Neural Computing and Applications (2022) 34:7991–8004
https://doi.org/10.1007/s00521-022-06902-5

ORIGINAL ARTICLE

Fake visual content detection using two-stream convolutional neural networks

Bilal Yousaf[1] · Muhammad Usama[2] · Waqas Sultani[1] · Arif Mahmood[1] · Junaid Qadir[3,4]

Received: 3 February 2021 / Accepted: 4 January 2022 / Published online: 20 January 2022

Abstract

Rapid progress in adversarial learning has enabled the generation of realistic-looking fake visual content. To distinguish between fake and real visual content, several detection techniques have been proposed. The performance of most of these techniques however drops off significantly if the test and the training data are sampled from different distributions. This motivates efforts towards improving the generalization of fake detectors. Since current fake content generation techniques do not accurately model the frequency spectrum of the natural images, we observe that the frequency spectrum of the fake visual data contains discriminative characteristics that can be used to detect fake content. We also observe that the information captured in the frequency spectrum is different from that of the spatial domain. Using these insights, we propose to complement frequency and spatial domain features using a two-stream convolutional neural network architecture called TwoStreamNet. We demonstrate the improved generalization of the proposed two-stream network to several unseen generation architectures, datasets, and techniques. The proposed detector has demonstrated significant performance improvement compared to the current state-of-the-art fake content detectors with the fusing of frequency and spatial domain streams also improving the generalization of the detector.

Keywords Deepfakes · Two-stream network · Frequency stream · Combination of discrete Fourier transform and discrete wavelet

1 Introduction

Recent technological advancements in artificial intelligence (AI) have led to various beneficial applications in vision, language, and speech processing. However, at the same time, the power of these technologies may be exploited by adversaries for illegal or harmful uses. For example, Deepfakes—a portmanteau of the terms "deep learning" and "fake"—may be used to produce or alter photo-realistic audio-visual content with the help of deep learning for an illegal or harmful purpose. Deepfake technology enables one to effectively synthesize realistic-looking fake audio or video of a real person speaking and performing in any arbitrary way [1]. The term Deepfake was first coined by a Reddit community for synthetically replacing the face of a person with the face of another person. The term expanded with time to include similar techniques such as Lip-Sync [2, 3], facial expression reenactment [4–6], full-body and background manipulation as well as audio synthesis [7–12].

✉ Junaid Qadir
jqadir@qu.edu.qa

Bilal Yousaf
msds18007@itu.edu.pk

Muhammad Usama
muhammadusama@lums.edu.pk

Waqas Sultani
waqas.sultani@itu.edu.pk

Arif Mahmood
arif.mahmood@itu.edu.pk

[1] Department of Computer Science, Information Technology University (ITU), Lahore, Pakistan

[2] Lahore University of Management Sciences (LUMS), Lahore, Pakistan

[3] Department of Computer Science and Engineering (CSE), College of Engineering, Qatar University, Doha, Qatar

[4] Department of Electrical Engineering, Information Technology University (ITU), Lahore, Pakistan

The rise of technology such as Deepfake has eroded the traditional confidence in the authenticity of audio and video as any digital content (audio, video, text) can be easily subverted using advanced deep learning techniques for synthesizing images trained on readily accessible public videos and images [13–16]. The gravity and urgency of the Deepfake threat can be gauged by noting that in recent times a CEO was scammed using Deepfake audio for $243,000 [17] and a fake video of the president of Gabon has resulted in a failed coup attempt. Other potential effects of the Deepfake threat include danger to journalism and democratic norms because elections can be manipulated and democratic discourse may be disrupted by creating fake speeches of contending leaders [1, 3]. Unfortunately, most of the current research focuses on creating and improving Deepfakes, and there is a lack of focus on reliable Deepfake detection. For instance, among those papers uploaded to arXiv in 2018, 902 papers focused on Generative Adversarial Networks (GANs), a common method for Deepfake generation, while only 25 papers related to anti-forgery related topics [16].

Recent research shows that neural networks can be used for detecting fake content [18–22]. These methods however require a large amount of fake and real training data to accurately learn the data distributions of both classes. The performance of these methods drops significantly on the unseen fake data if sampled from a different distribution or a different generation process. It is because the underlying model may over-fit the available training data and therefore lose its ability to generalize to unseen data. To enable the model to classify previously unseen data will require a large amount of data from the new distribution which may not always be available in such problems. Attackers and defenders are continuously improving their approaches and rolling out new attacks and defenses. Therefore, it may be very difficult to collect a large amount of fake data for new manipulation techniques. Ideally, for such scenarios, a fake content detector is needed that should be able to detect fake data without explicit training on that particular type of fake content.

Nataraj et al. [23] proposed to improve the detectors for fake images by using hand-crafted co-occurrence matrices as input features. They can produce good results on only one unseen test set, however, their approach did not perform well on other types [24]. Zhang et al. [25] discovered that classifiers do not generalize well between GAN models and proposed to use the Discrete Fourier Transform (DFT) spectrum of full images as an input to the deep learning models to detect fake images. In contrast, we propose to calculate DFT on 8x8 blocks of the images and to combine these with Discrete Wavelet Transform (DWT) features. Furthermore, instead of using only the information from the frequency domain, we also propose to combine the artifacts from the spatial domain and show through extensive experiments that this technique generalizes well on many unseen test sets.

In the current work, we propose a two-stream network for fake visual content detection. The first stream called 'Spatial Stream' detects the fake data employing RGB images while the second stream dubbed as 'Frequency Stream' utilizes a combination of DFT and DWT for discriminating fake and real visual content. The frequency stream exploits the fact that the distribution of the frequency spectrum of the fake visual data remains distinct from the distribution of the real data frequency spectrum. This is illustrated in Fig. 1, which shows the DFT-magnitude spectrum for a sample of real and fake images. It can be seen that the frequency spectrum has patterns that are different from that of real images. These differences are used to classify the fake versus real content. To elaborate the frequency spectrum differences further we have shown the average spectra of the fake and real images from 11 different GAN generators. Following the method used by [25], we used all the images available in our test set to calculate the spectrum on the high passed filtered images and then took the average (Fig. 2). Since the information captured by the frequency stream is different from the information captured by the spatial stream, both these streams complement each other, and fusing them can provide better performance and generalization to unseen fake data detection. To the best of our knowledge, this is the first work that studies cross-modal information fusion to improve fake content detection generalization.

The main contributions of this paper are summarized next.

