Improving Deepfake Detectors against Real-world Perturbations with Amplitude-Phase Switch Augmentation

Yuhang Lu, Ruizhi Luo, and Touradj Ebrahimi

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

ABSTRACT

In recent years, the remarkable progress in facial manipulation techniques has raised social concerns due to their potential malicious usage and has received considerable attention from both industry and academia. While current deep learning-based face forgery detection methods have achieved promising results, their performance often degrades drastically when they are tested in non-trivial situations under realistic perturbations. This paper proposes to leverage the information in the frequency domain, particularly the phase spectrum, to better differentiate between deepfakes and authentic images. Specifically, a new augmentation method called degradation-based amplitude-phase switch (DAPS) is proposed, which disregards the sensitive amplitude spectrum of a forged facial image and enforces the detection network to focus on phase components during the training process. Extensive evaluation results from a realistic assessment framework show that the proposed augmentation method significantly improves the robustness of two deepfake detectors analyzed and consistently outperform other augmentation approaches under various perturbations.

Keywords: Deepfake Detection, Data Augmentation, Robustness, Face Manipulation

1. INTRODUCTION

Recent studies have shown a rapid development of facial manipulation techniques enabling users to modify facial regions in an image or video and create a so-called "Deepfake". For example, current deep learning-driven generative models[1–3] are capable of changing the identities or modifying the facial attributes that are hardly distinguishable by human eyes. The forged image or video can be abused for malicious purposes, causing severe trust issues in society at large. Therefore, it is crucial to develop effective face forgery detection methods.

Nowadays, multiple datasets, benchmarks, and competitions[4–7] have been launched to assist the progress of developing more advanced deepfake detection methods. At the same time, a variety of methods[8–15, 15–17] have been proposed. Earlier studies mainly relied on hand-crafted features, while most of the current work adopts deep learning tools to tackle this challenge. Recent learning-based detection methods often leverage prior knowledge of certain face manipulation methods from specific databases and then mine the forgery clues using classical convolutional neural networks (CNNs) in a supervised manner. These approaches have achieved great success in some well-known datasets, such as FaceForensics++,[4] Celeb-DF,[5] and DFDC.[6]

Despite their excellent performance, most of the previous deepfake detection methods tend to suffer from overfitting problems. These methods often experience drastic performance drops when facing deepfakes created by unseen forgery techniques because they were trained on datasets created by specific face manipulation methods. Existing work[10, 15, 17] have attempted to address the generalization problem by exploiting some common artifacts shared by multiple datasets. For instance, Face X-ray[15] and SBIs[17] methods proposed to directly detect the blending artifacts instead of general forgery traces and managed to significantly improve the generalization ability. However, these methods are still susceptible to common perturbations because the blending artifacts can be easily corrupted by any processing operations such as compression. In fact, this is another challenge that commonly exists in the real world and has attracted little attention from researchers in deepfake detection. In more realistic situations, deepfake contents on social media can be post-processed by various image and video

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

processing operations such as resizing, compression, or stylization filters. Artifacts created by these operations can mask the forgery clues and mislead a deepfake detector, resulting in incorrect decisions. To the best of our acknowledgment, most of the current learning-based detection methods were developed under simple and constrained scenarios with less realistic face manipulation datasets. In this context, this paper aims at developing a deepfake detection method that is robust to more practical and realistic situations.

Many previous studies[18–23] suggest that it is easier to distinguish the forgery clues of a deepfake in the frequency domain by comparing to the normal frequency distributions of authentic images. Meanwhile, it has been shown[24] that the phase spectrum of the Fourier transform of an image is more resilient to disturbances than the corresponding amplitude spectrum. From another perspective, data augmentation techniques have been widely used in different vision tasks to enhance the generalization ability and robustness of a deep neural network. Inspired by the above, this paper proposes a new data augmentation method that aims at enforcing the neural network to focus on the phase component of the frequency distribution of the training data. More specifically, the paper brings the following contributions:

- A new data augmentation method called degradation-based amplitude-phase switch (DAPS) is conceived to improve the robustness of general deepfake detection methods under real-world conditions.
- Several classical data augmentation techniques have been adapted to the deepfake detection task to further compare with the DAPS augmentation method.
- Extensive experiments have been performed with a realistic image and video deepfake assessment framework that shows the proposed augmentation method brings significant improvement in the robustness of two deepfake detectors under consideration and consistently outperforms other augmentation approaches.

2. RELATED WORK

Over the past years, deepfake detection has gained significant attention in the scientific community due to its wide application and potential threat to public trust and has become an emerging research area. In recent years, various attempts have been made and remarkable performance achieved. In this section, current face forgery detection methods are reviewed from three aspects.

