Pattern Recognition Letters 179 (2024) 185–191

Contents lists available at ScienceDirect

Pattern Recognition Letters

journal homepage: www.elsevier.com/locate/patrec

Less is more: A minimalist approach to robust GAN-generated face detection

Tanusree Ghosh *, Ruchira Naskar

Department of Information Technology, Indian Institute of Engineering Science and Technology, Shibpur, 711103, India

ARTICLE INFO

Editor: Maria De Marsico

Keywords:
Fake image detection
Deepfake detection
GAN forensics
Digital image forensics
Synthetic image detection
GAN-face detection

ABSTRACT

Hyper-realistic images that are not differentiable from authentic images to regular viewers have become extremely easy to generate and highly accessible. Furthermore, the increasing pervasiveness of social media networks in our daily lives has facilitated the easy dissemination of fake news accompanied by such synthetic images. Hyper-realistic artificial face images are often illicitly used as profile pictures on social media sites, further using such profiles to spread fabricated information, resulting in social perils. Most available synthetic image detectors are challenging to implement in practical scenarios due to their high complexity and performance degradation for images from Online Social Networks (OSNs). In this work, we develop a deep learning-based lightweight synthetic image detector called Relative Chrominance Distance Network (*RCD-Net*). In this paper, we introduce the RCD image feature set for the first time, which gives a pair-wise chrominance component-based distance measure. To show its effectiveness, we explore multiple luminance-chrominance spaces. Compared to the state-of-the-art (SOTA), our model hugely reduces the network parameter requirements, making it incredibly lightweight. We also study the robustness of the proposed solution against common post-processing operations in the context of online social media networks. Experimental results prove that the proposed solution achieves SOTA performance at a much lower complexity than available solutions.

1. Introduction

Continuous breakthroughs in Generative Artificial Intelligence technologies have enabled the hassle-free generation of hyper-realistic images. Even in the last decade, synthetic multimedia creation used to be considered a highly skilled job. Naturally, access to the creation of such 'Fake' content was limited to the masses. In 2014, the invention of the Generative Adversarial Network (GAN) changed the whole scenario of synthetic multimedia creation [1]. In consecutive years, Generative AI has matured so much that synthetic multimedia can deceive human judgments of realism [2]. Variations of GAN models (e.g. StyleGAN, StyleGAN2 etc.) can produce hyper-realistic face images from random noise input [3–5].

Even though such technologies have been a boon to the entertainment industry for creating innocuous content, illicit uses of the same may create social turbulence. Moreover, the ever-increasing spread of the internet and social media in every stage of society has made disseminating fake news containing synthetic media accessible like never before. With freely accessible websites,[1] anyone can create synthetic multimedia. Widespread use of hyper-realistic artificial face images in profile pictures for fake profiles are being used to spread misinformation and political, economical propaganda.[2] A recent study has shown that the performance of regular viewers to identify fake images is poor (≈50%), which makes them vulnerable to being influenced by fake images online [6]. Hence, the paramount importance of developing accurate fake image detectors has emerged. Eventually, focus of digital forensics' community has shifted towards synthetic image detection from traditional camera source identification and forgery detection [7, 8].

The well-explainable approach to detect fake faces is physiological cues like facial asymmetries, colour inconsistencies in both iris etc. [9]. Shape and boundary inconsistencies in pupils have also been explored [10]. Such cues have some intrinsic advantages over others. Regular observers can use such artefacts to detect synthetic images with proper training and guidance. However, with the advancements in image generation, such artefacts are disappearing quickly, making the need for apparently 'invisible' cues. In this direction, two approaches are notable: 'Deep-Features' based detectors and 'Handcrafted Features' based detectors. Do et al. [11] uses a transfer-learning model VGG-Net with pre-trained weights from VGG-Face.

Modified 'Xception' [12] was proposed later [13]. Chen et al. [14] designed separate global and local feature extraction modules using CNN. However, detectors solely based on Deep-Learning models may

* Corresponding author.
E-mail addresses: 2021itp001.tanusree@students.iiests.ac.in (T. Ghosh), ruchira@it.iiests.ac.in (R. Naskar).

[1] https://thispersondoesnotexist.com/

[2] https://www.npr.org/2022/12/15/1143114122/ai-generated-fake-faces-have-become-a-hallmark-of-online-influence-operations

https://doi.org/10.1016/j.patrec.2024.02.017
Received 4 July 2023; Received in revised form 21 December 2023; Accepted 19 February 2024
Available online 22 February 2024
0167-8655/© 2024 Elsevier B.V. All rights reserved.

not work if such detector models are used as discriminators while generating GAN images. Deep learning models having the ability to extract features automatically fail to explain their detection performances.

From a different perspective, Frank et al. [15], Gragnaniello et al. [16] explored frequency domain artefacts. Wang et al. [17] explored anti-forensics for GAN-generated image detection.

On the other hand, handcrafted features-based detectors are mostly designed with custom features based on domain knowledge. Facial landmarks like eyes, lips, etc., have been used as features with Support Vector Machine as the classifier [18]. However, these handcrafted features heavily rely on the quality of generated images, so they are likely ineffective for newer-generation synthetic images due to their indistinguishability from real images.

Statistical properties of images have shown their strong ability to discriminate synthetic images due to the difference in image generation flow between camera-generated and GAN-generated images. Particularly, statistical differences in the colour domain are well-explored area [19–26]. Hence statistical properties in colour domain coupled with Deep Neural Network based classifiers has emerged as an effective solution to detect synthetic images with close to perfect detection performance. As our solution lies in this domain, notable related works will be discussed in a later section with details.