1 A novel two-stream architecture for fake visual content detection consisting of a Spatial Stream (SS) and a Frequency Stream (FS) is proposed. The SS learns the difference between the distributions of real and fake visual content in the spatial space using RGB images,

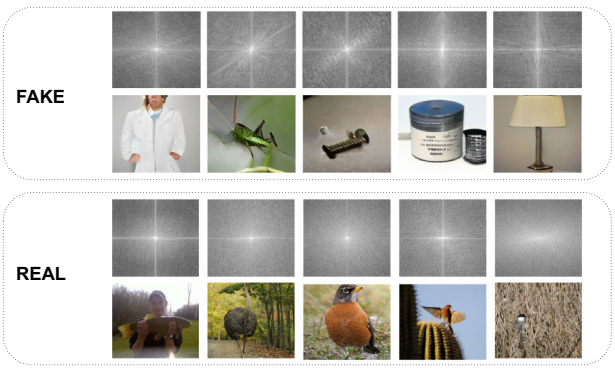

Fig. 1 DFT-magnitude spectrum for fake and real images has discriminative features which can be exploited for improved fake detection performance

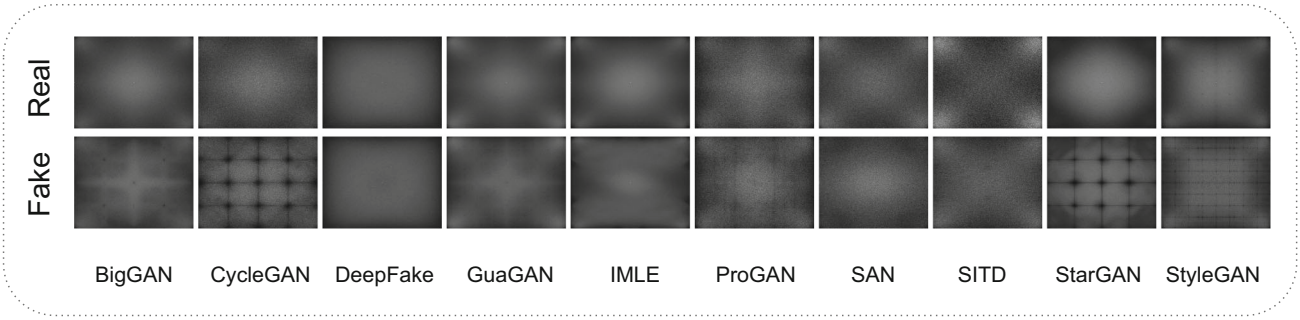

Fig. 2 We show that average spectrum (calculated on high-pass filtered image, similar to Zhang et al. [25]) for fake and real images have discriminative features which can be exploited for improved fake detection performance. Using the method of Zhang et al. [25], we first calculate the spectra of all the images in the test set and then take average of all

while the FS learns to discriminate between the distributions of real and fake content in the frequency domain. The coefficients of the stationary frequencies are captured using DFT, while the coefficients of spatially varying multi-scale frequencies are captured using Haar Wavelet transform. The spatial and frequency information complement each other and therefore their fusion improves fake visual content detection.

2 The proposed two-stream network comprising a frequency and a spatial domain stream has outperformed the state-of-the-art fake detection methods with a significant margin. A detailed analysis of the proposed approach is performed and we empirically demonstrate that the proposed approach is robust across different quality JPEG compression and blurriness artifacts.

In Sect. 2, we discuss the related work and cover the traditional image forensics techniques and the latest deep learning-based image forensics algorithms with a prime focus on generalization. In Sect. 3, we present our proposed methodology with pre-processing schemes, training, and testing procedures. In Sect. 4, we introduce the datasets used for evaluating and providing the results of our experiments. Section 5 critically evaluates the performance and the generalization of the proposed methodology by performing an ablation study. Finally, Sect. 6 concludes the paper and also points towards future directions.

2 Related work

In this section, we briefly review recent works needed to understand the state-of-the-art solutions in image forensics. We have divided this section into four subsections. We begin with a brief overview of the hand-crafted image forensic techniques followed by a discussion on deep-learning-based image forensic approaches. After that, we discuss methods that focus on improving generalization. Finally, we conclude the section by covering the state-of-the-art frequency-domain techniques that are specifically designed for image forensic applications.

2.1 Hand-crafted image forensics

A variety of methods are available in the literature for detecting traditional image manipulation techniques. Most of these manipulations are designed with the help of image editing tools. The traditional techniques make use of hand-crafted features to detect specific clues that are created as a result of different manipulations. For example, several blind noise estimation algorithms have been proposed to detect region splicing forgeries [26, 27]. Popescu et al. [28] detected the image forgeries by estimating the re-sampling in the images. Haodong et al. [29] integrated tampering possibility maps to improve forgery localization. Yuanfang et al. [30] identified potential artifacts in hue, saturation, dark and bright channels of fake colorized images, and developed detection methods based on histograms and feature encoding. Similarly, Peng et al. [31] used contact information of the standing objects and their supporting planes extracted from their reconstructed 3D poses to detect splicing forgeries. However, these techniques are unable to provide comparable performance to that of pixel-based methods in realistic situations. In recent works, learning-based techniques have become the preferred method compared to traditional image forensics for achieving state-of-the-art detection performance [32–35].

2.2 Deep learning based image forensics

Due to the success of deep learning in different fields, several researchers have recently leveraged deep learning approaches for fake visual content detection. Yan et al. [36] proposed an algorithm based on difference images (DIs) and illuminant map (IM) as feature extractors to detect re-colorized images. Quan et al. [37] designed a deep CNN network with two cascaded convolutional layers to detect

computer-generated images. McCloskey et al. [38] detected fake images by exploiting artifacts in the color cues, whereas Li et al. [39] used face warping artifacts for the forgery detection. Li et al. [40] noticed that eye blinking in fake videos is different from the natural videos and used this fact to expose the fake videos. Similarly, Yang et al. [41] have detected Deepfakes by identifying the inconsistent head poses. Recently, Afchar et al. [22] proposed two compact forgery detection networks (Meso-4 and MesoInception-4) in which forgery detection is done by analyzing the mesoscopic properties[1]. of Deepfake videos. Similarly, Nataraj et al. [23] have shown that features extracted from the co-occurrence matrix can help improve fake data detection. Wang et al. [42] proposed an anomaly detector-based approach that uses pre-trained face detectors as a feature extractor. Yang et al. [43] proposed the use of saliency maps to distinguish between real and fake images. Guo et al. [44] proposed a procedure for identifying fake face images by exploiting the GAN-generated artifacts in the Iris of the eye. Most of the aforementioned fake image detection techniques fail to distinguish between real and fake images if the visual data is sampled from a different distribution.

2.3 Methods focused on generalization

In this subsection, we briefly describe the fake detection approaches focused on generalization. Cozzolino et al. [45] proposed an autoencoder-based method to improve the performance of the model where learned weights are transferred for a different generation method. Zhang et al. [25] proposed a generalizable architecture named AutoGAN and evaluated its generalization ability on two types of generative networks. Xuan et al. [46] proposed that by using Gaussian Blur or Gaussian noise, one can destroy unstable low-level noise cues and force models to learn more intrinsic features to improve the generalization ability of the model. Similarly, Wang et al. [24] suggested that careful pre-and post-processing with data augmentation (such as blur and JPEG compression) improves the generalization ability. They have also shown improved fake detection results on multiple test sets by training on just one image generation network.

2.4 Frequency domain methods

Gueguen et al. [47] extracted features from the frequency domain to perform classification tasks on images. Ehrlich et al. [48] proposed an algorithm to convert the convolutional neural network (CNN) models from the spatial domain to the frequency domain. Xu et al. [49] proposed learning in the frequency domain and have shown that the performance of object detection and segmentation tasks gets improved in the frequency domain as compared to using the spatial RGB domain. Durall et al. [50] have shown that fake images have a difference in high-frequency coefficients compared to the natural images which he used for fake detection. Wang et al. [24] have shown that the artifacts in the frequency spectrum of fake images can be detected. Zhang et al. [25] proposed that if instead of raw pixels, frequency spectrum (2D-DCT on all 3 channels) is used as an input to the fake image detector, the performance of the detector improves. These frequency response base detectors target specific properties of the image generation process therefore, their performance degrades when fake images from unseen distributions are tested. In contrast to these existing methods, the proposed algorithm fuses information from the spatial domain and the frequency domain to achieve improved generalization. Also, we propose to fuse DFT with Wavelet Transform to improve the discrimination in the frequency domain. These innovations have resulted in significant improvement in fake content detection compared to the existing methods.