2.1 Deepfake Detection in Spatial Domain

With the recent advancement in deep learning, various methods have been proposed to address the challenge of face forgery detection. The majority of them exploit forgery clues in the spatial domain, such as RGB and HSV. Some approaches[11,13,16,25] detect deepfakes based on hand-crafted features, such as the inconsistency of head pose,[11] face expression,[13] eye blinking,[25] and lips movement.[16] Later on, the development of deep learning has enabled an effective extraction of deep representations from images and video. Some work mainly treated deepfake detection as a binary classification task and adapted the structure of existing neural networks to identify manipulated faces. For example, Zhou et al.[8] proposed to detect deepfakes with a two-stream neural network adapted from GoogLeNet.[26] MesoNet[9] designed a shallow neural network that comprises two inception modules and two convolution layers. Nguyen et al.[27] leveraged the capsule network[28] to detect face manipulation, which reduced the number of parameters while maintaining comparable performance to conventional convolutional neural networks (CNNs). Rössler et al.[4] demonstrated exceptional performance in detecting deepfakes created by various algorithms using the efficient XceptionNet.[29] It now serves as a popular baseline approach in benchmarks. Recent creative attempts in network structures have explored the usage of more advanced architectures, such as autoencoders,[12,14] EfficientNets,[17,30] as well as vision transformers,[31,32] and have further boosted the accuracy in detecting forged faces. In addition, a number of studies attempt to localize the manipulated regions in addition to performing the classification task. Some researchers[12,33–35] directly adopted multi-task learning to simultaneously detect deepfake and localize the modified areas, while more recent work[32,36,37] leveraged attention mechanism to jointly predict location information of manipulated regions.

Although these methods achieved sound performance at their times, they are incapable of obtaining high accuracy in more recent benchmarks and challenges, such as Deepfake Detection Challenge (DFDC)[6] and Trusted Media Challenge (TMC),[7] mainly due to the rapid progress in deepfake technologies.

2.2 Deepfake Detection in Frequency Domain

Frequency domain analysis is an important method in image and video processing and has been widely employed in vision tasks such as image classification[38] and super-resolution.[39, 40] Most such techniques convert the image from the spatial domain to the frequency domain with Discrete Fourier Transform (DFT), Discrete Cosine Transform (DCT), or Wavelet Transform (WT). Several studies[18–23] have proposed to resolve the deepfake detection task by analyzing the forgery clues in the frequency domain. Durall et al.[18] first attempted to extract frequency-domain information using DFT and average the amplitude of each frequency band to capture the abnormal information contained in forged images. F^3-Net[19] proposed a two-stream collaborative learning framework, which is composed of two frequency-aware branches. One extracts frequency information from images via DCT and the other branch analyzes the statistical discrepancy between real and forged images in the frequency domain. This approach achieves state-of-the-art performance on heavily compressed deepfake video. Frank et al.[20] conducted an in-depth frequency analysis on GAN-generated fake images via DFT and revealed obvious grid-like patterns in their frequency counterparts, which validated the potential of identifying forged faces in the frequency domain. Liu et al.[22] further verified the forgery clues in the frequency domain caused by the up-sampling operation in GANs and proposed to focus on the phase spectrum of the frequency components, which preserves more critical information for detection. Li et al.[21] integrated frequency transformation into a metric learning framework to learn more discriminative features for face forgery detection. Luo et al.[23] conducted a similar frequency analysis as the previous work[21] but took a complementary point of view, which focused on the high-frequency features to improve the generalization ability. Their method mainly extracted high-frequency noises at multiple scales for face forgery detection and achieved promising results in the cross-manipulation evaluation.

2.3 Data Augmentation

Data augmentation is a widely used technique to enhance the generalization ability and robustness of deep learning-based models. Common augmentation operations employed in vision tasks include random flipping, cropping, translation, etc. In addition to these traditional operations, more advanced augmentation techniques have been conceived specifically for vision models and have demonstrated outstanding performance in image classification tasks. In an earlier work, Cutout,[41] randomly square regions were masked out during the training and showed improvement in robustness and overall performance of a convolutional neural network in image classification tasks. Later on, Mixup[42] linearly combined two images and corresponding labels in the training batch. Cutmix[43] used a similar strategy of mixing different data by replacing a portion of an image with a portion of another image. AutoAugment[44] is a learnable approach that optimizes the selection of various augmentation operations. It iteratively discovers the best augmentation policies with reinforcement learning techniques. AugMix[45] utilized stochasticity and cascaded various augmentation operations and achieved state-of-the-art performance on ImageNet-C.[46] Other studies[47, 48] have explored augmenting training data with Gaussian noise and managed to improve the performance of an object classifier on corrupted images. In deepfake detection, researchers[49] proposed a realistic augmentation chain that managed to improve the robustness of a common deepfake detector under the attack of real-world perturbations.

3. PROPOSED METHOD

This section starts by describing the motivation for the proposed augmentation method. Then, two pivot steps of the augmentation pipeline, i.e. degradation-based augmentation chain and amplitude-phase switch are introduced.