However, we believe to combat the spreading of disinformation through OSNs, fake image detectors should satisfy these requirements:

- While uploading or downloading images or videos to/from OSNs, they undergo OSN-specific processing that induces some artefacts. Such artefacts can deceive detectors. As the exact post-processing steps involved in the post-processing of any OSN are unknown, it is a common practice to test detectors' performance with common post-processing operations like blurring, resizing, etc., that mimic operations of OSN on images. To effectively function in real-world applications, a reliable fake image detector should not only yield accurate detection performance on uncompressed raw images but also maintain its robustness when dealing with compressed and post-processed images, thereby ensuring consistent performance across various challenging image conditions.
- For the integration of fake image detectors into everyday use by general consumers, it is crucial to transition these technologies from a controlled laboratory setting to a more accessible format. Therefore, prioritizing the development of compact, lightweight and efficient models over the traditionally resource-intensive and complex Deep Neural Networks (DNNs) is a more pragmatic approach, ensuring broader usability and adaptability in various real-world contexts.

Although few previous research has addressed challenges associated with Online Social Networks (OSNs) and devised robust methodologies effective for post-processed images, the primary objective of these studies was to enhance detection accuracy. Consequently, the development of solutions optimized for both computational efficiency (in terms of space and time requirements) and practical applicability has not been thoroughly investigated. Typically, the performance of Deep Neural Network (DNN)-based classifiers improves with an increase in the number of parameters and larger feature representations, leading to a scenario where high-performance fake image detection models become encumbered with excessive parameters and prolonged operational times. Therefore, a critical need exists for achieving an optimal balance in practical synthetic image detectors: ensuring high detection accuracy for both uncompressed and compressed, post-processed images while simultaneously maintaining a lightweight architecture.

In this direction, our main contributions, presented in this work, are as follows:

- We propose four novel feature sets by carefully analysing their statistical relations in the chrominance and luminance domain. We propose the chrominance-based feature Relative Chrominance Distance (RCD), the minimal size feature that performs satisfactorily on synthetic image detection alone. Further, we integrate luminance-based features with them for performance improvement while slightly compromising on feature size. We introduce a luminance-based feature: High-Frequency-Residual (HFR), that exploits luminance components in the frequency domain. We further test the performance of all feature sets on five different colour domains.
- We propose a lightweight Deep-Learning based classifier that uses a very less number of parameters compared to state-of-the-art classifiers.
- Further, we study the effect of various post-processing operations (e.g. blurring, sharpening, JPEG compression, etc.) on the detection result of our model.

Fig. 1. (a)–(d) Camera-Generated Images from FFHQ dataset. (e)–(h) Synthetic Images from STYLEGAN2 dataset.

The rest of the paper is organized as follows: Section 2 briefly covers existing works in colour space-based fake image detection. Section 3 discusses our proposed methodology in detail. Section 4 covers experimental results and analysis. In Section 5, we show the results of the ablation study. Finally, Section 6 concludes the paper.

2. Related works

Real and synthetic image generation processes are significantly different. Camera-generated images go through a unique Colour Filter Array (CFA): Bayer's filter, to generate RGB channels in captured images. It captures a single colour (Red, Green, or Blue) per pixel and interpolates values (Demosaicing process) from neighbouring pixels to generate three channels per pixel. In contrast, GAN architecture takes a random latent vector as input and employs some convolution and deconvolution steps to expand the size of the random vector. Ultimately, the last layer collapses multi-channel feature maps into a three-channel colour image.

McCloskey and Albright [27] first explored colour clues for detecting GAN images. They analysed the inner workings of GAN and showed that the frequency of over-exposed pixels differentiates real and GAN-generated images. It is due to the 'Normalization' operation in GAN's generator that limits the generated intensities.

Inspired by the applications in the steganalysis domain, Nataraj et al. [19] use three co-occurrence matrices for three colour channels (R, G and B) as input features to a deep learning-based classifier. Barni et al. [20], Nowroozi and Mekdad [21] extend this method, adding three cross-channel co-occurrence matrices for three combinations: Red–Green, Green–Blue and Blue–Red channels. Even though they perform better than [19], their feature space is multiplied three times. He et al. [24] explored YCbCr, HSV and LAB colour spaces. They train multiple shallow CNNs with chrominance residuals, and the resulting deep representations of features from multi-channels were finally fed into a Random Forest (RF) classifier. On a similar track, Li et al. [23] extract features from chrominance components of HSV and YCbCr colour spaces. They calculate co-occurrence matrices on chrominance residuals to represent feature space. Finally, concatenated co-occurrence vectors are utilized with a deep-learning-based classifier.

Chen et al. [25] train a two-stream neural network utilizing chrominance and luminance information from RGB and YCbCr colour spaces. They also proposed a convolutional block attention module and a multi-layer feature aggregation module into the classic Xception model [12]

Buy me a coffee https://ko-fi.com/jurnalin https://dbook.eu.org

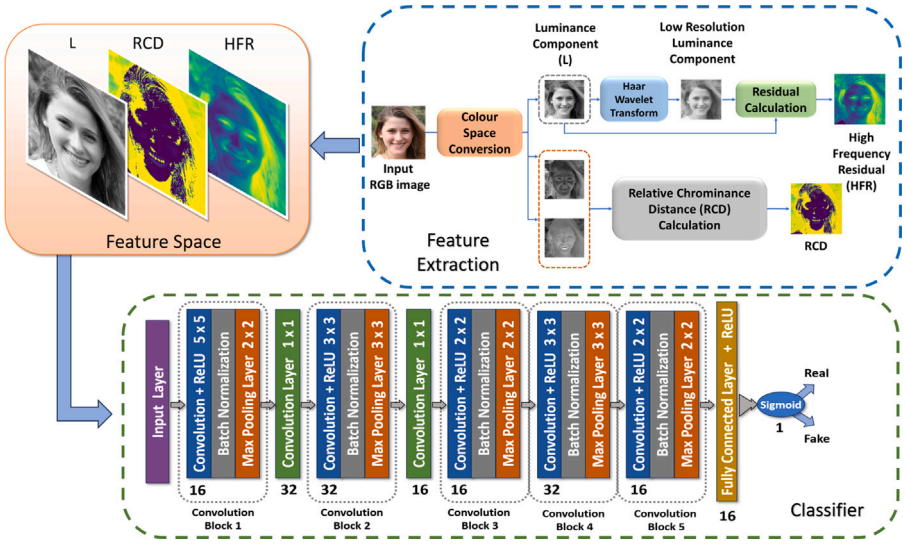


Fig. 2. Proposed workflow.

