VISVESVARAYA TECHNOLOGICAL UNIVERSITY,

BELAGAVI - 590018

A Technical Seminar Report on

Fake Colorized Image Detection

Submitted in partial fulfillment of the requirements as per VTU curriculum of

BACHELOR OF ENGINEERING

IN

COMPUTER SCIENCE & ENGINEERING

By

JYOTHI LAKSHMI **4AL16CS129**

Under the Guidance of

Dr.Mohideen S Badhusha

Senior Professor

DEPARTMENT OF COMPUTER SCIENCE & ENGINEERING

ALVA'S INSTITUTE OF ENGINEERING AND TECHNOLOGY

MOODBIDRI-574225, KARNATAKA

2019– 2020

ALVA'S INSTITUTE OF ENGINEERING AND TECHNOLOGY

MIJAR, MOODBIDRI D.K. -574225

KARNATAKA

DEPARTMENT OF COMPUTER SCIENCE & ENGINEERING

CERTIFICATE

This is to certify that **Jyothi Lakshmi, 4AL16CS129** has submitted Technical Seminar Report on " **Fake Colorized Image Detection** " for the VIII semester B.E. in Computer Science & Engineering during the year 2019-20

Dr.Mohideen S Badhusha	**Mr. Harish Kunder**	**Dr. Manjunath Kotari**
Seminar Guide	**Seminar Coordinator**	**Professor and Head**

ACKNOWLEDGEMENT

The satisfaction and euphoria that accompany a successful completion of any task would be incomplete without the mention of people who made it possible, success is the epitome of hard work and perseverance, but steadfast of all is encouraging guidance.

So, with gratitude I acknowledge all those whose guidance and encouragement served as beacon of light and crowned the effort with success.

I thank my seminar guide **Dr.Mohideen S Badhusha, Senior Professor** in Department of Computer Science & Engineering, who has been my source of inspiration. He has been especially enthusiastic in giving his valuable guidance and critical reviews.

The selection of Technical Seminar Topic as well as the timely completion is mainly due to the interest and persuasion of my seminar coordinator **Mr**. **Harish Kunder,** Associate Professor, Department of Computer Science & Engineering. I will remember his contribution for ever.

I sincerely thank, **Dr. Manjunath Kotari**, Professor and Head, Department of Computer Science & Engineering who has been the constant driving force behind the completion of the seminar.

I thank Principal **Dr.Peter Fernandes,** for his constant help and support throughout.

I am also indebted to **Management of Alva's Institute of Engineering and Technology, Mijar, Moodbidri** for providing an environment which helped me in completing the seminar.

Also, I thank all the teaching and non-teaching staff of Department of Computer Science & Engineering for the help rendered.

Finally I would like to thank my parents and friends whose encouragement and support was invaluable

JYOTHI LAKSHMI

4AL16CS129

TABLE OF CONTENTS

ABSTRACT

Image forensics aims to detect the manipulation of digital images. Currently, splicing detection, copy-move detection, and image retouching detection are attracting significant attention from researchers. However, image editing techniques develop over time. An emerging image editing technique is colorization, in which gray scale images are colorized with realistic colors. Unfortunately, this technique may also be intentionally applied to certain images to confound object recognition algorithms. To the best of our knowledge, no forensic technique has yet been invented to identify whether an image is colorized. It is observed that, compared with natural images, colorized images, which are generated by three state-of-the-art methods, possess statistical differences for the hue and saturation channels.

Besides, it is also observed statistical inconsistencies in the dark and bright channels, because the colorization process will inevitably affect the dark and bright channel values. Based on observations, i.e., potential traces in the hue, saturation, dark, and bright channels, it is proposed two simple yet effective detection methods for fake colorized images: Histogram-based fake colorized image detection and feature encoding-based fake colorized image detection. Experimental results demonstrate that both proposed methods exhibit a decent performance against multiple state-of-the-art colorization approaches.

LIST OF FIGURES

LIST OF TABLES

Chapter 1

INTRODUCTION

The rapid proliferation of image editing technologies has increased both the ease with which images can be manipulated and the difficulty in distinguishing between altered and natural images. In addition to the conventional image editing techniques such as splicing, copy-move and retouching, more image editing techniques, such as colorization and image generation, are proposed. Since these types of image editing techniques generate new content with/without references, it is denoted as the generative image editing techniques. Although image editing techniques can provide significant aesthetic or entertainment value, they may also be used with malicious intent. In general, various image editing approaches employ different mechanisms. Splicing and copy-move techniques usually manipulate part of the image and perform object-level changes. Image retouching techniques usually change the images via a variety of mechanisms. For example, contrast enhancement adjusts the contrast of the image, while image inpainting usually fills the holes in images according to the image content. Among the generative image editing techniques, image generation usually produces a meaningful image from a noise vector with/without some additional information such as text or a class label. Colorization, on the other hand, usually colorizes images with visually plausible colors, which may cause misjudgement when specific objects/scenes must be identified/tracked.

Fortunately, numerous image forensic technologies have been developed in the past decades. According to their mechanisms and applications, they can be categorized into two classes, active techniques and passive techniques. The active techniques usually refer to watermarking techniques, which embed authentication information in the to-be-protected images. When the integrities of these images demand verification, watermark extraction procedures are performed and the extracted watermarks are compared to the original watermark to detect forgeries. Since the active techniques require the watermark to be embedded prior to detection, the applications, in practice, are limited. In contrast, passive image forgery detection approaches, to which our proposed methods belong, usually detect the manipulations to the input images directly. Traditionally, passive image forgery (editing) detection techniques have mainly focused on splicing detection, copy-move detection and image retouching detection. It is observed that no method has yet been developed to detect the fake images generated by generative image editing techniques.

1.1 STATEMENT OF THE PROBLEM

If these images are examined by humans, the cost increases drastically as the number of to-be-examined images increases. Obviously, detection via human eyes is insufficient for the big data era. On the other hand, conventional image forgery detection techniques are designed with different assumptions that may not be appropriate for generative fake image detection. Therefore, generative fake image detection demands specific studies and designs. Among the different generative image forgery techniques, colorization already achieves excellent performances. As shown in Fig. 1.1, fake colorized images, which are generated by a state-of-the-art algorithm, are visually indistinguishable, if no ground-truth images exist for comparison. Therefore, the necessity to develop a scheme for fake colorized image detection increases rapidly. It is aimed to address this new problem by providing feasible solutions. Specifically, it has been proposed two simple yet effective schemes for detecting fake colorized images, which are generated by fully automatic colorization methods.

Fig.1.1. (a) Real Images. (b) Fake colorized Images.

1.2 OBJECTIVES

- It is observed that fake colorized images and their corresponding natural images exhibit statistical differences, which can be further utilized as detection traces, in both color channels and image prior. The color channels involved are the hue and saturation channels, while the exploited extreme channels prior is proposed.

- According to the statistical differences in the color channels and image priors, it is proposed that a fake colorized image detection scheme, named Histogram based Fake Colorized Image Detection (FCID-HIST), by proposing four detection features. Each feature calculates the most distinctive bin and the total variation of the normalized histogram distribution for hue, saturation, dark and bright channels, respectively.
- To better utilize the statistical information of the training images, it is considered by exploiting the divergences inside different moments of the data vectors and propose a fake colorized image detection scheme, named Feature Encoding based Fake Colorized Image Detection (FCID-FE), by modelling the created four-dimensional samples with a Gaussian mixture model (GMM) [10] and encoding the samples into Fisher feature vectors.
- The two proposed methods demonstrate a decent performance in various tests for detecting fake images generated by three state-of-the-art colorization methods.

1.3 NEED FOR THE STUDY

1.3.1 Review of Forgery Detection

Forgery detection has been investigated for decades. In general, forgery detection explores different characteristics of images and attempts to find traces to analyse. As mentioned above, most of the traditional forgery detection techniques can be categorized into three classes, copy-move detection, splicing detection and image retouching detection. Copy-move detection relies on identifying duplicated regions in a tampered image. Intuitively, these techniques tend to seek an appropriate feature in a certain domain, such that the detection can be performed via searching the most similar two units (such as patches). Different methods usually exploit different features. Reference [13] explores features in the frequency domain by dividing the image into overlapping blocks and detects the copy-move forgery via matching the quantized discrete cosine transform (DCT) coefficients. Reference [14] performs a rotation invariant detection based on the Fourier-Mellin transform. Reference [15] localizes the duplicated regions based on the Zernike moments, which exhibit the rotation invariance property, of small image blocks. Reference [15] reports decent results especially when the duplicated regions are smooth. Reference [16] employs the famous SIFT feature [17] to detect multiple duplicated regions and estimates the geometric transformation performed by the copy-move operation. Reference [18] presents a SIFT based detection method by matching the SIFT features via a broad first search neighbours clustering algorithm and distinguishing the duplicated origins from the tampered regions via CFA features. Reference [19]

introduces a hierarchical SIFT-based key point matching technique to solve a drawback of previous key point matching-based detection techniques, which tends to give poor performances when the copy-moved regions are small or smooth. Although copy-move detection technologies have been developed rapidly, they cannot be directly applied to the fake colorized image detection because no copy-move operations exist in the fake colorized images. Splicing detection usually detects the manipulated regions which originate from different source images. Different from copy-move detection, these approaches detect the tampered regions with various traces (features), which usually reveal the inconsistencies between the tampered regions and the unchanged regions. Currently, splicing detection can be classified into four categories, compression-based methods, camera-based methods, physics-based methods and geometry-based methods, according to their mechanisms. Compression-based methods assume that the spliced region and the original image have undergone different types of image compression and may exhibit different compression artifacts. For example, [20] considers the DCT coefficient distributions of each 8×8 block and computes the tampering probability. By considering the advantages and disadvantages of different block sizes, [21] constructs a multiscale scheme, employs the Benford's law at each level and fuses the results together to obtain a final localization map. Unfortunately, the compression-based methods are not appropriate for fake colorized image detection because the assumption may not always be valid.

Camera-based methods consider traces left on the image during the capturing process. Reference [22] detects the existence of the CFA artifacts, which are due to the demosaicking process in the CFA cameras, and thus obtains the localization map. Reference [23] exploits the photo-response nonuniformity noises (i.e., the sensor noises) of the camera to distinguish the tampered regions from the original ones. Reference [24] also considers the photo-response non-uniformity noises and a multiscale framework to conduct a multiscale analysis and detects small forgeries more accurately. Even if the camera-based methods can be employed to detect the fake colorized images, their robustness is incompetent because the sensor noises and the artifacts can easily be affected by noises and some common post-processing operations such as compression.

Physics-based methods perform detection based on different physics phenomenon inconsistencies. Reference [25] considers the blur type inconsistency between the spliced region and the original image to localize the tampered region. Reference [26] explores the illuminant-based transform spaces and combines different image descriptors, such as color, shape and texture, to detect forged regions. Since the fake colorized images to be examined in this paper are forged for the whole image, these inconsistencies cannot be utilized to detect the fake colorized images.

Geometry-based methods utilize the geometry information inside images for detection. Reference [27] explores detecting the compositions with the two-view geometrical constraints. Reference [28] considers the planar homographic in the test images and adopts graph-cut algorithm to obtain the final localization map. Unfortunately, since the geometrical characteristics are rarely manipulated in the fake colorized images, the geometry-based methods will also fail to detect the colorized images.

Image retouching detection usually considers that the original images are restored or enhanced. For example, [29] is designed to detect the inpainted images by considering the similarities, distances and number of identical pixels among different blocks. Reference [3] calculates the histograms and performs detection via the peak/gap artifacts induced from contrast enhancement. These techniques can hardly be applied to the new problem because their mechanisms are specially designed for their own assumptions.

The below table provides a summary of existing forensic techniques. Although many detection technologies have been developed, they are currently not directly applicable to the detection of images which are manipulated by generative methods. Specially designed techniques are necessary to address the detection of fake colorized images.

Table 1.1: SUMMARY OF THE EXISTING FAKE IMAGE DETECTION APPROACHES

Category	**Method**	**Core Mechanism**	**Potential result of detecting colorized images**
Copy-move detection	[13]	Quantized DCT coefficients	Not applicable
	[14]	Fourier-Mellin Transform	Not applicable
	[15]	Zernike moments	Not applicable
	[16]	SIFT feature	Not applicable
	[18]	SIFT and CFA features	Not applicable
	[19]	Hierarchical SIFT-based keypoint matching	Not applicable
Splicing detection	[20]	DCT coefficient distribution of each block	Not applicable
	[21]	Multiscale scheme based on Benford's law	Not applicable
	[22]	CFA artifacts	Possible but with low robustness
	[23]	PRNU noises	
	[24]	Multiscale scheme based on PRNU noioses	

	[25] [26] [27] [28]	Blur type inconsistency Illuminant-based transform spaces Two-view geometrical constraints Planar homographies	Possible but with low robustness Possible but with low robustness Not applicable Not applicable Not applicable Not applicable
Image retouching detection	[29] [3]	Block similarities and distances Peak/gap artifacts	Not applicable Not applicable

1.3.2 Review of Colorization

Colorization, a term describing the color adding process to grayscale images, was firstly introduced by Wilson Markle in 1970. However, this area began to develop rapidly in the 21st century. Colorization techniques can be categorized into the following types: scribble-based, example-based and fully automatic. Scribble-based methods are supervised techniques in which users begin assigning colors to pixels in the grayscale image. The milestone work [30], which assumes that the neighbouring pixels with similar intensities should have similar colors, is proposed at first. Various other approaches have been proposed in succession, such as [31], which constructs dictionaries for color and textures via sparse representation and colorizes the images accordingly.

Example-based methods [32], [33] usually require the users to supervise the system by providing reference color image(s) similar to the greyscale image. The system then transfers the colors in the reference color image(s) to the target greyscale image by searching for similar patterns/objects. The performances of these methods are dependent on the quality of the reference image(s). If the divergence between the greyscale image and the reference image(s) is high, the colorized result may be unsatisfactory. In contrast with the supervised approaches above, fully automatic methods require no supervision when performing the colorization task. Reference [34] trains a neural network and predicts the chrominance values by considering the pixel patch, DAISY and semantic features. Reference [35] colorizes the images by jointly utilizing the local and global priors with an end-to-end network. Reference [4] proposes a stateof-the-art approach, which exploits the hypercolumn to utilize both low-level and semantic representations, and colorizes the

images in the Hue-Chroma-Lightness (HCL) color space. Reference [36] calculates the statistical distributions of the chrominance information in the LAB space and introduces a classification-style colorization approach based on a deep network.

These techniques are briefly summarized in below table. Due to the high performances of the fully automatic colorization techniques, it is focused on the detection of the fake colorized images which are generated via following techniques.

Table 1.2: SUMMARY OF THE EXISTING COLORIZATION APPROACHES

Category	**Method**	**Core Mechanism**	**Side-information**
Scribble-based method	[30]	Neighboring pixels with similar intensities should have similar colors	User scribbles
	[31]	Construct color and texture dictionaries	User scribbles
Example-based method	[32]	Global optimization of colors at pixel-level	Reference color image
	[33]	Propogating the learned dictionaries	Reference color image
Fully automatic method	[34]	Network with pixel patch, DAISY and semantic features	None
	[35]	End-to-End network with local and global priors	None
	[4]	End-to-End network with Hypercolumn	None

Chapter 2

METHODOLOGY

The rapid progress in colorization technologies has enabled colorized images to be visually indistinguishable from natural images. State-of-the-art colorization methods are already capable of misleading human observers in the subjective tests. To distinguish the fake colorized images from the natural images, the study is made on the statistics of the fake colorized images, which are generated by three state-of-the-art methods and propose two simple yet effective detection schemes, FCID-HIST and FCID-FE.

2.1 Observations and Statistics

According to our observation, the colorized images tend to possess less saturated colors, and the colorization method favours some colors over others, though these differences are difficult to be visually perceived. Since the Hue-SaturationValue (HSV) color space separately represents the chrominance information in the hue and saturation channel, we calculate the normalized histograms (each containing 200 bins) of the hue and saturation channel in 15000 natural images and their corresponding fake colorized images, separately, as shown in Fig. 2.1.

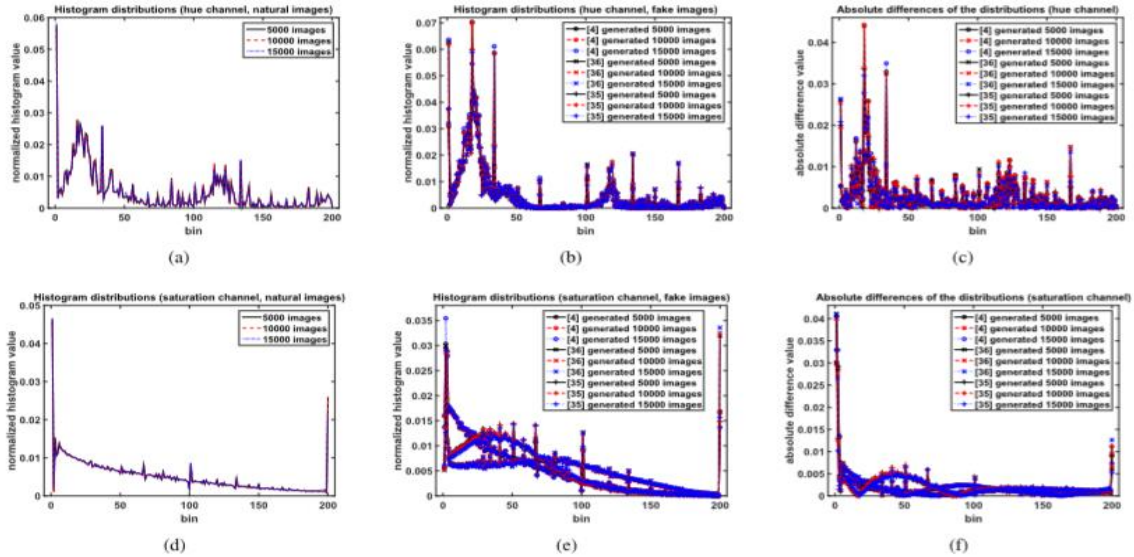


Fig.2.1 (a) Normalized histogram distribution of the hue channel (natural images). (b) Normalized histogram distribution of the hue channel (fake images). (c) Absolute differences of the distribution in (a) and (b). (d) Normalized histogram distribution of the saturationchannel (natural images). (e) Normalized histogram distribution of the saturationchannel (fake images). (f) Absolute differences of the distributions of the distributions in (d) and (e).

As shown in Fig. 2.1, the statistics of the natural and fake colorized images are different in both the hue and saturation channels, and there also exist statistical differences (especially for the peaks in the histograms) among the fake images generated by different colorization methods. For the hue channel, the histogram of the fake images tends to be smoother and possesses more significant peaks compared to the natural images. For the saturation channel, the histogram of the fake images also exhibits different peak values and variances compared to the histogram of the natural images. These statistics indicate that the fake images favour different colors and possess saturation differences compared to the natural images. Therefore, the natural and fake colorized images are statistically identifiable, though the fake colorized images seemed visually indistinguishable. In addition to the statistical differences in the color channels, differences also exist in some image priors because they are not considered explicitly in the colorization process even though the deep neural networks possess good generalization ability.