3.1 Motivation

The proposed augmentation method is inspired by the following key observations. Several studies[18–23] have shown that frequency-domain analysis is capable of capturing hidden forgery clues based on abnormal frequency distributions. In particular, the phase spectrum of a frequency-domain deepfake image is more sensitive to the up-sampling artifacts caused by the deepfake creation process when compared to the amplitude spectrum.[20] Moreover, it has been shown[24] that the phase spectrum of a signal in the frequency domain is more resilient than the corresponding magnitude spectrum. The latter is easily disturbed by perturbations such as noise or

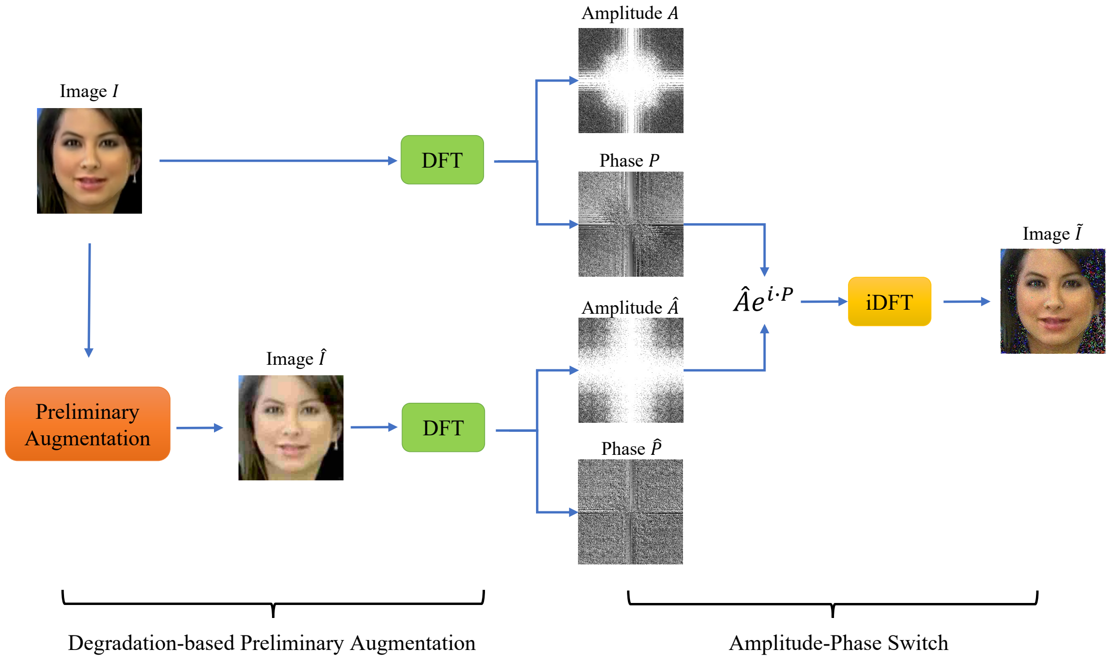

Figure 1: Workflow of the proposed degradation-based amplitude-phase switch (DAPS) augmentation method.

other artifacts from common processing operations. Therefore, forcing the detector to disregard the susceptible amplitude component and emphasize the resilient phase spectrum at the same time can potentially improve both the performance of the detector and its robustness against real-world disturbances. This paper proposes a new data augmentation technique, called degradation-based amplitude-phase switch (DAPS), for this purpose.

3.2 Degradation-based Amplitude-Phase Switch Augmentation

Figure 1 illustrates the proposed degradation-based amplitude-phase switch (DAPS) augmentation pipeline. Given an input image I, a preliminary augmentation is first applied to it and the output counterpart is denoted as $\hat{I}$. The pre-augmentation operations are designed in a way to simulate real-world perturbations. However, some of the subtle forgery clues could be hidden in the various artifacts produced by these processing operations and potentially impair the detection performance. To eliminate the disturbances of such artifacts, a second step of the DAPS augmentation aims at making the detection model concentrate more on the phase spectrum of the training data and less on the changes in the amplitude. For this purpose, Discrete Fourier Transform (DFT) is applied to the image pair $\{I, \hat{I}\}$, and then their amplitude spectrum is switched. The amplitude spectrum of the pre-augmented image and the phase spectrum of the original image is re-combined to generate a new training sample. Below, two pivot steps, i.e. degradation-based augmentation chain and amplitude-phase switch, of the proposed method are introduced in more details.

3.2.1 Degradation-based Augmentation Chain

The first step provides a preliminary data augmentation to the input data. Common data augmentation operations include a variety of geometric and color space transformations to enrich training data, such as translation, rotation, flipping, change of contrast, etc. More recent studies either concatenate multiple geometric transformations into augmentation chains[44, 45] or introduce random cut and paste to re-combine several input images.[41, 43] However, according to our experiments, these methods bring a somewhat limited impact on the robustness of deepfake detection.

