# Detecting Image Forgery Based On Color Phenomenology

Jamie Stanton
University of Dayton
Dayton, OH USA
stantonj1@udayton.edu

Keigo Hirakawa
University of Dayton
Dayton, OH USA
khirakawa1@udayton.edu

Scott McCloskey
Honeywell
Minneapolis, MN USA
Scott.McCloskey@honeywell.com

## Abstract

*We propose White Point-Illuminant Consistency (WPIC) algorithm that detects manipulations in images based on the phenomenology of color. Segmented regions of the image are converted to chromaticity coordinates and compared to the white point reported in the camera's EXIF file. In manipulated images, the chromaticity coordinates will have a shifted illuminant color relative to the EXIF-reported white point. Absent manipulation, chromaticity coordinates will be in agreement with the specified white point. We detect image manipulations using a convolutional neural network operating on a histogram of relevant statistics that indicate the white point shift. We verify this using a real world data set to demonstrate its effectiveness.*

## 1. Introduction

In digital media forensics, the ability to verify the provenance of images is important. With advances in image manipulation technology, it has become commonplace to see edited images on social media, the shear volume of which makes it impossible to authenticate the images' integrity by human experts. The goal of image forensics work is to identify manipulated images automatically [13, 15, 2, 5, 1, 7, 17].

In our work, we draw inspirations from methods in [2, 5, 1, 7, 17] to detect image manipulations based on color appearance. Specifically, "color" of an observed light stems from an interaction between the illumination and the reflectance of the object surface. The observed light, therefore, obeys certain rule or structure that is consistent with the physical phenomenology of color. One such structure is a dichromatic image model, where the specular and the diffuse components of the reflected light affect the the overall appearance of the object color in a predictable manner, proven useful in color constancy [10] [6] and computer graphics [14].

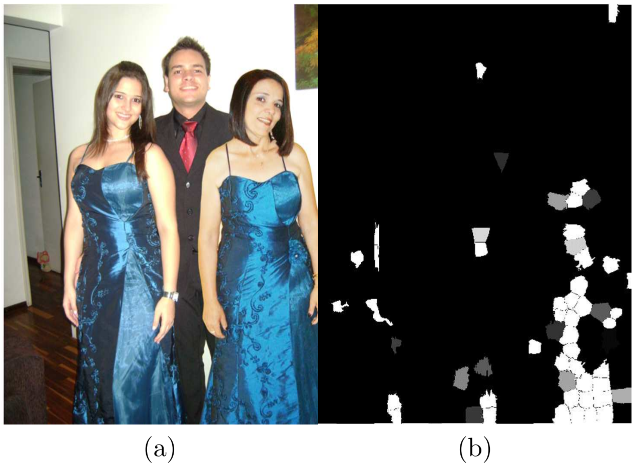
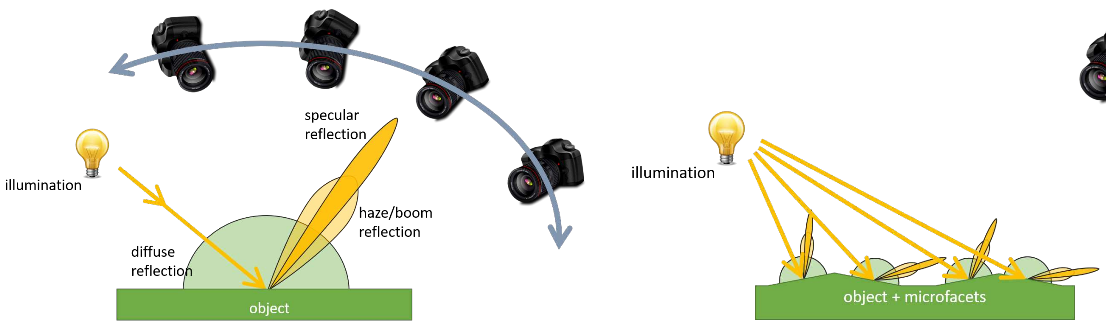


Figure 1. (a) A DSO-1 dataset image [5]. The woman on the right has been spliced in. (b) The proposed WPIC test statistics (shown as pixel intensities) computed over each superpixel strongly suggest that the white point and illuminant color over the spliced woman's dress are inconsistent.

