, two AI-based image compression techniques [60, 61] are also considered in this framework with multiple 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 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 the denoising operation, the blurriness 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 the DnCNN technique [62] is assessed in the 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 are 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 Realistic influencing factors for video deepfakes

Face forgeries by deepfake technology are spread over the Internet not only in the form of images but also as video. The processing operations and various video effects are very common on different social media, smartphone applications, and streaming platforms. Their impact on the accuracy of detection methods should not be neglected.

The framework includes seven categories of video processing operations with commonly used parameters. The illustrative example of testing data is shown in Fig. 2. The factors are also described in detail as follows.

Video compression: Similar to images, uncompressed raw video requires a large amount of storage space. Although lossless video compression codecs can perform at a compression factor of 5 to 12, a typical lossy compression video can achieve a much lower data rate while maintaining high visual quality. In fact, compression technologies for video provide the basis for the distribution of video worldwide. The potential deepfake video propagates among social networks after being compressed several times. However, the possible side effect of lossy compression artifacts on deep learning-based detectors has not been sufficiently studied. It is necessary to test the robustness of a deepfake detector on compressed authentic and deepfake video. In this context, the proposed assessment framework consists of test data compressed by H.264 codec using the FFMPEG toolbox with two constant rate factors, namely 23 and 40.

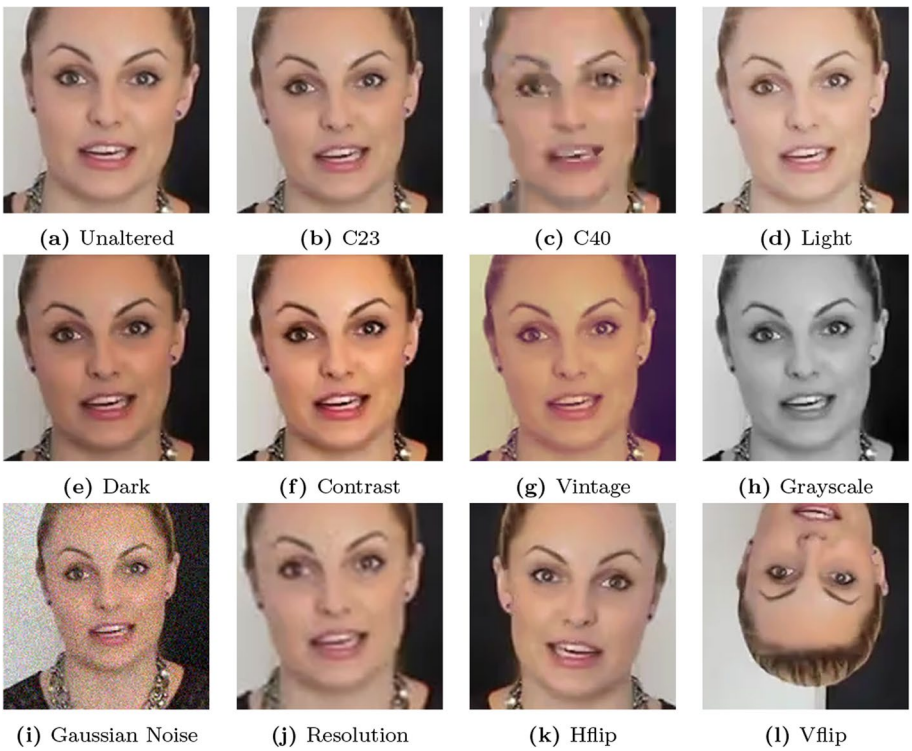

Fig. 2 Example of a typical video frame in the FFpp test set after applying different video processing operations. Some notations are explained as follows. C23 and C40: Video compression using H.264 codec with factors of 23 and 40. Light and Dark: Increase and decrease brightness. Resolution: Reduce video resolution. Hflip and Vflip: Horizontal and Vertical flip

Flip: Flipping a video horizontally describes the creation of a mirror video of the original footage. It is a very common video editing method that prevents video cuts from disorienting the viewer. But whether and to which extent the flipping operation can affect a deepfake detector has not been evaluated before. On the other hand, the vertical flipping operation is one of the easiest ways to fool a detector. In fact, most current detectors will not adjust or correct the face pose during preprocessing step. Hence, one can simply upload a flipped video to avoid being detected while it is still readable to a human.

Video filters: In recent years, video filters have become popular on social media. They are preset treatments included in many video editing apps, software, and social media platforms, providing easy access for users to alter the look of a video clip. Some common types of video filters include color filters, beauty filters, stylization filters, etc. The overall color palette of a deepfake video can be changed by a video filter on social media, making it an out-of-distribution sample from common deepfake databases. In the proposed assessment framework, two typical filters, 'Vintage' and 'Grayscale', are considered.

Brightness: Brightness is a measure of the overall lightness or darkness of a video. Adjusting the brightness of a video can affect the way that colors are perceived, as well as the visibility of details and textures. For example, increasing the brightness can make it easier to see details in shadows, while decreasing the brightness can obscure details in highlights. In real-world conditions, the brightness of a video is often adjusted to create a different sense of style of a video. The assessment framework takes this situation into consideration and measures the performance of a detector under different brightness conditions. More specifically, the 'Lighten' and 'Darken' commands in the FFMPEG toolbox are applied to the testing video, respectively.

Contrast: Contrast refers to the difference between the lightest and darkest areas of a video. Similar to brightness, adjusting contrast is one of the most common operations to change the visual appearance of a video. The 'Contrast' command in the FFMPEG toolbox is employed to increase the contrast of the testing video.

Noise: Similar to images, video noise is a common problem in video clips shot in low-light conditions or with small-sensor devices such as mobile phones. It often appears as annoying grains and artifacts in the video. Gaussian noise with a temporal variance but fixed strength is applied to the video data.

Resolution: Resolution refers to the number of pixels in a video. There is an important trade-off between the resolution and file size. Decreasing the resolution of a video will generally result in a low-quality video with fewer details to be displayed on the screen. But it can also reduce the file size, which makes it easier to store and share. On the other hand, the resolution change can also affect the ratio of width to height of the video. The performance of the deepfake detector when facing low-resolution or stretched video will be evaluated by the proposed framework.

3.3 Assessment methodology

Current deepfake detection algorithms are based on deep learning and rely heavily on the distribution of the training data. These methods are typically evaluated using a test dataset that is similar to the training sets. Some benchmarks, such as [8, 9], attempt to measure the performance of deepfake detectors under more realistic conditions by adding random perturbations to partial test data and mixing up with others. However, there

is no standard approach for determining the proportion or strength of these perturbations, which makes the results of these benchmarks more stochastic and less reliable. The assessment methodology proposed in this paper aims to more thoroughly measure the impact of various influencing factors, at different severity levels, on the performance of deepfake detection algorithms.

In this section, the principle and usage of our assessment framework are introduced in detail. First, the deepfake detector is trained on its original target datasets, such as FaceForensics++ [4]. The processing operations and corruptions in the framework are not applied to the training data. Then, as illustrated in Fig. 3, multiple copies of the test set are created, and each type of distortion at one specific severity level is applied to the copies independently. The standard test data together with different distorted data are fed to the deepfake detector respectively. Finally, the detector generates "real or fake" predictions. During the entire evaluation, the true positive rate (TPR) and false positive rate (FPR) are measured by constantly comparing the detector's predictions and the binary ground-truth labels. The ROC curve is plotted and the Area Under the Curve (AUC) score is reported as the final metric. An overall evaluation score can be obtained by averaging the scores from each distortion style and strength level to report the general performance of a tested detector. Besides, the computed metrics can also be grouped by each operation category to further analyze the robustness of one deepfake detector on a specific processing operation.

In addition, to relieve the burden on storage caused by the multiple copies of the 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 the TorchVison toolbox and can be easily integrated into the evaluation process.