The main insight in designing the preliminary data augmentation is motivated by typical perturbations that images and video are subject to in real-world conditions. In order to improve the resilience of deepfake detectors against realistic perturbations, this paper adopts a similar format as[45] to build an augmentation chain

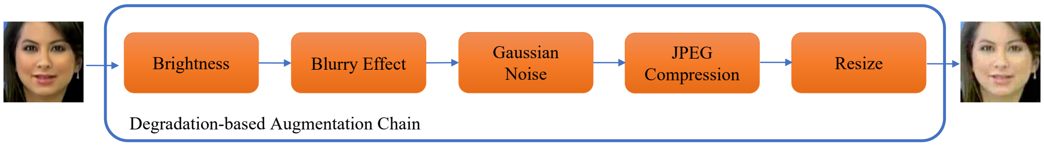

Figure 2: Preliminary augmentation chain based on realistic degradation processes.

but meanwhile selects multiple transformations that can simulate the common processing operations in the real world.

As shown in Figure 2, the input image is first modified by image enhancement operation and then convoluted with a Gaussian blurring kernel. Afterward, additive Gaussian noise is applied to the augmented data, followed by JPEG compression artifacts. At the end, the image is resized to a lower resolution to simulate loss of information and then it is up-sampled back to the original size to obtain the final augmented training data.

3.2.2 Amplitude-Phase Switch

The main objective of the second augmentation step is to emphasize the importance of the phase spectrum of input data during the training process. This section introduces amplitude-phase switch operation that enforces the deepfake detector to concentrate more on the phase than on the amplitude spectrum.

As illustrated in Figure 1, for each input image I and its corresponding pre-augmented counterpart $\hat{I}$, Discrete Fourier Transform (DFT) is applied to both and the corresponding frequency domain signals $\mathcal{F}(I)$ and $\mathcal{F}(\hat{I})$ are obtained through equations 1 and 2.

$$\mathcal{F}(I) = DFT(I) = \mathcal{A}e^{i\cdot\mathcal{P}}, \tag{1}$$

$$\mathcal{F}(\hat{I}) = DFT(\hat{I}) = \hat{\mathcal{A}}e^{i\cdot\hat{\mathcal{P}}}. \tag{2}$$

Afterward, the amplitude and phase spectrum of the two signals are switched. Specifically, the amplitude component of the pre-augmented data and the phase spectrum of the original data re-combine together to build a new signal $\hat{\mathcal{A}}e^{i\cdot\mathcal{P}}$ and the augmented training sample is obtained through the inverse Discrete Fourier Transform (iDFT) operation.

$$\tilde{I} = iDFT(\hat{\mathcal{A}}e^{i\cdot\mathcal{P}}), \tag{3}$$

where $\tilde{I}$ is the final augmented data used for training.

4. EXPERIMENTS AND RESULTS

This section first introduces the overall experimental setups and then presents the substantial experimental results to show the superiority of the proposed augmentation method.

4.1 Implementation Details

4.1.1 Datasets

This paper adopts the challenging FaceForensics++[4] dataset for experiments. FaceForensics++ comprises 1000 real video collected from YouTube. Each video was altered by four different types of face manipulation methods: Deepfake,[50] Face2Face,[51] FaceSwap,[52] and NeuralTextures.[53] The original dataset provides data in three quality levels, namely C0 (raw), C23 (light compression), and C40 (heavy compression). The experiments in this paper are mainly conducted with the raw quality of this dataset. As for the data preprocessing, 100 frames are first randomly extracted from every video for training purposes. Then, the dlib[54] face detector is applied to each video to extract and crop the face regions. Finally, the face images are resized into 300×300 pixels before feeding into the deepfake detection models.

Table 1: Frame-level AUC (%) scores of XceptionNet method tested on unaltered and distorted variants of FaceForensics++ test set via the realistic image deepfake assessment framework. DL-Comp is the abbreviation of deep learning-based compression operation.[57] Po-Gau Noise means Poisson-Gaussian noise. Resize refers to downsizing the image for certain scales.

Methods	Augmentation	Unaltered	JPEG				DL-Comp				Gaussian Noise				Po-Gau Noise
			95	60	30	AVG	High	Med	Low	AVG	5	10	30	AVG	
XceptionNet	No Aug	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
	Cutout[41]	99.63	85.50	56.79	54.41	65.57	52.14	48.60	48.60	49.78	51.24	50.20	50.20	50.55	50.67
	CutMix[43]	**99.80**	76.42	53.18	51.67	60.42	70.59	49.15	47.25	55.66	56.53	55.02	50.00	53.85	50.80
	AutoAugment[44]	99.53	98.83	75.32	65.18	79.78	89.21	50.93	53.24	64.46	92.76	72.00	54.28	73.01	63.00
	Augmix[45]	97.23	69.81	58.59	58.83	62.41	82.44	60.83	55.99	66.42	70.06	64.64	58.41	64.37	61.22
	DAPS	99.64	98.97	90.51	80.68	**90.05**	93.68	63.40	54.82	**70.63**	93.95	82.12	60.90	**78.99**	**74.09**