It has been exploited the recently proposed extreme channels prior (ECP), which consists of the dark channel prior (DCP) and the bright channel prior (BCP). Intuitively, DCP assumes that the dark channel of a natural image is close to zero, while BCP assumes that the bright channel of a natural image is close to 255. The dark channel Idc and bright channel Ibc of an image I are defined as shown by the following equations, respectively.

$$I_{dc}(x) = \min_{y \in \Omega(x)} \left(\min_{c_p \in (r,g,b)} I_{c_p}(y) \right),$$
$$I_{bc}(x) = \max_{y \in \Omega(x)} \left(\max_{c_p \in (r,g,b)} I_{c_p}(y) \right),$$

……Eq (2.1)

where x stands for the pixel location, Icp denotes a color channel of I and (x) represents the local patch centered at the location x. Note that the local patch sizes here are identical to the settings in [9]. By calculating the histograms of the dark channel and bright channel of 15000 natural images and their corresponding fake colorized images separately, Fig. 2.2 presents the expected differences, especially for the peak values, and supports our observations above.

2.2 FCID-HIST

By exploiting the existing statistical differences, we propose the Histogram based Fake Colorized Image Detection (FCID-HIST) method to detect fake colorized images. In FCID-HIST, four detection features, the hue feature Fh, the saturation feature Fs, the dark channel feature Fdc and the bright channel feature Fbc,

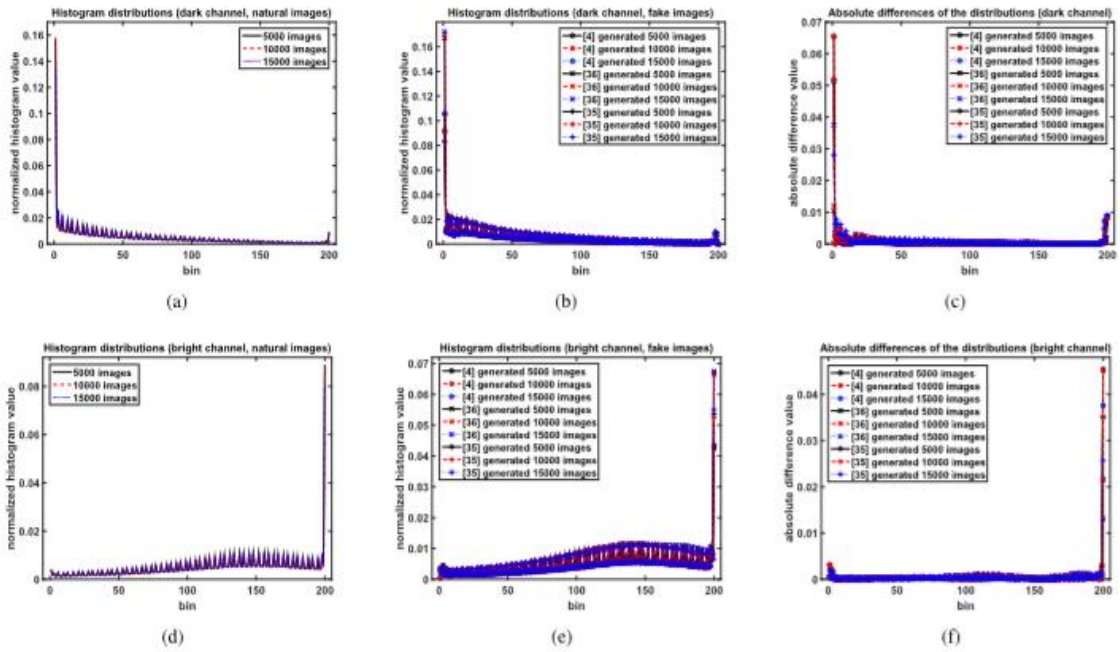


Fig.2.2. (a) Normalized histogram distribution of the dark channel (natural images). (b) Normalized histogram distribution of the dark channel (fake images). (c) Absolute differences of the distributions in (a) and (b). (d) Normalized histogram distribution of the bright channel (natural images). (e) Normalized histogram distributions of the bright channel (fake images). (f) Absolute differences of the distributions in (d) and (c).

are proposed to detect forgeries. The hue feature is constructed from the normalized hue channel histogram distributions. Let Kh be the total number of bins in each normalized hue channel histogram distribution. It is defined Disth,n and Disth,f as the normalized hue channel histogram distribution for the natural and fake training images, respectively, and Distα h as the corresponding histogram for the αth input image, which can be either a training or testing image. Intuitively, to differentiate the fake colorized images from the natural images, the distinctive features should reveal the largest divergences between the two types of images. (Note that the Euclidean distance is employed in this paper to calculate the divergences.)

Therefore, it is selected the most distinctive bin Distα h (υh), whose two corresponding bins in Disth,n and Disth, f give the largest divergence between the two histogram distributions, as part of the hue feature, as follows

$$F\alpha\ h\ (1) = Dist\alpha\ h\ (\upsilon h) \quad \text{.....Eq (2.2)}$$

where the index of the most distinctive bin υh for the hue channel is calculated via next equation.

$$\upsilon h = \text{argmax } x \, ||Disth,n(x) - Disth, f(x)||$$

$$= \text{argmax } x \, |Disth,n(x) - Disth, f(x)| \quad \text{.....Eq (2.3)}$$

The distributions Disth,n and Disth, f also vary differently with respect to the bins. It is taken into account for this difference in the hue feature by computing the first order derivative of the normalized hue channel histogram distribution

$$DistD\alpha\ h\ (l) = Dist\ \alpha\ h\ (l+1) - Dist\alpha\ h \quad \text{.....Eq (2.4)}$$

to capture the varying trend of the histogram distribution. This total variation is calculated as below equation shows.

$$F_h^{\alpha}(2) = \sum_{l=1}^{K_h - 1} |Dist D_h^{\alpha}(l)| \quad \text{.....Eq (2.5)}$$

The proposed hue feature is then formed by combining the above equations into a vector, as below equation demonstrates.

$$F_h^{\alpha} = [F_h^{\alpha}(1) \quad F_h^{\alpha}(2)] \quad \text{.....Eq (2.6)}$$

Similarly, the saturation feature Fα s , the dark channel feature Fα dc and the bright channel feature Fα bc can be constructed by utilizing the normalized histogram distributions (Dists,n, Dists,f), (Distdc,n, Distdc,f), and (Distbc,n, Distbc, f) for the saturation, bright, and dark channels of the training images respectively. In the same manner as Eq. 2.3, the indexes for the most distinctive bins υs, υdc and υbc can be calculated by below equation.

$$\upsilon_{c_h} = \underset{x}{\text{argmax}} |Dist_{c_h,n}(x) - Dist_{c_h,f}(x)|, \quad c_h = s, dc, bc \quad \text{.....Eq (2.7)}$$

Then, the most distinctive bins for each feature can be calculated via below equation.

$$F_{c_h}^{\alpha}(1) = Dist_{c_h}^{\alpha}\left(\underset{x}{\text{argmax}} |Dist_{c_h,n}(x) - Dist_{c_h,f}(x)|\right), \quad c_h = s, dc, bc \quad \text{.....Eq (2.8)}$$

$$F_{c_h}^{\alpha}(2) = \sum_{l=1}^{K_{c_h}-1} |Dist D_{c_h}^{\alpha}(l)|, \quad c_h = s, dc, bc$$

……..Eq (2.9)

where Distα ch represents the normalized ch channel histogram distribution of the αth input image. The total variation of each distribution is computed via below equation.

where Kch stands for the total number of bins in each normalized ch channel histogram distribution and DistDα ch denotes the first order derivative of the normalized ch channel histogram distribution. Then, the features are formed as shown in below equation.

$$F_{c_h}^{\alpha} = [F_{c_h,0}^{\alpha} \; F_{c_h,1}^{\alpha}], \quad c_h = s, dc, bc$$

…….Eq (2.10)

With all the features calculated, the final detection feature FαHIST for the αth input image can be constructed via below equation.

$$F_{HIST}^{\alpha} = [F_h^{\alpha} \; F_s^{\alpha} \; F_{dc}^{\alpha} \; F_{bc}^{\alpha}]$$

……Eq (2.11)

After the detection feature is calculated, FCID-HIST employs the supporting vector machine (SVM) for training and detecting the fake colorized images. The FCID-HIST algorithm is summarized as shown in Algorithm 1. For convenience, it is been used Kh = Ks = Kdc = Kbc .

2.3 FCID-FE

Although FCID-HIST gives a decent performance in the experiments, which are demonstrated in the latter section, these features may not fully utilize the statistical differences between the natural and fake colorized images because the distributions are modelled channel by channel. Therefore, it is proposed another scheme, Feature Encoding based Fake Colorized Image Detection (FCID-FE), to better exploit the statistical information by jointly modelling the data distribution and exploiting the divergences inside different moments of the distribution. Let Iβ h , Iβ s , Iβ dc and Iβ bc be the hue, saturation, dark and bright channels of a training image respectively, where β is the index of the training image. Then, it is created a training sample set via below equation.

$$\Phi\big((z-1)*i*j+(i-1)*j+j\big) = [I_h^{\beta}(i,j)\ I_s^{\beta}(i,j)\ I_{dc}^{\beta}(i,j)\ I_{bc}^{\beta}(i,j)]$$

……..Eq (2.12)

In contrast to the histogram modelling, FCID-FE models the sample data distribution G with a Gaussian mixture model (GMM) [10] as shown in the below equation.

$$G(\Phi|\Theta) = \sum_{n=1}^{N} \log p(\Phi_n|\Theta)$$

……Eq (2.13)

where N is the number of samples in stands for the parameter set of the constructed GMM and is defined in below equation.

$$\Theta = \omega_a,\quad \mu_a,\ \sigma_a,\ a = 1,\ldots,N_m,\ \sum_{n=1}^{N_m}\omega_a = 1$$

…… Eq (2.14)

where ωa represents the weight, μa stands for the mean value vector, σa denotes the covariance matrix and Nm is the number of Gaussian distributions in the distribution model. Then, the likelihood of n being modelled by the GMM can be represented by below equation.

$$p(\Phi_n|\Theta) = \sum_{m=1}^{N_m} \log \omega_m\, p_m(\Phi_m|\Theta)$$

…….. Eq (2.15)

By the following equation,

$$p_m(\Phi_m|\Theta) = \frac{\exp[-(1/2)(\Phi_m-\mu_a)^T\sigma_a^{-1}(\Phi_m-\mu_a)]}{(2\pi)^{N_v/2}|\sigma_a|^{1/2}}$$

…….. Eq(2.16)

where Nv denotes the number of dimensions of each sample vector. Then, GMM can be constructed by determining the parameter.

With the determined GMM, FCID-FE utilizes different moments of the distribution and encodes each subset of the sample vectors, which belongs to each training image, into training Fisher vectors [11] as expressed by Eq (2.17).

$$F_{FE}^{\beta} = [\frac{\lambda_1 \delta G(\Phi^{\beta}|\Theta)}{\delta \omega_a} \quad \frac{\lambda_2 \delta G(\Phi^{\beta}|\Theta)}{\delta \mu_{a,v}} \quad \frac{\lambda_3 \delta G(\Phi^{\beta}|\Theta)}{\delta \sigma_{a,v}}] \quad \text{…….. Eq (2.17)}$$

where v = 1, 2, ……Nv and λ1, λ2 and λ3

Training Stage:
Input: N_1 natural and fake colorized training images, the corresponding labels $L_{r,HIST}$, K_h, K_s, K_{dc}, K_{bc}, SVM parameters
Output: v_h, v_s, v_{dc}, v_{bc}, trained SVM classifier

Compute $Dist_{h,n}$, $Dist_{s,n}$, $Dist_{dc,n}$, $Dist_{bc,n}$
Compute $Dist_{h,f}$, $Dist_{s,f}$, $Dist_{dc,f}$, $Dist_{bc,f}$
Compute v_h, v_s, v_{dc}, v_{bc} ▷ refer to Eq. 4 and 7
for $i = 1$ to N_1 **do**
    Compute $Dist_h^i$, $Dist_s^i$, $Dist_{dc}^i$, $Dist_{bc}^i$
    Compute $F_h^i(1)$, $F_s^i(1)$, $F_{dc}^i(1)$, $F_{bc}^i(1)$ ▷ refer to Eq. 3 and 8
    Compute $F_h^i(2)$, $F_s^i(2)$, $F_{dc}^i(2)$, $F_{bc}^i(2)$ ▷ refer to Eq. 5 and 9
    Compute F_h^i, F_s^i , F_{dc}^i, F_{bc}^i ▷ refer to Eq. 6 and 10
    Compute F_{HIST}^i ▷ refer to Eq. 11
end for
Train SVM with F_{HIST}, $L_{r,HIST}$ and SVM parameters

Testing Stage:
Input: N_2 test images, K_h, K_s, K_{dc}, K_{bc}, v_h, v_s, v_{dc}, v_{bc}, trained SVM classifier
Output: Detection labels $L_{e,HIST}$

for $i = 1$ to N_2 **do**
    Compute $Dist_h^i$, $Dist_s^i$, $Dist_{dc}^i$, $Dist_{bc}^i$
    Compute $F_h^i(1)$, $F_s^i(1)$, $F_{dc}^i(1)$, $F_{bc}^i(1)$ ▷ refer to Eq. 3 and 8
    Compute $F_h^i(2)$, $F_s^i(2)$, $F_{dc}^i(2)$, $F_{bc}^i(2)$ ▷ refer to Eq. 5 and 9
    Compute F_h^i, F_s^i , F_{dc}^i, F_{bc}^i ▷ refer to Eq. 6 and 10
    Compute F_{HIST}^i ▷ refer to Eq. 11
    Obtain $L_{e,HIST}(i)$ with F_{HIST}^i and the trained SVM clasifier
end for

ALGORITHM 2.1

$$\lambda_1 = \left(N\left(\frac{1}{\omega_a} + \frac{1}{\omega_1}\right)\right)^{-1/2}$$
$$\lambda_2 = \left(\frac{N\omega_a}{(\sigma_{a,v})^2}\right)^{-1/2}$$
$$\lambda_3 = \left(\frac{2N\omega_a}{(\sigma_{a,v})^2}\right)^{-1/2}$$

…… Eq (2.18)

Training Stage:
Input: N_3 natural and fake colorized training images, the corresponding labels $L_{r,FE}$, SVM parameters
Output: Θ, trained SVM classifier

1: Create samples Φ ▷ refer to Eq. 12
2: Estimate GMM model Θ from Φ
3: **for** $i = 1$ to N_3 **do**
4: Encode Φ^i to F_{FE}^i with Θ ▷ refer to Eq. 17
5: **end for**
6: Train SVM with F_{FE}, $L_{r,FE}$ and SVM parameters

Testing Stage:
Input: N_4 test images, Θ, trained SVM classifier
Output: Detection labels $L_{e,FE}$

1: Create samples Φ ▷ refer to Eq. 12
2: **for** $i = 1$ to N_4 **do**
3: Encode Φ^i to F_{FE}^i with Θ ▷ refer to Eq. 17
4: Obtain $L_{e,FE}(i)$ with F_{FE}^i and the trained SVM clasifier
5: **end for**

ALGORITHM 2.2

Then, SVM is employed as the training classifier. For testing, FCID-FE will first construct the test sample set for each input image. Next, the existing GMM from the training dataset is employed to encode each test image into the Fisher vector. At last, FCID-FE classifies these feature vectors via the trained SVM. The algorithm of FCID-FE is summarized in Algorithm 2.3.

2.4 Experimental Results

In this section, the experimental setups, evaluation measurements, databases and results are introduced accordingly.

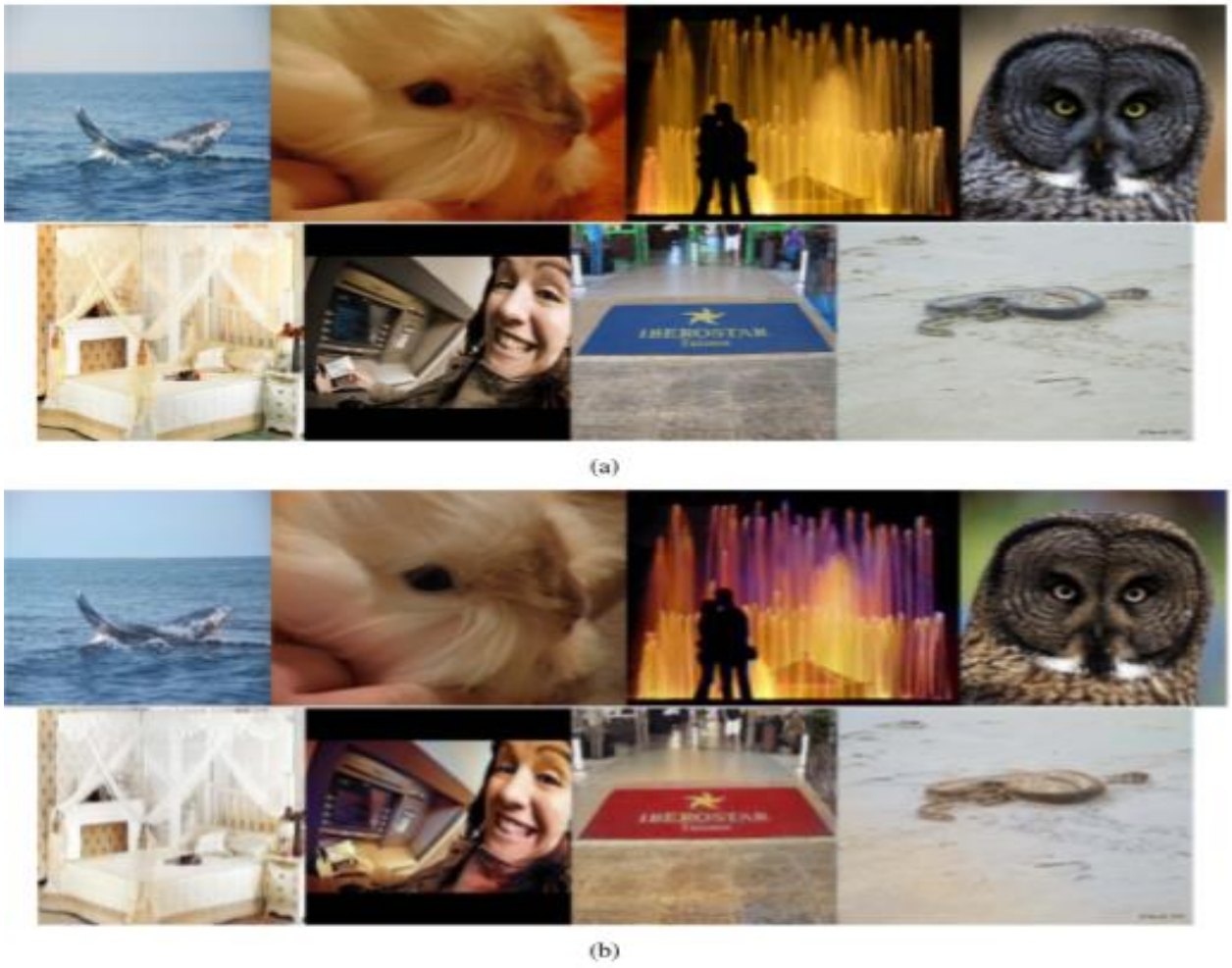


Fig.2.3. (a) Real Images. (b) Fake colorized Images.

