J. Vis. Commun. Image R. 95 (2023) 103876

Contents lists available at ScienceDirect

J. Vis. Commun. Image R.

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

Full Length Article

Detection of GAN generated image using color gradient representation ☆

Yun Liu [a,*], Zuliang Wan [a], Xiaohua Yin [a], Guanghui Yue [b], Aiping Tan [a,*], Zhi Zheng [c]

[a] College of Information, Liaoning University, Shenyang 110036, China
[b] School of Biomedical Engineering, Health Science Center, Shenzhen University, Shenzhen 518000, China
[c] Department of Electronic and Information Engineering, Beijing Jiaotong University, Beijing 100000, China

ARTICLE INFO

MSC:
41A05
41A10
65D05
65D17

Keywords:
Image generative model
Generative adversarial networks
Fake image identification

ABSTRACT

With the development of generative adversarial network (GANs) technology, the technology of GAN generates images has evolved dramatically. Distinguishing these GAN generated images is challenging for the human eye. Moreover, the GAN generated fake images may cause some behaviors that endanger society and bring great security problems to society. Research on GAN generated image detection is still in the exploratory stage and many challenges remain. Motivated by the above problem, we propose a novel GAN image detection method based on color gradient analysis. We consider the difference in color information between real images and GAN generated images in multiple color spaces, and combined the gradient information and the directional texture information of the generated images to extract the gradient texture features for GAN generated images detection. Experimental results on PGGAN and StyleGAN2 datasets demonstrate that the proposed method achieves good performance, and is robust to other various perturbation attacks.

1. Introduction

Lately, with the noteworthy advancement of deep learning and the development of compute equipment performance, image synthesis technology based on deep learning has evolved dramatically. Specifically, the Generative Adversarial Network (GAN) proposed in 2014 gave brought forth a clever way to deal with image synthesis [1]. Later, some advanced GANs, such as boundary equilibrium GAN (BEGAN) [2], progressive growing of GAN (PGGAN) [3], StyleGAN [4], and StyleGAN2 [5], have been shown great success in generating images in many areas, for example, image super-resolution [6,7], image translation [8], and image inpainting [9].

There are several typical GAN generated image shown in Fig. 1. It can concluded that it is really difficult for humans to distinguish which are fake faces and which are real faces through human eyes. Distinguish these GAN generated images is challenging for the human eye. If these generated GAN images used to fabricate fake news, or to defraud by forging personal information on social networks, it may greatly disrupt the social order, even affect the world, coming about in moral, lawful, and security issues. The potential risks of GAN Generated Image cannot be ignored and may have adverse effects on society. It is imminent to find the effective detection techniques of GAN Generated Image.

Motivated by the above problem, many detection algorithms have been proposed, and these algorithms can generally be divided into traditional methods and deep learning methods. Traditional methods [10–15] are constructed in light of hand-crafted image features extraction, like texture feature and structure features, which can present an explainable detection procedure. However, these methods fail to obtain satisfied generalization performance, and have high computational complexity. With the outstanding results of deep learning technology in image classification task, deep learning method arouse researchers' attention [16–31]. Compared to traditional method, deep learning methods yield less computational complexity, but lack interpretability. Besides, deep learning methods need amount of date to support their model, and have overfitting problem, which greatly affect the performance and limit their practical application. To this end, more endeavors should be ought to establish effective GAN generated image detection methods.

The HVS is very sensitive to gradients and textures, and a lot of information about the image is hidden in the textures and gradients of the image [32]. The gradient information of the real image may be different from the fake image. Besides, [15] studied statistical properties of residual domain in the chrominance components of the HSV and YCbCr color spaces and found the differences between the generated and the real images. [33] [15] [34] [35] [36] [37] [38] are also considered the importance of different color spaces for GAN face detection, which shows that choosing an appropriate color space is extremely important for GAN face detection. However, the color spaces used in these models are not the same, so it is very necessary to analyze

☆ This paper has been recommended for acceptance by Dr. Zicheng Liu.

* Corresponding authors.
E-mail addresses: yunliu@tju.edu.cn (Y. Liu), aipingtan@lnu.edu.cn (A. Tan).

https://doi.org/10.1016/j.jvcir.2023.103876
Received 27 May 2022; Received in revised form 29 May 2023; Accepted 10 June 2023
Available online 22 June 2023

Fig. 1. Several typical GAN generated image.

the effects of different color space on GAN images detection area, which is considered in our work. [35,39] pay attention to the traces generated by the upsampling process in the GAN generation process, which proves the importance of multi-scale features. Therefore, we propose a novel GAN images detection method based on color gradient analysis. Considering the important traces information left in color channels in generated images, we analysis the color feature difference between GAN image and real image, and found that the generated image has more gradient information disparities in HSV color channels. To upgrade the effective of the proposed model, we extract the local texture descriptor from two directional gradient channels to reflect the spatial texture information, which has powerful discriminative ability on generated image detection. Our model is robust to attacks such as image resizing, compression, blurring, and noise attacks. Furthermore, compared to the black-box nature of deep neural networks, our extracted features are more interpretable. Our main contributions are shown as follows:

- Considering the inherent trace left in color information, we studies the color information difference between real images and GAN generated images based on a variety of color spaces. By analysis the gradient information and the directional texture information of the generated images, we chose the HSV color space to extract two directional gradient texture features (the horizontal and the vertical gradient channels), which makes our model more interpretable.
- We propose a novel gradient-domain local direction number (GLDN) pattern descriptor to extract the gradient texture features. GLDN enhance the intensity change and structural information of the local information, which can effectively grasp the inherent trace of the generated images and significantly improve the performance of GAN image classification.
- We apply the Multi-GLDN histogram (MGLH) to quantify the gradient texture information of each GLDN map block and aggregates the classification information between natural images and GAN generated images into texture descriptors, which can greatly reduce the computational complexity and yield higher accuracy.

The remainder of this paper is organized as follows. Section 2 presents previous related literature on GAN generated images and some detection methods. Section 3 presents our proposed method. Section 4 presents the experimental results and analysis. Finally, Section 5 concludes the study.