Our hypothesis is that when a color image is manipulated, the manipulated color pixels become inconsistent with the color phenomenology. If a portion of an image is manipulated (e.g. image splice), it would appear as an outlier with respect to the dichromatic image model, while a global manipulation would result in a violation of the physical model entirely. Thus we develop a new classifier aimed at detecting inconsistencies stemming from image manipulations that affect the object color appearance. Specifically, we leverage the colorimetric techniques described in [16] and [9] that make use of the chromaticity coordinate representation to help decompose the specular and diffuse components of the object appearance. We can compare the specular component with the EXIF-reported white point (which is typically corrected to D65 standard illumination) to determine adherence with the dichromatic model.

In Section 2 we review the requisite mathematics to understand the dichromatic model and the white point. In Section 3, we analyze the subtle distinctions between



(a) Fresnel reflection (b) Microfacets

Figure 2. Illustration of Fresnel reflection model. (a) Contributions from specular, boom, and diffuse components depend on the incident and viewing angle of the object. (b) Microfacets randomize the surface normal. A stationary camera observes random specular/boom/diffuse combinations.

the notion of illuminant and the white point, and how this affects the way image splice attacks are carried out. In Section 4 we develop the proposed WPIC algorithm and verify the effectiveness experimentally in Section 5, before making concluding remarks in Section 6.

## 2. Background

### 2.1. Fresnel Reflection

The notion of "color" is a property of light as perceived by human eye and brain. It is a perceptual representation of light spectrum as encoded by three types of photoreceptors in eye called cones. In camera hardware, this information is typically encoded by red, green, and blue intensity values (RGB). While color perception is an important and active area of research [11], in this paper we focus on the color phenomenology, or the physics of light interacting with light which we infer based on RGB values recorded in a camera.

A camera captures an image by recording light reflecting objects illuminated by a light source (the "illuminant"). Known as Fresnel reflection, the spatial-spectral-angular appearance of an object is comprised of diffuse, specular, and haze/boom reflection. Thus the camera observation $\boldsymbol{O} = (O_R, O_G, O_B)^T \in \mathbb{R}^3$ is modeled as their convex combination [8]:

$$\begin{aligned} \boldsymbol{O} =& \alpha\boldsymbol{L} + \mathrm{diag}(\boldsymbol{L})\boldsymbol{R} + \mathrm{diag}(\boldsymbol{L})\,\mathrm{diag}(\boldsymbol{E})\boldsymbol{R} \\ =& \alpha\boldsymbol{L} + \underbrace{\mathrm{diag}(\boldsymbol{L})(\boldsymbol{I}+\mathrm{diag}(\boldsymbol{E}))\boldsymbol{R}}_{\boldsymbol{D}} \end{aligned} \tag{1}$$

where $\alpha \in [0,1]$ is a mixture weight that controls the contribution the contributions from diffuse, specular, and boom components; $\boldsymbol{I} \in \mathbb{R}^{3\times 3}$ is an identity matrix; $\boldsymbol{L} = (L_R, L_G, L_B)^T \in \mathbb{R}^3$ denote the RGB tristimulus values of illumination color; $\boldsymbol{R} = (R_R, R_G, R_B)^T \in [0,3]^3$ denote the RGB reflectance values of object diffuse color; and $\boldsymbol{E} = (E_R, E_G, E_B)^T \in [0,1]^3$ is the RGB reflectance values of an environment such that $\mathrm{diag}(\boldsymbol{L})\boldsymbol{E}$ is an ambient light also illuminating the diffuse object to yield the ambient response $\mathrm{diag}(\boldsymbol{L})\,\mathrm{diag}(\boldsymbol{E})\boldsymbol{R}$. In the subsequent discussion,

$$\boldsymbol{D} = \mathrm{diag}(\boldsymbol{L})(\boldsymbol{I}+\mathrm{diag}(\boldsymbol{E}))\boldsymbol{R} \tag{2}$$

is referred to as the all-encompassing "diffuse" component representing the Lambertian reflection. As illustrated by Figure 2(a), the value of $\alpha$ depend on the object material as well as the incident and viewing angles of the light onto the object surface.