Table 1
Chi-Square distance of multiple colour channels.

Dataset	R	G	B	L	A	B	RCD_{LAB}	Y	U	V	RCD_{YUV}
FFHQ & STYLEGAN2	116.00	112.12	103.03	111.45	138.36	83.96	**454.65**	110.33	89.02	77.92	**226.50**

for better feature handling. Recently, Qiao et al. [22] employed ten layered Cross-colour Spatial Co-occurrence matrices consisting of selected channels from RGB, HSV and YCbCr.

In this work, we simultaneously explore chrominance and luminance-based features. We also incorporate frequency-domain features for better stability and robustness along with a DNN classifier. Being substantially lightweight compared to earlier works, our methodology performs at par or better.

3. Proposed methodology

In this section, we illustrate the proposed framework for detecting GAN-generated images. The pipeline of the framework is shown in Fig. 2.

Our detection framework consists of two significant steps: Data preprocessing and binary classification. We convert RGB images into five colour spaces: LAB, YCbCr, HSL, LUV, and YUV. As discussed earlier, We use four features for each colour space. Finally, we use extracted features in a well-designed DNN to classify pristine and synthesized images.

3.1. Colour spaces in context of synthetic images

Colour images have two components: chrominance, which conveys colour information (hue and saturation), and luminance, which conveys brightness information. Synthetic images are usually represented in RGB colour spaces at the time of generation. Hence, they tend to mimic the properties of RGB colour space to look realistic. Biologically, human eyes are more sensitive to luminance change than chrominance change [28]. Therefore, visually undetectable artefacts may lie in chrominance components. Colour spaces like LAB, LUV, HLS, YCbCr, YUV have separate luminance and chrominance channels and they are easily convertible from RGB space by linear or non-linear equations [29].

3.2. Relative chrominance difference

Earlier works by He et al. [24], Li et al. [23], Chen et al. [25], Qiao et al. [22] considering colour spaces for synthetic image detection utilize only chrominance components as features for synthetic image detection. He et al. [24] consider six chrominance channels from different colour spaces as features and further processed them with high pass filter. On the other hand, Li et al. [23] utilize four channels for further processing in the residual domain. While both methods work well, we aim to represent chrominance information in a much smaller space. We propose **Relative Chrominance Distance** (RCD), calculated as the pixel-wise distance between chrominance components for any colour space.

For an image **I**, in Colour space 'X', its luminance channel is $\mathbf{I}_l$, and two chrominance channels are $\mathbf{I}_{c_1}$ and $\mathbf{I}_{c_2}$. Then the $\mathbf{RCD}_X(\mathbf{I})$ can be formulated as:

$$RCD_X(I) = |I_{c_1} - I_{c_2}| \tag{1}$$

For example, RCD for an image I in the LAB colour domain will be calculated as:

$$RCD_{LAB}(I) = |I_A - I_B| \tag{2}$$

To prove the effectiveness of our RCD feature, we select 500 images randomly from the FFHQ dataset to represent real images and 500 images randomly from the STYLEGAN2 dataset for synthetic images. Then, we extract colour channels (R, G, B, L, A, B, Y, U, V) for each image separately. For each channel, we calculate pixel-wise mean values. Then, we calculate the histogram of the resulting mean. We calculate the Chi-Square distance between obtained histogram distributions for both real and synthetic datasets.

It is considered that the more the Chi-Square distance, the better the separability between datasets. As shown in Table 1, We calculate the Chi-Square distance between both real and fake datasets for RGB (R, G, B), LAB (L, A, B), YUV (Y, U, V) colour spaces, RCD feature $\mathbf{RCD_{LAB}}$ on chrominance of LAB space, RCD feature $\mathbf{RCD_{YUV}}$ on chrominance of YUV space. From the obtained results, it is evident that the top two maximum distances are obtained by $\mathbf{RCD_{LAB}}$ and $\mathbf{RCD_{YUV}}$. Both RCD features have better performance than their respective chrominance and luminance channels. All distributions are visualized in Fig. 3.

3.3. Luminance

The post-processing and JPEG compression degrade the performance of fake image detectors. However, DNN-based classifiers utilizing perceptually uniform colour space YCbCr, can perform better than

Buy me a coffee https://ko-fi.com/jurnalin https://dbook.eu.org

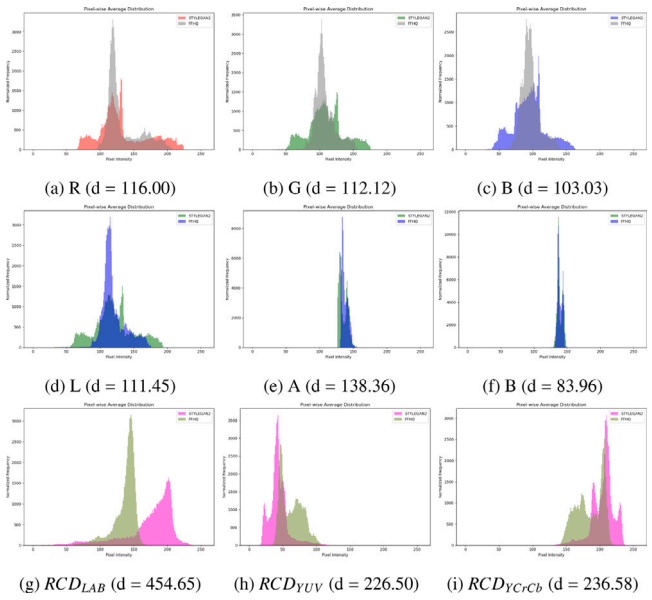


Fig. 3. Average Histogram distribution for Real and Fake images in different colour channels: First row for R3(a),G3(b),B3(c) channels. Second row for L3(d), A3(e), B3(f) channels. Third row for RCD features of LAB3(g), YUV3(h) and YCrCb3(i) colour spaces. Chi-square distances of respective cases are given.