2. Related works

2.1. GANs image generative models

Generative Adversarial Networks (GAN) were first proposed by Ian Goodfellow et al. in 2014 [1]. GAN was a novel machine learning framework at the time, which consisted of a generator and discriminator. The generator attempts to generate synthetic samples from a lot of real data such as images, and the discriminator attempts to accurately distinguish the synthetic samples and judge whether the synthetic samples come from the generator or the real data. The generator and discriminator further improve themselves by competing with each other. When the discriminator discovers some distinctions between the genuine samples and the generated samples, the generator changes its parameters to generate samples that are nearer to the genuine samples, while the discriminator improves its capacity to recognize the genuine samples and the generated samples again by changes its parameters. In an optimal case, after multiple iterations, the generator ultimately imitates the distribution data that cannot be distinguish by the discriminator. Based on the strong generation ability of GAN, one of the most well known uses of GANs is to generate images of faces. Many GANs are basically known for advanced generative adversarial models by showing their face generation results, like BEGAN [2], PGGAN [3], StyleGAN [4], DCGAN [40], EBGAN [41] and WGAN [42]. Furthermore, other GAN have exhibited impressive results in various domains other than human faces [43–46]. As GAN becomes more and more topical, a series of works around the structural improvement of GAN and the evaluation of generative results have become the focus of attention.

2.2. Fake image identification

Until now, various algorithms have been proposed for detecting GAN generated images, which are generally divided into the traditional method [10–15] and the deep learning method [16–31].

Before the presence of GAN images, numerous techniques were created to distinguish fake images exposed to altering and rebroadcasting. The vast majority of those strategies depend on specific tracking of fake image processing pipelines, for example, the JPEG compression errors during image recompression [47] and the image quality distortion that occurs in fake face images [48,49]. Some works at first use traditional image forensic methods to detect GAN generated images. However, such designated techniques are not appropriate for detecting GAN generated images. After that, some detection general methods are based on spectra information in the frequency domain or statistical features extracted from image textures. With the continuous development of deep learning technology, some works utilize deep neural networks to conduct the detection of GAN generated images [16–31]. However, these detection methods either have limited receptive fields, or lack the comprehension of global information. And most importantly, they lack the interpretability. These deep learning-based algorithms still remains to be improved. The specific overview is extended as follows.

2.2.1. Traditional methods

By considering the facial parts configuration distinction between the facial parts generated by GAN models and the real face, Yang et al. [50] apply the locations of the facial marker points to uncover GAN generated images. But with the development of GAN, GAN generated images solve the above configuration inconsistent problem, it is difficult to distinguish GAN generated images with nature images just by inconsistent configurations in face. Li et al. [38], by analyzing the differences between camera imaging and GAN generated images, extracted the statistics information from residual images of three different color components for GAN generated images identification. Tang et al. [10] extract the spectral correlation of natural color images base on DWT(discrete wavelet transform) and the standard correlation coefficient to detects GAN generated fake images. Chandrasegaran et al. [11] Focus on high frequency Fourier spectrum decay attributes to detects GAN generated fake images. Marra et al. [12] investigated whether GANs leave specific fingerprints on GAN generated images. S et al. [13] combined the frequency and spatial domains and extracted image quality features from them to detect GAN generated images,

which provide another way to solve this problem. Tao et al. [51] proposed an effective model by combining the texture and sensor noise based statistical features to detection GAN-generated face. Xia et al. [36] chosen to exploit facial textural disparities in multi-color channels to distinguish the real video from the DeepFake video. McCloskey and Albright [52] proposed a straightforward GAN generated image detector by measuring the frequency of saturated and under-exposed pixels in image. But this method is easily susceptible to noise and example attacks to detect natural and GAN generated images. Matern et al. [53] proposed a forensics method for GAN generated image detection by selecting visual artifacts in the global consistency of organs and the eye color. But the algorithm is less robust when faced with some common processes in image transmission, such as lossy compression and blurring.

Overall, these traditional intrinsic attributes-based algorithms yield low generalization capability and high computational complexity.

2.2.2. Deep learning methods

Many previous works use various architectures of CNN to detect fake images. Marra et al. [16] and Chen et al. [17] use improved Xception model for GAN generated image detection. Tariq et al. [18] proposed an ensemble-based neural network classifier called Shallow Convolutional Network (ShallowNet) to detect GAN generated fake images. Hsu et al. [19] proposed a two-streamed network to adopt contrastive loss to seek typical features. They apply the pairwise image information as the input of their two-streamed network to detect the fake and real images. In addition, many works add classification features between real and generated images to CNN. Based on the neural network and signal processing methods, Zhao et al. [20] proposed a deepfake detection network for extracting fused RGB features and texture information. He et al. [33] employed residual signals of chrominance components from multi color spaces to learn deep representations via the CNN for GAN-generated face detection. Nataraj et al. [54] proposed a deep convolutional neural network framework by extracting the co-occurrence matrices on three color channels in the pixel domain. Mo et al. [24] proposed a CNN-based algorithm that extracted features from the residuals domain, which centers around the image high-frequency components to identify GAN generated fake images. Yang et al. [21] proposes a method by using two intrinsic clues in the channel difference image and spectrum image view of the camera imaging process. Afchar et al. [25] presented two networks for the image-level detection of fake face contents in videos, which centers around the mesoscopic properties of images. Yang et al. [22] focused on texture artifacts in post-processing operations and enhanced it for forgery detection by using a guided filter with saliency map as a guide map. Guo et al. [23] proposed an adaptive manipulation trace extraction network that focuses on highlighting manipulation traces to detect GAN generated fake images. People usually pay more attention to some local areas when distinguishing between the real and GAN generated images. Restricted by the nature of CNN, these detection techniques have restricted receptive fields and lack the comprehension of global information. So some works pay attention to the global information. By introducing the local-to-global strategy to CNN, Quan et al. [26] proposed a new CNN architecture to conduct the fake image detection. Mi. et al. [28] introduced the powerful Self-Attention operation into the neural network to learn more about the global information rather than just focusing on the local information in the image. Chen et al. [17] introduced the feature pyramid network to improve the Xception model for locally GAN-generated face detection. Later, they proposed a dual-stream network by combining the luminance component and chrominance components, which achieve a better performance than the above Xception model [34]. Motivated by the above works, Peng et al. [55] proposed a new dual-stream network by considering spatial and frequency domains model. Gangan et al. [37] proposed a Multi-Colorspace fused EfficientNet model to obtain information from multiple color spaces. Among those methods, the generalization ability of these deep learning-based algorithms remains to be improved.