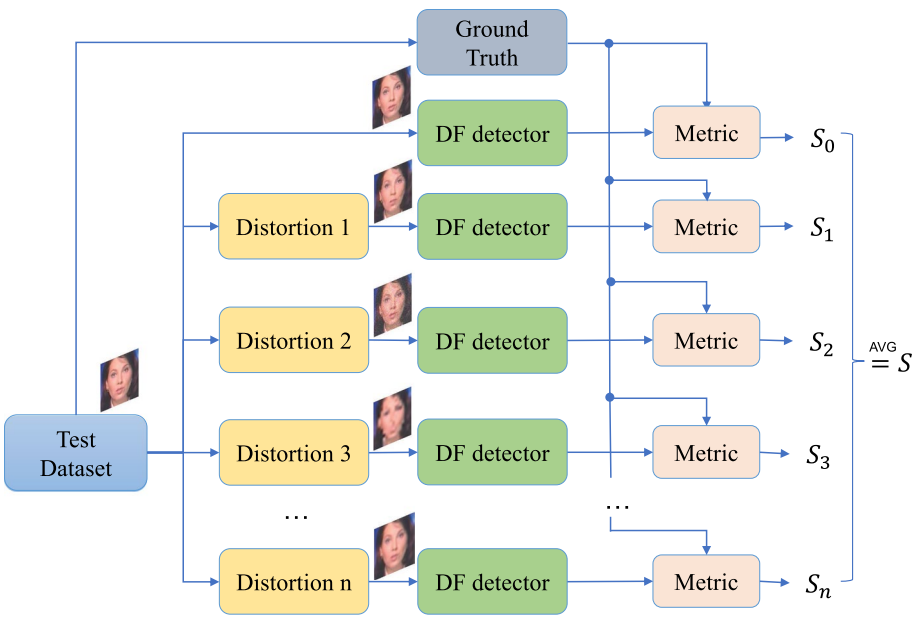

Fig. 3 Workflow of the proposed assessment framework. Distortions caused by processing operations are first applied to test data separately. The corresponding predictions by the deepfake detector are compared with the ground-truth label ("real or fake")

4 Stochastic degradation-based augmentation

To improve the ability of deepfake detection methods to handle realistic distortions and pre-processing operations, an effective data augmentation approach is proposed which leads to a robustness improvement.

Standard data augmentation methods often introduce geometric and color space transformation to enrich training data and improve the model generalization ability. But according to our experiments, this type of augmentation technique is less effective for deepfake detection under realistic conditions.

Motivated by a typical data acquisition and transmission pipeline in the real world, the stochastic degradation-based augmentation (SDAug) method is proposed. 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 the data degradation process, a carefully designed augmentation chain is conceived, which allows the training data to better resemble real-world conditions and further boosts the performance of deepfake detection methods.

Generally, 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 $\boldsymbol{n}$. In the end, ***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 the common data augmentation process, the SDAug method is implemented in a stochastic manner. The term 'stochastic' can be interpreted in the following two aspects. First, each aforementioned augmentation operation will occur with a certain probability in the augmentation chain. Second, each operation will use a random severity level for every frame. The realistic scenario is rather complex and does not necessarily consist 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, stochastic augmentation helps preserve more information from the original training data and therefore prevents 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 Experimental results

In this work, numerous experiments have been conducted to demonstrate the effectiveness and usage of the proposed assessment framework. The experimental setup will be described at the beginning of this section, followed by the substantial assessment results and analysis for both image and video scenarios. Then, the impact of three image compression technologies on deepfake detectors is further discussed as an example of the multiple applications of the framework. In the end, the effectiveness of the proposed augmentation technique is reported and analyzed.

5.1 Implementation details

5.1.1 Datasets

Two widely used face manipulation datasets are selected in this paper for extensive experimentation. For both datasets, there is a strict split up in the dataset suggested by the dataset provider and the video used for training will not appear in the validation and testing stages.

FaceForensics++ [4], denoted by FFpp, contains 1000 pristine and 4000 manipulated video generated by four different deepfake creation algorithms. In addition, 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*, *FFpp-C23*, and *FFpp-C40* 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, to provide a fair baseline, only uncompressed data are used for the final assessment.

Celeb-DFv2 [63] is another high-quality dataset, with 590 pristine celebrity video and 5639 fake video. The test data are selected as recommended by [63] while the rest are left for training purposes, where the training and validation sets were split into 80% and 20% accordingly.

5.1.2 Detection methods

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

Capsule-Forensics is a deepfake detection method based on a combination of capsule networks and CNNs. The capsule network was initially proposed by [31] to address some limitations of CNNs and it used a rather smaller amount of parameters than traditional CNN to train very deep neural networks. [11] employed the capsule network as a component in a deepfake detection pipeline for detecting manipulated images and video. This method achieved the best performance at that time in the FaceForensics++ dataset compared to its competing methods.

XceptionNet [30] is a popular CNN architecture in many computer vision tasks and has been used to detect face manipulations when it works as a classification network. Rössler et al. [4] first adopted it as a baseline in the FaceForensics++ benchmark along

with three other approaches. The detection system based on XceptionNet architecture was first pre-trained using ImageNet database [49] and then re-trained on a specific dataset for the deepfake detection task. It achieved excellent performance in the FaceForensics++ benchmark on both compressed and uncompressed contents and has become a popular baseline method for recent deepfake detection approaches.

SBIs [21] refers to a data synthetic method, Self-blended Images, which is specially designed for deepfake detection tasks. This method aims to generate hardly recognizable fake samples that contain common face forgery traces to encourage the model to learn more general and robust representations for face forgery detection. The overall detection system is based on a pre-trained deep classification network, EfficientNet-b4 [64]. After retraining with the SBIs technique, the detector demonstrates an impressive generalization ability to different unseen face manipulations and achieves the current state-of-the-art in cross-dataset settings. But its robustness to common image and video processing operations has not been measured.

UIA-VIT [39] detects face forgery using vision transformer technique. This approach jointly trains an end-to-end pipeline that both classifies the deepfake images and estimates the modification areas in an unsupervised manner. Overall, the UIA-VIT method focuses on intra-frame inconsistency without pixel-level annotations and achieves state-of-the-art performance regarding generalization ability.

5.1.3 Training details

The Capsule-Forensics, XceptionNet, and UIA-VIT methods are trained with Adam optimizer with $\beta_1 = 0.9$, $\beta_2 = 0.999$. Following the hyper-parameters suggested in the original paper, the Capsule-Forensics model is trained from scratch for 25 epochs with a learning rate of 5×10^{-4}, the XceptionNet model is trained for 10 epochs with a learning rate of 1×10^{-3}, and the UIA-VIT model is trained for 8 epochs with a learning rate of 3×10^{-5}. During training, 100 frames are randomly sampled from each video in the training set. For evaluation and testing, 32 frames are extracted from the video in the validation and test set. Extracted frames are pre-processed and cropped around the face regions using the dlib toolbox [65]. The face regions are finally resized into 300x300 pixels before feeding to the network.

The SBIs method has a different experimental setting from the previous three methods. It is retrained with SAM [66] optimizer for 100 epochs. The batch size and learning rate are set to 32 and 1×10^{-3}, respectively. During the training phase, only authentic high-quality video is used and the corresponding fake samples are created by their proposed self-blending method.

5.1.4 Performance metrics

During the evaluation, the Area Under Receiver Operating Characteristic Curve (AUC) is used as a metric in all experiments.

5.2 Assessment results on realistic image deepfakes

In this section, the performance of the Capsule-Forensics, XceptionNet, and UIA-VIT methods is measured when facing more realistic image deepfakes produced by the assessment framework. The three deepfake detectors are trained on the original

unaltered training sets of FFpp and Celeb-DFv2, respectively. The assessment framework further evaluates the performance of these detectors and summarizes the results as shown in Table 2 and Fig. 4.

In general, our findings draw the following conclusions. First, even mild real-world processing operations can have a noticeable negative impact on detection accuracy. The first two detectors present exceptional performance on unaltered FFpp and CelebDFv2 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. Although UIA-VIT is known for outstanding generalization ability, it also suffers from performance degradation in front of processing operations.