Methods	Augmentation	Gaussian Blur				Gamma Correction					Resize				Average
		3	7	11	AVG	0.1	0.75	1.3	2.5	AVG	x4	x8	x16	AVG	
XceptionNet	No Aug	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	63.89
	Cutout[41]	86.95	62.25	52.52	67.24	55.85	99.52	99.42	76.41	82.80	75.49	59.47	53.11	62.69	66.12
	Cutmix[43]	99.37	68.11	50.43	72.64	51.15	99.70	99.71	96.81	86.84	93.62	64.79	50.13	69.51	68.69
	AutoAugment[44]	99.05	62.35	50.62	70.67	97.65	99.55	99.52	99.40	**99.03**	91.55	63.74	54.34	69.88	77.42
	Augmix[45]	93.52	61.50	53.09	69.37	98.64	98.89	94.60	73.32	91.36	72.04	56.13	50.27	59.48	71.48
	DAPS	99.43	83.62	67.60	**83.55**	53.00	97.01	97.11	81.74	82.22	94.96	69.32	55.37	**73.22**	**81.55**

4.1.2 Detection Methods

To show the effectiveness of the proposed augmentation technique, it has been tested on two different face forgery detection methods.

XceptionNet[29] is a popular CNN architecture in many computer vision tasks. Roššler et al.[4] first utilized it to detect face manipulations in the FaceForensics++ benchmark. It achieves excellent results in identifying forged contents created by different manipulation methods and has become a popular baseline method for learning-based deepfake detection approaches

UIA-VIT[32] detects face forgery using the vision transformer technique. This approach jointly trains an end-to-end pipeline that both classifies the deepfake images and estimates the location 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.

4.1.3 Training Details

Following the hyper-parameters suggested in the original paper, the XceptionNet model is first pre-trained on ImageNet[55] and then fine tuned for 10 epochs with a learning rate 1×10^{-3}, while the UIA-VIT model is trained from scratch for 8 epochs with an initial learning rate of 3×10^{-5}, which is reduced when the validation accuracy arrives at plateau. Both methods are trained with Adam optimizer with $\beta_1 = 0.9, \beta_2 = 0.999$.

4.1.4 Evaluation Framework and Performance Metrics

This work mainly explores methods to improve the robustness of deepfake detection in real-world situations. Therefore, a realistic image and video assessment framework[56] has been employed for a fair measurement and comparison among different augmentation methods. In principle, the deepfake detectors are first trained on the originally targeted dataset as usual, but they are evaluated with multiple test data distorted by a variety of processing operations. Notably, the robustness of the detector against the following image processing operations will be measured: JPEG compression, learning-based compression,[57] noises, blurry effect, gamma correction operation, and low-resolution effect. Similarly, the video assessment framework measures detection performance when facing video compression, video filter, brightness and contrast changing, geometric flipping, low-resolution effect, and temporal noise.

During the evaluation, the frame-level Area Under the Receiver Operating Characteristic Curve (AUC) is adopted as the metric. An overall AUC score is reported by averaging the scores on different test sets, which reveals the robustness of the detector.

Table 2: Frame-level AUC (%) scores of UIA-VIT methods tested on unaltered and distorted variants of FaceForensics++ test set via the realistic image deepfake assessment framework. DL-Comp is the abbreviation of deep learning-based compression operation.[57] Po-Gau Noise means Poisson-Gaussian noise. Resize refers to downsizing the image for certain scales.

Methods	Augmentation	Unaltered	JPEG				DL-Comp				Gaussian Noise				Po-Gau Noise
			95	60	30	AVG	High	Med	Low	AVG	5	10	30	AVG	
UIA-VIT	No Aug	**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
	Cutout[41]	99.10	99.12	94.60	82.72	92.15	89.39	58.01	58.29	68.56	96.15	87.32	68.23	83.90	80.11
	CutMix[43]	99.33	99.29	93.21	79.07	90.52	51.77	59.61	56.02	55.80	94.26	83.08	65.44	80.93	76.69
	AutoAugment[44]	99.22	99.13	97.19	89.66	95.33	96.72	71.36	63.60	77.23	98.54	93.64	74.48	88.89	87.12
	Augmix[45]	99.20	99.14	96.31	87.31	94.25	96.28	67.07	63.79	75.71	98.10	91.59	74.80	88.16	85.86
	DAPS	98.61	98.43	97.29	94.55	**96.76**	97.54	88.64	73.17	**86.45**	98.21	96.65	81.45	**92.10**	**92.22**