3. The proposed method

GAN generated images may have some strong correlations between nearby pixels during the GAN image generation process, so we mainly focus on the pixel-level relationship difference between GAN generated fake images and natural images in the gradient domain, and propose a GAN generated image detection method based on gradient-domain LDN (GLDN). The whole framework is shown in Fig. 2. Given an input image, the image is first converted to a specific multiple color spaces to calculate the horizontal gradient map G_x and vertical gradient map G_y for each color channel. The LDN is then applied to extract richer texture features from the above two directional gradient map, namely GLDN maps. By dividing the GLDN maps into several small blocks, the local color gradient features histograms is obtained. Then, all the local feature histograms are concatenated to form the final feature vector. Finally, a classifier is used in this study to detect the image.

3.1. Gradient operator

Sobel operator is a classical gradient-based extract method, so chose the Sobel operator to extract gradient information.

Let the value of pixel (i, j) in the image be $P(i, j)$. Then the corresponding gradient can be presented as:

$$\nabla f_{(i,j)} = \begin{pmatrix} \partial_f/\partial_i \\ \partial_f/\partial_j \end{pmatrix} \tag{1}$$

And its vector modulus is:

$$\left|\nabla f_{i,j}\right| = \sqrt{(\partial_f/\partial_i)^2 + (\partial_f/\partial_j)^2} \tag{2}$$

The horizontal gradient vector and vertical gradient vector can be presented as:

$$G_i = \partial_f/\partial_i \tag{3}$$

$$G_j = \partial_f/\partial_j \tag{4}$$

Therefore, Eq. (2) can be rewritten as:

$$\left|\nabla f_{i,j}\right| \doteq |G_i| + |G_j| \tag{5}$$

In the Sobel gradient operator, for a certain pixel, the gradient is obtained by the weighted sum of the pixels in the 3×3 neighborhood, and the Sobel convolution kernels is presented as:

$$S_h = \begin{bmatrix} -1 & -2 & -1 \\ 0 & 0 & 0 \\ 1 & 2 & 1 \end{bmatrix} \tag{6}$$

$$S_v = \begin{bmatrix} -1 & 0 & 1 \\ -2 & 0 & 2 \\ -1 & 0 & 1 \end{bmatrix} \tag{7}$$

where S_h and S_v is horizontal and vertical template.

After convolving the Sobel operator kernel with the original image, the gradient map G_x and G_y can be written as:

$$G_x = f_{i+1,j-1} + 2f_{i+1,j} + f_{i+1,j+1} - f_{i-1,j-1} - 2f_{i-1,j} - f_{i-1,j+1} \tag{8}$$

$$G_y = f_{i-1,j+1} + 2f_{i,j+1} + f_{i+1,j+1} - f_{i-1,j-1} - 2f_{i,j-1} - f_{i+1,j-1} \tag{9}$$

3.2. Gradient-domain Local Directional Number Pattern (GLDN)

The LDN [56] is a texture descriptor that encodes the structural information and the intensity variations of the facial texture. Compared with other texture descriptors such as LBP [57], LDN shows better performance on image classification. So we combine LDN with the above two gradient maps to extract Gradient-domain LDN features. Take the horizontal gradient as an example, the compass masks are

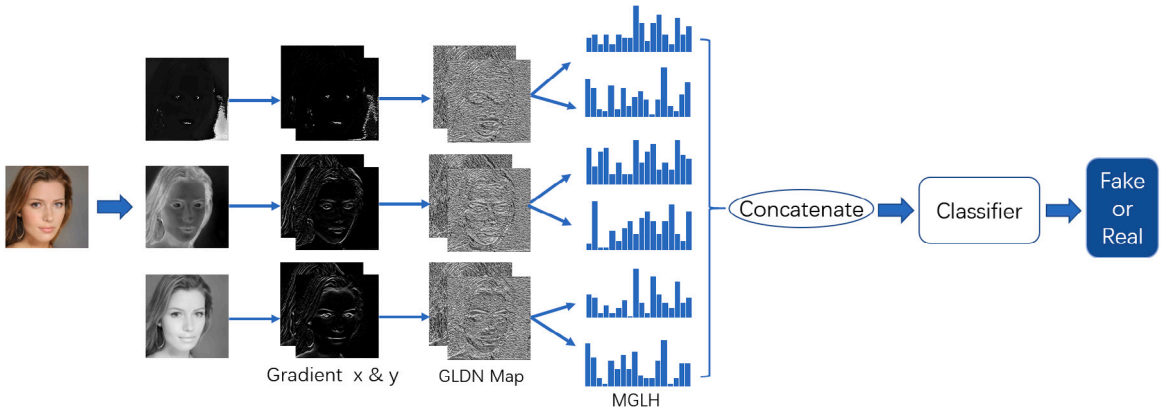

Fig. 2. The overall framework of the proposed method.

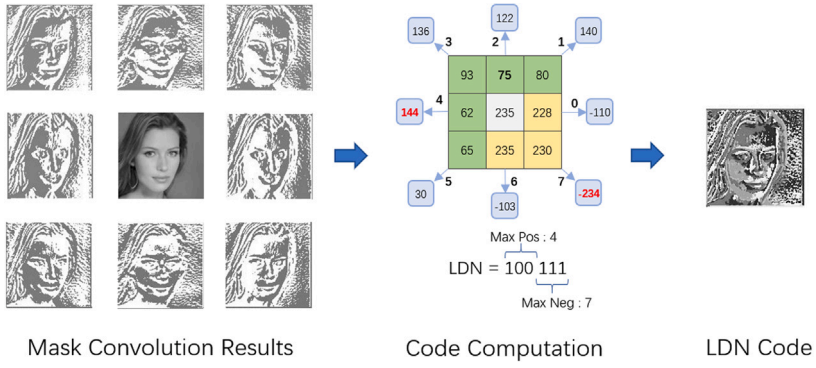

Fig. 3. LDN code computation.

convoluted with the horizontal gradient map to extract GLDN map, shown in Fig. 3.