Second, the Capsule-Forensics and XceptionNet methods are prone to be affected by different types of perturbation. When trained on the same high-quality dataset, the Capsule-Forensics method is generally more robust toward JPEG compression, synthetic noise, and gamma correction operation, while XceptionNet at times presents slightly better results that could be of statistical nature. The results from the assessment framework provide valuable guidance toward improving a specific deepfake detector. Moreover, among the considered influencing factors, noise and blurriness effects on images are the most prominent for deepfake detectors. The performance of both detectors deteriorates 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 based on the assessment results. When trained only with very high-quality data (*FFpp-Raw*), both the Capsule-Forensics and XceptionNet models will be extremely sensitive to nearly all kinds of realistic processing operations. On the contrary, training the model with relatively low-quality data slightly improves the robustness toward low-intensity processing operations and distortions, but with a cost on the original high-quality testing set. For example, both models trained with compressed data (*FFpp-C23*, *FFpp-Full*) show a higher AUC score on our realistic benchmark, but their performance on original unaltered data decreases by 0.5–1%. However, although training with compressed data slightly improves the robustness of UIA-VIT against compression and noise, it brings more negative impact when facing other processing operations.

5.3 Assessment results on realistic video deepfakes

In addition to images, the framework provides a comprehensive evaluation for the four detection methods, i.e. Capsule-Forensics, XceptionNet, SBIs, and UIA-VIT on video deepfakes under real-world conditions. Table 3 summarizes the performance of the four deepfake detection methods using the proposed realistic benchmark.

As a result, when trained with high-quality data, both the Capsule-Forensics and XceptionNet methods show a similar trend as in the previous image deepfake detection benchmark and perform poorly when facing pre-processed video deepfakes. The SBIs and UIA-VIT methods outperform the other two detectors and present relatively stable scores in front of most video processing operations, particularly those artifacts introduced by changing brightness or assigning video filters.

However, when the previous two methods are trained directly on compressed data, they maintain higher robustness toward multiple processing operations and even outperform the SBIs method, whose overall score even decreases by 0.66% instead. On the

Table 2 AUC (%) scores of the Capsule-Forensics, denoted as CapsuleNet, XceptionNet, and UIA-VIT methods tested on unaltered and distorted variants of FFpp and Celeb-DFv2 test set respectively. Raw, C23, and Full refer to different quality settings of the FFpp. DL-Comp refers to deep learning-based compression [60] and 'High' refers to the high-quality compressed image

Methods	**TrainSet**	**Unaltered**	**JPEG**				**DL-Comp**				**Gaussian Noise**				**Po-Gau Noise**
			95	**60**	**30**	**AVG**	**High**	**Med**	**Low**	**AVG**	**5**	**10**	**30**	**AVG**	
CapsuleNet	FFpp-Raw	99.20	97.91	76.48	59.60	78.00	55.24	54.50	50.92	53.55	61.80	59.73	51.26	57.60	55.63
	FFpp-C23	96.32	95.09	95.76	74.91	88.59	56.96	57.42	81.57	65.32	84.51	78.56	58.63	73.90	70.59
	FFpp-Full	94.52	94.95	92.18	84.50	90.54	86.83	60.98	55.69	67.83	89.03	78.54	57.95	75.17	64.87
	CelebDFv2	99.14	99.32	98.88	93.07	97.09	99.01	96.77	88.95	94.91	95.24	63.27	59.20	72.57	87.06
XceptionNet	FFpp-Raw	99.56	76.77	56.00	54.20	62.32	50.16	50.37	50.10	50.21	50.12	51.00	50.36	50.49	48.98
	FFpp-C23	98.64	98.01	95.99	82.77	92.26	96.11	56.25	55.71	69.36	71.35	55.84	50.87	59.35	51.48
	FFpp-Full	99.02	99.00	94.78	87.86	93.88	94.36	54.88	55.78	68.34	59.00	55.09	54.08	56.06	51.43
	CelebDFv2	99.73	99.78	99.59	97.76	99.04	96.23	90.23	75.46	87.31	94.85	69.87	52.50	72.41	86.87
UIA-VIT	FFpp-Raw	99.38	99.30	95.16	84.92	93.13	89.19	57.49	56.75	67.81	96.86	89.10	72.32	86.09	82.97
	FFpp-C23	97.99	97.94	96.86	91.92	95.57	88.72	65.32	59.32	71.12	97.02	93.05	74.50	88.19	85.37
Methods	**TrainSet**	**Gaussian Blur**				**Gamma Correction**					**Resize**				**Overall Average**
		3	**7**	**11**	**AVG**	**0.1**	**0.75**	**1.3**	**2.5**	**AVG**	**x4**	**x8**	**x16**	**AVG**	
CapsuleNet	FFpp-Raw	67.19	58.22	52.26	59.22	50.50	98.86	99.17	96.12	86.16	67.48	53.18	53.10	57.92	65.41
	FFpp-C23	85.21	53.94	52.04	63.73	52.08	95.06	96.72	92.91	84.19	79.33	64.62	50.33	64.76	73.41
	FFpp-Full	85.72	58.83	56.05	66.87	56.02	93.86	93.87	85.44	82.30	84.65	65.34	52.02	67.34	75.01
	CelebDFv2	99.01	91.04	77.52	89.19	79.60	98.52	99.40	94.62	93.04	89.22	66.98	61.94	72.71	86.58
XceptionNet	FFpp-Raw	68.76	55.61	50.70	58.36	54.66	98.66	99.57	70.45	80.84	68.60	55.80	50.45	58.28	60.08
	FFpp-C23	95.38	59.92	52.33	69.21	58.43	98.34	98.64	91.35	86.69	92.55	70.27	56.00	72.94	74.97
	FFpp-Full	96.36	70.51	54.50	73.79	51.38	98.91	98.84	88.91	84.51	93.47	75.30	60.55	76.44	75.50
	CelebDFv2	98.77	91.81	79.94	90.17	74.97	99.53	99.74	97.49	92.93	96.01	72.21	63.03	77.08	86.49
UIA-VIT	FFpp-Raw	98.81	86.71	72.62	86.05	57.35	99.05	99.04	89.14	86.15	98.44	87.14	61.37	82.32	85.49
	FFpp-C23	91.19	65.74	64.08	73.67	57.34	96.55	96.28	88.32	84.62	94.24	81.35	60.24	78.61	84.39

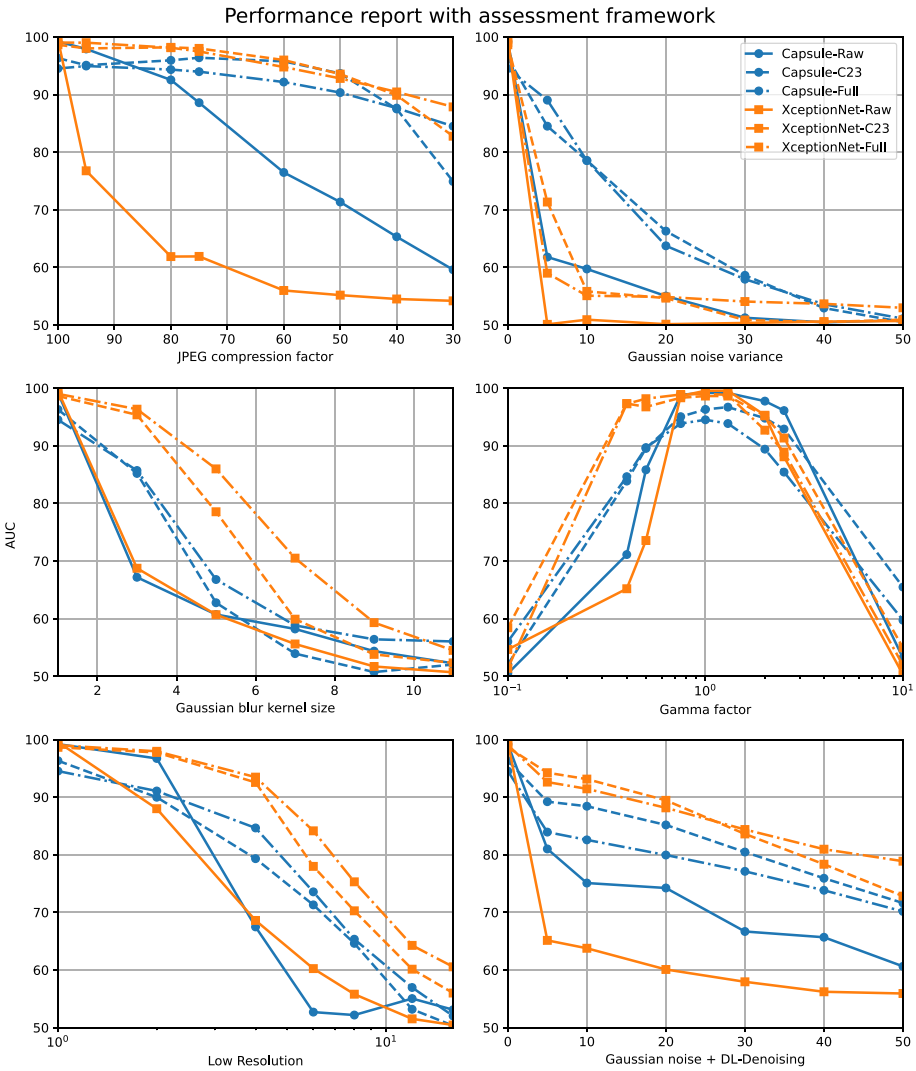

Fig. 4 Assessment results of two models trained on FFpp dataset. The suffixes of legends refer to the qualities of the training data. *Full* means using all available quality data for training

other hand, none of the three methods can properly classify video deepfakes processed by heavy compression, resolution reduction, or video noise.