3 Methodology

Improving the generalizability of a fake detection model is critical for its success in real-world applications where the fake content may be generated by unknown processes. We propose a generalizable fake detection model based on a two-stream convolutional network architecture shown in Fig. 3.

The proposed architecture is motivated by the excellent performance of two-stream networks in action recognition in videos. To the best of our knowledge, the proposed network performs quite well on both seen and unseen data and has outperformed existing state-of-the-art (SOTA) methods in a wide range of experiments as we shall discuss in later sections. Our proposed two-stream network is novel and such a combination of frequency stream and the spatial stream has not been proposed before. In the following, we discuss the RGB to YCbCr conversion, DFT, DWT, and the proposed architecture in more detail.

3.1 The RGB to YCbCr transformation

The three channels in RGB color space are correlated with each other. We consider an orthogonal color space for improved representation. In our experiments, we have used YCbCr that has performed better than RGB space. As recommended in previous research [51, 52], the following

[1] The eyes and mouth are determined as the mesoscopic features in the forgery detection in the Deepfake videos.

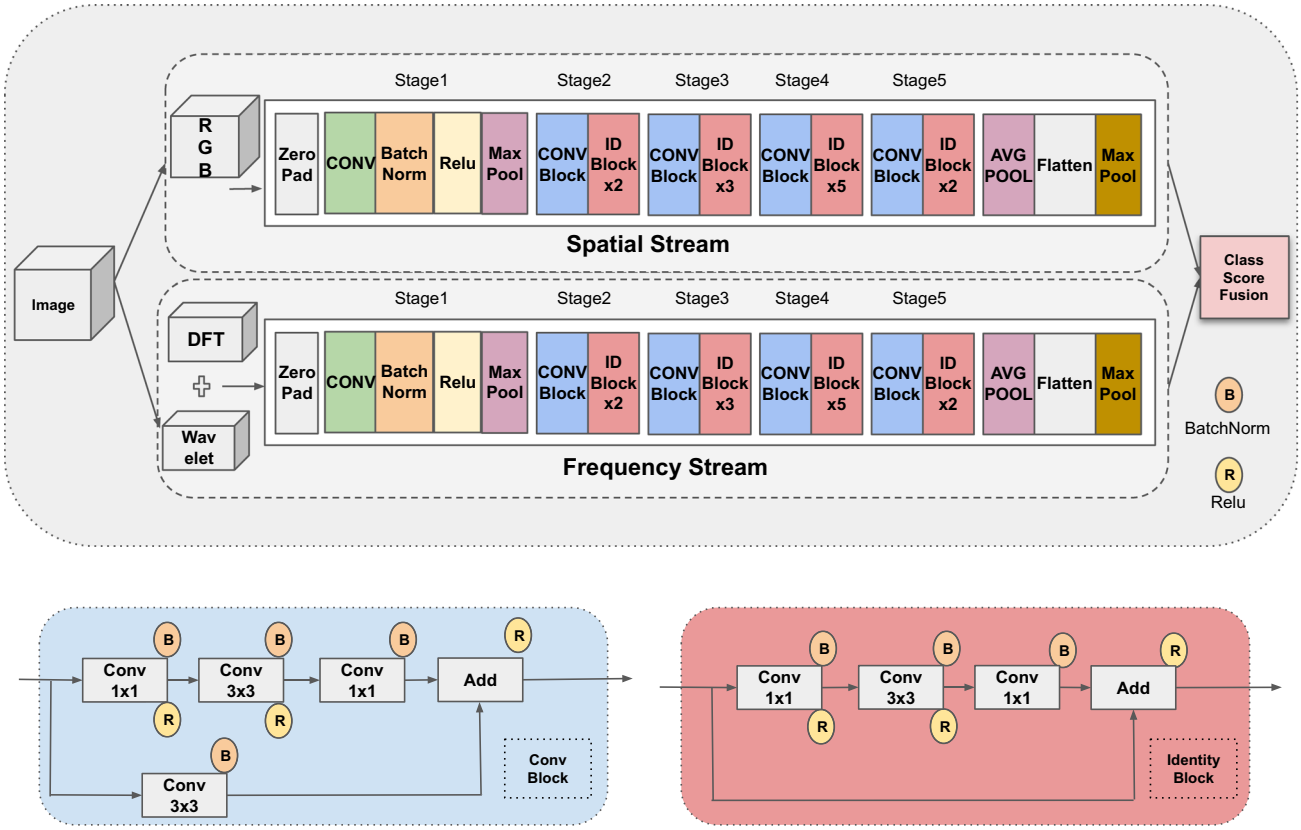

Fig. 3 Proposed two-stream convolutional neural network (TwoStreamNet). The two network streams capture spatial and frequency domain artifacts separately, and their outputs are fused at the end of the network to produce classification scores

formulas are used to convert from RGB to YCbCr color space:

$$Y = K_{\text{ry}}.R + K_{\text{gy}}.G + K_{\text{by}}.B, \text{ Cr} = B - Y, \text{ Cb} = \text{R} - \text{Y } K_{\text{ry}} + K_{\text{gy}} + K_{\text{by}} = 1, \quad (1)$$

where, $K_{\text{ry}}, K_{\text{gy}}$, and K_{by} are the coefficients for color conversion whose values are specified in Table 1 according to the standards. In our implementation, we used ITU601 [53].

3.2 A review of frequency domain transforms

To fully capture the frequency information from a YCbCr image, we compute DFT and DWT for each image.*Discrete Fourier Transform (DFT)*: Using DFT, one can decompose a signal into sinusoidal components of various frequencies ranging from 0 to maximum value possible based on the spatial resolution. For two dimensional data, i.e., images of size $W \times H$, the DFT can be computed using the following formula:

$$X_{w,h} = \sum_{n=0}^{W-1} \sum_{m=0}^{H-1} x_{w,h} e^{\frac{-i2\pi}{N}wn} e^{\frac{-j2\pi}{M}hm}, \quad (2)$$

where w is the horizontal spatial frequency, h is the vertical spatial frequency, $x_{w,h}$ is the pixel value at coordinates (w, h), and $X_{w,h}$ carries the magnitude and phase information of frequency at coordinates (w, h). *Discrete Wavelet Transform (DWT)*: Wavelet transform decomposes an image into four different sub-band images. High and low pass filters are applied at each row (column), and then they are down-sampled by 2 to get the high and low-frequency components of each row (column) separately. In this way, the original image is converted into four sub-band images: High-high (HH), High-low (HL), Low-high (LH), and Low-low (LL). Each sub-band image preserves different features: HH region preserves high-frequency components in both horizontal and vertical direction; HL preserves high-frequency components in the horizontal direction and low-frequency components in the vertical direction; LH preserves low-frequency components in the vertical direction and high-frequency components in the horizontal direction; and finally, LL preserves low-frequency components in the vertical direction and low-frequency components in the horizontal direction.

Table 1 Coefficients K_{ry} and K_{by} of color conversion from RGB to YCbCr

Reference standard	K_{ry}	K_{by}
[53] ITU601 / ITU-T 709 1250/50/2:1	0.299	0.114
[54] ITU709 / ITU-T 709 1250/60/2:1	0.2126	0.0722
[55] SMPTE 240M (1999)	0.212	0.087

3.3 Frequency stream

In this stream, two different types of the frequency spectrum are fused to get improved frequency domain representation which can better discriminate between the real and the fake visual content. An overview of the frequency spectrum fusion is shown in Fig. 4.