### 2.2. Dichromaticity

Let $\boldsymbol{M} \in \mathbb{R}^{3\times 3}$ denote transformation from camera-specific RGB color space ("$\boldsymbol{O}$") to device-independent XYZ ("$\boldsymbol{X}$") color spaces. Rewriting (1) yields:

$$\boldsymbol{X} = \boldsymbol{MWO} = \alpha\boldsymbol{MWL} + \boldsymbol{MWD} \tag{3}$$

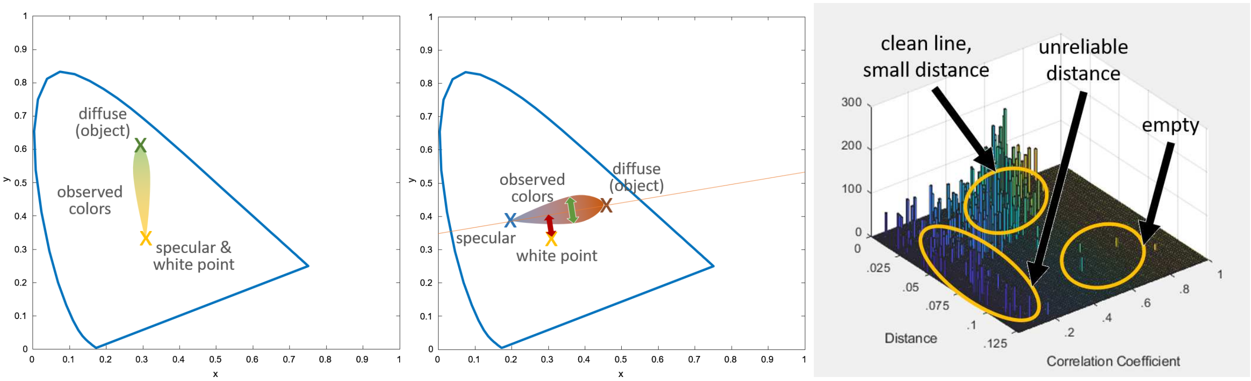
where $\boldsymbol{W} \in \mathbb{R}^{3\times 3}$ is a white balance matrix (see Section 2.3). Since "color" is invariant to the light intensity, the chromaticity coordinate representation normalizes $\boldsymbol{X}$:

$$\begin{aligned} \boldsymbol{x} =& \frac{\boldsymbol{X}}{\|\boldsymbol{X}\|_1} = \frac{\alpha\boldsymbol{MWL}+\boldsymbol{MWD}}{\|\alpha\boldsymbol{MWL}+\boldsymbol{MWD}\|_1} \\ =& \beta\underbrace{\frac{\boldsymbol{MWL}}{\|\boldsymbol{MWL}\|_1}}_{\boldsymbol{\ell}} + (1-\beta)\underbrace{\frac{\boldsymbol{MWD}}{\|\boldsymbol{MWD}\|_1}}_{\boldsymbol{d}}, \end{aligned} \tag{4}$$