2.4.1 Setups and Measurements

One implementation of SVM, the LIBSVM [38], is employed for classification and the RBF kernel is selected. The VLFeat software [39] is employed for GMM modelling and Fisher vector encoding. In this experiment, both the half total error rate (HTER) measurement and the receiver operating characteristic (ROC) curve (with the area under the curve (AUC) measurement) are

employed to evaluate the performances of the proposed methods. Denoting P, N, TPand TNas the positive samples, negative samples, true positive samples and true negative samples respectively, HTER is defined in below equation.

$$HTER = \frac{FPR + FNR}{2} = \frac{FP/(TN + FP) + FN/(TP + FN)}{2} \quad \text{……Eq (2.19)}$$

Note that the natural images and the fake colorized images are defined as the negative samples and the positive samples, respectively.

2.4.2 Databases

For a thorough evaluation of the proposed methods, different databases are employed/constructed for different experiments. The database D1 for parameter selection and validation by employing 10000 fake colorized images from the database ctest10k in [4] and their corresponding 10000 natural images from the ImageNet validation dataset are created.

The natural images in D1 include various types of images, such as animals, human, furniture and outdoor scenes. In addition to D1, different databases are also prepared to assess the performances of FCID-HIST and FCID-FE against different colorization methods. The database D2 consists of 2000 natural images randomly selected from the ImageNet validation dataset and their corresponding fake images, which are generated via [4]. The database D3 is constructed by randomly selecting 2000 fake colorized images from the results of [36] and 2000 corresponding natural images from the ImageNet validation dataset. The database D4, which contains 2000 natural images (randomly selected from the ImageNet validation dataset) and their corresponding generated fake images, is produced via employing the colorization approaches in [35]. Note that the selected natural images and their corresponding colorized images in D2-D4 are not overlapping with those in D1. Similarly, databases D5, D6 andD7 are constructed by randomly selecting 2000 natural images from the Oxford building dataset [41] and generating the corresponding colorized images with [4], [36], and [35], respectively. Note that the real images in the Oxford building dataset [41] contain various content provided by "Flickr".

2.4.3 Parameter Selection

Prior to evaluating the performances of FCID-HIST and FCID-FE against different colorization approaches, the optimal parameters of the proposed methods are tuned via experiments.

In the experiments, 1000 forged images and their corresponding natural images are randomly selected from database D1 to construct the parameter training (par-train) set, while another 1000 fake images and their corresponding natural images are selected from D1 to be the parameter testing (par-test) set. Note that the par-train set and the par-test set are not overlapping. Here, two parameters, c and g, which denote the cost and gamma in LIBSVM, are specifically tuned here via grid search. Tables 2.4.1 and 2.4.2 present the HTER results with different cs and gs for FCID-HIST and FCID-FE, respectively. As shown, FCID-HIST should select c =32, while FCID-FE should select c=2 for the parameter c. Since there exists multiple choices for g, for convenience and consistency, g =1/2 is selected for both FCID-HIST and FCID-FE.

Table 2.1

HTER OF FCID-HIST FOR DIFFERENT SVM PARAMETER SETTINGS (IN PERCENTAGE)

	g=1/64	g=1/32	g=1/16	g=1/8	g=1/4	g=1/2	g=1	g=2	g=4	g=8	g=16	g=32	g=64
c=1/64	24.30	24.30	24.30	24.30	24.30	24.30	24.30	24.05	23.20	23.40	23.95	24.60	26.15
c=1/32	24.30	24.30	24.30	24.30	24.30	24.30	24.30	24.05	23.20	23.40	23.95	24.60	26.15
c=1/16	24.30	24.30	24.30	24.30	24.30	24.30	24.30	24.05	23.20	23.40	23.95	24.60	26.15
c=1/8	24.30	24.30	24.30	24.30	24.30	24.30	24.30	24.05	23.20	23.40	23.95	24.60	26.15
c=1/4	24.30	24.30	24.30	24.30	24.30	24.30	24.30	24.05	23.20	23.40	23.95	24.60	26.15
c=1/2	24.30	24.30	24.30	24.30	24.30	24.30	24.30	24.05	23.20	23.40	23.95	24.60	26.15
c=1	24.30	24.30	24.30	24.30	24.30	24.30	24.30	24.05	23.20	23.40	23.95	24.60	26.15
c=2	23.55	23.55	23.55	23.55	23.55	23.55	23.55	22.65	22.65	23.05	23.50	24.45	26.55
c=4	22.90	22.90	22.90	22.90	22.90	22.90	22.90	22.15	22.20	22.80	23.90	25.15	27.80
c=8	22.30	22.30	22.30	22.30	22.30	22.30	22.30	22.20	22.50	22.85	23.70	25.95	28.55
c=16	22.00	22.00	22.00	22.00	22.00	22.00	22.00	21.75	21.90	23.25	24.40	26.75	29.10
c=32	**21.50**	**21.50**	**21.50**	**21.50**	**21.50**	**21.50**	**21.50**	21.65	22.15	24.20	24.95	27.75	30.55
c=64	21.65	21.65	21.65	21.65	21.65	21.65	21.65	22.15	22.50	24.10	25.75	28.30	31.00

Table 2.2

HTER OF FCID-FE FOR DIFFERENT SVM PARAMETER SETTINGS (IN PERCENTAGE)

	g=1/64	g=1/32	g=1/16	g=1/8	g=1/4	g=1/2	g=1	g=2	g=4	g=8	g=16	g=32	g=64
c=1/64	16.90	16.90	16.90	16.90	16.90	16.90	16.90	19.20	22.25	27.05	35.05	58.25	53.20
c=1/32	16.90	16.90	16.90	16.90	16.90	16.90	16.90	19.20	22.25	27.05	35.05	58.25	53.20
c=1/16	16.90	16.90	16.90	16.90	16.90	16.90	16.90	19.20	22.25	27.05	35.05	58.25	53.20
c=1/8	16.90	16.90	16.90	16.90	16.90	16.90	16.90	19.20	22.25	27.05	35.05	58.25	53.20
c=1/4	16.90	16.90	16.90	16.90	16.90	16.90	16.90	19.20	22.25	27.05	35.05	58.25	53.20
c=1/2	16.90	16.90	16.90	16.90	16.90	16.90	16.90	19.20	22.25	27.05	35.05	58.25	53.20
c=1	16.90	16.90	16.90	16.90	16.90	16.90	16.90	19.20	22.25	27.05	35.05	58.25	53.20
c=2	**16.65**	**16.65**	**16.65**	**16.65**	**16.65**	**16.65**	**16.65**	19.25	21.60	26.80	34.70	58.85	53.70
c=4	17.35	17.35	17.35	17.35	17.35	17.35	17.35	19.15	21.70	26.80	34.70	58.85	53.70
c=8	17.45	17.45	17.45	17.45	17.45	17.45	17.45	19.25	21.65	26.80	34.70	58.85	53.70
c=16	17.50	17.50	17.50	17.50	17.50	17.50	17.50	19.60	21.65	26.80	34.70	58.85	53.70
c=32	18.25	18.25	18.25	18.25	18.25	18.25	18.25	19.55	21.65	26.80	34.70	58.85	53.70
c=64	18.70	18.70	18.70	18.70	18.70	18.70	18.70	19.55	21.65	26.80	34.70	58.85	53.70

Next, is on the selection of the SVM threshold, which is important for the final classification step after the probabilities are estimated. In the test, the threshold varies from 0 to 1 with a step size of 0.01.

For each proposed method, a 10-fold cross threshold selection test is performed to obtain the optimal threshold by employing D1. Table 2.3 presents the optimal thresholds of each fold for FCID-HIST and FCID-FE.

Table 2.3

OPTIMAL THRESHOLD SELECTION OF FCID-HIST AND FCID-FE (THRESHOLD)

Method\Fold Number	**1**	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**9**	**10**
FCID-HIST	0.46	0.45	0.46	0.45	0.46	0.43	0.46	0.46	0.45	0.47
FCID-FE	0.5	0.51	0.52	0.43	0.47	0.54	0.45	0.5	0.49	0.51

Therefore, the optimal thresholds for FCID-HIST and FCID-FE, which are calculated via averaging the optimal thresholds of each fold, are 0.455 and 0.492, respectively. Note that the selected thresholds for FCID-HIST and FCID-FE will be employed in the subsequent experiments.

Since FCID-HIST exploits the histogram distributions to extract the detection features, the number of bins of the histograms Kcf, cf =h, s, dc, bc should be determined as well. Intuitively, when Kcf increases, part of the detection feature corresponding to the most distinctive bins may become less distinctive, while the rest of the detection feature corresponding to the total variations may capture more details and thus become more distinctive. To reveal the effects of Kcf, the partrain and par-test sets and the SVM parameters determined above are employed. In this test, Kcf, cf =h, s, dc, bc ranges from 200 to 260 with a step of 5. Besides, Kcf = 256, cf = h, s, dc, bc are also included.

As can be observed, there exists no obvious trends when Kcf varies. By considering the latter results demonstrated in Section 2.4-D, in which FCID-HIST gives unstable performances when the training dataset varies, it is believed that Kcf is not a deterministic aspect for the performances of FCID-HIST. Therefore, Kh = Ks = Kdc = Kbc are all set to be 200 for convenience.

2.4.4 Cross Validation

After the parameters are determined, the cross validations are performed on FCID-HIST and FCID-FE separately. Fig. 2.5 presents the cross validation results of FCID-HIST and FCID-FE.

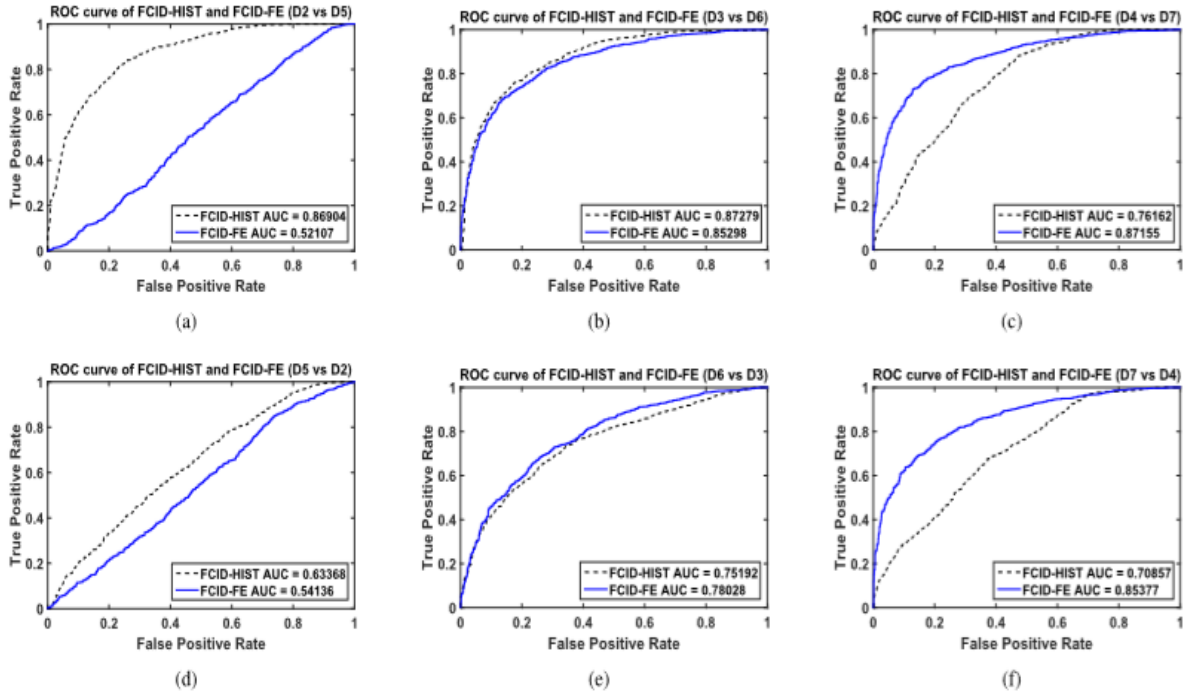


Fig.2.4. Detection results with different training sets

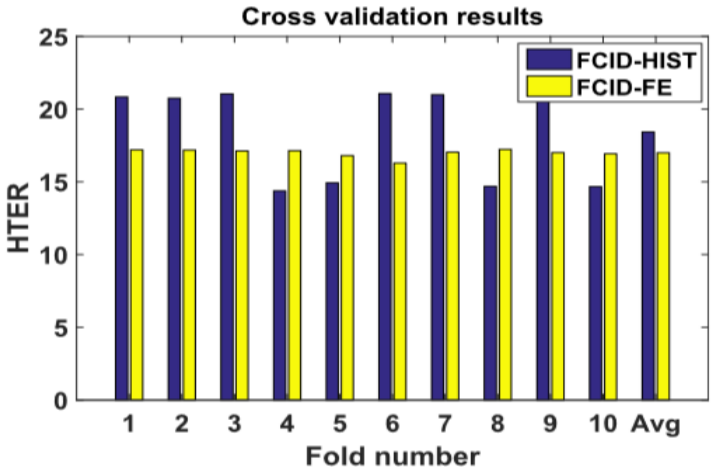


Fig.2.5. FCID-HIST and FCID-FE HTER results of 10-fold cross validation

As can be observed, both FCID-HIST and FCID-FE achieve a decent performance, where the average HTER of FCID-HIST is 18.423% and that of FCID-FE is 16.994%. Clearly, FCID-FE provides a slightly better performance compared to FCID-HIST. Note that FCID-HIST gives less consistent performances because the detection feature, especially the most distinctive bins, may vary for different training set. It indicates that the extracted handcrafted features in FCID-HIST possess less robustness compared to the moments-based features in FCID-FE.

The detection performances may be improved via exploring better and more consistent features in the future work. Compared to Figs. 2.6(a), 2.6(e) and 2.6(i), performance decreases when the training and testing datasets are from different databases, especially for FCID-HIST. These drops reveal that FCID-FE, which gives more consistent performances, models the statistical information of the images better compared to FCID-HIST.

2.4.5 Performance Evaluation

In the cross-validation tests, both FCID-HIST and FCID-FE performs decently. Here, a comprehensive performance evaluation for FCID-HIST and FCID-FE is performed with six additional databases D2, D3, D4, D5, D6 andD7. Since FCID-HIST and FCID-FE construct the feasible features automatically according to the training set, the proposed methods should be capable of detecting the fake images generated by different colorization methods, as long as the colorized images exhibit the observed differences. To demonstrate the performances of the detection methods against three latest colorization approaches, each of D2, D3 and D4 is equally divided into a training set and a testing set.

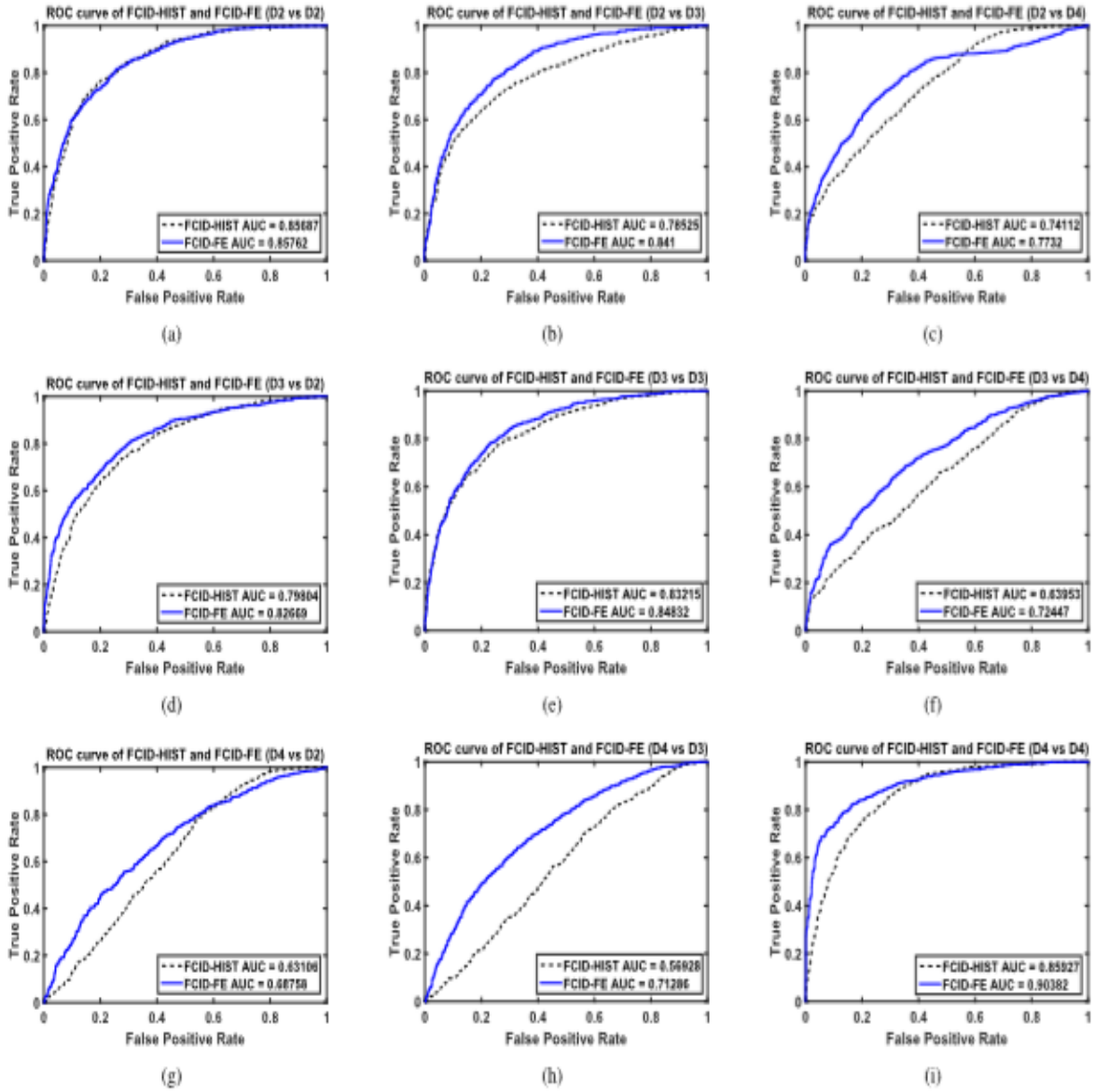


Fig.2.6 Detection results for the cross-validation method tests

The experiments are conducted in a manner that the training sets and testing sets may or may not originate from the identical databases, such that 9 experiments are performed to evaluate FCID-HIST and FCID-FE.