The three YCbCr channels are then transformed to the frequency domain using two different types of transformations, including DFT and DWT. Each channel is divided into a non-overlapping block of size 8×8 pixels and a transformation is applied on each block independently. The resulting coefficients are then concatenated back to obtain the arrays of the original image size. The output of the DFT converts one input channel into two output channels corresponding to real and imaginary coefficients. Similarly, the DWT converts one input channel into 4 output channels corresponding to low frequencies (LL), high and low frequencies (HL), high frequencies (HH), and low and high frequencies (LH). For three input channels (YCbCr), we obtain 18 output channels, 6 from DFT and 12 from DWT. All of these frequency output channels are concatenated to form 3D cubes of size $H \times W \times C$, where H is the height and W is the width of the image, and C=18 are the number of channels. We empirically observe that both DFT and DWT are necessary to capture essential information in the frequency domain at varying scales for improving the generalization ability of the proposed network.

3.4 Spatial stream

In this stream, RGB channels of the image are passed as input to the ResNet50 [56] as the classifier. RGB images are augmented in a special way using JPEG compression and Gaussian Blur as recommended by Wang et al. [24]. This stream is trained individually and plugged in the TwoStreamNet at the test time.

3.5 Two stream network architecture

The proposed two-stream network architecture is shown in Fig. 3. ResNet50 network is used as a backbone in both of the streams of the proposed architecture. Since the number of input channels in the frequency stream is larger as compared to the spatial stream, therefore the first layer of FS is accordingly modified. Both streams are independently trained, and the output of both streams is fused using the class probability averaging fusion method. In this fusion scheme, both streams contribute equally to the output to produce the final classification probability. The

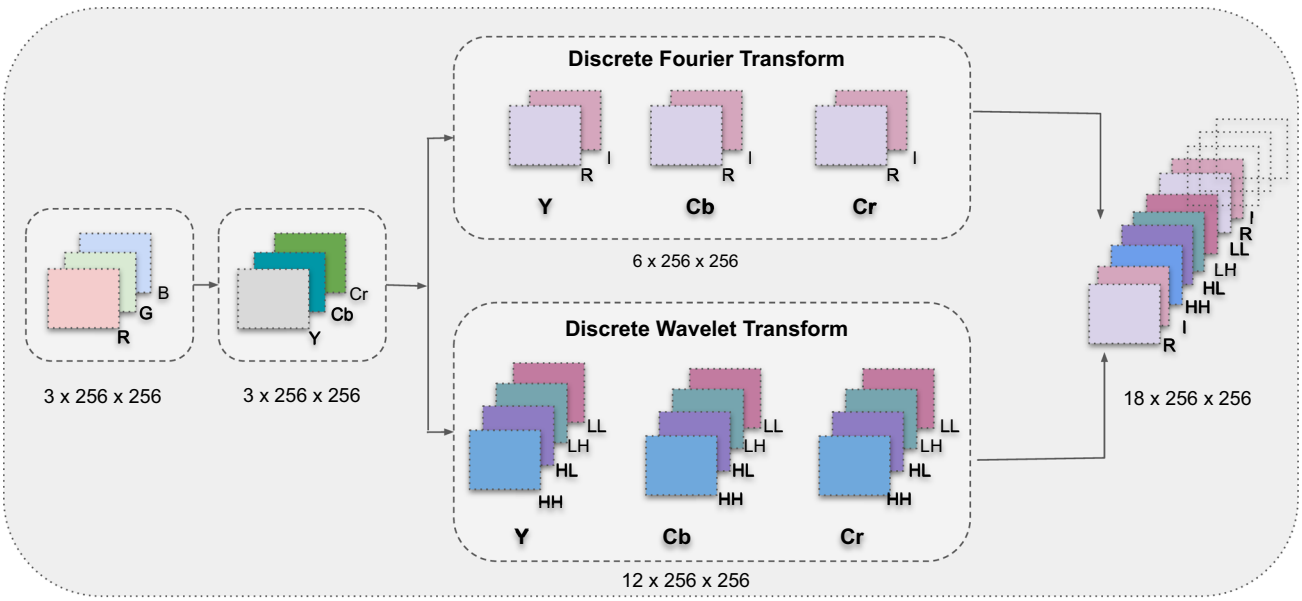

Fig. 4 Proposed pre-processing pipeline: the input image is first converted to YCbCr color space and then transformed to the frequency domain by applying DFT and Wavelet Transforms (DWT). After DFT, we get real (R) and imaginary (I) channels, and after WT we get four channels: HH, HL, LH, and LL. The resulting channels are concatenated to form 3D cubes which are then provided as input to the frequency stream for further processing

performance of the combined scores is significantly better than the performance of the individual streams.

4 Experiments and results

Training Dataset Following the protocol used by [24], the proposed two-stream network is trained using the fake images generated by ProGAN [57] and tested on the images generated by many other GANs. ProGAN has 20 different officially released models trained on different object categories of the LSUN dataset, which is a large-scale image dataset containing around one million labeled images for each of the 10 scene categories and 20 object categories[2] [58]. We choose 15 (airplane, bird, boat, bottle, bus, car, cat, chair, dog, horse, motorbike, person, sofa, train, and tv monitor) out of 20 models to create our validation and training set. We generated 10k fake images for training and 500 fake images for validation using each of the 15 models. For each of these 15 categories of fake images, we collect 10k of real images for training and 500 for validation randomly from the LSUN dataset [58]. In total, we have 300K training images and 15K validation images. For real images, we center crop the images equal to the size of the shorter edge and then resize the images to 256×256.

Testing Dataset Testing dataset images were generated using completely unseen generators as described in Table 4. To remain consistent with the current state of the art, the same generators are selected as that of [24]. The real images for testing purposes are obtained from the repository for each generator.

4.1 Implementation details

For training the FS, we use the Adam [71] optimizer with an initial learning rate of 0.0001, weight decay of 0.0005, and a batch size of 24. For all the training sets, we train the proposed network for 24 epochs. Large training data has helped the model to converge quickly. Lastly, we select the best model based on the validation set. While training each stream, data augmentation based on Gaussian blur and JPEG compression with 10% probability is used.

4.2 Evaluation metrics

We have used following metrics in our evaluation:

F1-Score F1-Score is the harmonic mean of precision and recall and is calculated below as:

$$F_1 = 2 \times \frac{\text{Precision} \times \text{Recall}}{\text{Precision} + \text{Recall}},$$

where Precision is the number of true positives divided by the summation of true positives and false positives and the recall is the number of true positive results divided by the number of all samples that should have been identified as positive.

Accuracy Accuracy is defined as the ratio of the correct predictions over the total number of predictions made. And is calculated as below:

$$Accuracy = \frac{TP + TN}{TP + TN + FP + FN}, \quad (4)$$

where TP, FP, TN and FN represents is the number of true positives, false positives, true negatives and false negative respectively.

4.3 Comparison with the existing state-of-the-art algorithms

We thoroughly evaluated the performance of the proposed method on the test dataset and compared it with the existing state-of-the-art [24]. We also compared the robustness analysis of our approach against some common real-world perturbations. In Table 2, we have shown a comparison of our results with the best results of Wang et al. [24] ([Blur+JPEG(0.1)]). Their results from their official web link[3]. Results show that our FS approach performs very well on the unseen manipulations and outperforms the state-of-the-art on several test sets while having competitive performance on the remaining. Results of the two-stream architecture demonstrate that our complete approach outperformed the state-of-the-art in almost all of the cases. Analysis of the results shows that when both spatial and frequency streams are combined into a two-stream architecture, they complement each other in a way that their combined accuracy is greater than any of them individually. This clearly shows that FS ConvNet has learned distinctive features that were not learned by SS ConvNet. Overall, the combination of FS and SS plays a vital role in improving the generalization ability of the fake image detectors.