The compass mask is created by using the derivative of a skewed Gaussian, which is robust to noise and illumination changes, shown as follows:

$$M_\sigma(x,y) = G'_\sigma(x+k,\ y) * G_\sigma(x,y) \tag{10}$$

where σ is the width of the Gaussian bell and x, y are location positions.$G_\sigma(x,y)$ is the Gaussian mask, shown as follows:

$$G_\sigma(x,y) = \frac{1}{2\pi\sigma^2}\exp(-\frac{x^2+y^2}{2\sigma^2}) \tag{11}$$

$*$ represents the convolution operation, G'_σ represents the derivative of G_σ with respect to x, and k represents the offset of the Gaussian. And use 1/4 of the mask diameter for this offset. Then, a set of masks $\{M^0_\sigma \cdots M^7_\sigma\}$ is generated by rotating M^0_σ compass mask 45° apart in eight different directions. The code computed by this mask with Gaussian parameter σ is denoted as LDN^G_σ.

In the coding scheme, the code GLDN is generated by combining the edge response and the dominant directional numbers of each mask $\{M^0 \cdots M^7\}$, the edge response image is extracted by convolving the compass masks with the gradient domain image. The whole process is shown in Fig. 3. The left side is the edge response image of eight masks, and the center is the edge response value of the corresponding mask of a pixel. Choose the prominent directional number from the positive and negative directions to encode the texture in this pixel neighborhood. The max positive directions in number 4 and the max negative directions in number 7 are chosen to encode the neighborhood texture, and then the final LDN code is obtained. To encode these prominent directional number, a fixed position is assigned to the top directional number. The top positive direction number is represented by the highest three most significant bits in the code, and the top negative direction number is represented by the least significant bits in the code. The GLDN code can be defined as:

$$GLDN(x,y) = 8i_{x,y} + j_{x,y} \tag{12}$$

where (x,y) is the position of the encoded pixel, $i_{x,y}$ is the maximum positive response directional number, and $j_{x,y}$ is the minimum negative response directional number, and they defined by:

$$i_{x,y} = arg_i \max\{\Pi^i(x,y) | 0 \le i \le 7\} \tag{13}$$

$$j_{x,y} = arg_j \min\{\Pi^j(x,y) | 0 \le j \le 7\} \tag{14}$$

where II^i is the convolution of the original image I, and the i th mask M^i, defined by:

$$II^i = I * M^i \tag{15}$$

3.3. Multi-GLDN histogram (MGLH)

To aggregate location information, we divide the image into N blocks $\{B^1 \cdots B^N\}$, then extract a histogram H^i from each block B^i. By creating the histogram H^i and using each code as a bin, we accumulate all the codes in the block in their respective bin by:

$$H^i(c) = \sum_{GLDN(x,y)=c} (x,y) \in R^i \quad v, \forall c \tag{16}$$

where (x,y) is the pixel position in the block B^i, and c is a GLDN code, GLDN(x,y) is the GLDN code for the position (x,y), and v is the accumulation value, usually the cumulative value is 1. Finally, the GLH is computed by concatenating those histograms:

$$GLH = \prod_{i=1}^{N} H^i \tag{17}$$

Table 1
Dataset description.

GAN type	Collection	Total
PGGAN	[3]	30 k
StyleGAN2	[5]	12 k
StyleGAN1	[4]	12 k
BigGAN	[45]	12 k
CycleGAN	[44]	12 k
StarGAN	[60]	12 k
GauGAN	[46]	12 k

The derivative-Gaussian mask can be freely resized. The resizing in the derivative Gaussian mask enables GLDN to capture different features in the image. Hence, the features at different scales are achieved by computing the $GLDN_{\sigma}^{G}$ code at n different σ_i, and using $GLDN_{\sigma}^{G}$ to calculate the histogram $H_{\sigma_i}^{i}$ of each σ_i according to Eq. (16) and concatenate them to merge the features at different resolutions. This is called the multi-GLDN histogram(MGLH), and it is computed by:

$$\mathrm{MGLH}_{\sigma_1 \cdots \sigma_n} = \prod_{j=1}^{N} \prod_{i=1}^{n} H_{\sigma_i}^{j} \tag{18}$$

where $\prod$ is the concatenation operation, $H_{\sigma_i}^{i}$ is the histogram of the $GLDN_{\sigma i}^{G}$ code at the B^j block, and n is the number used by σ.

3.4. Classifier

Ensemble classifier [58] consists of many base learners independently trained on different sets of images. Each single base learner is a simple classifier built on the feature space that randomly selected subspace. For an example in a given test set, the final decision of the classifier is computed by aggregating the results of all the individual base learners. Ensemble classifier is better at handling high-dimensional feature vectors and complex large training sets compared to other classifiers such as SVM [59]. [15] uses the Ensemble classifier as classifier for handcrafted feature-based methods and achieved good results. Motivated by the above reasons, the ensemble classifier is used as the classifier in this paper. The parameters were set as defaults used in [58].

4. Experiments

4.1. Experimental setup

4.1.1. Data collection

For this experiment, real images are selected from the CelebA [29] and CelebA-HQ [3] datasets, and the GAN generated fake images are generated by PGGAN [3] and StyleGAN2 [5]. To improve the quality of the dataset and make the fake images of the fake image dataset more diverse, images generated by various other GANs are used: StyleGAN [4], CycleGAN [44], BigGAN [45], GauGAN [46], and StarGAN [60]. We choose as much as possible the open sourced GAN and the corresponding released generated image dataset as the experimental objects to reduce the possible errors of the images generated by different GANs after training. Table 1 shows the source and number of generated image datasets. The GAN type is represented in the first column. The second column shows the source of generated fake images. The last column shows the quantity used in this experiment.