As can be observed from Tables 2.4-2.5 and Fig. 2.5, the proposed methods can successfully detect different fake images which are generated from different state-of-the-art colorization approaches, when the training and testing datasets are from the identical or different databases. Besides, FCID-FE gives more accurate detection results compared to FCID-HIST in most situations.

Table 2.4

HTER OF FCID-HIST FOR DIFFERENT DATABASES (IN PERCENTAGE)

Training\Testing	*D*2([4])	*D*3([36])	*D*4([35])
*D*2([4])	22.50	28.00	33.95
*D*3([36])	26.95	24.45	41.85
*D*4([35])	38.15	43.55	22.35

Table 2.5

HTER OF FCID-FE FOR DIFFERENT DATABASES (IN PERCENTAGE)

Training\Testing	*D*2([4])	*D*3([36])	*D*4([35])
*D*2([4])	22.30	23.65	31.70
*D*3([36])	25.10	22.85	34.25
*D*4([35])	38.50	36.15	17.30

Next, the cross-dataset tests are performed. The natural images in D2, D3 andD4, originating from the ImageNet validation dataset [40], and images in the D5, D6 and D7, originating from the Oxford building dataset [41], are employed to perform the cross-dataset tests. Similar to D2, D3 andD4, D5, D6 andD7 are all equally divided into training and testing sets. By pairing the databases in which the colorized images are generated from the same colorization method, three database pairs, D2 and D5, D3 andD6, D4 andD7, are obtained. For each pair of databases, the cross-dataset tests are performed by employing one database's training set and the other one's testing set, and vice versa. The experimental results of the cross-dataset tests are introduced in Tables 2.6-2.7 and Fig. 2.4.4.

Table 2.6

HTER OF FCID-HIST FOR CROSS-DATASET

TESTS (TRAINING VS TESTING)

D2([4]) vs. *D5*([4])	*D3*([36]) vs. *D6*([36])	*D4*([35]) vs. *D7*([35])
22.85	21.50	30.95
D5([4]) vs. *D2*([4])	*D6*([36]) vs. *D3*([36])	*D7*([35]) vs. *D4*([35])
43.45	30.75	36.60

Table 2.7

HTER OF FCID-FE FOR CROSS-DATASET

TESTS (TRAINING VS TESTING)

D2([4]) vs. *D5*([4])	*D3*([36]) vs. *D6*([36])	*D4*([35]) vs. *D7*([35])
51.40	22.70	20.20
D5([4]) vs. *D2*([4])	*D6*([36]) vs. *D3*([36])	*D7*([35]) vs. *D4*([35])
49.80	30.25	23.15

As shown, although the performance somewhat decreases, both methods still successfully differentiates between the colorized and natural images, and FCID-HIST again gives fewer stable performances compared to FCID-FE, with the exception of the D2 andD5 pair. The unsatisfactory performances for the D2 and D5 pair may be due to the different image content in different image datasets (D2 from the ImageNet dataset and D5 from the Oxford building dataset), which induces different statistical distributions. Since the proposed methods, especially FCID-FE, rely on extracting the detection features from the entire distributions, the classifier, which is trained by either D2 or D5, may fail to correctly classify certain images in the other one.

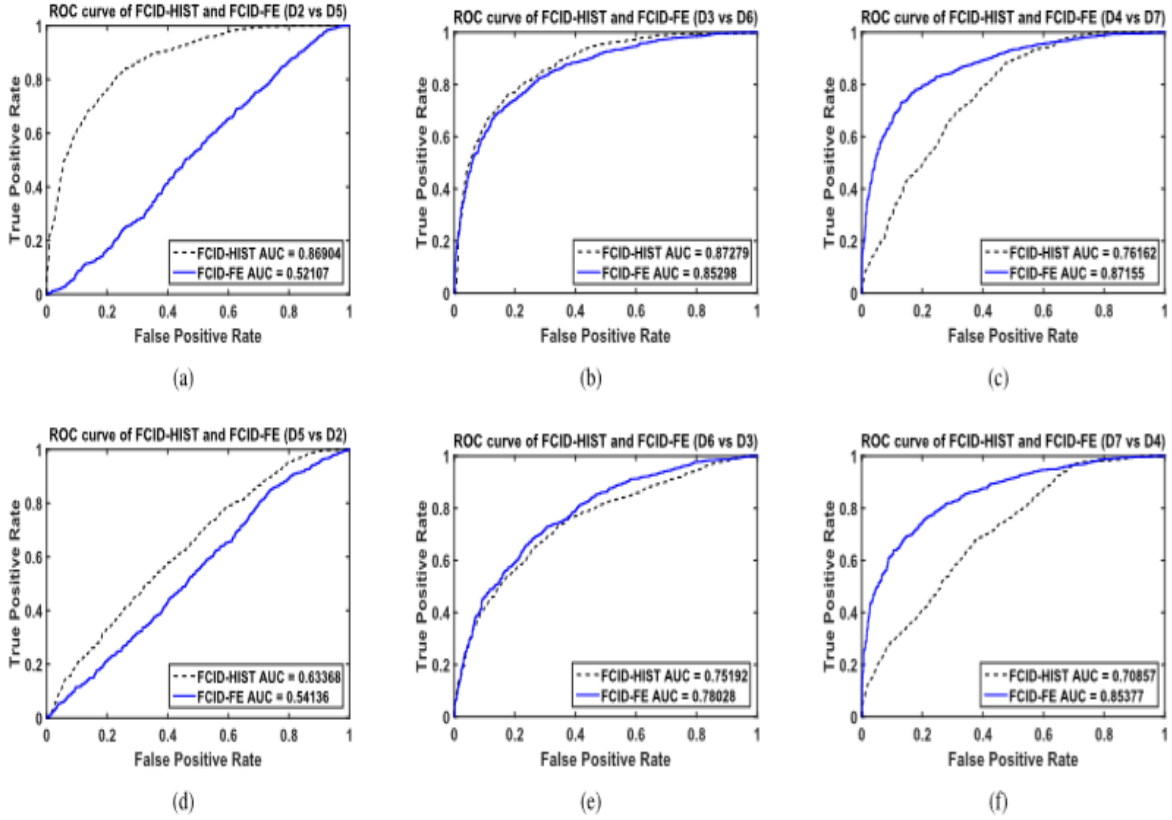


Fig.2.7. Detection results with different training vs testing sets.

Chapter 3

CONCLUSION

It is aimed to address a new problem in the field of fake image detection: fake colorized image detection. It is observed that fake colorized images and their corresponding natural images possess statistical differences in the hue, saturation, dark and bright channels. It is proposed two simple yet effective schemes, FCID-HIST and FCID-FE, to resolve this detection problem. FCID-HIST exploits the most distinctive bins and total variations of the normalized histogram distributions and creates features for detection, while FCID-FE models the data samples with GMM and creates Fisher vectors for better utilizing the statistical differences. It evaluates the performances of the proposed methods by selecting parameters for FCID-HIST and FCID-FE and detecting different fake images generated by state-of-the-art colorization approaches. The results demonstrate that both FCID-HIST and FCID-FE perform decently against different colorization approaches and FCID-FE provides more consistent and superior performances compared to FCID-HIST in most of the tests. Although the proposed FCID-HIST and FCID-FE give decent performances in the experiments, it is only a preliminary investigation, and there are many directions

Chapter 4

FUTURE STUDIES

For future studies that require further exploration. As the results indicate, the performance of the current methods sometimes degrades obviously when the training images and the testing images are generated from different colorization methods or different datasets, thus blind fake colorized image detection features and methods may be developed in the future by studying the common characteristics of different colorization methods. Moreover, better feature encoding approaches can be considered for improving performance, as well as the optimization of the detection features and parameters to improve the custom features constructed.

References

1) H. Farid, “Exposing digital forgeries from JPEG ghosts,” IEEE Trans. Inf. Forensics Security, vol. 4, no. 1, pp. 154–160, Mar. 2009.
2) J. Li, X. Li, B. Yang, and X. Sun, “Segmentation-based image copymove forgery detection scheme,” IEEE Trans. Inf. Forensics Security, vol. 10, no. 3, pp. 507–518, Mar. 2015.
3) G. Cao, Y. Zhao, R. Ni, and X. Li, “Contrast enhancement-based forensics in digital images,” IEEE Trans. Inf. Forensics Security, vol. 9, no. 3, pp. 515–525, Mar. 2014.
4) G. Larsson, M. Maire, and G. Shakhnarovich, “Learning representations for automatic colorization,” in Proc. Eur. Conf. Comput. Vis. (ECCV), 2016, pp. 577–593.
5) I. J. Goodfellow et al., “Generative adversarial nets,” in Proc. Adv. Neural Inf. Process. Syst. (NIPS), 2014, pp. 2672–2680.
6) F. Huang, X. Qu, H. J. Kim, and J. Huang, “Reversible data hiding in JPEG images,” IEEE Trans. Circuits Syst. Video Technol., vol. 26, no. 9, pp. 1610–1621, Sep. 2016.
7) J. Yin, R. Wang, Y. Guo, and F. Liu, “An adaptive reversible data hiding scheme for JPEG images,” in Proc. Int. Workshop Digit.-Forensics Watermarking, 2016, pp. 456–469.
8) J. Wang, S. Lian, and Y.-Q. Shi, “Hybrid multiplicative multiwatermarking in DWT domain,” Multidimensional Syst. Signal Process., vol. 28, no. 2, pp. 617–636, 2017.
9) Y. Yan, W. Ren, Y. Guo, R. Wang, and X. Cao, “Image deblurring via extreme channels prior,” in Proc. IEEE Int. Conf. Comput. Vis. Pattern Recognit. (CVPR), Jul. 2017, pp. 6978–6986.
10) J. D. R. Farquhar, S. Szedmak, H. Meng, and J. Shawe-Taylor, “Improving ‘bag-of-keypoints’ image categorization,” Dept. Electron. Comput. Sci., Univ. Southampton, Southampton, U.K., Tech. Rep., 2005.
11) F. Perronnin and C. Dance, “Fisher kernels on visual vocabularies for image categorization,” in Proc. IEEE Int. Conf. Comput. Vis. Pattern Recognit. (CVPR), Jun. 2007, pp. 1–8.
12) M. Ali Qureshi and M. Deriche, “A bibliography of pixel-based blind image forgery detection techniques,” Signal Process., Image Commun., vol. 39, pp. 46–74, Nov. 2015.
13) A. J. Fridrich, B. D. Soukal, and J. Lukáš, “Detection of copy-move forgery in digital images,” in Proc. Digit. Forensic Res. Workshop, 2003.
14) W. Li and N. Yu, “Rotation robust detection of copy-move forgery,” in Proc. IEEE Int. Conf. Image Process. (ICIP), Sep. 2010, pp. 2113–2116.
15) S.-J. Ryu, M. Kirchner, M.-J. Lee, and H.-K. Lee, “Rotation invariant localization of duplicated image regions based on zernike moments,” IEEE Trans. Inf. Forensics Security, vol. 8, no. 8, pp. 1355–1370, Aug. 2013.

16) I. Amerini, L. Ballan, R. Caldelli, A. Del Bimbo, and G. Serra, "A SIFT-based forensic method for copy–move attack detection and transformation recovery," IEEE Trans. Inf. Forensics Security, vol. 6, no. 3, pp. 1099–1110, Sep. 2011.

17) D. G. Lowe, "Distinctive image features from scale-invariant keypoints," Int. J. Comput. Vis., vol. 60, no. 2, pp. 91–110, 2004.

18) L. Liu, R. Ni, Y. Zhao, and S. Li, "Improved SIFT-based copy-move detection using BFSN clustering and CFA features," in Proc. IEEE Int. Conf. Intell. Inf. Hiding Multimedia Signal Process. (IIH-MSP), Aug. 2014, pp. 626–629.

19) Y. Li and J. Zhou, "Image copy-move forgery detection using hierarchical feature point matching," in Proc. Asia–Pacific Signal Inf. Process. Assoc. Annu. Summit Conf. (APSIPA ASC), Dec. 2016, pp. 1–4.

20) T. Bianchi and A. Piva, "Image forgery localization via block-grained analysis of JPEG artifacts," IEEE Trans. Inf. Forensics Security, vol. 7, no. 3, pp. 1003–1017, Jun. 2012.

21) P. Korus and J. Huang, "Multi-scale fusion for improved localization of malicious tampering in digital images," IEEE Trans. Image Process., vol. 25, no. 3, pp. 1312–1326, Mar. 2016.

22) P. Ferrara, T. Bianchi, A. D. Rosa, and A. Piva, "Image forgery localization via fine-grained analysis of CFA artifacts," IEEE Trans. Inf. Forensics Security, vol. 7, no. 5, pp. 1566–1577, Oct. 2012.

23) G. Chierchia, G. Poggi, C. Sansone, and L. Verdoliva, "A BayesianMRF approach for PRNU-based image forgery detection," IEEE Trans. Inf. Forensics Security, vol. 9, no. 4, pp. 554–567, Apr. 2014.

24) P. Korus and J. Huang, "Multi-scale analysis strategies in PRNU-based tampering localization," IEEE Trans. Inf. Forensics Security, vol. 12, no. 4, pp. 809–824, Apr. 2017.

25) K. Bahrami, A. C. Kot, L. Li, and H. Li, "Blurred image splicing localization by exposing blur type inconsistency," IEEE Trans. Inf. Forensics Security, vol. 10, no. 5, pp. 999–1009, May 2015.

26) T. Carvalho, F. A. Faria, H. Pedrini, R. D. S. Torres, and A. Rocha, "Illuminant-based transformed spaces for image forensics," IEEE Trans. Inf. Forensics Security, vol. 11, no. 4, pp. 720–733, Apr. 2016.

27) W. Zhang, X. Cao, Z. Feng, J. Zhang, and P. Wang, "Detecting photographic composites using two-view geometrical constraints," in Proc. IEEE Int. Conf. Multimedia Expo (ICME), Jun./Jul. 2009, pp. 1078–1081.

28) W. Zhang, X. Cao, Y. Qu, Y. Hou, H. Zhao, and C. Zhang, "Detecting and extracting the photo composites using planar homography and graph cut," IEEE Trans. Inf. Forensics Security, vol. 5, no. 3, pp. 544–555, Sep. 2010.

29) D. T. Trung, A. Beghdadi, and M.-C. Larabi, “Blind inpainting forgery detection,” in Proc. IEEE Global Conf. Signal Inf. Process. (GlobalSIP), Dec. 2014, pp. 1019–1023.
30) A. Levin, D. Lischinski, and Y. Weiss, “Colorization using optimization,” ACM Trans. Graph., vol. 23, no. 3, pp. 689–694, 2004.
31) J. Pang, O. C. Au, K. Tang, and Y. Guo, “Image colorization using sparse representation,” in Proc. IEEE Int. Conf. Acoust., Speech Signal Process. (ICASSP), May 2013, pp. 1578–1582.
32) G. Charpiat, M. Hofmann, and B. Schölkopf, “Automatic image colorization via multimodal predictions,” in Proc. Eur. Conf. Comput. Vis. (ECCV), 2008, pp. 126–139.
33) X. Chen, J. Li, D. Zou, and Q. Zhao, “Learn sparse dictionaries for edit propagation,” IEEE Trans. Image Process., vol. 25, no. 4, pp. 1688–1698, Apr. 2016.
34) Z. Cheng, Q. Yang, and B. Sheng, “Deep colorization,” in Proc. IEEE Int. Conf. Comput. Vis. (ICCV), Dec. 2015, pp. 415–423.
35) S. Iizuka, E. Simo-Serra, and H. Ishikawa, “Let there be color!: Joint end-to-end learning of global and local image priors for automatic image colorization with simultaneous classification,” ACM Trans. Graph., vol. 35, no. 4, p. 110, 2016.
36) R. Zhang, P. Isola, and A. A. Efros, “Colorful image colorization,” in Proc. Eur. Conf. Comput. Vis. (ECCV), 2016, pp. 649–666.
37) K. He, J. Sun, and X. Tang, “Single image haze removal using dark channel prior,” IEEE Trans. Pattern Anal. Mach. Intell., vol. 33, no. 12, pp. 2341–2353, Dec. 2011.
38) C.-C. Chang and C.-J. Lin, “LIBSVM: A library for support vector machines,” ACM Trans. Intell. Syst. Technol., vol. 2, no. 3, 2011, Art. no. 27.
39) VLFeat. Accessed: Jan. 19, 2017. [Online]. Available: http://VLFeat.org
40) O. Russakovsky et al., “ImageNet large scale visual recognition challenge,” Int. J. Comput. Vis., vol. 115, no. 3, pp. 211–252, Dec. 2015.
41) J. Philbin, O. Chum, M. Isard, J. Sivic, and A. Zisserman, “Object retrieval with large vocabularies and fast spatial matching,” in Proc. IEEE Int. Conf. Comput. Vis. Pattern Recognit. (CVPR), Jun. 2007, pp. 1–8.

Fake Colorized Image Detection

Yuanfang Guo, *Member, IEEE*, Xiaochun Cao, *Senior Member, IEEE*, Wei Zhang, *Member, IEEE*, and Rui Wang, *Member, IEEE*

***Abstract*— Image forensics aims to detect the manipulation of digital images. Currently, splicing detection, copy-move detection, and image retouching detection are attracting significant attention from researchers. However, image editing techniques develop over time. An emerging image editing technique is colorization, in which grayscale images are colorized with realistic colors. Unfortunately, this technique may also be intentionally applied to certain images to confound object recognition algorithms. To the best of our knowledge, no forensic technique has yet been invented to identify whether an image is colorized. We observed that, compared with natural images, colorized images, which are generated by three state-of-the-art methods, possess statistical differences for the hue and saturation channels. Besides, we also observe statistical inconsistencies in the dark and bright channels, because the colorization process will inevitably affect the dark and bright channel values. Based on our observations, i.e., potential traces in the hue, saturation, dark, and bright channels, we propose two simple yet effective detection methods for fake colorized images: Histogram-based fake colorized image detection and feature encoding-based fake colorized image detection. Experimental results demonstrate that both proposed methods exhibit a decent performance against multiple state-of-the-art colorization approaches.**

***Index Terms*— Image forgery detection, fake colorized image detection, hue, saturation, ECP.**

Manuscript received March 30, 2017; revised October 25, 2017, December 7, 2017, and January 22, 2018; accepted February 8, 2018. Date of publication February 16, 2018; date of current version April 4, 2018. This work was supported in part by the National Key Research and Development Program of China under Grant 2016YFC0801004, in part by the National Natural Science Foundation of China under Grant 61332012, Grant U1636214, and Grant 61602463, in part by the Beijing Natural Science Foundation under Grant 4172068, in part by the Fundamental Theory and Cutting Edge Technology Research Program of the Institute of Information Engineering, CAS, under Grant Y7Z0381102, in part by the Foundation of Science and Technology on Information Assurance Laboratory under Grant KJ-17-006, and in part by the CCF-Tencent Open Research Fund. The associate editor coordinating the review of this manuscript and approving it for publication was Prof. Anthony T. Ho. *(Corresponding author: Xiaochun Cao.)*

Y. Guo is with the State Key Laboratory of Information Security, Institute of Information Engineering, Chinese Academy of Sciences, Beijing 100093, China, and also with the Science and Technology on Information Assurance Laboratory, Beijing 100072, China (e-mail: guoyuanfang@iie.ac.cn; eeandyguo@connect.ust.hk).