In addition to benchmarking overall performance, the assessment framework also provides the means to analyze the behavior of a method under one specific realistic situation and help reveal the mechanism behind it. For instance, it is interesting to observe that, regardless of the training data, the SBIs method is more robust to geometric transformation than the other two and retains a good ability to accurately classify a vertically flipped video. It is because the SBIs method is based on local forgery traces instead of the global inconsistency on the face.

While the generalization problem is well-explored by synthetic data-based methods, how to improve robustness toward processing operations and distortions which exist in the real world is still an open question. This paper provides a systematic benchmarking approach that helps reveal the drawbacks of general deepfake detectors. For instance,

Table 3 AUC (%) scores of four selected deepfake detection methods on the distorted variants of the FFpp test set that are subject to different video processing operations. The notations C23 and C40 here refer to the two different compression rates using AVC/H.264 codec. The notation Resolution refers to reducing video resolution by a specific scale

Methods	TrainSet	Compression		Brightness		Grayscale	Contrast	Flipping		Resolution		Gaussian Noise	Vintage Filter	Overall Average
		C23	C40	Increase	Decrease			Horizontal	Vertical	x2	x4			
CapsuleNet	FFpp-Raw	77.97	54.14	73.31	70.62	68.38	69.31	73.13	63.20	65.43	56.99	54.14	72.94	66.63
XceptionNet		69.49	55.70	65.92	66.40	65.51	65.32	65.26	57.36	57.23	55.90	50.50	66.90	61.79
SBIs		90.43	76.27	86.38	86.47	86.27	85.94	85.98	79.28	76.35	63.62	71.52	86.54	81.25
UIA-VIT		93.82	71.56	91.10	88.55	88.91	89.18	89.02	77.74	79.78	72.72	71.50	87.11	83.42
CapsuleNet	FFpp-C23	95.61	66.03	93.27	92.31	87.43	91.55	91.98	71.49	80.28	67.56	53.50	88.86	81.66
XceptionNet		98.34	70.71	97.07	96.65	93.17	96.34	96.20	66.82	83.42	72.03	51.04	94.99	84.73
SBIs		91.71	75.43	87.63	86.51	87.31	87.40	86.84	81.22	75.40	64.31	57.06	86.28	80.59
UIA-VIT		96.82	75.94	95.08	94.58	92.79	95.49	95.47	81.00	88.59	78.80	75.42	93.12	88.59
CapsuleNet	FFpp-C40	82.64	78.33	80.22	80.77	79.30	52.78	78.64	61.53	76.88	71.91	78.41	75.82	74.77
XceptionNet		83.25	80.69	80.85	82.83	80.65	51.74	81.39	55.70	80.62	74.99	71.30	78.43	75.20
SBIs		83.00	70.66	76.49	78.15	77.49	62.82	77.34	69.14	67.04	56.42	76.67	76.50	72.64
UIA-VIT		86.98	83.86	85.13	85.73	79.77	86.22	86.12	68.66	84.88	77.74	78.93	82.98	82.25

although the SBIs method demonstrates a good generalization ability in cross-dataset experiments in their paper [21], our assessment framework shows that it is susceptible to some common perturbations in the real world, such as video compression, video noise, and low resolution.

5.4 Impact of different image coding algorithms

The assessment framework additionally provides means to measure the impact of a specific type of processing operation on the performance of a deepfake detector. For instance, image compression operation is almost inevitable during the distribution of a fake image. Meanwhile, AI-based compression technologies have become increasingly popular and are often capable of obtaining relatively smaller bitstreams. However, it is unknown to which extent the learning-based compression algorithms will affect the deepfake detection methods comparonventional JPEG compression.

In this section, a detailed comparison has been made between JPEG compression and two popular AI-based image compression methods, denoted by *bmshj* [60] and *hific* [61], respectively. In detail, the Capsule-Forensics and XceptionNet methods are first trained on uncompressed data. Afterward, their performance on different compressed data is evaluated using the framework and is then reported in Fig. 5. As a result, the image compression operation generally brings more negative impact to XceptionNet than to the Capsule-Forensics method. The latter obtains relatively high AUC scores when the test data are compressed by JPEG with high compression factors. Although the *bmshj*-based compression method is capable of achieving lower bitrates than JPEG, it brings significant negative impact to both detectors, whose predictions are close to random guess regardless of the select compression factor. On the contrary, both tested detectors are more robust to test data compressed using *hific* codec than using JPEG operation or *bmshj* codec, even with extremely low bitrates. The results reported in this section imply that *hific* codec introduces fewer adversarial artifacts, which can interrupt the functionality of other AI-based detectors.

5.5 Experimental results with augmentation

Table 4 shows the evaluation results of the Capsule-Forensics and XceptionNet methods trained on the unaltered FFpp dataset together with the proposed augmentation

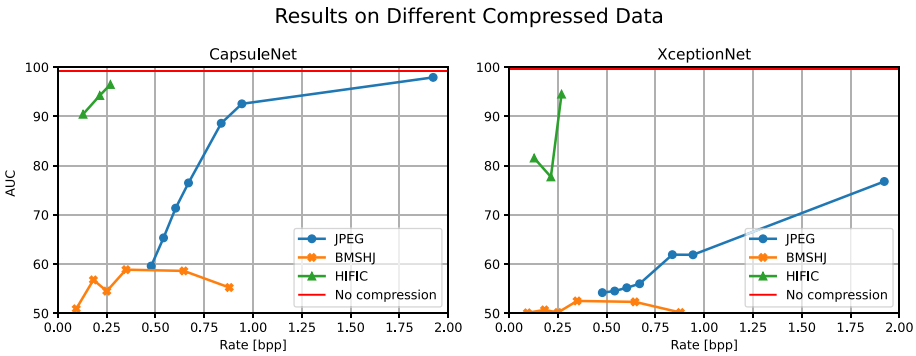

Fig. 5 Detection performance on data compressed by conventional and AI-based coding algorithms

Table 4 AUC (%) scores of cores of the Capsule-Forensics, denoted as CapsuleNet, and XceptionNet methods tested on unaltered and distorted variants of FFpp