where $\|\cdot\|_p$ is an $L^p$ norm, and

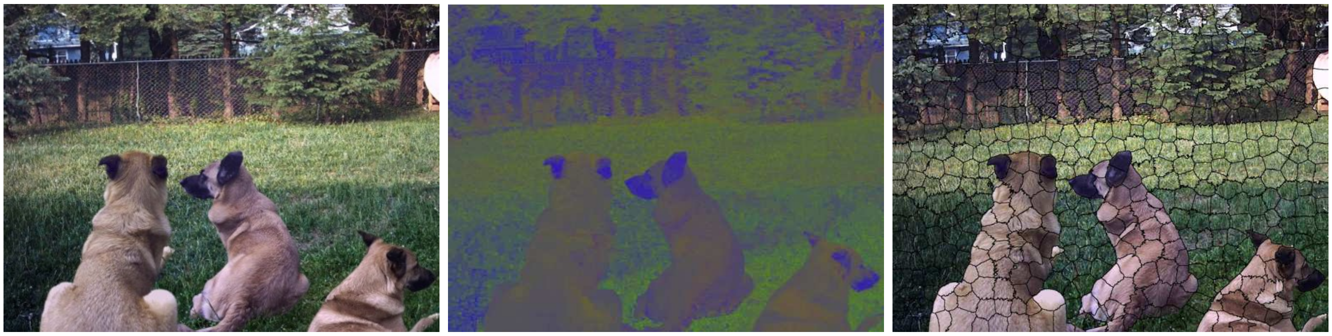
$$\beta = \frac{\alpha\|\boldsymbol{MWL}\|_1}{\|\alpha\boldsymbol{MWL}+\boldsymbol{MWD}\|_1}. \tag{5}$$

The significance is that the chromaticity coordinate $\boldsymbol{x}$ of the recorded light is a convex combination of the illuminant chromaticity $\boldsymbol{\ell} := \frac{\boldsymbol{MWL}}{\|\boldsymbol{MWL}\|_1} \in [0,1]^3$ and



(a) (b) (c)

Figure 3. Conceptual illustration of the test statistics. (a) A scatter plot of chromaticity coordinates from pixels belonging to the same microfaceted object. The line formed by the scatter plot intersects the illuminant/white point (yellow ×) and the diffuse (green ×) colors. (b) A scatter plot of pixels that have been manipulated. The illuminant/specular (blue ×) no longer coincides with the white point. The test statistics is the Eucledian distance $\varepsilon$ between the white point (yellow ×) and the line formed by the scatter plot (see red arrow); and the Pearson product-moment correlation coefficient $\rho$ of the scatter plot (see green arrow). The test statistics is computed for each superpixel. (c) Illuminant error histogram (IET) is a two-dimensional histogram of $\varepsilon$ and $\rho$, accumulated over all superpixels. If the image is authentic, there is a large concentration of "small $\varepsilon$, high $\rho$." The distance $\varepsilon$ is unreliable when $\rho$ is small. IET is the input into our classifier (CNN).



(a) Input image (b) Chromaticity coordinates (b) Superpixel

Figure 4. (a) Example input image. (b) Chromaticity coordinate normalize by intensity, leaving only color variation. Mapped to RGB for visualization. (c) Superpixels group spatially neighboring pixels of similar appearance.

the diffuse chromaticity $\boldsymbol{d} := \frac{\boldsymbol{MWD}}{\|\boldsymbol{MWD}\|_1} \in [0,1]^3$. Thus the chromaticity coordinate of the observed light lays on a line segment between $\boldsymbol{\ell}$ and $\boldsymbol{d}$, as shown in Figure 3(a). This property is known as the dichromaticity.

Recall that $\alpha \in [0,1]$ depend on the incident/viewing angle of the light relative to the object surface. In the real world, the object surface areas are never as idealized as shown in Figure 2(a). Microfacets (small discontinuities across a surface) cause the the surface normal of the object to have a randomized angles, as in Figure 2(b). Thus $\alpha \in [0,1]$ and $\beta \in [0,1]$ are stochastic—a stationary camera observes random specular/boom/diffuse combination as determined by a *random* convex combination of the illuminant color $\boldsymbol{\ell} \in [0,1]^3$ and the diffuse color $\boldsymbol{d} \in [0,1]^3$ (analogous to rotating a camera to various angles, as in Figure 2(a)). Consider the scatter plot of chromaticity coordinates $\boldsymbol{x} = \beta\boldsymbol{\ell} + (1-\beta)\boldsymbol{d}$ within one single object, as shown in Figure 3(a). Thanks to randomized $\beta$ values, a line formed by the cluster of points intersects $\boldsymbol{\ell}$. This colorimetric property in (4) has been exploited in white balance methods in [16, 9] and computer graphics [14]. We leverage (4) in the proposed image forensics method, described below.