4.1.2. Implementation details

In this paper, the ensemble classifier [58] is used as the classifier for this method. The parameters of the classifier were set to the default values used in [58]. We selected CelebA-HQ, PGGAN, CelebA and StyleGAN2 datasets to conduct our experiments.

To compare as fair as possible, we refer to the experimental setup of previous models to conduct the experimental comparison. For the PGGAN dataset, same as [28], the image is resized to 256 × 256 and compressed using lossy JPEG compression with quality 95. 15k real images from CelebA HQ and 15k GAN generated fake images by PGGAN is randomly selected as the training set and the testing set, and the ratio of training set and testing set is 1:1. In all, 7,500 real faces and 7,500 GAN generated faces are randomly selected as the training set, and the others as testing set. For the StyleGAN2 dataset, same as [10], we choose 5k real faces from CelebA and 5k StyleGAN2 generated faces as the training set, and 1k real face and 1k StyleGAN2 generated face images as the testing set. And four common perturbation attacks are selected to evaluate the effectiveness and robustness of our model, including compression, blurring, resizing, and adding noise.

Table 2
Comparison with different color spaces (Accuracy, %).

Color Space	RGB	YCbCr	HSV
Accuracy	97.6	96.3	**99.4**

4.1.3. Evaluation metrics

In this paper, we use four common metrics, Accuracy, Recall, Precision, and AUC(area under curve of receiver operating characteristics) to evaluate the performance of our method. Accuracy indicates the proportion of all predictions that are correct in the total number of experimental cases, precision indicates the proportion of correct predictions that are positive to all positive predictions, and recall rate indicates the proportion of correct predictions that are positive to all actual positives. AUC is defined as the area under the ROC(receiver operating characteristic curve), and the value of AUC is not greater than 1, which is used to evaluate the performance against the four perturbation attacks. For all metrics, higher values represent better classification performance of the method.

4.2. Detection performance

Since GAN generally imitate fake images in the RGB domain, they focus more on imitating the features of real images in the RGB domain while ignoring the features in other color domains. The distinction between the real image and the fake image might be more distinct in other variety of color domains. Therefore, we analyze several commonly used color spaces to detect images generated by GAN: RGB, HSV, and YCbCr. The results on different color spaces are shown in Table 2. From the results, it can be found that the ACC results of the model based on HSV are higher than the other color space, so we chose to extract feature vectors based on the HSV color space in our paper to make the experimental comparison.

To evaluate the performance of our proposed method, experiments are conducted on GAN generated human face image recognition. Eight previous works are applied to make the comparison, and five of their results are derived from work [28], and work [15,36,37] is computed based on the code released in their work. Among them. Works [24–28,37] are built based on CNN model, and work [15] is proposed based on the statistics information in residual images of three different color components. Work [36] is proposed based on facial textural disparities in multi-color channels to detect GAN generated images. The experimental settings are completely referring to the work [28].

The comparison results are shown in Table 3, with the best performing models highlighted in bold. From the experimental results, it can be concluded that the proposed GAN generated face image detection method achieves the best detection accuracy, which proves the effectiveness of the combination of image gradient information and color information. The performance of our model exceeds the deep learning method and traditional method, which not only is interpretable but also has good performance.

To further evaluate the effectiveness of the proposed method, we also conducted experiments on data sets CelebA and StyleGAN2. The

Table 3
Comparison with other methods in PGGAN (Accuracy, %).

Test Sets	[24]	[25]	[26]	[27]	[28]	[15]	[36]	[37]	our
None Post Processes	97.7	96.9	96.8	87.7	99.3	96.6	98.7	99.2	**99.4**

Table 4
Comparison with other methods in STYLEGAN2.

	Precision	Recall	Accuracy
AutoGAN	75.7%	66.3%	72.5%
FakeSpotter	91.2%	92.4%	91.9%
FID	98.5%	98.5%	98.5%
our	**99.8%**	**99.7%**	**99.8%**

experimental settings refer to the article [10], and we choose the same number of images and preprocessing method. The comparison between our experimental results and previous work is shown in Table 4, and the best-performed models are highlighted in bold. FID [10] applies discrete wavelet transform (DWT) to RGB color space components separately to detect GAN generated images. AutoGAN [39] focuses on the artifacts in GAN generated images and uses a classifier based on a deep neural network to detect fake images. FakeSpotter [61] discovers AI-generated fake faces by monitoring neuron behavior to detect fake images. The experimental results show that our proposed method outperforms other works, which proves that the gradient based HSV components are more effective and interpretable in fake image detection.

4.3. Robustness analysis

To prove the robustness of the proposed model, eight advanced detection methods [15,24–28,36,37] are adopted to conduct the experimental comparison. CelebA-HQ and PGGAN datasets are processed by referring to [28]: GF and MF represent Gaussian filter and median filter respectively, the following numbers represent the size of the filter, JPEG represents lossy quality compression, and the value of compression quality ranges from 80 to 95. The experimental results are shown in Table 5, with the best-performing models highlighted in bold. The results show that the proposed method has satisfactory performance except for median filtering. Since median filtering changes the relationship between pixels in a finite neighborhood, and it affects the top positive direction number and top negative direction number of pixels, so the performance of our method on median filtering is a little worse than previous works. Although work [28] performed best among these models on MF5 and MF7, it was not the best in the other situations. In summary, our method has effectiveness and robustness in fake image detection and can be effectively used in various real-life environments.

Referring to [10], under these several perturbation attacks to images, we conduct experiments to evaluate the effectiveness of our method and use AUC as the performance indicator. Specifically, the compression quality is set between 0 and 100. Blur indicates that the Gaussian blur is utilized in the experimental image. The size of the Gaussian kernel is (3, 3), and the value of the standard deviation of the Gaussian kernel controls the intensity of the blur. In resizing, different scale factors are selected to control the size of images. The Gaussian additive noise is utilized for adding noise to images, and the variance controls the intensity of noise. As shown in Fig. 4, the results of several works for the above perturbation attacks with different types and intensities are shown. Compared with other works, our method performs better in compression and Resizing. The AUC scores of our method fluctuate in a stable range under different intensities of perturbation attack, which further demonstrates the effectiveness and robustness of our model.