Methods	TrainScheme	Unaltered	JPEG				DL-Comp				Gaussian Noise				Po-Gau Noise
			95	60	30	AVG	High	Med	Low	AVG	5	10	30	AVG	
CapsuleNet	FFpp-Raw	**99.20**	97.91	76.48	59.60	78.00	55.24	54.50	50.92	53.55	61.80	59.73	51.26	57.60	55.63
	FFpp-Full	94.52	94.95	92.18	84.50	90.54	86.83	60.98	55.69	67.83	89.03	78.54	57.95	75.17	64.87
	FFpp-Augmix	98.6	**98.68**	79.67	57.62	78.66	71.10	53.51	51.61	58.74	75.11	61.93	56.80	64.61	59.19
	FFpp-DAug	93.06	92.90	91.24	88.90	91.01	90.35	**81.96**	**70.00**	**80.77**	92.50	88.78	79.99	87.09	86.63
	FFpp-SDAug	98.16	97.97	**96.36**	**94.08**	**96.14**	**93.81**	71.41	59.74	74.99	**97.05**	**93.89**	**83.51**	**91.48**	**87.06**
XceptionNet	FFpp-Raw	**99.56**	76.77	56.00	54.20	62.32	50.16	50.37	50.10	50.21	50.12	51.00	50.36	50.49	51.02
	FFpp-Full	99.02	**99.00**	94.78	87.86	93.88	94.36	54.88	55.78	68.34	59.00	55.09	54.08	56.06	51.43
	FFpp-Augmix	99.15	87.12	63.38	59.58	70.03	77.86	62.07	55.76	65.23	77.37	68.45	56.99	67.60	62.00
	FFpp-DAug	89.51	89.47	89.27	89.00	89.25	89.49	**88.71**	**86.16**	88.12	89.43	89.30	88.22	88.98	88.97
	FFpp-SDAug	98.44	98.25	**97.36**	**96.12**	**97.24**	**98.03**	87.76	82.74	**89.51**	**97.37**	**95.88**	**91.71**	**94.99**	**94.57**
Methods	**TrainScheme**	**Gaussian Blur**				**Gamma Correction**					**Resize**				**Overall Average**
		3	**7**	**11**	**AVG**	**0.1**	**0.75**	**1.3**	**2.5**	**AVG**	**x4**	**x8**	**x16**	**AVG**	
CapsuleNet	FFpp-Raw	67.19	58.22	52.26	59.22	50.50	**98.86**	**99.17**	96.12	86.16	67.48	52.18	53.10	57.59	65.35
	FFpp-Full	85.72	58.83	56.05	66.87	56.02	93.86	93.87	85.44	82.30	84.65	65.34	52.02	67.34	75.01
	FFpp-Augmix	90.86	54.08	50.67	65.20	**76.17**	98.57	98.42	94.53	**91.92**	89.62	61.39	50.58	67.20	71.06
	FFpp-DAug	91.79	86.00	79.95	85.91	67.39	92.40	93.13	91.83	86.19	88.42	77.06	55.22	73.57	84.09
	FFpp-SDAug	**96.86**	**90.32**	**80.31**	**89.16**	60.17	97.68	98.18	**96.91**	88.24	**93.54**	**79.22**	**58.05**	**76.94**	**86.16**
XceptionNet	FFpp-Raw	68.76	55.61	50.70	58.36	54.66	98.66	**99.57**	70.45	80.84	68.60	55.80	50.45	58.28	60.08
	FFpp-Full	96.36	70.51	54.50	73.79	51.38	98.91	98.84	88.91	84.51	93.47	75.30	60.55	76.44	75.50
	FFpp-Augmix	90.45	62.58	53.00	68.68	**93.45**	**99.33**	98.32	85.87	**94.24**	64.64	54.57	50.00	56.40	70.36
	FFpp-DAug	89.22	87.62	85.28	87.37	69.08	89.42	89.35	87.74	83.90	88.31	81.30	63.89	77.83	85.91
	FFpp-SDAug	**98.31**	**97.35**	**94.51**	**96.72**	80.48	98.25	98.44	**97.75**	93.73	**97.30**	**86.26**	**67.14**	**83.57**	**92.63**

The suffix +DAug denotes that the model is trained with the proposed augmentation chain but without the stochastic manner. The suffix +SDAug denotes that the model is trained with the stochastic degradation-based augmentation technique. In this table, Bold font denotes the highest score

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. For example, previous experiments show that the detectors are more vulnerable to synthetic noises and blurry effects. The sub-figures in Figs. 6 and 7 further illustrate the impact of increasing the severity of these distortions on the two detection methods. 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 Table 4 and the last two sub-figures in Figs. 6 and 7, both detectors are much more robust toward learning-based compression, low-resolution effects, and other mixed distortions. A similar observation is obtained from the video deepfake assessment framework, see Tables 5 and 6. Although these video processing operations are not present in the proposed augmentation chain, the SDAug technique brings performance improvement to the Capsule-Forensics and XceptionNet methods on nearly all kinds of processed video deepfakes.

To compare with conventional augmentation methods based on geometric and color space transformation, the well-known Augmix [67] augmentation technique is evaluated under the same realistic assessment framework. This method generates multiple augmentation chains that work in parallel by randomly applying transformations to the training data. As a result, Augmix brings limited improvements to the robustness of the detector compared to SDAug, see Table 4. Its overall performance is even worse than simply training with low-quality data, which implies that the traditional data augmentation method is less practical when facing real-world distortions.

To show the effectiveness of the stochastic mechanism, an extra model has been trained using the same degradation-based augmentation chain but without randomness, which means the input data will be processed by all the augmentation operations with a fixed strength level. The corresponding experiment results are also reported in Table 4 and Figs. 6, 7, denoted as **+DAug**. As a result, the models trained with DAug are able to improve the performance on multiple processed data but the AUC scores degrade heavily on the original unmodified data. In comparison, the model trained with SDAug shows more significant robustness improvement and meanwhile maintains high performance on original high-quality data.

Finally, cross-dataset evaluations have been conducted for the Capsule-Forensics and XceptionNet methods to evaluate the generalization ability of those models trained with the proposed augmentation technique. First, the two detectors are trained on the FFpp dataset but tested on the Celeb-DFv1 and Celeb-DFv2 test sets for frame-level AUC scores. The two methods obtain very low scores on the new dataset. In comparison, the proposed augmentation scheme brings a noticeable performance improvement for both detectors on new datasets, showing its capability to improve the generalization ability on unseen forensic face contents. Moreover, we conduct more cross-manipulation experiments on FaceForensics++ which consists of four types of manipulations,

Table 5 AUC (%) scores of three selected deepfake detection methods trained with the SDAug augmentation method on the distorted variants of the FFpp test set

Methods	TrainSet	Compression		Brightness		Grayscale	Contrast	Flipping		Resolution		Gaussian Noise	Vintage Filter	Overall Average
		C23	C40	Increase	Decrease			Horizontal	Vertical	x2	x4			
CapsuleNet	FFpp- Raw	77.97	54.14	73.31	70.62	68.38	69.31	73.13	63.20	65.43	56.99	54.14	72.94	66.63
+SDAug		92.76	72.32	89.56	89.50	88.17	89.93	89.40	62.61	78.55	74.93	81.86	85.21	82.90
XceptionNet		69.49	55.70	65.92	66.40	65.51	65.32	65.26	57.36	57.23	55.90	50.50	66.90	61.79
+SDAug		94.89	80.53	93.36	93.05	92.64	92.98	92.32	57.65	88.42	81.60	88.91	90.47	87.24
SBIs		90.43	76.27	86.38	86.47	86.27	85.94	85.98	79.28	76.35	63.62	71.52	86.54	81.25
+SDAug		89.31	76.60	85.55	86.24	85.41	84.76	85.48	78.67	77.06	64.31	77.13	86.11	81.39