RGB space [25]. For certain post-processing operations like Gaussian noise, chrominance components consist of better features than luminance, whereas for JPEG images, the luminance feature offers better detection performance.

However, as we aim to combat fake images on social media, where most images are distributed in JPEG format, along with chrominance, we consider luminance (L) as one of the components in the feature set.

3.4. High-frequency residual of luminance

Due to the intrinsic operation of upsampling or deconvolution during GAN image synthesis, generated images acquire frequency artefacts. Frequency residual artefact-based fake video detection has recently been explored [30].

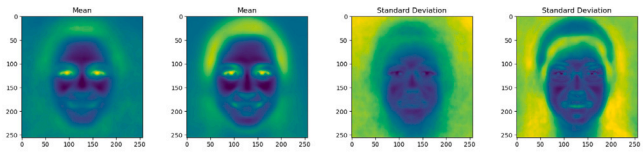
Motivated by this study, we propose a high-frequency residual on the luminance channel as a feature. We utilize *Discrete Wavelet Transform (DWT)* for this purpose. DWT is well-known in computer vision applications for its ability to decouple multi-level frequencies. In this work, we use Haar Wavelet Transform, one of the oldest and simplest DWT methods, to calculate high-frequency residuals on the luminance channel. Given an image, DWT transform decouples frequency components recursively at each level. Given any image, I, at level-1, four frequency components are calculated, Three high-frequency components: HH, HL, LH and one low-frequency component: LL. Filters for 2 × 2 Haar-transformations are: $f_{LL} = \frac{1}{2}[\begin{smallmatrix}1&1\\1&1\end{smallmatrix}]$, $f_{LH} = \frac{1}{2}[\begin{smallmatrix}1&1\\-1&-1\end{smallmatrix}]$, $f_{HL} = \frac{1}{2}[\begin{smallmatrix}1&-1\\1&-1\end{smallmatrix}]$, and $f_{HH} = \frac{1}{2}[\begin{smallmatrix}1&-1\\-1&1\end{smallmatrix}]$ The LL, LH, HL, and HH components are calculated as follows: $LL = f_{LL} \times I, LH = f_{LH} \times I, HL = f_{HL} \times I, HH = f_{HH} \times I$.

For any given image $\mathbf{I}$, we first extract its luminance channel: $\mathbf{I_L}$. Then, we apply the DWT-Haar transform as discussed above and obtain $I_{L_{LL}}$, $I_{L_{LH}}$, $I_{L_{HL}}$ and $I_{L_{HH}}$.

To obtain our proposed feature High-Frequency Residual (HFR), we first use Bilinear Interpolation on $I_{L_{LL}}$ to make it the same shape as $\mathbf{I_L}$. Then we obtain HFR for $\mathbf{I}$ as follows:

$$\mathrm{HFR}_X(I) = \mathrm{I}_L - \mathrm{I}_{L_{LL}} \tag{3}$$

To support our claim of the effectiveness of the HFR feature for discriminating Real and Synthetic images, we calculate the HFR matrix on the luminance channel; L (from LAB space), for 500 randomly chosen STYLEGAN2 images and 500 randomly chosen real images from the FFHQ dataset. It is pertinent to mention that considered STYLEGAN2 images are generated from the FFHQ dataset. Hence, images from both datasets look much similar, as shown in Fig. 1. We calculate pixel-wise mean and standard deviation for each dataset. As shown in Fig. 4, the mean and standard deviation values of both types of images differ significantly.



(a) HFR mean for Real Images (b) HFR mean for Fake Images (c) HFR SD for Real Images (d) HFR SD for Fake Images

Fig. 4. Mean and standard deviation of L channel HFR for Real and Fake images.

3.5. Selection of features

In earlier sections, we have discussed three features, namely: Relative Chrominance Distance (RCD), Luminance channel (L) and High-Frequency Residual (HFR). As discussed earlier, we believe RCD feature that has a feature space of shape $n \times n$ for raw RGB image of shape $n \times n \times 3$ (*'n' represent no. of pixels*) can alone discriminate real and synthetic images. To the best of our knowledge, RCD is the minimal feature yet on similar solutions utilizing *Colour domain based features + DNN based classifier*. While RCD is designed to work in extreme cases where optimum solution size is the only focus, we analyse three more feature set combinations to enhance RCD's performance. Table 2 shows details of all feature set sizes (Set1: RCD, Set2: RCD+L, Set3: RCD+HFR, Set4: RCD+L+HFR). In our experiments, all the above-mentioned features are used in normalized form.

3.6. Deep-learning based classifier

In this section, we describe our proposed DNN-based classifier. Our network comprises five regular Convolution Blocks containing *Convolutional layer*, *Batch Normalization* and *Max Pool layer* with different filter sizes and numbers. We employ one special convolution block, having filter size 1 × 1 twice. This particular filter is used for feature fusion across multiple layers. As we use features from multiple domains simultaneously, we incorporate this convolutional block for better feature learning. The input layer tensor size varies according to the choice of features, as shown in Table 2. Feature sets, except set 1, use feature tensor comprising concatenated features. The detailed architecture, including filter size and the number of filters for each layer, is shown in Fig. 2. At the last layer, we use one single neuron with 'sigmoid' activation function.