Methods	Augmentation	Gaussian Blur				Gamma Correction					Resize				Average
		3	7	11	AVG	0.1	0.75	1.3	2.5	AVG	x4	x8	x16	AVG	
UIA-VIT	No Aug	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
	Cutout[41]	97.79	83.08	69.63	83.50	55.17	98.55	98.96	88.61	85.32	97.25	83.33	60.52	80.37	84.13
	Cutmix[43]	97.56	46.14	39.03	60.91	48.59	98.88	99.05	86.76	83.32	97.40	85.98	62.35	81.91	78.68
	AutoAugment[44]	98.52	89.34	80.22	89.36	84.40	99.16	99.12	98.49	**95.29**	97.65	87.04	66.36	83.68	89.51
	Augmix[45]	98.21	79.29	66.69	81.40	84.85	99.09	99.13	97.41	95.12	97.00	84.14	66.69	82.61	87.79
	DAPS	98.17	95.31	90.23	**94.57**	77.95	98.36	98.15	93.63	92.02	96.59	87.01	70.32	**84.64**	**92.17**

4.2 Experimental Results

This section provides experimental results assessed under two typical scenarios, i.e. image and video deepfakes in realistic situations. More specifically, the selected XceptionNet and UIA-VIT models are first trained on the unaltered FaceForensics++ (Raw) dataset with different data augmentation techniques. Their performance is then measured by two realistic assessment frameworks[56] to compare the robustness improvement brought by each augmentation method.

4.2.1 Results on Realistic Image Deepfakes

The realistic image assessment framework reports the robustness of two deepfake detectors trained with different augmentation techniques under the attack of various image processing operations and summarizes the results as shown in Table 1 and 2.

First, while the state-of-the-art UIA-VIT detector obtains better performance than XceptionNet when facing different perturbations, it still suffers from certain types of corruption such as learning-based compression artifacts, noises, and low-resolution effects.

Second, the performance of four classical augmentation techniques in the image classification field has been investigated. Due to the domain gap between the image classification and deepfake detection tasks, the augmentation techniques that are well-known in the former task are not necessarily effective in the latter. For example, it is shown that randomly cutting out image patches[41] from the training data leads to few improvements to the robustness of XceptionNet method while harming the performance of the UIA-VIT model. AutoAugment[44] and Augmix[45] are two similar approaches that build up sequential augmentation chains with classic 2D geometric and color-space transformations. Both of them are able to bring marginal improvements to the robustness of the two detectors, particularly under the impact of gamma correction operations.

More importantly, the proposed DAPS augmentation method achieves considerably higher scores than classical data augmentation techniques under the attack from most perturbations. For example, the UIA-VIT model trained with DAPS augmentation is significantly more robust to corruptions caused by learning-based compression, noises, or blurry operation.

4.2.2 Results on Realistic Video Deepfakes

Besides the image scenario, this work also provides a comprehensive evaluation of the augmentation methods under the impact of various video processing operations and the results have been summarized in Table 3.

The results from the video deepfake assessment framework verify that the Cutout and Cutmix augmentation methods are not suitable for robust deepfake detection tasks, although they are effective in image classification

Table 3: Frame-level AUC (%) scores of XceptionNet and UIA-VIT methods tested on unaltered and distorted variants of FaceForensics++ test set via the realistic video deepfake assessment framework.

Methods	Augmentation	Compression		Brightness		Contrast Increasing	Grayscale Filter	Vintage Filter	Flipping		Resolution		Gaussian Noise	Average
		C23	C40	↑	↓				Horizontal	Vertical	x2	x4		
XceptionNet	No Aug	66.49	55.70	65.92	66.40	65.32	65.51	66.90	65.26	**57.36**	57.23	55.90	50.50	61.54
	Cutout[41]	66.88	54.84	63.98	64.44	63.98	62.18	63.97	63.93	52.60	55.78	54.59	50.18	59.78
	Cutmix[43]	64.22	53.54	61.16	62.34	64.07	58.20	61.50	60.71	54.12	55.64	55.84	51.28	58.55
	AutoAugment[44]	83.13	58.32	77.73	77.52	**78.95**	**83.45**	76.33	76.06	52.41	62.32	53.97	55.17	69.61
	Augmix[45]	63.87	56.17	62.50	62.02	61.38	54.23	63.72	61.34	51.24	56.98	55.37	55.95	58.73
	DAPS	**85.68**	**65.20**	**80.76**	**81.26**	78.63	83.23	**79.43**	**80.51**	56.41	**71.29**	**64.05**	**62.28**	**74.06**
UIA-VIT	No Aug	93.82	71.56	91.10	88.55	89.18	88.91	87.11	89.02	**77.74**	79.78	72.72	71.50	83.42
	Cutout[41]	92.72	70.21	89.97	87.28	88.27	88.98	85.71	87.20	76.48	77.42	70.88	68.17	81.94
	Cutmix[43]	93.19	68.27	88.54	87.36	87.40	90.48	86.19	87.40	76.73	77.03	68.72	64.42	81.31
	AutoAugment[44]	**95.53**	73.99	92.63	92.34	92.30	89.82	88.85	92.29	77.04	81.11	73.71	74.42	85.34
	Augmix[45]	95.03	76.06	91.44	91.47	92.06	**91.03**	87.93	90.88	76.03	82.52	74.22	74.36	85.25
	DAPS	94.87	**80.46**	**93.09**	**92.47**	**92.36**	**91.03**	**91.11**	**92.66**	72.15	**87.35**	**82.50**	**80.32**	**87.53**