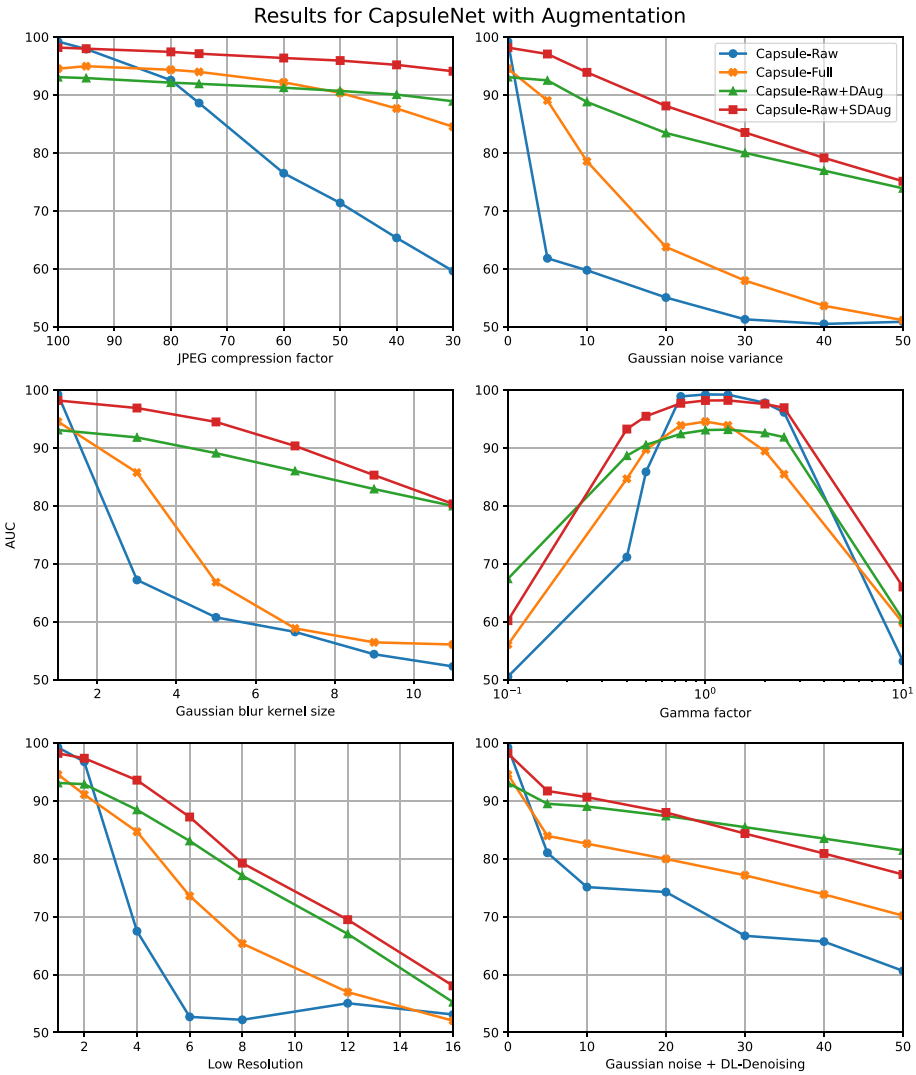

Fig. 6 Performance comparison between models trained on FFpp-Raw only and trained with the proposed augmentation method

namely DeepFakes, Face2Face, FaceSwap, and NeuralTextures. In specific, the XceptionNet model is trained on one type of manipulation and is tested on the remaining three. The results demonstrated in Fig. 8 show that the model trained with SDAug consistently achieves superior generalization performance.

5.6 Limitations and Future Work

The experiments carried out in this paper are mainly limited to video deepfakes or standard-quality image deepfakes. The detection of HD single-image deepfakes created by completely different methods, such as GANs, has not been evaluated with the proposed assessment framework. Although preliminary explorations have been done by previous work [68], there have been more advanced techniques recently to create HD single-image deepfakes, not only by GANs but also by Diffusion Models, and corresponding

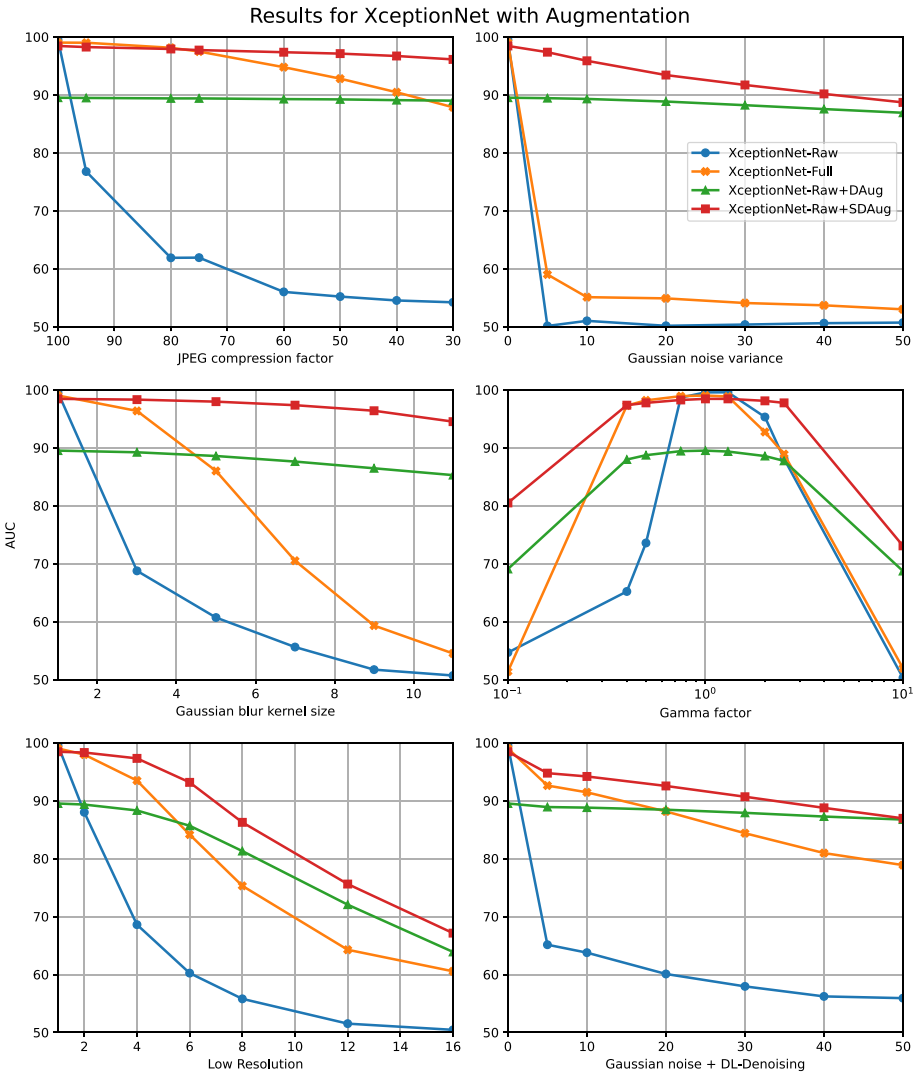

Fig. 7 Performance comparison between models trained on FFpp-Raw only and trained with the proposed augmentation method

Table 6 Cross-manipulation evaluation on Celeb-DFv1 and Celeb-DFv2 (AUC(%)) after training on FFpp dataset

Deepfake Detector	Augmentation Method	FFpp	Celeb-DFv1	Celeb-DFv2
Capsule	No Aug	99.20	43.36	54.39
	Augmix	98.66	53.45	58.65
	DAug	93.51	71.35	68.39
	SDAug	97.82	**74.84**	**71.86**
XceptionNet	No Aug	99.56	39.35	50.00
	Augmix	99.15	50.27	53.04
	DAug	78.64	64.79	62.81
	SDAug	98.44	**80.67**	**73.88**

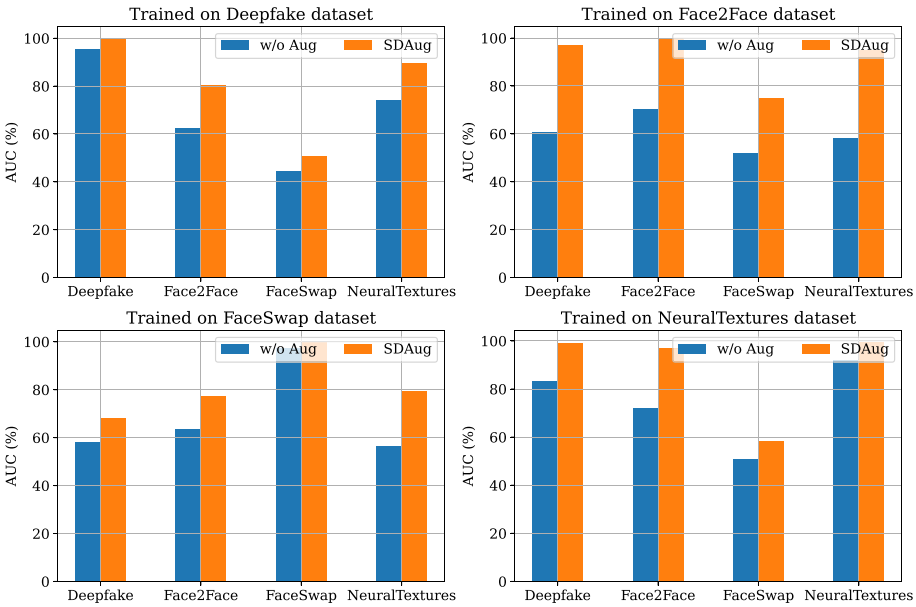

Fig. 8 Cross-manipulation experiments on FaceForensic++ (Raw) dataset with XceptionNet trained on four different types of manipulated dataset separately, namely Deepfake, Face2Face, FaceSwap, NeuralTextures. AUC (%) scores are compared between the XceptionNet model trained with or without the SDAug technique

detection methods. It would be interesting to extend the assessment framework to be able to study the robustness of state-of-the-art HD image deepfake detectors.