4. Experimental results

4.1. Dataset and classifier settings

We select 20,000 random pristine face images from FFHQ dataset [3] and 20,000 random synthetic images from StyleGAN2 dataset [4]. Primarily we divide these 40,000 images as follows: 28,000 for training, 8000 for validation and 4000 for testing. Here we consider image size 256 × 256 for all experiments. We use the 'Adam' optimizer and 'Binary Cross-Entropy' loss function for all experiments. We train our classifier for 40 epochs for each case. We use a Learning Rate(LR) scheduler for efficient convergence. For the first ten epochs, LR is set to 1×10^{-3}; for the subsequent ten epochs, LR is multiplied by 1×10^{-1}; while for the last 20 epochs, LR is multiplied by 1×10^{-3}.

Buy me a coffee https://ko-fi.com/jurnalin https://dbook.eu.org

Table 2
Detection performance of proposed models.

Colour domain	Method	Accuracy (%)	AUC score	Feature space
LAB	RCD	95.34	99.20	256 × 256 × 1
	RCD+L	98.39	99.81	256 × 256 × 2
	RCD+HFR	95.88	99.77	256 × 256 × 2
	RCD+L+HFR	**98.86**	99.91	256 × 256 × 3
LUV	RCD	95.00	98.37	256 × 256 × 1
	RCD+L	98.14	99.22	256 × 256 × 2
	RCD+HFR	98.21	99.79	256 × 256 × 2
	RCD+L+HFR	**98.59**	99.85	256 × 256 × 3
HLS	RCD	98.14	99.76	256 × 256 × 1
	RCD+L	97.68	99.45	256 × 256 × 2
	RCD+HFR	97.92	99.65	256 × 256 × 2
	RCD+L+HFR	**99.20**	99.99	256 × 256 × 3
YCbCr	RCD	96.20	99.19	256 × 256 × 1
	RCD+L	98.55	99.35	256 × 256 × 2
	RCD+HFR	97.58	99.80	256 × 256 × 2
	RCD+L+HFR	**98.59**	99.89	256 × 256 × 3
YUV	RCD	96.63	99.45	256 × 256 × 1
	RCD+L	98.07	99.53	256 × 256 × 2
	RCD+HFR	98.11	99.69	256 × 256 × 2
	RCD+L+HFR	**98.50**	99.99	256 × 256 × 3

4.2. Performance analysis

We use two metrics to evaluate the performance of our proposed solutions: Accuracy and average AUC score. As discussed in the earlier section, we evaluate the performance of our proposed feature sets with five colour spaces: LAB and LUV (Both belong to XYZ colour space representation family), HLS, YCrCb and YUV (Both have similar colour space representation). RCD is the basic feature with the least number of feature size (256 × 256 × 1), and RCD+HFR+L has the largest feature size (256 × 256 × 3). As shown in Table 2, all our experiments obtain detection accuracy greater than 95%. Though, for each colour space, the RCD feature alone gives the least accuracy (still not less than 95%), the fact is to be noted that RCD has the least feature space. Feature sets 2 and 3 have acceptable performances in all spaces. In all five colour spaces, feature spaces 2 and 3 obtain better performance than feature set 1 but worse than feature set 4, which indicates that Feature L and HFR add discriminating information to minimal feature RCD. As expected, Feature Set 4 (RCD+L+HFR) obtains the best result for each colour space.

However, to prove the effectiveness of our proposed feature set, we use Feature Set 4 with XceptionNet [12] and obtain 99% accuracy. The average performance of our best model (HLS+Feature 4) is compared with other SOTA solutions in Table 3. Among them, Qiao et al. [22], Nowroozi and Mekdad [21], Nataraj et al. [19], Chen et al. [25], Li et al. [23] use Colour domain-based features, whereas Frank et al. [15] use frequency domain-based feature. Number of parameters used by Nowroozi and Mekdad [21], Nataraj et al. [19] is ≈ 22 Million, Xception based models use more than 20 Million and Qiao et al. [22] use more than 1 Million parameters in their proposed CNN classifier. Our proposed solutions RCD-Net with minimal feature set RCD and RCD+L+HFR settings use ≈ $19k$ and ≈ $20k$ parameters respectively . It must be noted that even though our models have close detection performance to them, our model uses **0.001%** of parameters used in CNN models by Nowroozi and Mekdad [21], Nataraj et al. [19] for the DNN classifier. Our proposed RCD-Net with the minimal feature set (RCD) obtains the best performance of 98.14% accuracy on HLS space. However, Co-occurrence feature-based solution [19] use (256 × 256 × 3) feature space, Cross-co-occurrence feature-based solutions [20,21] use (256 × 256 × 6) features, Cross-Color Spatial Co-Occurrence feature based solution [22] use (256 × 256 × 10) feature space.

4.3. Robustness of detectors

To check the robustness of our solutions, we consider the worst-performing RCD feature and the best-performing (RCD+L+HFR) feature models from all five colour spaces. We perform common post-processing operations on all images of the test set and tested on our models along with SOTA models [19,20,22]. As shown in Table 4, we marked the best performance for each operation in bold and the second best performance underlined. Feature set 4 on YCrCb has the highest average accuracy. Even though feature set 1 models did not perform as well as feature set 4 models, their performance is still acceptable, with the best average accuracy of 94.19% on YUV colour space. However, for each feature set, average accuracy scores are so close that presenting them as a range is much more convenient. For Feature Set 1, the average accuracy lies in the range of 92.61 to 94.19, whereas Feature Set 4 has an average accuracy range of 96.17 to 97.62. Both the feature sets beat SOTAs' performance.