X. Cao is with the State Key Laboratory of Information Security, Institute of Information Engineering, Chinese Academy of Sciences, Beijing 100093, China, and also with the School of Cyber Security, University of Chinese Academy of Sciences, Beijing 100049, China (e-mail: caoxiaochun@iie.ac.cn).

W. Zhang is with JD AI Research, Beijing 100105, China, and was with the State Key Laboratory of Information Security, Institute of Information Engineering, Chinese Academy of Sciences, Beijing 100093, China (e-mail: wzhang.cu@gmail.com).

R. Wang is with the State Key Laboratory of Information Security, Institute of Information Engineering, Chinese Academy of Sciences, Beijing 100093, China (e-mail: wangrui@iie.ac.cn).

Color versions of one or more of the figures in this paper are available online at http://ieeexplore.ieee.org.

Digital Object Identifier 10.1109/TIFS.2018.2806926

I. Introduction

THE rapid proliferation of image editing technologies has increased both the ease with which images can be manipulated and the difficulty in distinguishing between altered and natural images. In addition to the conventional image editing techniques such as splicing [1], copy-move [2] and retouching [3], more image editing techniques, such as colorization [4] and image generation [5], are proposed. Since these types of image editing techniques generate new content with/without references, we denote them as the generative image editing techniques.

Although image editing techniques can provide significant aesthetic or entertainment value, they may also be used with malicious intent. In general, various image editing approaches employ different mechanisms. Splicing and copy-move techniques usually manipulate part of the image and perform object-level changes. Image retouching techniques usually change the images via a variety of mechanisms. For example, contrast enhancement adjusts the contrast of the image, while image inpainting usually fills the holes in images according to the image content. Among the generative image editing techniques, image generation usually produces a meaningful image from a noise vector with/without some additional information such as text or a class label. Colorization, on the other hand, usually colorizes images with visually plausible colors, which may cause misjudgment when specific objects/scenes must be identified/tracked.

Fortunately, numerous image forensic technologies have been developed in the past decades. According to their mechanisms and applications, they can be categorized into two classes, active techniques and passive techniques. The active techniques usually refer to watermarking techniques [6]–[8], which embed authentication information in the to-be-protected images. When the integrities of these images demand verification, watermark extraction procedures are performed and the extracted watermarks are compared to the original watermark to detect forgeries. Since the active techniques require the watermark to be embedded prior to detection, the applications, in practice, are limited.

In contrast, passive image forgery detection approaches, to which our proposed methods belong, usually detect the manipulations to the input images directly. Traditionally, passive image forgery (editing) detection techniques have mainly focused on splicing detection [1], copy-move detection [2] and image retouching detection [3]. To the best of our knowledge, no method has yet been developed to detect the fake images generated by generative image editing

techniques. If these images are examined by humans, the cost increases drastically as the number of to-be-examined images increases. Obviously, detection via human eyes is insufficient for the big data era. On the other hand, conventional image forgery detection techniques are designed with different assumptions that may not be appropriate for generative fake image detection. Therefore, generative fake image detection demands specific studies and designs.

Among the different generative image forgery techniques, colorization already achieves excellent performances. As shown in Fig. 1, fake colorized images, which are generated by a state-of-the-art algorithm [4], are visually indistinguishable, if no ground-truth images exist for comparison. Therefore, the necessity to develop a scheme for fake colorized image detection increases rapidly. In this paper, we aim to address this new problem by providing feasible solutions. Specifically, we propose two simple yet effective schemes for detecting fake colorized images, which are generated by fully automatic colorization methods. The contributions are summarized as follows:

1: We observe that fake colorized images and their corresponding natural images exhibit statistical differences, which can be further utilized as detection traces, in both color channels and image prior. The color channels involved are the hue and saturation channels, while the exploited extreme channels prior is proposed in our recent work [9].
2: According to the statistical differences in the color channels and image priors, we propose a fake colorized image detection scheme, named Histogram based Fake Colorized Image Detection (FCID-HIST), by proposing four detection features. Each feature calculates the most distinctive bin and the total variation of the normalized histogram distribution for hue, saturation, dark and bright channels, respectively.
3: To better utilize the statistical information of the training images, we consider exploiting the divergences inside different moments of the data vectors and propose a fake colorized image detection scheme, named Feature Encoding based Fake Colorized Image Detection (FCID-FE), by modeling the created four-dimensional samples with a Gaussian mixture model (GMM) [10] and encoding the samples into Fisher feature vectors [11].
4: In the experiments, the two proposed methods demonstrate a decent performance in various tests for detecting fake images generated by three state-of-the-art colorization methods.

The rest of the paper is organized as follows. Section II presents the necessary background. Section III introduces the proposed work. Section IV describes the experimental results in various tests and analyzes the proposed methods. Finally, Section V summarizes the paper and discusses future work.

II. Background

In this section, conventional forgery detection techniques and colorization techniques are reviewed accordingly.

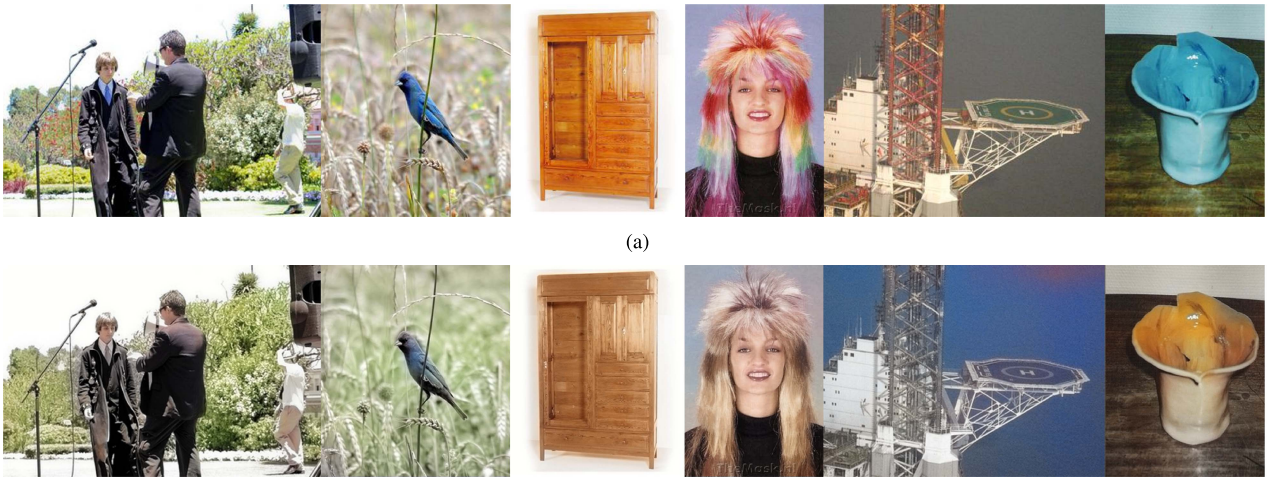
A. Review of Forgery Detection

Forgery detection [12] has been investigated for decades. In general, forgery detection explores different characteristics of images and attempts to find traces to analyze. As mentioned above, most of the traditional forgery detection techniques can be categorized into three classes, copy-move detection, splicing detection and image retouching detection.

Copy-move detection relies on identifying duplicated regions in a tampered image. Intuitively, these techniques tend to seek an appropriate feature in a certain domain, such that the detection can be performed via searching the most similar two units (such as patches). Different methods usually exploit different features. Reference [13] explores features in the frequency domain by dividing the image into overlapping blocks and detects the copy-move forgery via matching the quantized discrete cosine transform (DCT) coefficients. Reference [14] performs a rotation invariant detection based on the Fourier-Mellin transform. Reference [15] localizes the duplicated regions based on the Zernike moments, which exhibit the rotation invariance property, of small image blocks. Reference [15] reports decent results especially when the duplicated regions are smooth. Reference [16] employs the famous SIFT feature [17] to detect multiple duplicated regions and estimates the geometric transformation performed by the copy-move operation. Reference [18] presents a SIFT based detection method by matching the SIFT features via a broad first search neighbors clustering algorithm and distinguishing the duplicated origins from the tampered regions via CFA features. Reference [19] introduces a hierarchical SIFT-based keypoint matching technique to solve a drawback of previous keypoint matching based detection techniques, which tends to give poor performances when the copy-moved regions are small or smooth. Although copy-move detection technologies have been developed rapidly, they cannot be directly applied to the fake colorized image detection because no copy-move operations exist in the fake colorized images.

Splicing detection usually detects the manipulated regions which originate from different source images. Different from copy-move detection, these approaches detect the tampered regions with various traces (features), which usually reveal the inconsistencies between the tampered regions and the unchanged regions. Currently, splicing detection can be classified into four categories, compression-based methods, camera-based methods, physics-based methods and geometry-based methods, according to their mechanisms.

Compression-based methods assume that the spliced region and the original image have undergone different types of image compression and may exhibit different compression artifacts. For example, [20] considers the DCT coefficient distributions of each 8×8 block and computes the tampering probability. By considering the advantages and disadvantages of different block sizes, [21] constructs a multiscale scheme, employs the Benford's law at each level and fuses the results together to obtain a final localization map. Unfortunately, the compression-based methods are not appropriate for fake colorized image detection because the assumption may not always be valid.



(a)

(b)

Fig. 1. (a) Real images. (b) Fake colorized images.

TABLE I

SUMMARY OF THE EXISTING FAKE IMAGE DETECTION APPROACHES

Category	Method	Core mechanism	Potential result of detecting colorized images
Copy-move detection	[13]	Quantized DCT coefficients	Not applicable
	[14]	Fourier-Mellin Transform	Not applicable
	[15]	Zernike moments	Not applicable
	[16]	SIFT feature	Not applicable
	[18]	SIFT & CFA features	Not applicable
	[19]	Hierarchical SIFT-based keypoint matching	Not applicable
Splicing detection	[20]	DCT coefficient distributions of each block	Not applicable
	[21]	Multiscale scheme based on Benford's law	Not applicable
	[22]	CFA artifacts	Possible but with low robustness
	[23]	PRNU noises	Possible but with low robustness
	[24]	Multiscale scheme based on PRNU noises	Possible but with low robustness
	[25]	Blur type inconsistency	Not applicable
	[26]	Illuminant-based transform spaces	Not applicable
	[27]	Two-view geometrical constraints	Not applicable
	[28]	Planar homographies	Not applicable
Image retouching detection	[29]	Block similarities and distances	Not applicable
	[3]	Peak/gap artifacts	Not applicable

Camera-based methods consider traces left on the image during the capturing process. Reference [22] detects the existence of the CFA artifacts, which are due to the demosaicking process in the CFA cameras, and thus obtains the localization map. Reference [23] exploits the photo-response non-uniformity noises (i.e., the sensor noises) of the camera to distinguish the tampered regions from the original ones. Reference [24] also considers the photo-response non-uniformity noises and a multiscale framework to conduct a multiscale analysis and detects small forgeries more accurately. Even if the camera-based methods can be employed to detect the fake colorized images, their robustness is incompetent because the sensor noises and the artifacts can easily be affected by noises and some common post-processing operations such as compression.

Physics-based methods perform detection based on different physics phenomenon inconsistencies. Reference [25] considers the blur type inconsistency between the spliced region and the original image to localize the tampered region. Reference [26] explores the illuminant-based transform spaces and combines different image descriptors, such as color, shape and texture, to detect forged regions. Since the fake colorized images to be examined in this paper are forged for the whole image, these inconsistencies cannot be utilized to detect the fake colorized images.

Geometry-based methods utilize the geometry information inside images for detection. Reference [27] explores detecting the compositions with the two-view geometrical constraints. Reference [28] considers the planar homographies in the test images and adopts graph-cut algorithm to obtain the final localization map. Unfortunately, since the geometrical characteristics are rarely manipulated in the fake colorized images, the geometry-based methods will also fail to detect the colorized images.

Image retouching detection usually considers that the original images are restored or enhanced. For example, [29] is designed to detect the inpainted images by considering the similarities, distances and number of identical pixels among different blocks. Reference [3] calculates the histograms and performs detection via the peak/gap artifacts induced from contrast enhancement. These techniques can hardly be applied to the new problem because their mechanisms are specially designed for their own assumptions.

Table I provides a summary of existing forensic techniques. Although many detection technologies have been developed, they are currently not directly applicable to the detection

TABLE II
SUMMARY OF THE EXISTING COLORIZATION APPROACHES

Category	Method	Core mechanism	Side-information
Scribble-based method	[30]	Neighboring pixels with similar intensities should have similar colors	User scribbles
	[31]	Construct color and texture dictionaries	User scribbles
Example-based method	[32]	Global optimization of colors at pixel-level	Reference color image
	[33]	Propagating the learned dictionaries	Reference color image
Fully automatic method	[34]	Network with pixel patch, DAISY and semantic features	None
	[35]	End-to-End network with local and global priors	None
	[4]	End-to-End network with Hypercolumn	None
	[36]	Classification-style colorization in LAB space	None

of images which are manipulated by generative methods. Specially designed techniques are necessary to address the detection of fake colorized images.

B. Review of Colorization

Colorization, a term describing the color adding process to grayscale images, was firstly introduced by Wilson Markle in 1970. However, this area began to develop rapidly in the 21st century. Colorization techniques can be categorized into the following types: scribble-based, example-based and fully automatic.

Scribble-based methods are supervised techniques in which users begin assigning colors to pixels in the grayscale image. The milestone work [30], which assumes that the neighboring pixels with similar intensities should have similar colors, is proposed at first. Various other approaches have been proposed in succession, such as [31], which constructs dictionaries for color and textures via sparse representation and colorizes the images accordingly.

Example-based methods [32], [33] usually require the users to supervise the system by providing reference color image(s) similar to the greyscale image. The system then transfers the colors in the reference color image(s) to the target greyscale image by searching for similar patterns/objects. The performances of these methods are dependent on the quality of the reference image(s). If the divergence between the greyscale image and the reference image(s) is high, the colorized result may be unsatisfactory.

In contrast with the supervised approaches above, fully automatic methods require no supervision when performing the colorization task. Reference [34] trains a neural network and predicts the chrominance values by considering the pixel patch, DAISY and semantic features. Reference [35] colorizes the images by jointly utilizing the local and global priors with an end-to-end network. Reference [4] proposes a state-of-the-art approach, which exploits the hypercolumn to utilize both low-level and semantic representations, and colorizes the images in the Hue-Chroma-Lightness (HCL) color space. Reference [36] calculates the statistical distributions of the chrominance information in the LAB space and introduces a classification-style colorization approach based on a deep network.

These techniques are briefly summarized in Table II. Due to the high performances of the fully automatic colorization techniques, we focus on the detection of the fake colorized images which are generated via these techniques in this paper.

III. METHODOLOGY

The rapid progress in colorization technologies has enabled colorized images to be visually indistinguishable from natural images. State-of-the-art colorization methods are already capable of misleading human observers in the subjective tests [36]. To distinguish the fake colorized images from the natural images, we study the statistics of the fake colorized images, which are generated by three state-of-the-art methods [4], [36], [35], and propose two simple yet effective detection schemes, FCID-HIST and FCID-FE.

A. Observations and Statistics

According to our observation, the colorized images tend to possess less saturated colors, and the colorization method favors some colors over others, though these differences are difficult to be visually perceived. Since the Hue-Saturation-Value (HSV) color space separately represents the chrominance information in the hue and saturation channel, we calculate the normalized histograms (each containing 200 bins) of the hue and saturation channel in 15000 natural images and their corresponding fake colorized images, separately, as shown in Fig. 2.

As shown in Fig. 2, the statistics of the natural and fake colorized images are different in both the hue and saturation channels, and there also exist statistical differences (especially for the peaks in the histograms) among the fake images generated by different colorization methods. For the hue channel, the histogram of the fake images tends to be more smooth and possesses more significant peaks compared to the natural images. For the saturation channel, the histogram of the fake images also exhibits different peak values and variances compared to the histogram of the natural images. These statistics indicate that the fake images favor different colors and possess saturation differences compared to the natural images. Therefore, the natural and fake colorized images are statistically identifiable, though the fake colorized images seemed visually indistinguishable.