On the other hand, although the proposed augmentation technique is in general very helpful in improving the robustness of deepfake detectors when facing various real-world image and video processing operations, some limitations have been observed from the previous results report. First of all, the augmentation chain is hand-designed and the selection of hyperparameters might not be optimal. The proposed augmentation chain could be improved by conducting a parameter search with AutoML technology. Second, according to Table 5, the augmentation method generally provides limited help for SBIs method, because SBIs is entirely based on synthetic data and the augmentation can possibly corrupt the manually designed forgery traces. It could be promising to incorporate our proposed augmentation operations into the forgery data synthesis process to further improve the robustness of detectors based on synthetic forgery data.

6 Conclusion

Most of the current deepfake detection methods are designed to be as high performing as possible on specific benchmarks. But it has been shown that current assessment and ranking approaches employed in related benchmarks are less reliable and insightful. In this work, a more systematic performance assessment approach is proposed for deepfake detectors in realistic situations. To show the necessity and usage of the assessment framework, extensive experiments have been performed, where the robustness of four popular deepfake detectors is reported and analyzed. Furthermore, motivated by the assessment results, a new data augmentation chain based on a natural data degradation process has been conceived and shown to significantly improve the model's robustness

against distortions from various image and video processing operations. The effectiveness and limitations of the proposed augmentation method have been also discussed in detail.

Abbreviations

AUC	Area under receiver operating characteristic curve
DAug	Degradation-based augmentation
DCNNs	Deep convolutional neural networks
DFDC	Deepfake detection challenge
FFpp	FaceForensicsplusplus
GANs	Generative adversarial networks
SDAug	Stochastic degradation-based augmentation
TMC	Trusted media challenge

Acknowledgements
No additional acknowledgments.

Author contributions
All authors participated in the design of the analytics, performance measures, experiments, and writing of the manuscript. All authors read and approved the final manuscript.

Funding
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.

Availability of data and materials
The datasets used and/or analyzed during the current study are available from the corresponding author on reasonable request.

Declarations

Ethics approval and consent to participate
Not applicable

Consent for publication
Not applicable

Competing interests
The authors declare that they have no competing interests.

Received: 4 January 2023 Accepted: 11 January 2024
Published online: 13 February 2024

References

1. T. Karras, T. Aila, S. Laine, J. Lehtinen, Progressive growing of gans for improved quality, stability, and variation. arXiv preprint (2017)
2. T. Karras, S. Laine, T. Aila, A style-based generator architecture for generative adversarial networks. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 4401–4410 (2019)
3. T. Karras, S. Laine, M. Aittala, J. Hellsten, J. Lehtinen, T. Aila, Analyzing and improving the image quality of stylegan. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 8110–8119 (2020)
4. A. Rössler, D. Cozzolino, L. Verdoliva, C. Riess, J. Thies, M. Nießner, FaceForensics++: Learning to detect manipulated facial images. In: International Conference on Computer Vision (ICCV) (2019)
5. J. Thies, M. Zollhöfer, M. Nießner, Deferred neural rendering: image synthesis using neural textures. ACM Trans. Graph. **38**(4), 1–12 (2019)
6. Y. Nirkin, Y. Keller, T. Hassner, Fsgan: Subject agnostic face swapping and reenactment. In: Proceedings of the IEEE/CVF International Conference on Computer Vision, pp. 7184–7193 (2019)
7. E. Zakharov, A. Shysheya, E. Burkov, V. Lempitsky, Few-shot adversarial learning of realistic neural talking head models. In: Proceedings of the IEEE/CVF International Conference on Computer Vision, pp. 9459–9468 (2019)
8. B. Dolhansky, J. Bitton, B. Pflaum, J. Lu, R. Howes, M. Wang, C.C. Ferrer, The deepfake detection challenge dataset **2006**, 07397 (2020)
9. L. Jiang, R. Li, W. Wu, C. Qian, C.C. Loy, DeeperForensics-1.0: a large-scale dataset for real-world face forgery detection (2020). 2001.03024
10. W. Chen, B. Chua, S. Winkler, AI Singapore trusted media challenge dataset. arXiv preprint arXiv:2201.04788 (2022)
11. H.H. Nguyen, J. Yamagishi, I. Echizen, Use of a capsule network to detect fake images and videos. ArXiv (2019)
12. H. Zhao, W. Zhou, D. Chen, T. Wei, W. Zhang, N. Yu, Multi-attentional deepfake detection. 2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2185–2194 (2021)

13. H. Liu, X. Li, W. Zhou, Y. Chen, Y. He, H. Xue, W. Zhang, N. Yu, Spatial-phase shallow learning: rethinking face forgery detection in frequency domain. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 772–781 (2021)
14. Y. Qian, G. Yin, L. Sheng, Z. Chen, J. Shao, Thinking in frequency: Face forgery detection by mining frequency-aware clues. In: European Conference on Computer Vision, pp. 86–103 (2020). Springer
15. J. Li, H. Xie, J. Li, Z. Wang, Y. Zhang, 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, pp. 6458–6467 (2021)
16. Y. Luo, Y. Zhang, J. Yan, W. Liu, Generalizing face forgery detection with high-frequency features. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 16317–16326 (2021)
17. A. Khodabakhsh, R. Ramachandra, K. Raja, P. Wasnik, C. Busch, Fake face detection methods: Can they be generalized? In: 2018 International Conference of the Biometrics Special Interest Group (BIOSIG), pp. 1–6 (2018). IEEE
18. X. Xuan, B. Peng, W. Wang, J. Dong, On the generalization of GAN image forensics. In: Chinese Conference on Biometric Recognition, pp. 134–141 (2019). Springer
19. A. Haliassos, K. Vougioukas, S. Petridis, M. Pantic, Lips don't lie: A generalisable and robust approach to face forgery detection. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 5039–5049 (2021)
20. M. Kim, S. Tariq, S.S. Woo, Fretal: Generalizing deepfake detection using knowledge distillation and representation learning. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 1001–1012 (2021)
21. K. Shiohara, T. Yamasaki, Detecting deepfakes with self-blended images. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), pp. 18720–18729 (2022)
22. S.F. Dodge, L. Karam, Understanding how image quality affects deep neural networks. 2016 Eighth International Conference on Quality of Multimedia Experience (QoMEX), 1–6 (2016)
23. M. Mehdipour Ghazi, H. Kemal Ekenel, A comprehensive analysis of deep learning based representation for face recognition. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops, pp. 34–41 (2016)
24. K. Grm, V. Štruc, A. Artiges, M. Caron, H.K. Ekenel, Strengths and weaknesses of deep learning models for face recognition against image degradations. IET Biom. **7**(1), 81–89 (2018)
25. Y. Lu, T. Ebrahimi, A novel assessment framework for learning-based deepfake detectors in realistic conditions. In: Applications of Digital Image Processing XLV, vol. 12226, pp. 207–217 (2022). SPIE
26. S. Agarwal, H. Farid, Y. Gu, M. He, K. Nagano, H. Li, Protecting world leaders against deep fakes. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) Workshops (2019)
27. X. Yang, Y. Li, S. Lyu, Exposing deep fakes using inconsistent head poses. ICASSP 2019 - 2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), 8261–8265 (2019)
28. T. Jung, S. Kim, K. Kim, Deepvision: deepfakes detection using human eye blinking pattern. IEEE Access **8**, 83144–83154 (2020). https://doi.org/10.1109/ACCESS.2020.2988660
29. P. Zhou, X. Han, V.I. Morariu, L.S. Davis, Two-Stream Neural Networks for Tampered Face Detection. In: 2017 IEEE Conference on Computer Vision and Pattern Recognition Workshops (CVPRW), pp. 1831–1839 (2017). https://doi.org/10.1109/CVPRW.2017.229. ISSN: 2160-7516
30. F. Chollet, Xception: Deep learning with depthwise separable convolutions. 2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 1800–1807 (2017)
31. S. Sabour, N. Frosst, G.E. Hinton, Dynamic routing between capsules. Adv. Neural Inf. Process.**30** (2017)
32. D. Güera, E.J. Delp, Deepfake video detection using recurrent neural networks. In: 2018 15th IEEE International Conference on Advanced Video and Signal Based Surveillance (AVSS), pp. 1–6 (2018). IEEE
33. E. Sabir, J. Cheng, A. Jaiswal, W. AbdAlmageed, I. Masi, P. Natarajan, Recurrent convolutional strategies for face manipulation detection in videos. Interfaces (GUI) **3**(1), 80–87 (2019)
34. I. Masi, A. Killekar, R.M. Mascarenhas, S.P. Gurudatt, W. AbdAlmageed, Two-branch recurrent network for isolating deepfakes in videos. In: European Conference on Computer Vision, pp. 667–684 (2020). Springer
35. H.H. Nguyen, F. Fang, J. Yamagishi, I. Echizen, Multi-task learning for detecting and segmenting manipulated facial images and videos. In: 2019 IEEE 10th International Conference on Biometrics Theory, Applications and Systems (BTAS), pp. 1–8 (2019). IEEE
36. M. Du, S. Pentyala, Y. Li, X. Hu, Towards generalizable deepfake detection with locality-aware autoencoder. In: Proceedings of the 29th ACM International Conference on Information & Knowledge Management, pp. 325–334 (2020)
37. D.M. Montserrat, H. Hao, S.K. Yarlagadda, S. Baireddy, R. Shao, J. Horvath, E. Bartusiak, J. Yang, D. Guera, F. Zhu, E.J. Delp, Deepfakes detection with automatic face weighting. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) Workshops (2020)
38. Y. Zheng, J. Bao, D. Chen, M. Zeng, F. Wen, Exploring temporal coherence for more general video face forgery detection. In: Proceedings of the IEEE/CVF International Conference on Computer Vision, pp. 15044–15054 (2021)
39. W. Zhuang, Q. Chu, Z. Tan, Q. Liu, H. Yuan, C. Miao, Z. Luo, N. Yu, UIA-VIT: unsupervised inconsistency-aware method based on vision transformer for face forgery detection. In: European Conference on Computer Vision, pp. 391–407 (2022). Springer
40. H. Dang, F. Liu, J. Stehouwer, X. Liu, A.K. Jain, On the detection of digital face manipulation. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 5781–5790 (2020)
41. T. Saikia, C. Schmid, T. Brox, Improving robustness against common corruptions with frequency biased models. In: Proceedings of the IEEE/CVF International Conference on Computer Vision, pp. 10211–10220 (2021)
42. L. Li, J. Bao, T. Zhang, H. Yang, D. Chen, F. Wen, B. Guo, Face x-ray for more general face forgery detection. 2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 5000–5009 (2020)
43. S. Seferbekov, https://github.com/selimsef/dfdc_deepfake_challenge
44. Z. Hanqing, C. Hao, Z. Wenbo, https://github.com/cuihaoleo/kaggle-dfdc
45. A. Davletshin, https://github.com/NTech-Lab/deepfake-detection-challenge