While uncompressed image formats like PNG or BMP are preferred for experimental purposes, those are rarely used in real-world scenarios. Here, we check our best and worst-performing models' performance on JPEG images and compare them with SOTA models. We converted our test set images into JPEG format with five quality factors: 90, 80, 70, 60, 50. Table 5 shows the detailed result. For JPEG compressed images, the Feature 4 model in YCbCr works best, outperforming SOTA models. Detectors' performance usually degrades for JPEG images because, during JPEG compression, image statistics are entirely changed

Table 3
Performance Comparison with SOTA. Metric: Accuracy(%).

Method	Co-Net [19]	Cross Co-Net [21]	CSC-Net [22]	Improved Xception [25]	DCT+Classifier [15]	Adjacency Matrix [31]	XceptionNet + Our Feature Set	RCD-Net (Feature RCD)	RCD-Net (Feature L+RCD+HFR)	Co-Occurrence Matrix + Our Classifier	RGB image + Our Classifier
Detection accuracy	96.13	99.30	99.95	97.70	98.58	99.05	99.00	98.14	99.20	68.50	98.75

Table 4
Testing on post-processing operations. Metric: Accuracy (%).

Operations	Parameters	RCD-Net (LAB)		RCD-Net (LUV)		RCD-Net (HLS)		RCD-Net (YCbCr)		RCD-Net (YUV)		CSC-Net	Cross-Co-Net	Co-Net
		RCD	RCD+L+HFR	RCD	RCD+L+HFR	RCD	RCD+L+HFR	RCD	RCD+L+HFR	RCD	RCD+L+HFR			
Median filter	3 x 3	93.10	97.50	93.53	98.22	95.75	98.44	94.17	97.52	96.08	96.88	**99.35**	85.13	81.48
	5 x 5	89.41	96.06	90.30	**96.65**	77.88	95.63	90.63	96.45	92.86	95.31	93.80	83.65	75.98
Gaussian noise	1.0	95.39	97.95	94.20	98.19	98.02	**98.66**	95.06	98.29	96.83	97.50	94.43	93.68	76.35
	2.0	95.61	98.04	94.37	98.21	97.90	**98.61**	95.59	98.14	96.85	97.82	74.25	96.80	76.73
CLAHE	–	95.36	97.52	90.88	**98.14**	90.61	95.66	84.70	96.16	82.94	90.55	94.70	50.32	51.43
Average blurring	3 x 3	93.60	97.77	93.59	**97.95**	92.70	97.79	93.58	97.57	95.46	96.50	97.30	86.90	93.68
	5 x 5	89.38	95.00	91.17	93.30	60.92	93.35	87.87	**97.59**	91.40	94.59	82.68	76.63	88.23
Gamma correction	0.8	95.70	97.78	92.99	97.82	97.79	**98.49**	95.01	98.26	96.26	97.59	95.08	83.15	82.28
	0.9	95.46	98.14	93.15	98.09	97.89	**98.61**	95.24	98.44	96.59	97.67	98.00	90.98	87.23
	1.2	95.44	98.09	92.58	98.12	97.80	**98.26**	95.29	98.09	96.78	97.25	96.90	85.53	87.20
Resizing	0.5	92.61	97.02	92.60	**97.78**	82.55	97.40	92.24	97.35	94.07	96.16	79.80	92.47	57.93
Average	–	93.73	97.35	92.61	97.51	89.98	97.35	92.67	**97.62**	94.19	96.17	91.48	84.11	78.04

Buy me a coffee https://ko-fi.com/jurnalin https://dbook.eu.org

Table 5
Robustness testing of models against JPEG compression. Metric: Accuracy(%).

Quality factor	RCD-Net(LAB)		RCD-Net(LUV)		RCD-Net(HLS)		RCD-Net(YCbCr)		RCD-Net(YUV)		CSC-Net	Cross Co-Net	Co-Net
	RCD	RCD+L+HFR	RCD	RCD+L+HFR	RCD	RCD+L+HFR	RCD	RCD+L+HFR	RCD	RCD+L+HFR			
90	79.40	89.83	85.30	97.89	91.84	97.37	90.43	**98.14**	92.59	96.99	98.13	95.58	95.90
80	68.80	82.91	73.49	97.97	87.57	96.63	88.24	**97.99**	89.35	96.58	97.40	94.93	88.65
70	62.08	76.69	67.63	97.84	84.15	96.08	86.09	**97.89**	85.81	96.50	97.23	95.00	83.68
60	58.18	71.99	62.00	97.67	80.13	95.24	84.40	**97.82**	83.06	96.65	96.83	95.65	94.23
50	56.87	68.58	59.42	97.37	76.22	94.57	81.67	**97.84**	80.08	96.92	96.90	96.20	79.13

Table 6
Performance on Cross-Dataset.

Dataset		Accuracy (%)
FFHQ vs. StyleGAN2		98.59
Class-wise accuracy	FFHQ	98.60
	StyleGAN2	98.59
CelebA vs. StyleGAN2		99.11
Class-wise accuracy	CelebA	99.95
	StyleGAN2	98.13

Table 7
Detection accuracy (%) of component features.

	RCD	L	HFR	L+HFR	RCD+L+HFR
Accuracy (%)	96.20	98.59	97.87	97.80	98.59

by DCT and chroma subsampling operations. However, it converts images to the YCbCr domain as a part of the process. Probably, due to this, features learned on YCbCr domain still work better than other colour spaces.