In addition to the statistical differences in the color channels, differences also exist in some image priors because they are not considered explicitly in the colorization process even though the deep neural networks possess good generalization ability. In this paper, we exploit our recently proposed extreme channels prior (ECP) [9], which consists of the dark channel prior (DCP) [37] and the bright channel prior (BCP). Intuitively, DCP assumes that the dark channel of a natural image

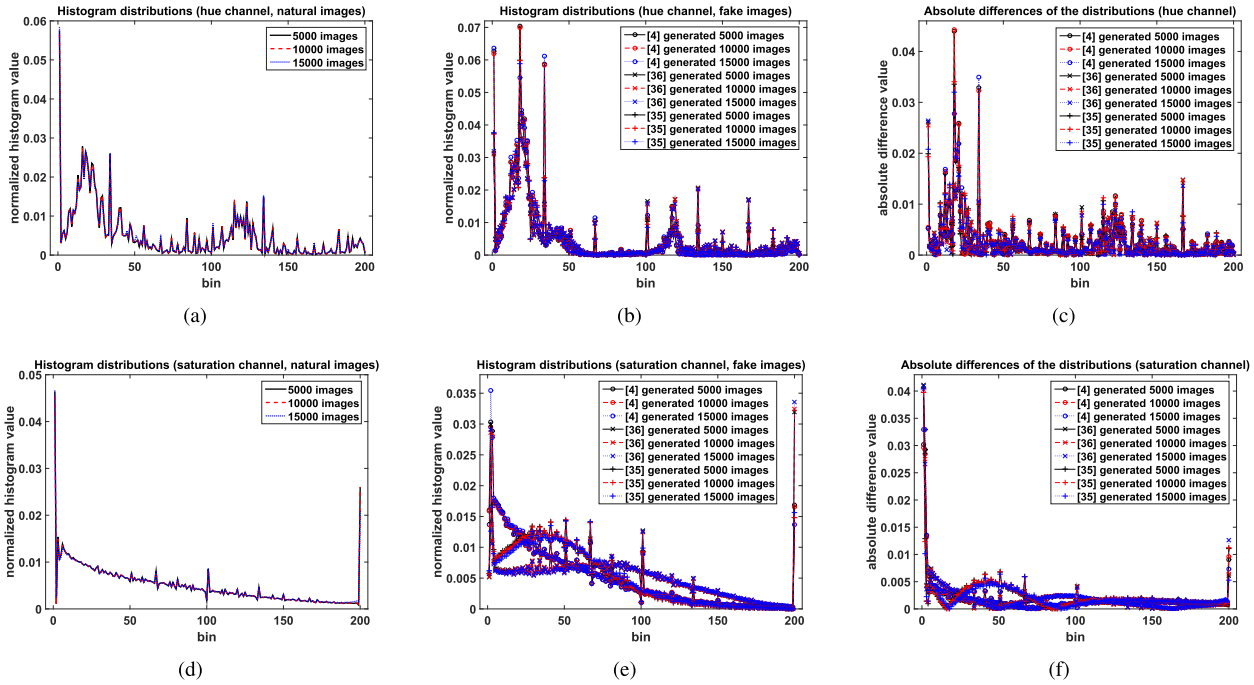


Fig. 2. (a) Normalized histogram distribution of the hue channel (natural images). (b) Normalized histogram distribution of the hue channel (fake images). (c) Absolute differences of the distributions in (a) and (b). (d) Normalized histogram distribution of the saturation channel (natural images). (e) Normalized histogram distribution of the saturation channel (fake images). (f) Absolute differences of the distributions in (d) and (e).

is close to zero, while BCP assumes that the bright channel of a natural image is close to 255. The dark channel I_{dc} and bright channel I_{bc} of an image I are defined as shown by Eqs. 1 and 2, respectively.

$$I_{dc}(x) = \min_{y\in\Omega(x)}\Big(\min_{c_p\in(r,g,b)} I_{c_p}(y)\Big), \tag{1}$$

$$I_{bc}(x) = \max_{y\in\Omega(x)}\Big(\max_{c_p\in(r,g,b)} I_{c_p}(y)\Big), \tag{2}$$

where x stands for the pixel location, I_{c_p} denotes a color channel of I and $\Omega(x)$ represents the local patch centered at the location x. Note that the local patch sizes here are identical to the settings in [9].

By calculating the histograms of the dark channel and bright channel of 15000 natural images and their corresponding fake colorized images separately, Fig. 3 presents the expected differences, especially for the peak values, and supports our observations above.

B. FCID-HIST

By exploiting the existing statistical differences, we propose the Histogram based Fake Colorized Image Detection (FCID-HIST) method to detect fake colorized images.

In FCID-HIST, four detection features, the hue feature F_h, the saturation feature F_s, the dark channel feature F_{dc} and the bright channel feature F_{bc}, are proposed to detect forgeries.

The hue feature is constructed from the normalized hue channel histogram distributions. Let K_h be the total number of bins in each normalized hue channel histogram distribution. We define $Dist_{h,n}$ and $Dist_{h,f}$ as the normalized hue channel histogram distribution for the natural and fake training images, respectively, and $Dist_h^\alpha$ as the corresponding histogram for the αth input image, which can be either a training or testing image.

Intuitively, to differentiate the fake colorized images from the natural images, the distinctive features should reveal the largest divergences between the two types of images. (Note that the Euclidean distance is employed in this paper to calculate the divergences.) Therefore, we select the most distinctive bin $Dist_h^\alpha(\upsilon_h)$, whose two corresponding bins in $Dist_{h,n}$ and $Dist_{h,f}$ give the largest divergence between the two histogram distributions, as part of the hue feature, as follows

$$F_h^\alpha(1) = Dist_h^\alpha(\upsilon_h) \tag{3}$$

where the index of the most distinctive bin υ_h for the hue channel is calculated via Eq. 4.

$$\begin{aligned}\upsilon_h &= \operatorname*{argmax}_x ||Dist_{h,n}(x) - Dist_{h,f}(x)||_2 \\ &= \operatorname*{argmax}_x |Dist_{h,n}(x) - Dist_{h,f}(x)|\end{aligned} \tag{4}$$

The distributions $Dist_{h,n}$ and $Dist_{h,f}$ also vary differently with respect to the bins. We account for this difference in the hue feature by computing the first order derivative of the normalized hue channel histogram distribution $DistD_h^\alpha(l) = Dist_h^\alpha(l+1) - Dist_h^\alpha(l)$ to capture the varying trend of the histogram distribution. This total variation is calculated as Eq. 5 shows.

$$F_h^\alpha(2) = \sum_{l=1}^{K_h-1} |DistD_h^\alpha(l)| \tag{5}$$

The proposed hue feature is then formed by combining Eq. 3 with Eq. 5 into a vector, as Eq. 6 demonstrates.

$$F_h^\alpha = [F_h^\alpha(1) \quad F_h^\alpha(2)] \tag{6}$$

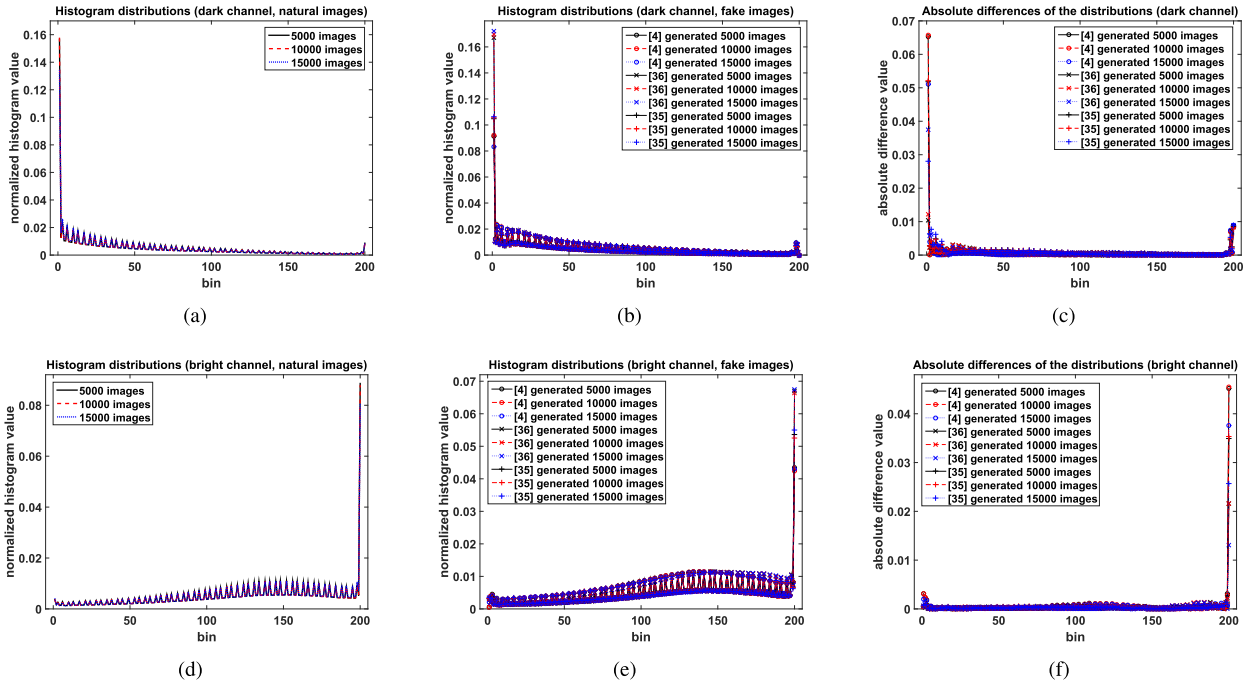


Fig. 3. (a) Normalized histogram distribution of the dark channel (natural images). (b) Normalized histogram distribution of the dark channel (fake images). (c) Absolute differences of the distributions in (a) and (b). (d) Normalized histogram distribution of the bright channel (natural images). (e) Normalized histogram distribution of the bright channel (fake images). (f) Absolute differences of the distributions in (d) and (e).

Similarly, the saturation feature F_s^α, the dark channel feature F_{dc}^α and the bright channel feature F_{bc}^α can be constructed by utilizing the normalized histogram distributions ($Dist_{s,n}$, $Dist_{s,f}$), ($Dist_{dc,n}$, $Dist_{dc,f}$), and ($Dist_{bc,n}$, $Dist_{bc,f}$) for the saturation, bright, and dark channels of the training images respectively.

In the same manner as Eq. 4, the indexes for the most distinctive bins v_s, v_{dc} and v_{bc} can be calculated by Eq. 7.

$$v_{c_h} = \underset{x}{\operatorname{argmax}} |Dist_{c_h,n}(x) - Dist_{c_h,f}(x)|, \quad c_h = s, dc, bc \tag{7}$$

Then, the most distinctive bins for each feature can be calculated via Eq. 8.

$$F_{c_h}^\alpha(1) = Dist_{c_h}^\alpha\left(\underset{x}{\operatorname{argmax}} |Dist_{c_h,n}(x) - Dist_{c_h,f}(x)|\right), \quad c_h = s, dc, bc \tag{8}$$

where $Dist_{c_h}^\alpha$ represents the normalized c_h channel histogram distribution of the αth input image.

The total variation of each distribution is computed via Eq. 9.

$$F_{c_h}^\alpha(2) = \sum_{l=1}^{K_{c_h}-1} |DistD_{c_h}^\alpha(l)|, \quad c_h = s, dc, bc \tag{9}$$

where K_{c_h} stands for the total number of bins in each normalized c_h channel histogram distribution and $DistD_{c_h}^\alpha$ denotes the first order derivative of the normalized c_h channel histogram distribution.

Then, the features are formed as shown in Eq. 10.

$$F_{c_h}^\alpha = [F_{c_h,0}^\alpha \; F_{c_h,1}^\alpha], \quad c_h = s, dc, bc \tag{10}$$

With all the features calculated, the final detection feature F_{HIST}^α for the αth input image can be constructed via Eq. 11.

$$F_{HIST}^\alpha = [F_h^\alpha \; F_s^\alpha \; F_{dc}^\alpha \; F_{bc}^\alpha] \tag{11}$$

After the detection feature is calculated, FCID-HIST employs the supporting vector machine (SVM)[38] for training and detecting the fake colorized images. The FCID-HIST algorithm is summarized as shown in Algorithm 1. For convenience, we let $K_h = K_s = K_{dc} = K_{bc}$ in this paper.

C. FCID-FE

Although FCID-HIST gives a decent performance in the experiments, which are demonstrated in the latter section, these features may not fully utilize the statistical differences between the natural and fake colorized images because the distributions are modeled channel by channel. Therefore, we propose another scheme, Feature Encoding based Fake Colorized Image Detection (FCID-FE), to better exploit the statistical information by jointly modeling the data distribution and exploiting the divergences inside different moments of the distribution.

Let I_h^β, I_s^β, I_{dc}^β and I_{bc}^β be the hue, saturation, dark and bright channels of a training image respectively, where β is the index of the training image. Then, we create a training sample set Φ via Eq. 12.

$$\Phi\big((z-1)*i*j + (i-1)*j + j\big) = [I_h^\beta(i,j) \; I_s^\beta(i,j) \; I_{dc}^\beta(i,j) \; I_{bc}^\beta(i,j)] \tag{12}$$

In contrast to the histogram modeling, FCID-FE models the sample data distribution G with a Gaussian mixture model

Algorithm 1 FCID-HIST

Training Stage:

Input: N_1 natural and fake colorized training images, the corresponding labels $L_{r,HIST}$, K_h, K_s, K_{dc}, K_{bc}, SVM parameters

Output: υ_h, υ_s, υ_{dc}, υ_{bc}, trained SVM classifier

Compute $Dist_{h,n}$, $Dist_{s,n}$, $Dist_{dc,n}$, $Dist_{bc,n}$
Compute $Dist_{h,f}$, $Dist_{s,f}$, $Dist_{dc,f}$, $Dist_{bc,f}$
Compute υ_h, υ_s, υ_{dc}, υ_{bc} ▷ refer to Eq. 4 and 7
for $i = 1$ to N_1 **do**
    Compute $Dist_h^i$, $Dist_s^i$, $Dist_{dc}^i$, $Dist_{bc}^i$
    Compute $F_h^i(1)$, $F_s^i(1)$, $F_{dc}^i(1)$, $F_{bc}^i(1)$ ▷ refer to Eq. 3 and 8
    Compute $F_h^i(2)$, $F_s^i(2)$, $F_{dc}^i(2)$, $F_{bc}^i(2)$ ▷ refer to Eq. 5 and 9
    Compute F_h^i, F_s^i , F_{dc}^i, F_{bc}^i ▷ refer to Eq. 6 and 10
    Compute F_{HIST}^i ▷ refer to Eq. 11
end for
Train SVM with F_{HIST}, $L_{r,HIST}$ and SVM parameters

Testing Stage:

Input: N_2 test images, K_h, K_s, K_{dc}, K_{bc}, υ_h, υ_s, υ_{dc}, υ_{bc}, trained SVM classifier

Output: Detection labels $L_{e,HIST}$

for $i = 1$ to N_2 **do**
    Compute $Dist_h^i$, $Dist_s^i$, $Dist_{dc}^i$, $Dist_{bc}^i$
    Compute $F_h^i(1)$, $F_s^i(1)$, $F_{dc}^i(1)$, $F_{bc}^i(1)$ ▷ refer to Eq. 3 and 8
    Compute $F_h^i(2)$, $F_s^i(2)$, $F_{dc}^i(2)$, $F_{bc}^i(2)$ ▷ refer to Eq. 5 and 9
    Compute F_h^i, F_s^i , F_{dc}^i, F_{bc}^i ▷ refer to Eq. 6 and 10
    Compute F_{HIST}^i ▷ refer to Eq. 11
    Obtain $L_{e,HIST}(i)$ with F_{HIST}^i and the trained SVM clasifier
end for

(GMM) [10] as shown in Eq. 13.

$$G(\Phi|\Theta) = \sum_{n=1}^{N} \log p(\Phi_n|\Theta) \tag{13}$$

where N is the number of samples in Φ, Θ stands for the parameter set of the constructed GMM and Θ is defined in Eq. 14.

$$\Theta = \omega_a, \quad \mu_a, \ \sigma_a, \ a = 1, \ldots, N_m, \ \sum_{n=1}^{N_m} \omega_a = 1 \tag{14}$$

where ω_a represents the weight, μ_a stands for the mean value vector, σ_a denotes the covariance matrix and N_m is the number of Gaussian distributions in the distribution model.

Then, the likelihood of Φ_n being modeled by the GMM Θ can be represented by Eq. 15.

$$p(\Phi_n|\Theta) = \sum_{m=1}^{N_m} \log \omega_m p_m(\Phi_m|\Theta) \tag{15}$$

Algorithm 2 FCID-FE

Training Stage:

Input: N_3 natural and fake colorized training images, the corresponding labels $L_{r,FE}$, SVM parameters

Output: Θ, trained SVM classifier

Create samples Φ ▷ refer to Eq. 12
Estimate GMM model Θ from Φ
for $i = 1$ to N_3 **do**
    Encode Φ^i to F_{FE}^i with Θ ▷ refer to Eq. 17
end for
Train SVM with F_{FE}, $L_{r,FE}$ and SVM parameters

Testing Stage:

Input: N_4 test images, Θ, trained SVM classifier

Output: Detection labels $L_{e,FE}$

Create samples Φ ▷ refer to Eq. 12
for $i = 1$ to N_4 **do**
    Encode Φ^i to F_{FE}^i with Θ ▷ refer to Eq. 17
    Obtain $L_{e,FE}(i)$ with F_{FE}^i and the trained SVM clasifier
end for

where $p_m(\Phi_m|\Theta)$ is defined by Eq. 16.

$$p_m(\Phi_m|\Theta) = \frac{\exp\left[-(1/2)(\Phi_m - \mu_a)^T \sigma_a^{-1}(\Phi_m - \mu_a)\right]}{(2\pi)^{N_\upsilon/2}|\sigma_a|^{1/2}} \tag{16}$$

where N_υ denotes the number of dimensions of each sample vector. Then, GMM can be constructed by determining the parameter set Θ.

With the determined GMM, FCID-FE utilizes different moments of the distribution and encodes each subset Φ^β of the sample vectors, which belongs to each training image, into training Fisher vectors [11] as expressed by Eq. 17.

$$F_{FE}^\beta = \left[\frac{\lambda_1 \delta G(\Phi^\beta|\Theta)}{\delta \omega_a} \ \frac{\lambda_2 \delta G(\Phi^\beta|\Theta)}{\delta \mu_{a,\upsilon}} \ \frac{\lambda_3 \delta G(\Phi^\beta|\Theta)}{\delta \sigma_{a,\upsilon}}\right] \tag{17}$$

where $\upsilon = 1, 2, \ldots, N_\upsilon$ and λ_1, λ_2 and λ_3 are defined in Eqs. 18-20.

$$\lambda_1 = \left(N\left(\frac{1}{\omega_a} + \frac{1}{\omega_1}\right)\right)^{-1/2} \tag{18}$$

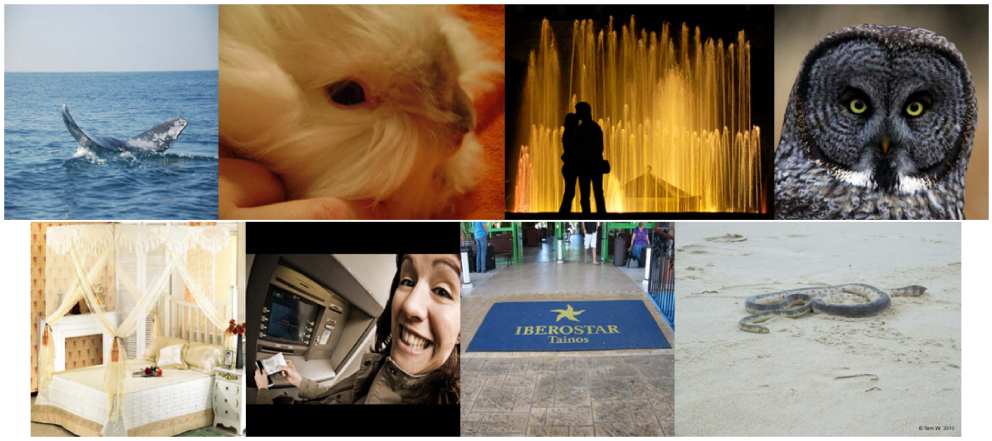
$$\lambda_2 = \left(\frac{N\omega_a}{(\sigma_{a,\upsilon})^2}\right)^{-1/2} \tag{19}$$

$$\lambda_3 = \left(\frac{2N\omega_a}{(\sigma_{a,\upsilon})^2}\right)^{-1/2} \tag{20}$$

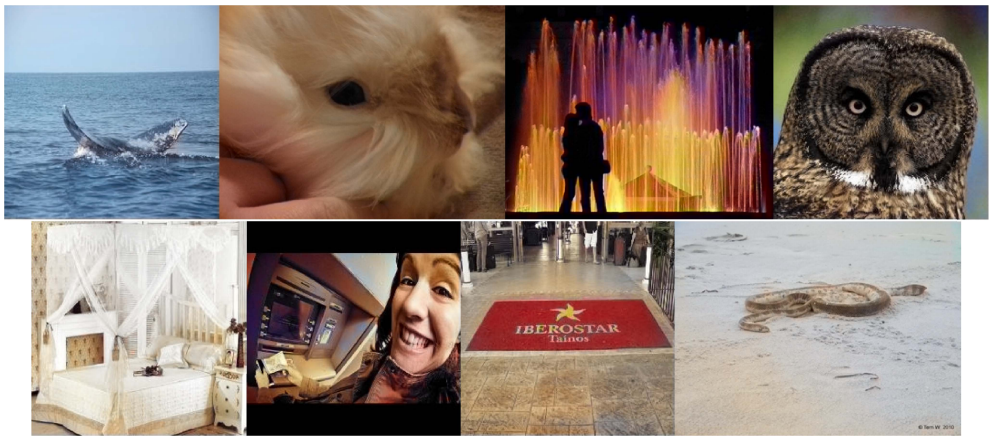
Then, SVM is employed as the training classifier. For testing, FCID-FE will first construct the test sample set for each input image via Eq. 12. Next, the existing GMM from the training dataset is employed to encode each test image into the Fisher vector with Eq. 17. At last, FCID-FE classifies these feature vectors via the trained SVM. The algorithm of FCID-FE is summarized in Algorithm 2.