In addition, to verify the universality and stability of our proposed method on other common GANs generated image datasets, we compare our model with FID and AutoGAN models on 5 different GAN-generated image datasets, and the training and testing datasets are randomly selected at ratio of 5:1. Take the StyleGAN image dataset as an example, it includes 6,000 real faces from CelebA and 6,000 generated faces by StyleGAN model, of which 5,000 real faces and 5,000 generated faces are randomly selected as the training set, and the others as testing set. The experimental settings of other 4 GAN-generated image datasets are the same as that of StyleGAN image dataset. The indicator for performance evaluation is the average precision score, and the experimental results are given in Table 6, and the best-performed model is highlighted in boldface. We can see that the results of our model performance are all higher than 96, and present a very stable performance. Our model yields the best on styleGAN, BigGAN, and GauGAN, and ranks second on CycleGAN and StarGAN. Although the AutoGAN model performed best on CycleGAN and StarGAN, it has the worst results on styleGAN and GauGAN, and second worse results on BigGAN. Overall, our proposed model has stable performance and better results.

4.4. Discussion

Considering that HVS is very sensitive to gradients, and many important traces information hidden in the chrominance components in the generated, we choose the color domain information and the gradient information of the image as clues to detect GAN image. Considering that multi-scale traces are generated in the GAN generation process, we apply the Multi-GLDN histogram (MGLH) to quantify the gradient texture information of each GLDN map block and aggregates the classification information between natural images and GAN generated images into texture descriptors, which can greatly reduce the computational complexity and yield higher accuracy.

The method proposed in this paper achieves satisfactory results in detecting GAN generated images and is also robust to several common perturbation attacks. In addition, experiments on multiple GAN generated image datasets also show the availability of our method, which proves the reasonable of combining multiple color channels and gradient information. However, since median filtering changes the relationship between pixels in a finite neighborhood, and it affects the top positive direction number and top negative direction number of pixels, so the performance of our method on median filtering is a little worse than previous works, which shown in Table 5. And due to the extraction of features in the image block, the final feature vector has a longer dimension.

5. Conclusions

With the development and application of artificial intelligence in recent years, GAN has solved many problems and brought many benefits to human beings. But at the same time, it also brings a lot of disadvantages. GAN can also be used to generate fakes to deceive humans, which poses a potential threat to society. In this study, an EGLND-based GAN generated fake image detection method is proposed by considering the gradient information and color features. We perform extensive experiments on two advanced GAN generated image datasets, and GAN generated images under different perturbations. From the experimental results, our proposed method is effective in detecting GAN generated fake images and is robust to several common perturbation attacks. In addition, the analysis of real and fake image differences

Table 5
Robustness test (Accuracy, %).

Test Sets	[24]	[25]	[26]	[27]	[28]	[15]	[36]	[37]	our
None Post Processes	97.7	96.9	96.8	87.7	99.3	96.6	98.7	99.2	**99.4**
GF3	13.0	90.9	67.0	80.3	**99.2**	96.5	98.6	98.0	**99.2**
GF5	2.9	62.7	79.0	78.9	**99.1**	95.4	98.6	98.2	**99.1**
MF5	48.7	62.2	79.9	84.3	**99.1**	95.7	98.5	98.7	98.2
MF7	25.7	62.6	75.1	85.2	**99.1**	94.4	98.4	98.3	97.1
JPEG95	97.6	96.8	96.8	87.6	99.2	96.6	98.5	98.2	**99.5**
JPEG92	85.7	84.9	97.1	71.7	97.7	94.6	98.1	97.5	**98.9**
JPEG89	95.1	91.3	96.7	77.8	97.8	91.2	97.7	97.1	**98.6**
JPEG86	96.7	82.9	96.0	82.6	97.6	89.5	97.1	96.8	**98.2**
JPEG83	95.4	86.9	95.4	81.5	96.5	88.2	96.5	95.5	**97.6**
JPEG80	92.9	87.4	94.2	80.8	95.6	87.6	96.9	95.3	**97.1**