Even though the Feature Set 4 model in YCbCr space was not the best in synthetic image detection compared to other methods, their robustness to post-processed and compressed image detection is impressive. Moreover, their performance degradation in such complex scenarios is lesser than SOTA models, which proves their stability. Hence, we consider Feature Set 4 model on YCbCr domain as our best model. In general, all models with Feature Set 4 has less performance degradation while testing in complex scenario compared to other SOTA models. A reason for this is the fusion of features from multiple domains (Frequency, Luminance, Chrominance).

4.4. Test on cross-domain dataset

We further evaluate the performance of the proposed model for a test dataset consisting of images from cross-domains. In our investigation, we include images from StyleGAN2 [4], FFHQ [3] and CelebA [32] datasets for test, while using solely FFHQ-StyleGAN2 dataset in training the model. Specifically, we adopt 2000 real images from CelebA and 2000 synthetic images from StyleGAN2. Our solution achieves an average detection accuracy of **99.11%** in the cross-domain test. This marks a considerably good generalization capability of the model. Detailed domain-wise detection performance results are reported in Table 6.

5. Ablation study

To show the significance of each component in our proposed feature set RCD+L+HFR, we perform an ablation study on our best solution, RCD+L+HFR, on the YCbCr colour space, as shown in Table 7, that achieves the best performance on post-processed images.

Further, we evaluate their performance for post-processed images. Results are shown in Table 8. It is evident that HFR, being the minimal feature set, performs at par with RCD+L+HFR feature set. However, in practical scenarios, blur is more often used for face images to enhance the smoothness of skin, compared to other methods. As shown in Fig. 5, with the increase in average blur kernel parameter value, the performance of HFR drops drastically compared to RCD+L+HFR.

Table 8
RCD Net components on post-processing operation.

Operations	Parameter	L	HFR	L+HFR	RCD+L+HFR
Median filter	3 x 3	98.61	96.88	98.19	97.52
	5 x 5	90.80	96.90	95.81	96.45
Gaussian noise	1.0	98.69	97.89	98.46	98.29
	2.0	98.66	97.84	98.41	98.14
CLAHE	–	98.46	96.33	97.87	96.16
Average blurring	3 x 3	97.92	97.07	97.82	97.57
	5 x 5	76.02	96.30	91.15	97.59
Gamma correction	0.8	98.56	97.59	98.14	98.26
	0.9	98.66	97.64	98.39	98.44
	1.2	98.61	97.69	98.29	98.09
Resizing	0.5	96.85	96.97	97.32	97.35
Average	–	95.62	97.23	97.25	97.62

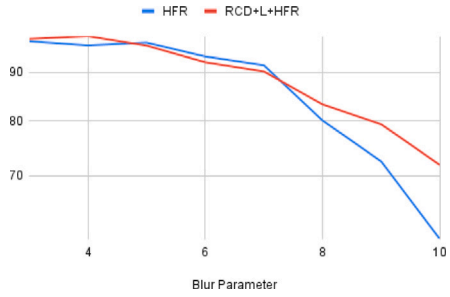


Fig. 5. Comparison of HFR and RCD+L+HFR for Average Blur.

Hence, it can be stated that RCD+L+HFR makes a strong feature set that is much more robust to post-processing attacks and JPEG compression that mimic real OSN scenario.

6. Concluding remarks

As discussed earlier, with the tremendous growth of Generative AI in image synthesis, it is hard to believe what we see on the OSNs. To prevent those synthetic images from manipulating regular users' perception of reality, synthetic image detectors must be democratized outside lab environments.

To this end, we propose three ideas. Firstly, a novel feature, RCD is introduced. Secondly, the fusion of colour space-related statistical features and frequency domain features is explored. Thirdly, we propose a lightweight DNN classifier with very few parameters. We tested our methods on five different colour spaces. Through extensive experiments, we prove the effectiveness and robustness of our methods against standard post-processing and compression scenarios. We achieve state-of-the-art performance in synthetic image detection.

Buy me a coffee https://ko-fi.com/jurnalin https://dbook.eu.org

Moreover, the performance of our proposed solution degrades significantly less in the post-processing scenario compared to other SOTA solutions.

From our experiments, we can conclude that careful designing of features and custom classifiers can attain great results in much less complexity, which is paramount for present-day scenarios.

CRediT authorship contribution statement

Tanusree Ghosh: Conceptualization, Data curation, Methodology, Writing – original draft, Writing – review & editing. **Ruchira Naskar:** Conceptualization, Data curation, Methodology, Supervision, Writing – original draft, Writing – review & editing.

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Data availability

Data will be made available on request.

References

[1] I. Goodfellow, J. Pouget-Abadie, M. Mirza, B. Xu, D. Warde-Farley, S. Ozair, A. Courville, Y. Bengio, Generative adversarial networks, Commun. ACM 63 (11) (2020) 139–144.

[2] S.D. Bray, S.D. Johnson, B. Kleinberg, Testing human ability to detect deepfake images of human faces, 2022, arXiv preprint arXiv:2212.05056.

[3] T. Karras, S. Laine, T. Aila, A style-based generator architecture for generative adversarial networks, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2019, pp. 4401–4410.

[4] T. Karras, S. Laine, M. Aittala, J. Hellsten, J. Lehtinen, T. Aila, Analyzing and improving the image quality of stylegan, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2020, pp. 8110–8119.

[5] T. Karras, M. Aittala, S. Laine, E. Härkönen, J. Hellsten, J. Lehtinen, T. Aila, Alias-free generative adversarial networks, Adv. Neural Inf. Process. Syst. 34 (2021) 852–863.

[6] S. Nightingale, S. Agarwal, E. Härkönen, J. Lehtinen, H. Farid, Synthetic faces: how perceptually convincing are they? J. Vis. 21 (9) (2021) 2015.

[7] P. Yang, R. Ni, Y. Zhao, W. Zhao, Source camera identification based on content-adaptive fusion residual networks, Pattern Recognit. Lett. 119 (2019) 195–204.