IV. Experimental Results

In this section, the experimental setups, evaluation measurements, databases and results are introduced accordingly.



(a)



(b)

Fig. 4. (a) Real Images. (b) Fake colorized images.

A. Setups and Measurements

In this paper, one implementation of SVM, the LIBSVM [38], is employed for classification and the RBF kernel is selected. The VLFeat software [39] is employed for GMM modeling and Fisher vector encoding.

In our experiments, both the half total error rate ($HTER$) measurement and the receiver operating characteristic (ROC) curve (with the area under the curve (AUC) measurement) are employed to evaluate the performances of the proposed methods. Denoting P, N, TP and TN as the positive samples, negative samples, true positive samples and true negative samples respectively, $HTER$ is defined in Eq. 21.

$$HTER = \frac{FPR + FNR}{2} = \frac{FP/(TN + FP) + FN/(TP + FN)}{2} \quad (21)$$

Note that the natural images and the fake colorized images are defined as the negative samples and the positive samples, respectively.

B. Databases

For a thorough evaluation of the proposed methods, different databases are employed/constructed for different experiments. We create the database $D1$ for parameter selection and validation by employing 10000 fake colorized images from the database ctest10k in [4] and their corresponding 10000 natural images from the ImageNet validation dataset [40]. The natural images in $D1$ include various types of images, such as animals, human, furniture and outdoor scenes.

In addition to $D1$, different databases are also prepared to assess the performances of FCID-HIST and FCID-FE against different colorization methods. The database $D2$ consists of 2000 natural images randomly selected from the ImageNet validation dataset and their corresponding fake images, which are generated via [4]. The database $D3$ is constructed by randomly selecting 2000 fake colorized images from the results of [36] and 2000 corresponding natural images from the ImageNet validation dataset. The database $D4$, which contains 2000 natural images (randomly selected from the ImageNet validation dataset) and their corresponding generated fake images, is produced via employing the colorization approaches in [35]. Note that the selected natural images and their corresponding colorized images in $D2$-$D4$ are not overlapping with those in $D1$.

Similarly, databases $D5$, $D6$ and $D7$ are constructed by randomly selecting 2000 natural images from the Oxford building dataset [41] and generating the corresponding colorized images with [4], [36], and [35], respectively. Note that the real images in the Oxford building dataset [41] contain various content provided by "Flickr".

Some examples from the databases are shown in Figs. 1 and 4.

TABLE III
HTER OF FCID-HIST FOR DIFFERENT SVM PARAMETER SETTINGS (IN PERCENTAGE)

	g=1/64	g=1/32	g=1/16	g=1/8	g=1/4	g=1/2	g=1	g=2	g=4	g=8	g=16	g=32	g=64
c=1/64	24.30	24.30	24.30	24.30	24.30	24.30	24.30	24.05	23.20	23.40	23.95	24.60	26.15
c=1/32	24.30	24.30	24.30	24.30	24.30	24.30	24.30	24.05	23.20	23.40	23.95	24.60	26.15
c=1/16	24.30	24.30	24.30	24.30	24.30	24.30	24.30	24.05	23.20	23.40	23.95	24.60	26.15
c=1/8	24.30	24.30	24.30	24.30	24.30	24.30	24.30	24.05	23.20	23.40	23.95	24.60	26.15
c=1/4	24.30	24.30	24.30	24.30	24.30	24.30	24.30	24.05	23.20	23.40	23.95	24.60	26.15
c=1/2	24.30	24.30	24.30	24.30	24.30	24.30	24.30	24.05	23.20	23.40	23.95	24.60	26.15
c=1	24.30	24.30	24.30	24.30	24.30	24.30	24.30	24.05	23.20	23.40	23.95	24.60	26.15
c=2	23.55	23.55	23.55	23.55	23.55	23.55	23.55	22.65	22.65	23.05	23.50	24.45	26.55
c=4	22.90	22.90	22.90	22.90	22.90	22.90	22.90	22.15	22.20	22.80	23.90	25.15	27.80
c=8	22.30	22.30	22.30	22.30	22.30	22.30	22.30	22.20	22.50	22.85	23.70	25.95	28.55
c=16	22.00	22.00	22.00	22.00	22.00	22.00	22.00	21.75	21.90	23.25	24.40	26.75	29.10
c=32	**21.50**	**21.50**	**21.50**	**21.50**	**21.50**	**21.50**	**21.50**	21.65	22.15	24.20	24.95	27.75	30.55
c=64	21.65	21.65	21.65	21.65	21.65	21.65	21.65	22.15	22.50	24.10	25.75	28.30	31.00

TABLE IV
HTER OF FCID-FE FOR DIFFERENT SVM PARAMETER SETTINGS (IN PERCENTAGE)

	g=1/64	g=1/32	g=1/16	g=1/8	g=1/4	g=1/2	g=1	g=2	g=4	g=8	g=16	g=32	g=64
c=1/64	16.90	16.90	16.90	16.90	16.90	16.90	16.90	19.20	22.25	27.05	35.05	58.25	53.20
c=1/32	16.90	16.90	16.90	16.90	16.90	16.90	16.90	19.20	22.25	27.05	35.05	58.25	53.20
c=1/16	16.90	16.90	16.90	16.90	16.90	16.90	16.90	19.20	22.25	27.05	35.05	58.25	53.20
c=1/8	16.90	16.90	16.90	16.90	16.90	16.90	16.90	19.20	22.25	27.05	35.05	58.25	53.20
c=1/4	16.90	16.90	16.90	16.90	16.90	16.90	16.90	19.20	22.25	27.05	35.05	58.25	53.20
c=1/2	16.90	16.90	16.90	16.90	16.90	16.90	16.90	19.20	22.25	27.05	35.05	58.25	53.20
c=1	16.90	16.90	16.90	16.90	16.90	16.90	16.90	19.20	22.25	27.05	35.05	58.25	53.20
c=2	**16.65**	**16.65**	**16.65**	**16.65**	**16.65**	**16.65**	**16.65**	19.25	21.60	26.80	34.70	58.85	53.70
c=4	17.35	17.35	17.35	17.35	17.35	17.35	17.35	19.15	21.70	26.80	34.70	58.85	53.70
c=8	17.45	17.45	17.45	17.45	17.45	17.45	17.45	19.25	21.65	26.80	34.70	58.85	53.70
c=16	17.50	17.50	17.50	17.50	17.50	17.50	17.50	19.60	21.65	26.80	34.70	58.85	53.70
c=32	18.25	18.25	18.25	18.25	18.25	18.25	18.25	19.55	21.65	26.80	34.70	58.85	53.70
c=64	18.70	18.70	18.70	18.70	18.70	18.70	18.70	19.55	21.65	26.80	34.70	58.85	53.70

C. Parameter Selection

Prior to evaluating the performances of FCID-HIST and FCID-FE against different colorization approaches, the optimal parameters of the proposed methods are tuned via experiments. In the experiments, 1000 forged images and their corresponding natural images are randomly selected from database $D1$ to construct the parameter training (par-train) set, while another 1000 fake images and their corresponding natural images are selected from $D1$ to be the parameter testing (par-test) set. Note that the par-train set and the par-test set are not overlapping.

Here, two parameters, c and g, which denote the cost and gamma in LIBSVM, are specifically tuned here via grid search. Tables III and IV present the $HTER$ results with different cs and gs for FCID-HIST and FCID-FE, respectively. As shown, FCID-HIST should select $c = 32$, while FCID-FE should select $c = 2$ for the parameter c. Since there exists multiple choices for g, for convenience and consistency, $g = 1/2$ is selected for both FCID-HIST and FCID-FE in the rest of this paper.

Next, we study the selection of the SVM threshold, which is important for the final classification step after the probabilities are estimated. In the test, the threshold varies from 0 to 1 with a step size of 0.01. For each proposed method, a 10-fold cross threshold selection test is performed to obtain the optimal threshold by employing $D1$. Table V presents the optimal thresholds of each fold for FCID-HIST and FCID-FE. Therefore, the optimal thresholds for FCID-HIST and FCID-FE,

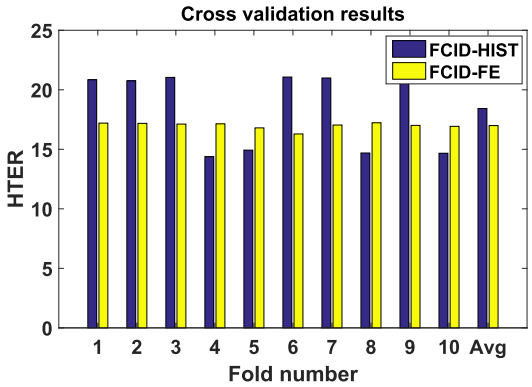


Fig. 5. FCID-HIST and FCID-FE HTER results of 10-fold cross validation.

which are calculated via averaging the optimal thresholds of each fold, are 0.455 and 0.492, respectively. Note that the selected thresholds for FCID-HIST and FCID-FE will be employed in the subsequent experiments.

Since FCID-HIST exploits the histogram distributions to extract the detection features, the number of bins of the histograms $K_{c_f}, c_f = h, s, dc, bc$ should be determined as well. Intuitively, when K_{c_f} increases, part of the detection feature corresponding to the most distinctive bins may become less distinctive, while the rest of the detection feature corresponding to the total variations may capture more details and thus become more distinctive. To reveal the effects of K_{c_f}, the par-train and par-test sets and the SVM parameters determined

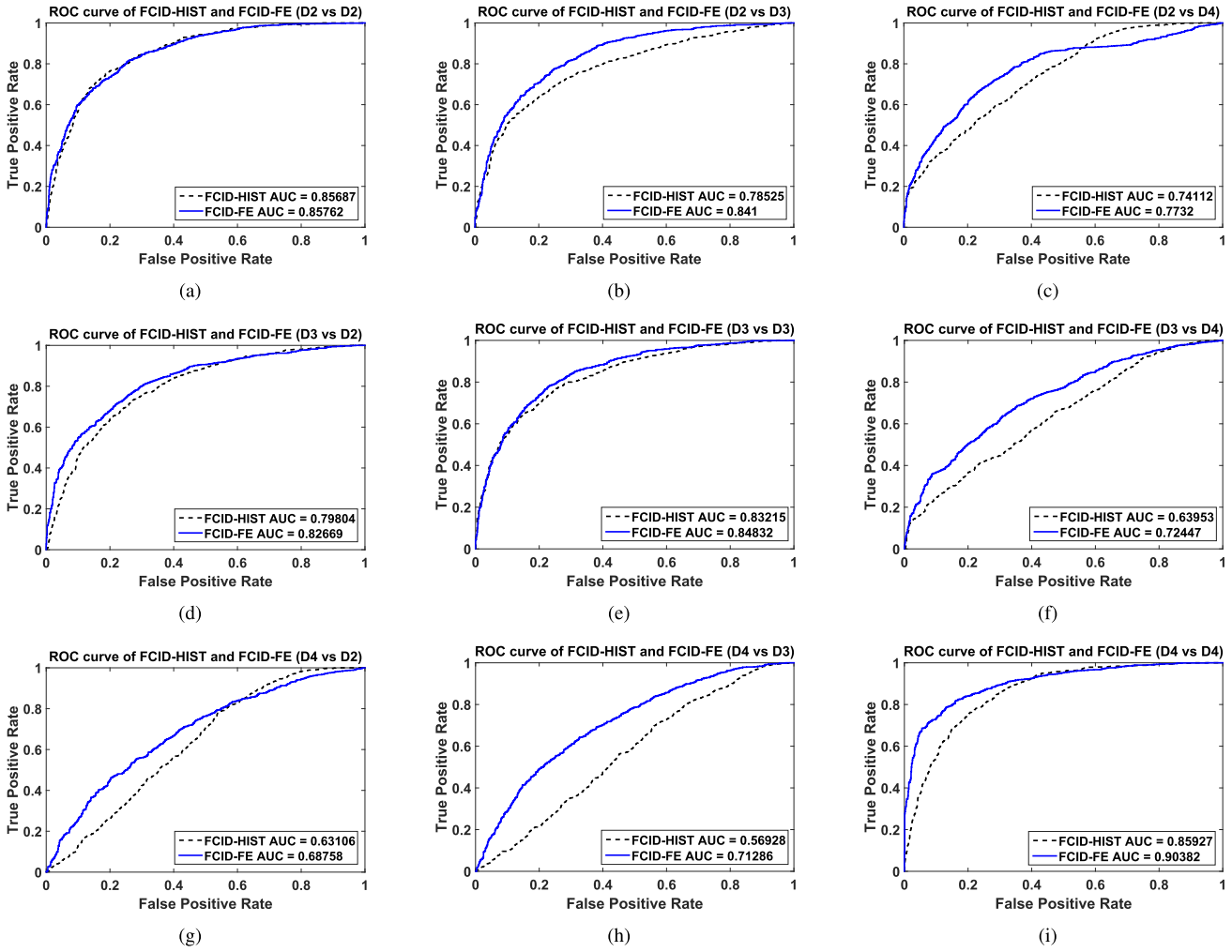


Fig. 6. Detection results for the cross colorization method tests. (a) $D2([4])$ vs $D2([4])$. (b) $D2([4])$ vs $D3([36])$. (c) $D2([4])$ vs $D4([35])$. (d) $D3([36])$ vs $D2([4])$. (e) $D3([36])$ vs $D3([36])$. (f) $D3([36])$ vs $D4([35])$. (g) $D4([35])$ vs $D2([4])$. (h) $D4([35])$ vs $D3([36])$. (i) $D4([35])$ vs $D4([35])$.

TABLE V
OPTIMAL THRESHOLD SELECTION OF FCID-HIST AND FCID-FE (THRESHOLD)

Method\Fold Number	1	2	3	4	5	6	7	8	9	10
FCID-HIST	0.46	0.45	0.46	0.45	0.46	0.43	0.46	0.46	0.45	0.47
FCID-FE	0.5	0.51	0.52	0.43	0.47	0.54	0.45	0.5	0.49	0.51

above are employed. In this test, $K_{c_f}, c_f = h, s, dc, bc$ ranges from 200 to 260 with a step of 5. Besides, we also include $K_{c_f} = 256, c_f = h, s, dc, bc$. As can be observed from Table VI, there exists no obvious trends when K_{c_f} varies. By considering the latter results demonstrated in Section IV-D, in which FCID-HIST gives unstable performances when the training dataset varies, we believe that K_{c_f} is not a deterministic aspect for the performances of FCID-HIST. Therefore, $K_h = K_s = K_{dc} = K_{bc}$ are all set to be 200 for convenience in this paper.

D. Cross Validation

After the parameters are determined, the cross validations are performed on FCID-HIST and FCID-FE separately. Fig. 5 presents the cross validation results of FCID-HIST and FCID-FE. As can be observed, both FCID-HIST and FCID-FE achieve a decent performance, where the average HTER of FCID-HIST is 18.423% and that of FCID-FE is 16.994%. Clearly, FCID-FE provides a slightly better performance compared to FCID-HIST. Note that FCID-HIST gives less consistent performances because the detection feature, especially the most distinctive bins, may vary for different training set. It indicates that the extracted handcrafted features in FCID-HIST possess less robustness compared to the moments-based features in FCID-FE. The detection performances may be improved via exploring better and more consistent features in the future work.

E. Performance Evaluation

In the cross validation tests, both FCID-HIST and FCID-FE performs decently. Here, a comprehensive performance

TABLE VI
THE EFFECTS OF K_{cf} IN FCID-HIST (HTER, IN PERCENTAGE)

K_{cf}	200	205	210	215	220	225	230	235	240	245	250	255	256	260
HTER	21.50	21.75	23.00	20.95	21.05	21.10	20.80	20.10	20.95	21.30	20.15	19.65	20.90	19.90

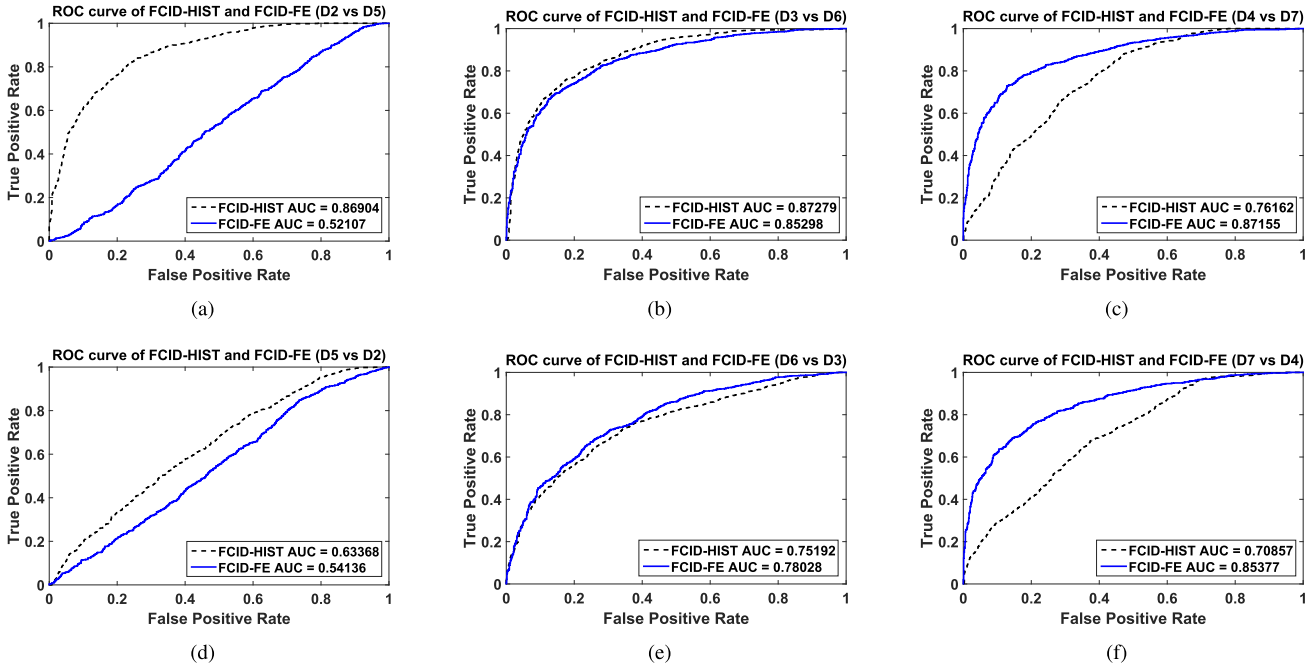


Fig. 7. Detection results with different training vs testing sets. (a) $D2$([4]) vs $D5$([4]). (b) $D3$([36]) vs $D6$([36]). (c) $D4$([35]) vs $D7$([35]). (d) $D5$([4]) vs $D2$([4]). (e) $D6$([36]) vs $D3$([36]). (f) $D7$([35]) vs $D4$([35]).

evaluation for FCID-HIST and FCID-FE is performed with six additional databases $D2$, $D3$, $D4$, $D5$, $D6$ and $D7$.