In Table 3 we compare our method to four different models [Cyc-Im, Cyc-Spec, Auto-Im, Auto-Spec] of Zhang et al. [25]. They released four models with two kinds of variations, first they used two datasets generated using two different GAN architectures CycleGAN and AutoGAN referenced as (Cyc and Auto) in Table 3, and as a second variation, they passed images as input to one model and frequency spectrum in the other model referenced as (Im and Spec) in Table 3. Results of our approach show that our

[2] https://www.yf.io/p/lsun.

[3] https://peterwang512.github.io/CNNDetection/.

Table 2 Comparison of the proposed Frequency Stream (FS) and two-stream network with the state-of-the-art method [24] using average accuracy. Best results of Wang et al. [24] with data augmentation using blur and JPEG (0.1) are reported where 0.1 mean JPEG compression is applied on 10% images. The same augmentation is also used in the proposed approaches. Both our approach and that of Wang et al. are trained using ProGAN only and tested on the data generated by 12 unseen generation processes mentioned in the top row

Metrics	Method	StarGAN	StyleGAN	SITD	BigGAN	StyleGAN2	CycleGAN	Whichfaceisreal	SAN	Deepfake	GuaGAN	CRN	IMLE
Accuracy	Wang et al. [24] [Spatial Stream]	91.7	87.1	90.3	70.2	84.4	**85.2**	83.6	**53.5**	50.5	78.9	**86.3**	**86.2**
	Frequency stream (Ours)	**96.67**	**89.36**	81.11	**72.08**	**91.83**	77.53	80.15	49.32	*68.90*	70.91	55.06	55.06
	Two stream (Ours)	**96.32**	**88.90**	**97.22**	**72.85**	**87.43**	84.09	**87.50**	50.23	**55.00**	**79.64**	77.75	77.75
F1-Score (Fake)	Wang et al. [24] [Spatial Stream]	91.31	85.19	89.91	61.13	81.53	84.2	81.92	3.56	12.83	75.45	**87.93**	**87.88**
	Frequency stream (Ours)	**97.63**	**88.21**	83.57	**71.71**	**91.14**	78.75	82.08	**28.39**	**61.13**	72.65	68.99	68.99
	Two stream (Ours)	**96.21**	**87.52**	**97.27**	**66.61**	**85.63**	**84.98**	**87.15**	4.39	**17.95**	**77.34**	81.79	81.79
F1-Score (Real)	Wang et al. [24] [Spatial Stream]	92.14	88.56	90.62	75.81	86.50	86.08	85.0	**66.67**	68.28	81.5	**84.13**	**84.08**
	Frequency stream (Ours)	**97.71**	**90.32**	77.78	72.43	**94.42**	76.97	77.76	60.78	**74.08**	68.93	18.41	18.41
	Two stream (Ours)	**96.43**	**90.01**	**97.18**	**77.13**	**88.84**	**86.20**	**87.83**	66.36	**69.00**	**81.5**	71.39	71.39

Table 3 Comparison of the proposed two-stream network with Zhang et al. [25] using accuracy. We compare the proposed network with 4 models released by [25], each one was trained using one of two image sources CycleGAN (Cyc) and AutoGAN (Auto), as well as one of two image representations images (Im) and spectrum (Spec). The blue text shows the same training and testing data

Method	Star GAN	Style GAN	SITD	Big GAN	Style GAN2	Cycle GAN	Which face is real	SAN	Deep fake	Gua GAN	CRN	IMLE
Zhang et al. [25] [Cyc-Im]	84.14	50.42	53.3	50.67	49.63	99.73	50	57.3	50.02	53.56	54.7	59.25
Zhang et al. [25] [Cyc-Spec]	**100**	50	50	50.375	51.84	100	50	50	50.08	50.46	50	49.98
Zhang et al. [25] [Auto-Im]	99.79	49.61	50.83	47.75	49.97	**98.64**	49.95	**62.55**	**60.53**	49.26	72.21	62.44
Zhang et al. [25] [Auto-Spec]	99.97	49.97	50	51.58	54.31	98.59	50	50.23	50.08	50.79	53.79	53.98
Two stream (Ours)	96.32	**88.90**	**97.22**	**72.85**	**87.43**	84.09	**87.50**	50.23	55.00	**79.64**	**77.75**	**77.75**