[8] D. Freire-Obregon, F. Narducci, S. Barra, M. Castrillon-Santana, Deep learning for source camera identification on mobile devices, Pattern Recognit. Lett. 126 (2019) 86–91.

[9] U.A. Ciftci, I. Demir, L. Yin, Fakecatcher: Detection of synthetic portrait videos using biological signals, IEEE Trans. Pattern Anal. Mach. Intell. (2020).

[10] H. Guo, S. Hu, X. Wang, M.-C. Chang, S. Lyu, Eyes tell all: Irregular pupil shapes reveal GAN-generated faces, in: ICASSP 2022-2022 IEEE International Conference on Acoustics, Speech and Signal Processing, ICASSP, IEEE, 2022, pp. 2904–2908.

[11] N.-T. Do, I.-S. Na, S.-H. Kim, Forensics face detection from GANs using convolutional neural network, ISITC 2018 (2018) 376–379.

[12] F. Chollet, Xception: Deep learning with depthwise separable convolutions, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2017, pp. 1251–1258.

[13] B. Chen, X. Ju, B. Xiao, W. Ding, Y. Zheng, V.H.C. de Albuquerque, Locally GAN-generated face detection based on an improved xception, Inform. Sci. 572 (2021) 16–28.

[14] B. Chen, W. Tan, Y. Wang, G. Zhao, Distinguishing between natural and GAN-generated face images by combining global and local features, Chin. J. Electron. 31 (1) (2022) 59–67.

[15] J. Frank, T. Eisenhofer, L. Schönherr, A. Fischer, D. Kolossa, T. Holz, Leveraging frequency analysis for deep fake image recognition, in: International Conference on Machine Learning, PMLR, 2020, pp. 3247–3258.

[16] D. Gragnaniello, D. Cozzolino, F. Marra, G. Poggi, L. Verdoliva, Are GAN generated images easy to detect? A critical analysis of the state-of-the-art, in: 2021 IEEE International Conference on Multimedia and Expo, ICME, IEEE, 2021, pp. 1–6.

[17] Y. Wang, X. Ding, Y. Yang, L. Ding, R. Ward, Z.J. Wang, Perception matters: Exploring imperceptible and transferable anti-forensics for GAN-generated fake face imagery detection, Pattern Recognit. Lett. 146 (2021) 15–22.

[18] X. Yang, Y. Li, H. Qi, S. Lyu, Exposing GAN-synthesized faces using landmark locations, in: Proceedings of the ACM Workshop on Information Hiding and Multimedia Security, 2019, pp. 113–118.

[19] L. Nataraj, T.M. Mohammed, S. Chandrasekaran, A. Flenner, J.H. Bappy, A.K. Roy-Chowdhury, B. Manjunath, Detecting GAN generated fake images using co-occurrence matrices, 2019, arXiv preprint arXiv:1903.06836.

[20] M. Barni, K. Kallas, E. Nowroozi, B. Tondi, CNN detection of GAN-generated face images based on cross-band co-occurrences analysis, in: 2020 IEEE International Workshop on Information Forensics and Security, WIFS, IEEE, 2020, pp. 1–6.

[21] E. Nowroozi, Y. Mekdad, Detecting high-quality GAN-generated face images using neural networks, Big Data Anal. Intell. Syst. Cyber Threat Intell. (2023) 235–252.

[22] T. Qiao, Y. Chen, X. Zhou, R. Shi, H. Shao, K. Shen, X. Luo, CSC-net: Cross-color spatial co-occurrence matrix network for detecting synthesized fake images, IEEE Trans. Cogn. Dev. Syst. (2023).

[23] H. Li, B. Li, S. Tan, J. Huang, Identification of deep network generated images using disparities in color components, Signal Process. 174 (2020) 107616.

[24] P. He, H. Li, H. Wang, Detection of fake images via the ensemble of deep representations from multi color spaces, in: 2019 IEEE International Conference on Image Processing, ICIP, IEEE, 2019, pp. 2299–2303.

[25] B. Chen, X. Liu, Y. Zheng, G. Zhao, Y.-Q. Shi, A robust GAN-generated face detection method based on dual-color spaces and an improved xception, IEEE Trans. Circuits Syst. Video Technol. 32 (6) (2021) 3527–3538.

[26] Z. Xia, T. Qiao, M. Xu, N. Zheng, S. Xie, Towards DeepFake video forensics based on facial textural disparities in multi-color channels, Inform. Sci. 607 (2022) 654–669.

[27] S. McCloskey, M. Albright, Detecting GAN-generated imagery using color cues, 2018, arXiv preprint arXiv:1812.08247.

[28] J.M. Kasson, W. Plouffe, An analysis of selected computer interchange color spaces, ACM Trans. Graph. 11 (4) (1992) 373–405.

[29] N.A. Ibraheem, M.M. Hasan, R.Z. Khan, P.K. Mishra, Understanding color models: a review, ARPN J. Sci. Technol. 2 (3) (2012) 265–275.

[30] B. Wang, X. Wu, Y. Tang, Y. Ma, Z. Shan, F. Wei, Frequency domain filtered residual network for deepfake detection, Mathematics 11 (4) (2023) 816.

[31] T. Pevnỳ, P. Bas, J. Fridrich, Steganalysis by subtractive pixel adjacency matrix, in: Proceedings of the 11th ACM Workshop on Multimedia and Security, 2009, pp. 75–84.

[32] Z. Liu, P. Luo, X. Wang, X. Tang, Deep learning face attributes in the wild, in: Proceedings of the IEEE International Conference on Computer Vision, 2015, pp. 3730–3738.

Buy me a coffee https://ko-fi.com/jurnalin https://dbook.eu.org