tasks. It is because both methods comprise a random-cutting operation, which can occasionally destroy some consistent forgery clues and make the training process much more difficult. AutoAugment method achieves comparable results with our method when facing perturbations in color space and the models trained with this augmentation technique are more robust to brightness and contrast changes and video filters. Notably, our proposed DAPS method introduces the most significant improvements to the robustness of two tested deepfake detectors. It outperforms other augmentation approaches, particularly in the cases of heavy compression, low resolution, and temporal noise.

It is also interesting to note that none of the augmentation techniques can improve the detection accuracy while facing a vertically flipped deepfake video, even if AutoAugment and Augmix methods contain similar geometric transformations in their augmentation chain. One needs to design a specific detection algorithm to resolve this problem, such as correcting the rotation of the video before conducting deepfake detection.

5. CONCLUSION

This paper provides a detailed review about the advantage of frequency-domain analysis in deepfake detection. A new data augmentation method, DAPS, is proposed to emphasize the resilient phase spectrum and disregard the susceptible amplitude components while training the detection model. The effectiveness of the conceived augmentation technique is evaluated by a realistic assessment framework, which significantly improves the robustness of two deepfake detection methods against real-world perturbations.

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] Karras, T., Aila, T., Laine, S., and Lehtinen, J., "Progressive growing of gans for improved quality, stability, and variation," *arXiv preprint arXiv:1710.10196* (2017).

[2] Brock, A., Donahue, J., and Simonyan, K., "Large scale gan training for high fidelity natural image synthesis," *arXiv preprint arXiv:1809.11096* (2018).

[3] Karras, T., Laine, S., and Aila, T., "A style-based generator architecture for generative adversarial networks," in [*Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*], 4401–4410 (2019).

[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] 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).

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

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

[8] 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.

[9] Afchar, D., Nozick, V., Yamagishi, J., and Echizen, I., "Mesonet: a compact facial video forgery detection network," in [*2018 IEEE international workshop on information forensics and security (WIFS)*], 1–7, IEEE (2018).

[10] Li, Y. and Lyu, S., "Exposing deepfake videos by detecting face warping artifacts," *arXiv preprint arXiv:1811.00656* (2018).

[11] 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).

[12] Nguyen, H. H., Fang, F., Yamagishi, J., and Echizen, I., "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)*], 1–8, IEEE (2019).

[13] 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).

[14] Du, M., Pentyala, S., Li, Y., and Hu, X., "Towards generalizable deepfake detection with locality-aware autoencoder," in [*Proceedings of the 29th ACM International Conference on Information & Knowledge Management*], 325–334 (2020).

[15] Li, L., Bao, J., Zhang, T., Yang, H., Chen, D., Wen, F., and Guo, B., "Face x-ray for more general face forgery detection," *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)* , 5000–5009 (2020).

[16] Haliassos, A., Vougioukas, K., Petridis, S., and Pantic, M., "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*], 5039–5049 (2021).

[17] Shiohara, K. and Yamasaki, T., "Detecting deepfakes with self-blended images," in [*Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*], 18720–18729 (June 2022).

[18] Durall, R., Keuper, M., Pfreundt, F.-J., and Keuper, J., "Unmasking deepfakes with simple features," *arXiv preprint arXiv:1911.00686* (2019).

[19] 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).

[20] Frank, J., Eisenhofer, T., Schönherr, L., Fischer, A., Kolossa, D., and Holz, T., "Leveraging frequency analysis for deep fake image recognition," in [*International conference on machine learning*], 3247–3258, PMLR (2020).

[21] 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).

[22] 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).

[23] 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).

[24] Chen, G., Peng, P., Ma, L., Li, J., Du, L., and Tian, Y., "Amplitude-phase recombination: Rethinking robustness of convolutional neural networks in frequency domain," in [*Proceedings of the IEEE/CVF International Conference on Computer Vision*], 458–467 (2021).

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

[26] Szegedy, C., Liu, W., Jia, Y., Sermanet, P., Reed, S., Anguelov, D., Erhan, D., Vanhoucke, V., and Rabinovich, A., "Going deeper with convolutions," in [*Proceedings of the IEEE conference on computer vision and pattern recognition*], 1–9 (2015).