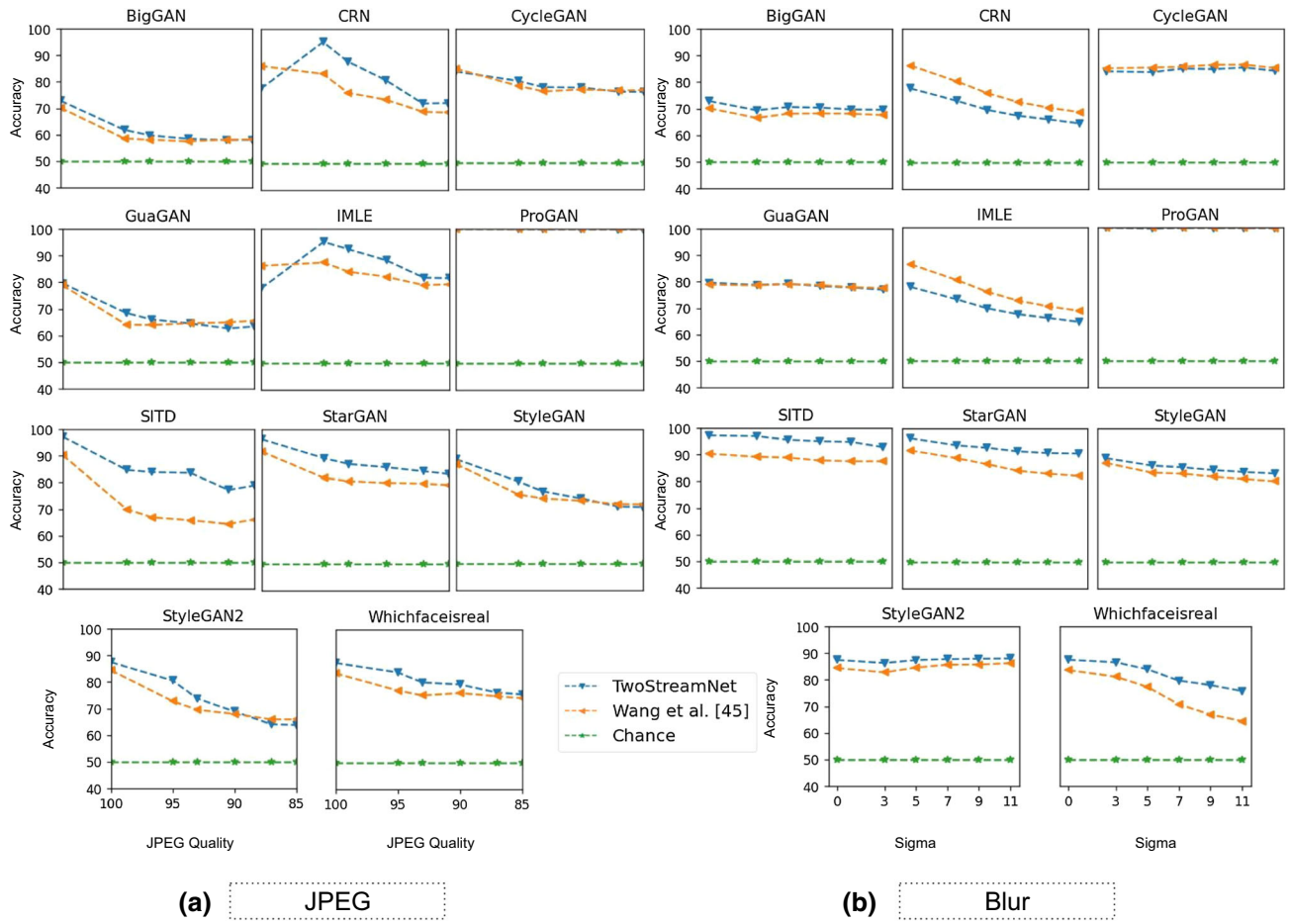

Fig. 5 Robustness comparison of the proposed algorithm with Wang et al. [24] for Gaussian Blur and JPEG Compression artifacts. In most experiments, the proposed two-stream net. We apply Gaussian Blur and JPEG Compression of different sizes on the test sets and measure the effect on the accuracy of our model. Our model performs near to the best for all the cross-modal datasets even when a large blurring effect is applied. Results show that our proposed solution is more robust as compared to the state of the art

two-stream architecture outperformed all models of Zhang et al. 3 in almost all of the test sets (Fig. 5).

In Fig. 6A, we have shown samples of the fake images which are misclassified by the state-of-the-art and are correctly classified by our proposed two-stream approach. These results demonstrate the ability of the proposed approach to detect high-quality fake images which are even very hard to discriminate by humans. Figure 6B shows the fake images which are misclassified by both Wang et al. [24] and us.

Robustness Analysis In real-world settings, fake images may undergo several post-processing operations like compression, smoothness, etc. Therefore, we have evaluated the performance of the proposed model on the images which are post-processed using JPEG compression and Gaussian blur. Specifically, we apply Gaussian blur with different standard deviations including [3, 5, 7, 9, 11] and JPEG compression with JPEG image quality factor of [85, 87, 90, 92, 95]. Results in Fig. 5 show that our approach is robust to common perturbations. For most of the models, the proposed approach significantly outperformed the state-of-the-art method at varying blur levels. Similarly, the proposed approach also performed better than the state-of-the-art methods for a wide range of JPEG compression.

5 Ablation study

In this section, we thoroughly validate the different components of the proposed approach by performing an ablation study (Table 4).

5.1 Combining DFT and DWT

As shown in Fig. 3, we propose to combine DWT and DFT for better feature representation and robust fake content detection. To verify the effectiveness of using both transformations, while keeping all the experimental settings the same, we experimented with DFT and DWT separately. After training for 20 epochs, the best epoch is chosen based


BigGAN
GuaGAN
SAN
CRN
SITD
IMLE
StarGAN
StyleGAN
CycleGAN
WhichFaceIsReal
StyleGAN2
Deepfake
(a)
(b)

◀**Fig. 6** **A** Examples of fake images correctly detected by the proposed two-stream network, however, misclassified by Wang et al. [24]; **B** Examples of fake image misclassified by both our proposed method and that of Wang et al. [24]

Table 4 Details of the testing dataset

Dataset	Real images	Fake images
StarGAN [59]	1999	1999
StyleGAN [60]	5991	5991
SITD [61]	180	180
BigGAN [62]	2000	2000
StyleGAN2 [63]	7988	7988
CycleGAN [64]	1321	1321
Whichfaceisreal [65]	1000	1000
GauGAN [66]	5000	5000
Deepfake [67]	2698	2707
CRN [68]	6382	6382
IMLE [69]	6382	6382
SAN [70]	219	219

on validation data accuracy. The results shown in Table 5 demonstrate that a combination of DFT and DWT is essential to produce robust feature representation for fake image detection.

5.2 Effect of block size

We study the effect of using different block sizes instead of computing DFT over the whole image. In Table 6, we have shown results of computing DFT on the block size of 8×8, 16×16, 32×32, and 256×256 (Full-Image size). Note that the block size experiments are performed while keeping identical experimental settings. Results demonstrate that 8×8 block size has consistently outperformed other block sizes. Therefore, transforming the image to the frequency domain using 8×8 blocks for DFT is more effective for fake image detection.

Table 5 Evaluation of DFT and DWT combination for fake image detection. Percentage accuracy is reported for the full image using only DFT, only DWT, and the combination DFT + DWT

Dataset	DFT	DWT	DFT + DWT
StarGAN [59]	78.81%	79.34%	**97.67%**
StyleGAN [60]	61.45%	69.11%	**87.1%**
SITD [61]	83.61%	49.72%	**90.3%**
BigGAN [62]	64.92%	66.92%	**72.08%**
StyleGAN2 [63]	68.02%	59.54%	**91.83%**
CycleGAN [64]	65.31%	55.94%	**77.53%**
Whichfaceisreal [65]	39.85%	77.40%	**80.15%**
GauGAN [66]	50.91%	67.22%	**70.91%**
Deepfake [67]	56.39%	51.88%	**68.90%**
CRN [68]	61.75%	**83.85%**	55.06%
IMLE [69]	47.82%	**79.90%**	55.06%
SAN [70]	**69.18%**	46.80%	49.32%

Table 6 Fake image detection accuracy variation by varying block sizes for DFT transform. The block size 8×8 has produced best results and is therefore used in our experiments

Dataset	8 x 8	16 x 16	32 x 32	'Full-image'
StarGAN [59]	**94**.62%	77.74%	51.05%	78.81%
StyleGAN [60]	**90**.65%	68.93%	46.47%	61.45%
SITD [61]	**86**.1%	55.28%	75.56%	83.61%
BigGAN [62]	**69**.42%	63.12%	48.25%	64.92%
StyleGAN2 [63]	**92**.23%	65.26%	53.26%	68.02%
CycleGAN [64]	**79**.96%	67.97%	36.68%	65.31%
Whichfaceisreal [65]	**81**.50%	63.25%	41.95%	39.85%
GauGAN [66]	**67**.11%	54.74%	57.95%	50.91%
Deepfake [67]	**60**.61%	51.90%	57.35%	56.39%
CRN [68]	51.43%	50.44%	57.90%	**61**.75%
IMLE [69]	51.50%	50.53%	**66**.28%	47.82%
SAN [70]	46.58%	51.83%	**72**.15%	69.18%

5.3 Effect of color-space

We evaluate the effectiveness of converting images into YCbCr color space before frequency transformations. We performed two experiments using the same settings to compare the performance of RGB with YCbCr color space. Results in Table 7 show that converting an image to YCbCr colorspace adds more discriminative features in the frequency domain and helps in better fake image detection.

5.4 Limitations

The computation of DFT and DWT is computationally expensive. Therefore to implement it for fake content detection in real-time applications using video data may be a potential limitation. However, this limitation can be overcome by using parallel computation of DFT and DWT. Figure 6B shows failure cases of the proposed algorithm which are the result of non-discriminative frequency domain features. Please note that these fake images are high quality and very hard to discriminate, even for humans.