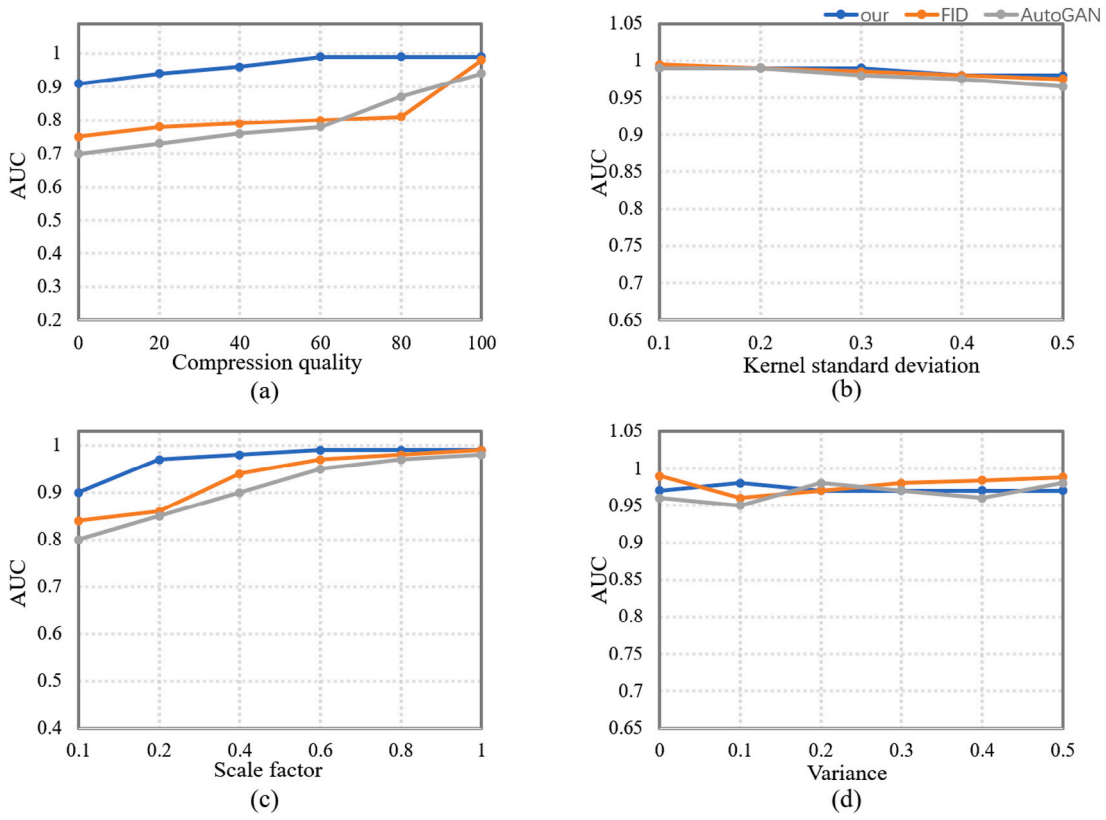

Fig. 4. Four perturbation attacks under different intensities. (a) Compression. (b) Blur. (c) Resizing. (d) Noise.

Table 6
Test results on different GAN datasets (Average Precision, %).

	StyleGAN	BigGAN	CycleGAN	StarGAN	GauGAN
FID	85.33	75.10	96.19	99.70	91.22
AutoGAN	68.60	84.90	**100.00**	**100.00**	61.00
our	**99.75**	**96.85**	98.33	**99.95**	**99.58**

during image imaging can be applied to other GAN generated images. The study of forgery detection is the foundation, and it is important to avoid the risk of artificial intelligence by establishing a strong defense mechanism. Next, we will try our best to find a better general method to detect GAN generated fake images among different GANs.

Declaration of competing interest

We declare that we have no financial and personal relationships with other people or organizations that can inappropriately influence our work, there is no professional or other personal interest of any nature or kind in any product, service and/or company that could be construed as influencing the position presented in, or the review of, the manuscript entitled.

Data availability

No data was used for the research described in the article.

Acknowledgments

This work is supported by the National Natural Science Foundation of China under Grant 61901205, and Liaoning Province Natural Science Foundation, China under Grant 2023-MS-139.

References

[1] I. Goodfellow, J. Pouget-Abadie, M. Mirza, B. Xu, D. Warde-Farley, S. Ozair, A. Courville, Y. Bengio, Generative adversarial nets, Adv. Neural Inf. Process. Syst. 27 (2014).

[2] D. Berthelot, T. Schumm, L. Metz, Began: Boundary equilibrium generative adversarial networks, 2017, arXiv preprint arXiv:1703.10717.

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

[4] 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.

[5] 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.

[6] C. Ledig, L. Theis, F. Huszár, J. Caballero, A. Cunningham, A. Acosta, A. Aitken, A. Tejani, J. Totz, Z. Wang, et al., Photo-realistic single image super-resolution using a generative adversarial network, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2017, pp. 4681–4690.

[7] Z. Wang, B. Chen, H. Zhang, H. Liu, Variational probabilistic generative framework for single image super-resolution, Signal Process. 156 (2019) 92–105.

[8] M.-Y. Liu, T. Breuel, J. Kautz, Unsupervised image-to-image translation networks, Adv. Neural Inf. Process. Syst. 30 (2017).

[9] S. Iizuka, E. Simo-Serra, H. Ishikawa, Globally and locally consistent image completion, ACM Trans. Graph. (ToG) 36 (4) (2017) 1–14.

[10] G. Tang, L. Sun, X. Mao, S. Guo, H. Zhang, X. Wang, Detection of GAN-synthesized image based on discrete wavelet transform, Secur. Commun. Netw. 2021 (2021).

[11] K. Chandrasegaran, N.-T. Tran, N.-M. Cheung, A closer look at fourier spectrum discrepancies for cnn-generated images detection, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2021, pp. 7200–7209.

[12] F. Marra, D. Gragnaniello, L. Verdoliva, G. Poggi, Do gans leave artificial fingerprints? in: 2019 IEEE Conference on Multimedia Information Processing and Retrieval, MIPR, IEEE, 2019, pp. 506–511.

[13] S. Kiruthika, V. Masilamani, Image quality assessment based fake face detection, Multimedia Tools Appl. (2022).

[14] J. Deng, X. Zhang, H. Chen, L. Wu, BGT: A blind image quality evaluator via gradient and texture statistical features, Signal Process., Image Commun. 96 (2021) 116315.

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

[16] F. Marra, D. Gragnaniello, D. Cozzolino, L. Verdoliva, Detection of gan-generated fake images over social networks, in: 2018 IEEE Conference on Multimedia Information Processing and Retrieval, MIPR, IEEE, 2018, pp. 384–389.

[17] 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.

[18] S. Tariq, S. Lee, H. Kim, Y. Shin, S.S. Woo, Detecting both machine and human created fake face images in the wild, in: Proceedings of the 2nd International Workshop on Multimedia Privacy and Security, 2018, pp. 81–87.

[19] C.-C. Hsu, Y.-X. Zhuang, C.-Y. Lee, Deep fake image detection based on pairwise learning, Appl. Sci. 10 (1) (2020) 370.

[20] L. Zhao, M. Zhang, H. Ding, X. Cui, MFF-net: Deepfake detection network based on multi-feature fusion, Entropy 23 (12) (2021) 1692.

[21] Y. Yu, R. Ni, W. Li, Y. Zhao, Detection of AI-manipulated fake faces via mining generalized features, ACM Trans. Multimed. Comput., Commun. Appl. (TOMM) 18 (4) (2022) 1–23.

[22] J. Yang, S. Xiao, A. Li, G. Lan, H. Wang, Detecting fake images by identifying potential texture difference, Future Gener. Comput. Syst. 125 (2021) 127–135.

[23] Z. Guo, G. Yang, J. Chen, X. Sun, Fake face detection via adaptive manipulation traces extraction network, Comput. Vis. Image Underst. 204 (2021) 103170.

[24] H. Mo, B. Chen, W. Luo, Fake faces identification via convolutional neural network, in: Proceedings of the 6th ACM Workshop on Information Hiding and Multimedia Security, 2018, pp. 43–47.