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

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

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

[30] Montserrat, D. M., Hao, H., Yarlagadda, S. K., Baireddy, S., Shao, R., Horvath, J., Bartusiak, E., Yang, J., Guera, D., Zhu, F., and Delp, E. J., "Deepfakes detection with automatic face weighting," in [*Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) Workshops*], (June 2020).

[31] Zheng, Y., Bao, J., Chen, D., Zeng, M., and Wen, F., "Exploring temporal coherence for more general video face forgery detection," in [*Proceedings of the IEEE/CVF International Conference on Computer Vision*], 15044–15054 (2021).

[32] Zhuang, W., Chu, Q., Tan, Z., Liu, Q., Yuan, H., Miao, C., Luo, Z., and Yu, N., "Uia-vit: Unsupervised inconsistency-aware method based on vision transformer for face forgery detection," in [*European Conference on Computer Vision*], 391–407, Springer (2022).

[33] Bappy, J. H., Roy-Chowdhury, A. K., Bunk, J., Nataraj, L., and Manjunath, B., "Exploiting spatial structure for localizing manipulated image regions," in [*Proceedings of the IEEE international conference on computer vision*], 4970–4979 (2017).

[34] Salloum, R., Ren, Y., and Kuo, C.-C. J., "Image splicing localization using a multi-task fully convolutional network (mfcn)," *Journal of Visual Communication and Image Representation* **51**, 201–209 (2018).

[35] Songsri-in, K. and Zafeiriou, S., "Complement face forensic detection and localization with faciallandmarks," *arXiv preprint arXiv:1910.05455* (2019).

[36] 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).

[37] 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).

[38] Franzen, F., "Image classification in the frequency domain with neural networks and absolute value dct," in [*Image and Signal Processing: 8th International Conference, ICISP 2018, Cherbourg, France, July 2-4, 2018, Proceedings 8*], 301–309, Springer (2018).

[39] Huang, H., He, R., Sun, Z., and Tan, T., "Wavelet-srnet: A wavelet-based cnn for multi-scale face super resolution," in [*Proceedings of the IEEE international conference on computer vision*], 1689–1697 (2017).

[40] Li, J., You, S., and Robles-Kelly, A., "A frequency domain neural network for fast image super-resolution," in [*2018 International Joint Conference on Neural Networks (IJCNN)*], 1–8, IEEE (2018).

[41] DeVries, T. and Taylor, G. W., "Improved regularization of convolutional neural networks with cutout," *arXiv preprint arXiv:1708.04552* (2017).

[42] Zhang, H., Cisse, M., Dauphin, Y. N., and Lopez-Paz, D., "mixup: Beyond empirical risk minimization," in [*International Conference on Learning Representations*], (2018).

[43] Yun, S., Han, D., Oh, S. J., Chun, S., Choe, J., and Yoo, Y., "Cutmix: Regularization strategy to train strong classifiers with localizable features," in [*Proceedings of the IEEE/CVF international conference on computer vision*], 6023–6032 (2019).

[44] Cubuk, E. D., Zoph, B., Mane, D., Vasudevan, V., and Le, Q. V., "Autoaugment: Learning augmentation policies from data," *arXiv preprint arXiv:1805.09501* (2018).

[45] Hendrycks, D., Mu, N., Cubuk, E. D., Zoph, B., Gilmer, J., and Lakshminarayanan, B., "Augmix: A simple data processing method to improve robustness and uncertainty," in [*International Conference on Learning Representations*], (2019).

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

[47] Ford, N., Gilmer, J., Carlini, N., and Cubuk, E. D., "Adversarial examples are a natural consequence of test error in noise," in [*ICML*], (2019).

[48] Rusak, E., Schott, L., Zimmermann, R. S., Bitterwolf, J., Bringmann, O., Bethge, M., and Brendel, W., "Increasing the robustness of dnns against image corruptions by playing the game of noise," (2020).

[49] Lu, Y. and Ebrahimi, T., "A novel assessment framework for learning-based deepfake detectors in realistic conditions," in [*Applications of Digital Image Processing XLV*], **12226**, 207–217, SPIE (2022).

[50] https://github.com/deepfakes/faceswap.

[51] Thies, J., Zollhofer, M., Stamminger, M., Theobalt, C., and Nießner, M., "Face2face: Real-time face capture and reenactment of rgb videos," in [*Proceedings of the IEEE conference on computer vision and pattern recognition*], 2387–2395 (2016).

[52] https://www.github.com/MarekKowalski/FaceSwap.

[53] Thies, J., Zollhöfer, M., and Nießner, M., "Deferred neural rendering: Image synthesis using neural textures," *Acm Transactions on Graphics (TOG)* **38**(4), 1–12 (2019).

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

[55] 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).

[56] Lu, Y. and Ebrahimi, T., "Assessment framework for deepfake detection in real-world situations," *arXiv preprint arXiv:2304.06125* (2023).

[57] 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).