Table 7 The compassion of fake detection performance using RGB and YCbCr Colorspace. YCbCr color space has performed better than RGB color space

Dataset	RGB	YCbCr
StarGAN [59]	66.83%	**78.81%**
StyleGAN [60]	59.32%	**61.45%**
SITD [61]	**87.50%**	83.61%
BigGAN [62]	**72.97%**	64.92%
StyleGAN2 [63]	60.42%	**68.02%**
CycleGAN [64]	**75.89%**	65.31%
Whichfaceisreal [65]	47.80%	39.85%
GauGAN [66]	**60.92%**	50.91%
Deepfake [67]	56.02%	**56.32%**
CRN [68]	58.83%	**61.75%**
IMLE [69]	48.24%	47.82%
SAN [70]	66.89%	**69.18%**

6 Conclusions

This paper addresses the problem of fake image detection. For this purpose, a two-stream network is proposed consisting of a spatial stream and a frequency stream. The proposed method generalizes to unseen fake image generator distributions much better than the current state-of-the-art approaches. The proposed method is also found to be more robust to the common image perturbations including blur and JPEG compression artifacts. The improved performance is leveraged by combining two types of frequency domain transformations, namely, Discrete Fourier Transform (DFT) and Discrete Wavelet Transform (DWT). Both transformations are applied upon YCbCr color-space, and different frequency domain channels are concatenated to discriminate fake images from the real ones. By exploiting the differences between the real and the fake image frequency responses, improved fake detection performance is achieved. In the future, we aim to extend this work for fake video and audio detection.

Funding The authors did not receive support from any organization for the submitted work.

Declarations

Conflict of interest We wish to confirm that there are no known conflicts of interest associated with this publication.

Ethical approval We confirm that the manuscript has been read and approved by all named authors and that there are no other persons who satisfied the criteria for authorship but are not listed. We further confirm that the order of authors listed in the manuscript has been approved by all of us. We understand that the Corresponding Author is the sole contact for the Editorial process (including Editorial Manager and direct communications with the office).

References

1. Chesney B, Citron D (2019) Deep fakes: a looming challenge for privacy, democracy, and national security. Calif L Rev 107:1753
2. Kumar R, Sotelo J, Kumar K, de Brébisson A, Bengio Y (2017) Obamanet: photo-realistic lip-sync from text. arXiv preprint arXiv:1801.01442
3. Suwajanakorn S, Seitz SM, Kemelmacher-Shlizerman I (2017) Synthesizing Obama: learning lip sync from audio. ACM Trans Graph (TOG) 36(4):95
4. Thies J, Zollhofer M, Stamminger M, Theobalt C, Nießner M (2016) Face2face: real-time face capture and reenactment of rgb videos. In: Proceedings of the IEEE conference on computer vision and pattern recognition, pp 2387–2395
5. Thies J, Zollhöfer M, Nießner M, Valgaerts L, Stamminger M, Theobalt C (2015) Real-time expression transfer for facial reenactment. ACM Trans Graph 34(6):183–191
6. Wiles O, Koepke A, Zisserman A (2018) X2face: a network for controlling face generation using images, audio, and pose codes. In: Proceedings of the European conference on computer vision (ECCV), pp 670–686
7. Chan C, Ginosar S, Zhou T, Efros AA (2019) Everybody dance now. In: Proceedings of the IEEE international conference on computer vision, pp 5933–5942
8. Cai H, Bai C, Tai YW, Tang CK (2018) Deep video generation, prediction and completion of human action sequences. In: Proceedings of the European conference on computer vision (ECCV), pp 366–382
9. Esser P, Haux J, Milbich T et al (2018) Towards learning a realistic rendering of human behavior. In: Proceedings of the European conference on computer vision (ECCV)
10. Wang Y, Skerry-Ryan RJ, Stanton D, Wu Y, Weiss RJ, Jaitly N, Yang Z, Xiao Y, Chen Z, Bengio S et al (2017)Tacotron: towards end-to-end speech synthesis. arXiv preprint arXiv:1703.10135
11. Arik SO, Chen J, Peng K, Ping W, Zhou Y (2018) Neural voice cloning with a few samples. In: Advances in neural information processing systems, pp 10019–10029
12. Tachibana H, Uenoyama K, Aihara S (2018) Efficiently trainable text-to-speech system based on deep convolutional networks with guided attention. In: 2018 IEEE international conference on acoustics, speech and signal processing (ICASSP, pp 4784–4788. IEEE
13. From porn to 'game of thrones': how deepfakes and realistic-looking fake videos hit it big. https://www.businessinsider.com/deepfakes-explained-the-rise-of-fake-realistic-videos-online-2019-6. Accessed 07 Dec 2020
14. Lee D (2018)Fake porn' has serious consequences. https://www.bbc.com/news/technology-42912529. Accessed 07 Dec 2020
15. Cole S (2018) Gfycat's AI solution for fighting deepfakes isn't working. https://www.vice.com/en_us/article/ywe4qw/gfycat-spotting-deepfakes-fake-ai-porn. Accessed 07 Dec 2020
16. Patrini G, Cavalli F, Henry A (2018) The state of deepfakes: reality under attack, annual report v.2.3. https://deeptracelabs.com/archive/. Accessed 07 Dec 2020
17. Damiani J (2019) A voice deepfake was used to scam a CEO out of 243,000\$. https://www.forbes.com/sites/jessedamiani/2019/09/03/a-voice-deepfake-was-used-toscam-a-ceo-out-of-243000/. Accessed 07 Dec 2020
18. Bayar B, Stamm MC (2016)A deep learning approach to universal image manipulation detection using a new convolutional