[25] D. Afchar, V. Nozick, J. Yamagishi, I. Echizen, Mesonet: a compact facial video forgery detection network, in: 2018 IEEE International Workshop on Information Forensics and Security, WIFS, IEEE, 2018, pp. 1–7.

[26] W. Quan, K. Wang, D.-M. Yan, X. Zhang, Distinguishing between natural and computer-generated images using convolutional neural networks, IEEE Trans. Inf. Forensics Secur. 13 (11) (2018) 2772–2787.

[27] S.-Y. Wang, O. Wang, R. Zhang, A. Owens, A.A. Efros, Cnn-generated images are surprisingly easy to spot... for now, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2020, pp. 8695–8704.

[28] Z. Mi, X. Jiang, T. Sun, K. Xu, Gan-generated image detection with self-attention mechanism against gan generator defect, IEEE J. Sel. Top. Sign. Proces. 14 (5) (2020) 969–981.

[29] 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.

[30] 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.

[31] 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.

[32] J. Deng, X. Zhang, H. Chen, L. Wu, BGT: A blind image quality evaluator via gradient and texture statistical features, Signal Process., Image Commun. 96 (2021) 116315.

[33] 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.

[34] 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.

[35] 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.

[36] 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.

[37] M.P. Gangan, K. Anoop, V. Lajish, Distinguishing natural and computer generated images using multi-colorspace fused EfficientNet, J. Inf. Secur. Appl. 68 (2022) 103261.

[38] H. Li, B. Li, S. Tan, J. Huang, Detection of deep network generated images using disparities in color components, 2018, 1808, pp. 1–13.

[39] X. Zhang, S. Karaman, S.-F. Chang, Detecting and simulating artifacts in gan fake images, in: 2019 IEEE International Workshop on Information Forensics and Security, WIFS, IEEE, 2019, pp. 1–6.

[40] A. Radford, L. Metz, S. Chintala, Unsupervised representation learning with deep convolutional generative adversarial networks, 2015, arXiv preprint arXiv:1511.06434.

[41] J. Zhao, M. Mathieu, Y. LeCun, Energy-based generative adversarial network, 2016, arXiv preprint arXiv:1609.03126.

[42] M. Arjovsky, S. Chintala, L. Bottou, Wasserstein generative adversarial networks, in: International Conference on Machine Learning, PMLR, 2017, pp. 214–223.

[43] P. Isola, J.-Y. Zhu, T. Zhou, A.A. Efros, Image-to-image translation with conditional adversarial networks, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2017, pp. 1125–1134.

[44] J.-Y. Zhu, T. Park, P. Isola, A.A. Efros, Unpaired image-to-image translation using cycle-consistent adversarial networks, in: Proceedings of the IEEE International Conference on Computer Vision, 2017, pp. 2223–2232.

[45] A. Brock, J. Donahue, K. Simonyan, Large scale GAN training for high fidelity natural image synthesis, 2018, arXiv preprint arXiv:1809.11096.

[46] T. Park, M.-Y. Liu, T.-C. Wang, J.-Y. Zhu, Semantic image synthesis with spatially-adaptive normalization, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2019, pp. 2337–2346.

[47] W. Luo, J. Huang, G. Qiu, JPEG error analysis and its applications to digital image forensics, IEEE Trans. Inf. Forensics Secur. 5 (3) (2010) 480–491.

[48] J. Galbally, S. Marcel, Face anti-spoofing based on general image quality assessment, in: 2014 22nd International Conference on Pattern Recognition, IEEE, 2014, pp. 1173–1178.

[49] D. Wen, H. Han, A.K. Jain, Face spoof detection with image distortion analysis, IEEE Trans. Inf. Forensics Secur. 10 (4) (2015) 746–761.

[50] 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.

[51] T. Fu, M. Xia, G. Yang, Detecting GAN-generated face images via hybrid texture and sensor noise based features, Multimedia Tools Appl. 81 (18) (2022) 26345–26359.

[52] S. McCloskey, M. Albright, Detecting GAN-generated imagery using saturation cues, in: 2019 IEEE International Conference on Image Processing, ICIP, IEEE, 2019, pp. 4584–4588.

[53] F. Matern, C. Riess, M. Stamminger, Exploiting visual artifacts to expose deepfakes and face manipulations, in: 2019 IEEE Winter Applications of Computer Vision Workshops, WACVW, IEEE, 2019, pp. 83–92.

[54] L. Nataraj, T.M. Mohammed, B. Manjunath, S. Chandrasekaran, A. Flenner, J.H. Bappy, A.K. Roy-Chowdhury, Detecting GAN generated fake images using co-occurrence matrices, Electron. Imaging 2019 (5) (2019) 532–1.

[55] C. Peng, T. Sun, Z. Mi, L. Yao, et al., Manipulated faces detection with adaptive filter, Secur. Commun. Netw. 2022 (2022).

[56] A.R. Rivera, J.R. Castillo, O.O. Chae, Local directional number pattern for face analysis: Face and expression recognition, IEEE Trans. Image Process. 22 (5) (2012) 1740–1752.
[57] C. Shan, S. Gong, P.W. McOwan, Facial expression recognition based on local binary patterns: A comprehensive study, Image Vis. Comput. 27 (6) (2009) 803–816.
[58] J. Kodovsky, J. Fridrich, V. Holub, Ensemble classifiers for steganalysis of digital media, IEEE Trans. Inf. Forensics Secur. 7 (2) (2011) 432–444.
[59] C.-C. Chang, C.-J. Lin, LIBSVM: A library for support vector machines, ACM Trans. Intell. Syst. Technol. 2 (3) (2011) 1–27.
[60] Y. Choi, M. Choi, M. Kim, J.-W. Ha, S. Kim, J. Choo, Stargan: Unified generative adversarial networks for multi-domain image-to-image translation, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2018, pp. 8789–8797.
[61] R. Wang, F. Juefei-Xu, L. Ma, X. Xie, Y. Huang, J. Wang, Y. Liu, Fakespotter: A simple yet robust baseline for spotting AI-synthesized fake faces, in: Proceedings of the Twenty-Ninth International Joint Conference on Artificial Intelligence, IJCAI '20, 2021.