Since FCID-HIST and FCID-FE construct the feasible features automatically according to the training set, the proposed methods should be capable of detecting the fake images generated by different colorization methods, as long as the colorized images exhibit the observed differences. TO demonstrate the performances of the detection methods against three latest colorization approaches [4], [36], [35], each of $D2$, $D3$ and $D4$ is equally divided into a training set and a testing set.

The experiments are conducted in a manner that the training sets and testing sets may or may not originate from the identical databases, such that 9 experiments are performed to evaluate FCID-HIST and FCID-FE. As can be observed from Tables VII-VIII and Fig. 6, the proposed methods can successfully detect different fake images which are generated from different state-of-the-art colorization approaches, when the training and testing datasets are from the identical or different databases. Besides, FCID-FE gives more accurate detection results compared to FCID-HIST in most situations. Compared to Figs. 6(a), 6(e) and 6(i), performance decreases when the training and testing datasets are from different databases, especially for FCID-HIST. These drops reveal that FCID-FE, which gives more consistent performances, models the statistical information of the images better compared to FCID-HIST.

Next, the cross dataset tests are performed. The natural images in $D2$, $D3$ and $D4$, originating from the ImageNet validation dataset [40], and images in the $D5$, $D6$ and $D7$, originating from the Oxford building dataset [41], are employed to perform the cross dataset tests.

TABLE VII
HTER OF FCID-HIST FOR DIFFERENT DATABASES (IN PERCENTAGE)

Training\Testing	$D2$([4])	$D3$([36])	$D4$([35])
$D2$([4])	22.50	28.00	33.95
$D3$([36])	26.95	24.45	41.85
$D4$([35])	38.15	43.55	22.35

TABLE VIII
HTER OF FCID-FE FOR DIFFERENT DATABASES (IN PERCENTAGE)

Training\Testing	$D2$([4])	$D3$([36])	$D4$([35])
$D2$([4])	22.30	23.65	31.70
$D3$([36])	25.10	22.85	34.25
$D4$([35])	38.50	36.15	17.30

Similar to $D2$, $D3$ and $D4$, $D5$, $D6$ and $D7$ are all equally divided into training and testing sets. By pairing the databases in which the colorized images are generated from the same colorization method, three database pairs, $D2$ and $D5$, $D3$ and $D6$, $D4$ and $D7$, are obtained. For each pair of databases, the cross-dataset tests are performed by employing one database's training set and the other one's testing set, and vice versa. The experimental results of the cross dataset tests are introduced in Tables IX-X and Fig. 7. As shown,

TABLE IX
HTER OF FCID-HIST FOR CROSS-DATASET TESTS (TRAINING VS. TESTING)

$D2$([4]) vs. $D5$([4])	$D3$([36]) vs. $D6$([36])	$D4$([35]) vs. $D7$([35])
22.85	21.50	30.95
$D5$([4]) vs. $D2$([4])	$D6$([36]) vs. $D3$([36])	$D7$([35]) vs. $D4$([35])
43.45	30.75	36.60

TABLE X
HTER OF FCID-FE FOR CROSS-DATASET TESTS (TRAINING VS. TESTING)

$D2$([4]) vs. $D5$([4])	$D3$([36]) vs. $D6$([36])	$D4$([35]) vs. $D7$([35])
51.40	22.70	20.20
$D5$([4]) vs. $D2$([4])	$D6$([36]) vs. $D3$([36])	$D7$([35]) vs. $D4$([35])
49.80	30.25	23.15

although the performance somewhat decreases, both methods still successfully differentiates between the colorized and natural images, and FCID-HIST again gives less stable performances compared to FCID-FE, with the exception of the $D2$ and $D5$ pair. The unsatisfactory performances for the $D2$ and $D5$ pair may be due to the different image content in different image datasets ($D2$ from the ImageNet dataset and $D5$ from the Oxford building dataset), which induces different statistical distributions. Since the proposed methods, especially FCID-FE, rely on extracting the detection features from the entire distributions, the classifier, which is trained by either $D2$ or $D5$, may fail to correctly classify certain images in the other one.

In summary, these results indicate that colorization induces statistical differences in the hue, saturation, dark and bright channels, and demonstrate the robustness of our proposed methods against different colorization methods and across different datasets.

V. CONCLUSION AND DISCUSSION

In this paper, we aimed to address a new problem in the field of fake image detection: fake colorized image detection. We observed that fake colorized images and their corresponding natural images possess statistical differences in the hue, saturation, dark and bright channels. We proposed two simple yet effective schemes, FCID-HIST and FCID-FE, to resolve this detection problem. FCID-HIST exploits the most distinctive bins and total variations of the normalized histogram distributions and creates features for detection, while FCID-FE models the data samples with GMM and creates Fisher vectors for better utilizing the statistical differences. We evaluate the performances of the proposed methods by selecting parameters for FCID-HIST and FCID-FE and detecting different fake images generated by state-of-the-art colorization approaches. The results demonstrate that both FCID-HIST and FCID-FE perform decently against different colorization approaches and FCID-FE provides more consistent and superior performances compared to FCID-HIST in most of the tests.

Although the proposed FCID-HIST and FCID-FE give decent performances in the experiments, this paper is only a preliminary investigation, and there are many directions for future studies that require further exploration. As our results indicate, the performance of our current methods sometimes degrades obviously when the training images and the testing images are generated from different colorization methods or different datasets, thus blind fake colorized image detection features and methods may be developed in the future by studying the common characteristics of different colorization methods. Moreover, better feature encoding approaches can be considered for improving performance, as well as the optimization of the detection features and parameters to improve the custom features constructed in this study.

REFERENCES

[1] H. Farid, "Exposing digital forgeries from JPEG ghosts," *IEEE Trans. Inf. Forensics Security*, vol. 4, no. 1, pp. 154–160, Mar. 2009.

[2] J. Li, X. Li, B. Yang, and X. Sun, "Segmentation-based image copy-move forgery detection scheme," *IEEE Trans. Inf. Forensics Security*, vol. 10, no. 3, pp. 507–518, Mar. 2015.

[3] G. Cao, Y. Zhao, R. Ni, and X. Li, "Contrast enhancement-based forensics in digital images," *IEEE Trans. Inf. Forensics Security*, vol. 9, no. 3, pp. 515–525, Mar. 2014.

[4] G. Larsson, M. Maire, and G. Shakhnarovich, "Learning representations for automatic colorization," in *Proc. Eur. Conf. Comput. Vis. (ECCV)*, 2016, pp. 577–593.

[5] I. J. Goodfellow *et al.*, "Generative adversarial nets," in *Proc. Adv. Neural Inf. Process. Syst. (NIPS)*, 2014, pp. 2672–2680.

[6] F. Huang, X. Qu, H. J. Kim, and J. Huang, "Reversible data hiding in JPEG images," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 26, no. 9, pp. 1610–1621, Sep. 2016.

[7] J. Yin, R. Wang, Y. Guo, and F. Liu, "An adaptive reversible data hiding scheme for JPEG images," in *Proc. Int. Workshop Digit.-Forensics Watermarking*, 2016, pp. 456–469.

[8] J. Wang, S. Lian, and Y.-Q. Shi, "Hybrid multiplicative multi-watermarking in DWT domain," *Multidimensional Syst. Signal Process.*, vol. 28, no. 2, pp. 617–636, 2017.

[9] Y. Yan, W. Ren, Y. Guo, R. Wang, and X. Cao, "Image deblurring via extreme channels prior," in *Proc. IEEE Int. Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jul. 2017, pp. 6978–6986.

[10] J. D. R. Farquhar, S. Szedmak, H. Meng, and J. Shawe-Taylor, "Improving 'bag-of-keypoints' image categorization," Dept. Electron. Comput. Sci., Univ. Southampton, Southampton, U.K., Tech. Rep., 2005.

[11] F. Perronnin and C. Dance, "Fisher kernels on visual vocabularies for image categorization," in *Proc. IEEE Int. Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2007, pp. 1–8.

[12] M. Ali Qureshi and M. Deriche, "A bibliography of pixel-based blind image forgery detection techniques," *Signal Process., Image Commun.*, vol. 39, pp. 46–74, Nov. 2015.

[13] A. J. Fridrich, B. D. Soukal, and J. Lukáš, "Detection of copy-move forgery in digital images," in *Proc. Digit. Forensic Res. Workshop*, 2003.

[14] W. Li and N. Yu, "Rotation robust detection of copy-move forgery," in *Proc. IEEE Int. Conf. Image Process. (ICIP)*, Sep. 2010, pp. 2113–2116.

[15] S.-J. Ryu, M. Kirchner, M.-J. Lee, and H.-K. Lee, "Rotation invariant localization of duplicated image regions based on zernike moments," *IEEE Trans. Inf. Forensics Security*, vol. 8, no. 8, pp. 1355–1370, Aug. 2013.

[16] I. Amerini, L. Ballan, R. Caldelli, A. Del Bimbo, and G. Serra, "A SIFT-based forensic method for copy–move attack detection and transformation recovery," *IEEE Trans. Inf. Forensics Security*, vol. 6, no. 3, pp. 1099–1110, Sep. 2011.

[17] D. G. Lowe, "Distinctive image features from scale-invariant keypoints," *Int. J. Comput. Vis.*, vol. 60, no. 2, pp. 91–110, 2004.

[18] L. Liu, R. Ni, Y. Zhao, and S. Li, "Improved SIFT-based copy-move detection using BFSN clustering and CFA features," in *Proc. IEEE Int. Conf. Intell. Inf. Hiding Multimedia Signal Process. (IIH-MSP)*, Aug. 2014, pp. 626–629.

[19] Y. Li and J. Zhou, "Image copy-move forgery detection using hierarchical feature point matching," in *Proc. Asia–Pacific Signal Inf. Process. Assoc. Annu. Summit Conf. (APSIPA ASC)*, Dec. 2016, pp. 1–4.

[20] T. Bianchi and A. Piva, "Image forgery localization via block-grained analysis of JPEG artifacts," *IEEE Trans. Inf. Forensics Security*, vol. 7, no. 3, pp. 1003–1017, Jun. 2012.

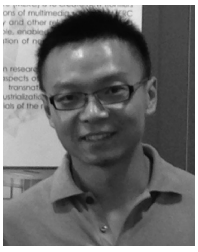
[21] P. Korus and J. Huang, "Multi-scale fusion for improved localization of malicious tampering in digital images," *IEEE Trans. Image Process.*, vol. 25, no. 3, pp. 1312–1326, Mar. 2016.
[22] P. Ferrara, T. Bianchi, A. D. Rosa, and A. Piva, "Image forgery localization via fine-grained analysis of CFA artifacts," *IEEE Trans. Inf. Forensics Security*, vol. 7, no. 5, pp. 1566–1577, Oct. 2012.
[23] G. Chierchia, G. Poggi, C. Sansone, and L. Verdoliva, "A Bayesian-MRF approach for PRNU-based image forgery detection," *IEEE Trans. Inf. Forensics Security*, vol. 9, no. 4, pp. 554–567, Apr. 2014.
[24] P. Korus and J. Huang, "Multi-scale analysis strategies in PRNU-based tampering localization," *IEEE Trans. Inf. Forensics Security*, vol. 12, no. 4, pp. 809–824, Apr. 2017.
[25] K. Bahrami, A. C. Kot, L. Li, and H. Li, "Blurred image splicing localization by exposing blur type inconsistency," *IEEE Trans. Inf. Forensics Security*, vol. 10, no. 5, pp. 999–1009, May 2015.
[26] T. Carvalho, F. A. Faria, H. Pedrini, R. D. S. Torres, and A. Rocha, "Illuminant-based transformed spaces for image forensics," *IEEE Trans. Inf. Forensics Security*, vol. 11, no. 4, pp. 720–733, Apr. 2016.
[27] W. Zhang, X. Cao, Z. Feng, J. Zhang, and P. Wang, "Detecting photographic composites using two-view geometrical constraints," in *Proc. IEEE Int. Conf. Multimedia Expo (ICME)*, Jun./Jul. 2009, pp. 1078–1081.
[28] W. Zhang, X. Cao, Y. Qu, Y. Hou, H. Zhao, and C. Zhang, "Detecting and extracting the photo composites using planar homography and graph cut," *IEEE Trans. Inf. Forensics Security*, vol. 5, no. 3, pp. 544–555, Sep. 2010.
[29] D. T. Trung, A. Beghdadi, and M.-C. Larabi, "Blind inpainting forgery detection," in *Proc. IEEE Global Conf. Signal Inf. Process. (GlobalSIP)*, Dec. 2014, pp. 1019–1023.
[30] A. Levin, D. Lischinski, and Y. Weiss, "Colorization using optimization," *ACM Trans. Graph.*, vol. 23, no. 3, pp. 689–694, 2004.
[31] J. Pang, O. C. Au, K. Tang, and Y. Guo, "Image colorization using sparse representation," in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process. (ICASSP)*, May 2013, pp. 1578–1582.
[32] G. Charpiat, M. Hofmann, and B. Schölkopf, "Automatic image colorization via multimodal predictions," in *Proc. Eur. Conf. Comput. Vis. (ECCV)*, 2008, pp. 126–139.
[33] X. Chen, J. Li, D. Zou, and Q. Zhao, "Learn sparse dictionaries for edit propagation," *IEEE Trans. Image Process.*, vol. 25, no. 4, pp. 1688–1698, Apr. 2016.
[34] Z. Cheng, Q. Yang, and B. Sheng, "Deep colorization," in *Proc. IEEE Int. Conf. Comput. Vis. (ICCV)*, Dec. 2015, pp. 415–423.
[35] S. Iizuka, E. Simo-Serra, and H. Ishikawa, "Let there be color!: Joint end-to-end learning of global and local image priors for automatic image colorization with simultaneous classification," *ACM Trans. Graph.*, vol. 35, no. 4, p. 110, 2016.
[36] R. Zhang, P. Isola, and A. A. Efros, "Colorful image colorization," in *Proc. Eur. Conf. Comput. Vis. (ECCV)*, 2016, pp. 649–666.
[37] K. He, J. Sun, and X. Tang, "Single image haze removal using dark channel prior," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 33, no. 12, pp. 2341–2353, Dec. 2011.
[38] C.-C. Chang and C.-J. Lin, "LIBSVM: A library for support vector machines," *ACM Trans. Intell. Syst. Technol.*, vol. 2, no. 3, 2011, Art. no. 27.
[39] *VLFeat*. Accessed: Jan. 19, 2017. [Online]. Available: http://VLFeat.org
[40] O. Russakovsky *et al.*, "ImageNet large scale visual recognition challenge," *Int. J. Comput. Vis.*, vol. 115, no. 3, pp. 211–252, Dec. 2015.
[41] J. Philbin, O. Chum, M. Isard, J. Sivic, and A. Zisserman, "Object retrieval with large vocabularies and fast spatial matching," in *Proc. IEEE Int. Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2007, pp. 1–8.

Yuanfang Guo received the B.S. and Ph.D. degrees from The Hong Kong University of Science and Technology, Hong Kong, in 2009 and 2015, respectively. He is currently an Assistant Professor with the State Key Laboratory of Information Security, Institute of Information Engineering, Chinese Academy of Sciences, Beijing, China. His research interests include image/video watermarking, data hiding, forensics, compression, restoration, and enhancement.

Xiaochun Cao received the B.S. and M.S. degrees in computer science from Beihang University, Beijing, China, and the Ph.D. degree in computer science from the University of Central Florida, Orlando, FL, USA. After graduation, he spent about three years at ObjectVideo Inc., as a Research Scientist. From 2008 to 2012, he was a Professor with Tianjin University, Tianjin, China.

He is currently a Professor with the Institute of Information Engineering, Chinese Academy of Sciences, Beijing, China. He has authored and co-authored over 170 journal and conference papers.

Prof. Cao is a fellow of the IET. His dissertation was nominated for the University of Central Florida's university-level Outstanding Dissertation Award. In 2004 and 2010, he was a recipient of the Piero Zamperoni Best Student Paper Award at the International Conference on Pattern Recognition. He is on the editorial boards of the IEEE TRANSACTIONS ON IMAGE PROCESSING and the IEEE TRANSACTIONS ON CIRCUITS AND SYSTEMS FOR VIDEO TECHNOLOGY.



Wei Zhang received the B.Eng. and M.Eng. degrees from Tianjin University, Tianjin, China, in 2008 and 2010, respectively, and the Ph.D. degree from the Department of Computer Science, City University of Hong Kong, Kowloon, Hong Kong, in 2015. He was a Visiting Scholar with the DVMM Group, Columbia University, New York, NY, USA, in 2014. He is currently a Research Scientist with JD AI Research. His research interests include large-scale visual instance search and mining, multimedia, and digital forensic analysis.

Rui Wang received the B.S. degree from Tsinghua University, Beijing, China, in 2005, and the Ph.D. degree from the Graduate University of Chinese Academy of Sciences, Beijing, China, in 2011. She is currently a Professor with the State Key Laboratory of Information Security, Institute of Information Engineering, Chinese Academy of Sciences, Beijing, China. Her research interests include image/video processing, image retrieval, and object detection.