layer. In: Proceedings of the 4th ACM workshop on information hiding and multimedia security, pp 5–10
19. Cozzolino D, Poggi G, Verdoliva L (2017) Recasting residual-based local descriptors as convolutional neural networks: an application to image forgery detection. In: Proceedings of the 5th ACM workshop on information hiding and multimedia security, pp 159–164
20. Rahmouni N, Nozick V, Yamagishi J, Echizen I (2017) Distinguishing computer graphics from natural images using convolution neural networks. In: 2017 IEEE workshop on information forensics and security (WIFS), pp 1–6. IEEE
21. Rössler A, Cozzolino D, Verdoliva L, Riess C, Thies J, Nießner M (2018) FaceForensics: a large-scale video dataset for forgery detection in human faces. arXiv preprint arXiv:1803.09179
22. Afchar D, Nozick V, Yamagishi J, Echizen I (2018) MesoNet: a compact facial video forgery detection network. In: 2018 IEEE international workshop on information forensics and security (WIFS), pp 1–7. IEEE
23. Nataraj L, Mohammed TM, Manjunath BS, Chandrasekaran S, Flenner A, Bappy JH, Roy Chowdhury AK (2019) generated fake images using co-occurrence matrices. Electron Imaging 2019(5):532–541
24. Wang SY, Wang O, Zhang R, Owens A, Efros AA (2020) Cnn-generated images are surprisingly easy to spot... for now. In: Proceedings of the IEEE conference on computer vision and pattern recognition, vol 7
25. Zhang X, Karaman S, Chang SF (2019) Detecting and simulating artifacts in gan fake images. In: 2019 IEEE international workshop on information forensics and security (WIFS), pp 1–6
26. Agarwal S, Farid H (2017) Photo forensics from jpeg dimples. In: 2017 IEEE workshop on information forensics and security (WIFS), pp 1–6. IEEE
27. Lyu S, Pan X, Zhang X (2014) Exposing region splicing forgeries with blind local noise estimation. Int J Comput Vis 110(2):202–221
28. Popescu AC, Farid H (2005) Exposing digital forgeries by detecting traces of resampling. IEEE Trans Signal Process 53(2):758–767
29. Li H, Luo W, Qiu X, Huang J (2017) Image forgery localization via integrating tampering possibility maps. IEEE Trans Inf Forensics Secur 12(5):1240–1252
30. Guo Y, Cao X, Zhang W, Wang R (2018) Fake colorized image detection. IEEE Trans Inf Forensics Secur 13(8):1932–1944
31. Peng B, Wang W, Dong J, Tan T (2018) Image forensics based on planar contact constraints of 3d objects. IEEE Trans Inf Forensics Secur 13(2):377–392
32. Zhou B, Khosla A, Lapedriza A, Oliva A, Torralba A (2016) Learning deep features for discriminative localization. In: Proceedings of the IEEE conference on computer vision and pattern recognition, pp 2921–2929
33. Huh M, Liu A, Owens A, Efros AA. (2018) Fighting fake news: image splice detection via learned self-consistency. In: Proceedings of the European conference on computer vision (ECCV), pp 101–117
34. Cozzolino D, Poggi G, Verdoliva L (2015) Splicebuster: a new blind image splicing detector. In: 2015 IEEE international workshop on information forensics and security (WIFS), pp 1–6. IEEE
35. Yuan Rao , Jiangqun Ni (2016) A deep learning approach to detection of splicing and copy-move forgeries in images. In: 2016 IEEE international workshop on information forensics and security (WIFS), pp 1–6. IEEE
36. Yan Y, Ren W, Cao X (2019) Recolored image detection via a deep discriminative model. IEEE Trans Inf Forensics Secur 14(1):5–17
37. Quan W, Wang K, Yan D, Zhang X (2018) Distinguishing between natural and computer-generated images using convolutional neural networks. IEEE Trans Inf Forensics Secur 13(11):2772–2787
38. McCloskey S, Albright M (2018) Detecting gan-generated imagery using color cues. arXiv preprint arXiv:1812.08247
39. Li Y, Lyu S (2018) Exposing deepfake videos by detecting face warping artifacts. arXiv preprint arXiv:1811.00656
40. Li Y, Chang MC, Lyu S (2018) In ictu oculi: exposing AI created fake videos by detecting eye blinking. In: 2018 IEEE international workshop on information forensics and security (WIFS), pp 1–7. IEEE
41. Yang X, Li Y, Lyu S (2019) Exposing deep fakes using inconsistent head poses. In: ICASSP 2019-2019 IEEE international conference on acoustics, speech and signal processing (ICASSP), pp 8261–8265. IEEE
42. Wang R, Juefei-Xu F, Ma L, Xie X, Huang Y, Wang J, Liu Y (2019) Fakespotter: a simple baseline for spotting ai-synthesized fake faces. arXiv preprint arXiv:1909.06122
43. Yang J, Xiao S, Li A, Lan G, Wang H (2021) Detecting fake images by identifying potential texture difference. Future Gener Comput Syst 125:127–135
44. Guo H, Hu S, Wang X, Chang MC, Lyu S (2021) Robust attentive deep neural network for exposing gan-generated faces. arXiv preprint arXiv:2109.02167
45. Cozzolino D, Thies J, Rössler A, Riess C, Nießner M, Verdoliva L (2018) Forensictransfer: weakly-supervised domain adaptation for forgery detection. arXiv preprint arXiv:1812.02510
46. Xuan X, Peng B, Wang W, Dong J (2019) On the generalization of GAN image forensics. In: Chinese conference on biometric recognition. Springer, pp 134–141
47. Gueguen L, Sergeev A, Kadlec B, Liu R, Yosinski J (2018) Faster neural networks straight from JPEG. In: Advances in neural information processing systems, pp 3933–3944
48. Ehrlich M, Davis LS (2019) Deep residual learning in the jpeg transform domain. In: Proceedings of the IEEE international conference on computer vision, pp 3484–3493
49. Xu K, Qin M, Sun F, Wang Y, Chen YK, Ren F (2020) Learning in the frequency domain. arXiv preprint arXiv:2002.12416
50. Durall R, Keuper M, Pfreundt FJ, Keuper J (2019) Unmasking deepfakes with simple features. arXiv preprint arXiv:1911.00686
51. Recommendation ITU-R (2011) Studio encoding parameters of digital television for standard 4: 3 and wide-screen 16: 9 aspect ratios
52. Radiocommunication ITU (2002) Parameter values for the HDTV standards for production and international programme exchange. Recommendation ITU-R BT, pp 709–5
53. Recommendation ITU-R BT.601-5 (1982-1995)
54. Recommendation ITU-R BT.709-5 (1990-2002)
55. Society of Motion Picture and Television Engineers SMPTE 240M-1999 "Television-Signal Parameters-1125-Line High-Definition Production". http://www.smpte.org
56. He K, Zhang X, Ren S, Sun J (2016) Deep residual learning for image recognition. In: Proceedings of the IEEE conference on computer vision and pattern recognition, pp 770–778
57. Karras T, Aila T, Laine S, Lehtinen J (2017) Progressive growing of GANs for improved quality, stability, and variation. arXiv preprint arXiv:1710.10196
58. Yu F, Seff A, Zhang Y, Song S, Funkhouser T, Xiao J (2015) LSUN: construction of a large-scale image dataset using deep learning with humans in the loop. arXiv preprint arXiv:1506.03365
59. Choi Y, Choi M, Kim M, Ha JW, Kim S, Choo J (2018) Stargan: unified generative adversarial networks for multi-domain image-to-image translation. In: Proceedings of the IEEE conference on computer vision and pattern recognition, pp 8789–8797

60. Karras T, Laine S, Aila T (2019) A style-based generator architecture for generative adversarial networks. In: Proceedings of the IEEE conference on computer vision and pattern recognition, pp 4401–4410
61. Chen C, Chen Q, Xu J, Koltun V (2018) Learning to see in the dark. In: Proceedings of the IEEE conference on computer vision and pattern recognition, pp 3291–3300
62. Brock A, Donahue J, Simonyan K (2018) Large scale GAN training for high fidelity natural image synthesis. In: International conference on learning representations
63. Karras T, Laine S, Aittala M, Hellsten J, Lehtinen J, Aila T (2019) Analyzing and improving the image quality of StyleGAN. arXiv preprint arXiv:1912.04958
64. Zhu JY, Park T, Isola P, Efros AA (2017) Unpaired image-to-image translation using cycle-consistent adversarial networks. In: Proceedings of the IEEE international conference on computer vision, pp 2223–2232
65. Which face is real? https://www.whichfaceisreal.com/. Accessed 07 Dec 2020
66. Park T, Liu MY, Wang TC, Zhu JY (2019) Semantic image synthesis with spatially-adaptive normalization. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition (CVPR)
67. Rossler A, Cozzolino D, Verdoliva L, Riess C, Thies J, Nießner M.(2019) FaceForensics++: learning to detect manipulated facial images. In: Proceedings of the IEEE international conference on computer vision, pp 1–11
68. Chen Q, Koltun V (2017) Photographic image synthesis with cascaded refinement networks. In: Proceedings of the IEEE international conference on computer vision, pp 1511–1520
69. Li K, Zhang T, Malik J (2019) Diverse image synthesis from semantic layouts via conditional IMLE. In: Proceedings of the IEEE international conference on computer vision, pp 4220–4229
70. Dai T, Cai J, Zhang Y, Xia ST, Zhang L (2019) Second-order attention network for single image super-resolution. In: Proceedings of the IEEE conference on computer vision and pattern recognition, pp 11065–11074
71. Kingma DP, Ba J (2014) Adam: a method for stochastic optimization. arXiv preprint arXiv:1412.6980