46. S. Jing, S. Huafeng, Y. Zhenfei, F. Zheng, Y. Guojun, C. Siyu, N. Ning, L. Yu, https://github.com/Siyu-C/RobustForensics
47. H. James, P. Ian, https://github.com/jphdotam/DFDC/
48. D. Hendrycks, T. Dietterich, Benchmarking neural network robustness to common corruptions and perturbations. Proceedings of the International Conference on Learning Representations (2019)
49. J. Deng, W. Dong, R. Socher, L.-J. Li, K. Li, L. Fei-Fei, Imagenet: A large-scale hierarchical image database. In: 2009 IEEE Conference on Computer Vision and Pattern Recognition, pp. 248–255 (2009). https://doi.org/10.1109/CVPR.2009.5206848
50. C. Michaelis, B. Mitzkus, R. Geirhos, E. Rusak, O. Bringmann, A.S. Ecker, M. Bethge, W. Brendel, Benchmarking robustness in object detection: Autonomous driving when winter is coming. arXiv preprint arXiv:1907.07484 (2019)
51. C. Kamann, C. Rother, Benchmarking the robustness of semantic segmentation models. 2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) (2020). https://doi.org/10.1109/cvpr42600.2020.00885
52. C. Sakaridis, D. Dai, L. Van Gool, Acdc: The adverse conditions dataset with correspondences for semantic driving scene understanding. In: Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), pp. 10765–10775 (2021)
53. S. Karahan, M.K. Yildirum, K. Kirtac, F.S. Rende, G. Butun, H.K. Ekenel, How image degradations affect deep cnn-based face recognition? In: 2016 International Conference of the Biometrics Special Interest Group (BIOSIG), pp. 1–5 (2016). IEEE
54. Y. Lu, L. Barras, T. Ebrahimi, A novel framework for assessment of deep face recognition systems in realistic conditions. In: 2022 10th European Workshop on Visual Information Processing (EUVIP), pp. 1–6 (2022). https://doi.org/10.1109/EUVIP53989.2022.9922840
55. F.A. Petitcolas, R.J. Anderson, M.G. Kuhn, Attacks on copyright marking systems. In: International Workshop on Information Hiding, pp. 218–238 (1998). Springer
56. R. Cogranne, Q. Giboulot, P. Bas, Alaska# 2: Challenging academic research on steganalysis with realistic images. In: 2020 IEEE International Workshop on Information Forensics and Security (WIFS), pp. 1–5 (2020). IEEE
57. A. Foi, M. Trimeche, V. Katkovnik, K. Egiazarian, Practical Poissonian-gaussian noise modeling and fitting for single-image raw-data. IEEE Trans. Image Process. **17**(10), 1737–1754 (2008). https://doi.org/10.1109/TIP.2008.2001399
58. T. Marciniak, A. Chmielewska, R. Weychan, M. Parzych, A. Dabrowski, Influence of low resolution of images on reliability of face detection and recognition. Multimed. Tools Appl. **74** (2013). https://doi.org/10.1007/s11042-013-1568-8
59. P. Li, L. Prieto, D. Mery, P.J. Flynn, On low-resolution face recognition in the wild: comparisons and new techniques. IEEE Trans. Inf. Forensics Secur. **14**, 2000–2012 (2019)
60. J. Ballé, D. Minnen, S. Singh, S.J. Hwang, N. Johnston, Variational image compression with a scale hyperprior. In: International Conference on Learning Representations (2018)
61. F. Mentzer, G.D. Toderici, M. Tschannen, E. Agustsson, High-fidelity generative image compression. Adv. Neural. Inf. Process. Syst. **33**, 11913–11924 (2020)
62. K. Zhang, W. Zuo, Y. Chen, D. Meng, L. Zhang, Beyond a gaussian denoiser: Residual learning of deep CNN for image denoising. IEEE Trans. Image Process. **26**(7), 3142–3155 (2017). https://doi.org/10.1109/TIP.2017.2662206
63. Y. Li, X. Yang, P. Sun, H. Qi, S. Lyu, Celeb-df: A large-scale challenging dataset for deepfake forensics. In: IEEE Conference on Computer Vision and Patten Recognition (CVPR) (2020)
64. M. Tan, Q. Le, Efficientnet: Rethinking model scaling for convolutional neural networks. In: International Conference on Machine Learning, pp. 6105–6114 (2019). PMLR
65. D.E. King, DLIB-ML: a machine learning toolkit. J. Mach. Learn. Res. **10**, 1755–1758 (2009)
66. P. Foret, A. Kleiner, H. Mobahi, B. Neyshabur, Sharpness-aware minimization for efficiently improving generalization. arXiv preprint arXiv:2010.01412 (2020)
67. D. Hendrycks, N. Mu, E.D. Cubuk, B. Zoph, J. Gilmer, B. Lakshminarayanan, Augmix: A simple data processing method to improve robustness and uncertainty. In: International Conference on Learning Representations (2019)
68. J. Sabel, F. Johansson, On the robustness and generalizability of face synthesis detection methods. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 962–971 (2021)

Publisher's Note