### 2.3. White Point and White Balance

The illuminant color $\boldsymbol{L} = (L_R, L_G, L_B)^T \in \mathbb{R}^3$ varies from scene to scene, affecting the observed light $\boldsymbol{O} = (O_R, O_G, O_B)^T$. Drawing on the fact that the human perception of color is (approximately) invariant to the illuminant color, digital cameras perform "white balance" to compensate for the illuminant color.

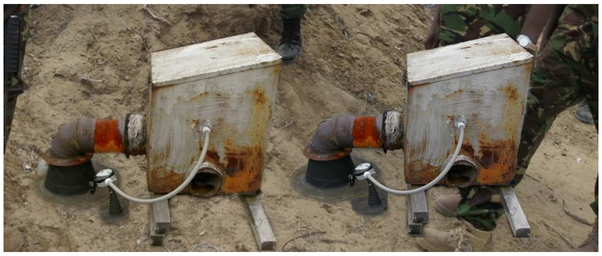


Figure 5. Example of image splice. The objects on the left and right were spliced with and without color transformation (by $\boldsymbol{Q}$), respectively. The color transformation is needed to match the ambient light better.

Specifically, white balance maps the illuminant color $\boldsymbol{\ell} \in [0,1]^3$ to a white point $\boldsymbol{w} \in [0,1]^3$, or a reference white/neutral in a color representation. White point is a quantity that is determined by the camera system and reported by the EXIF file in digital images.

Mathematically, the matrix $\boldsymbol{W} \in \mathbb{R}^{3\times 3}$ is chosen by the white balance algorithm internal to the camera to map the illuminant color $\boldsymbol{WL}$ and illuminant chromaticity $\boldsymbol{\ell} = \frac{\boldsymbol{MWL}}{\|\boldsymbol{MWL}\|_1}$ to the white point $\boldsymbol{w}$ (typically the D65). Hence in an authentic image, the illuminant color coincides with the white point (i.e. $\boldsymbol{w} = \boldsymbol{\ell}$).

## 3. Color Issues In Image Splicing

Because the white balance "standerdizes" the illumination color to the white point, it may seem as though image splicing is easy—after white balance, objects represented in the image *should* appear to be invariant to indoor and outdoor lighting, for example. In reality, the objects taken in differing environments do not match in color appearance. To understand why this is the case, consider the effect that white balance has on the diffuse component $\boldsymbol{D}$:

$$\boldsymbol{WD} = \boldsymbol{W}\,\mathrm{diag}(\boldsymbol{L})\boldsymbol{R} + \boldsymbol{W}\,\mathrm{diag}(\boldsymbol{L})\,\mathrm{diag}(\boldsymbol{E})\boldsymbol{R}. \tag{6}$$

By design, $\boldsymbol{W}\,\mathrm{diag}(\boldsymbol{L})$ has a standard white point color. However, white balance does *not* map the ambient light $\boldsymbol{W}\,\mathrm{diag}(\boldsymbol{L})\,\mathrm{diag}(\boldsymbol{E})$ to the white point $\boldsymbol{w}$.

Denote by $\boldsymbol{E}$ and $\boldsymbol{E}'$ the donor and the probe image environments, respectively. Then the white balanced donor object $\boldsymbol{WD}$ spliced into a new scene would appear "out of place" because human eye expects to see $\boldsymbol{W}\,\mathrm{diag}(\boldsymbol{L})\,\mathrm{diag}(\boldsymbol{E}')\boldsymbol{R}$ instead of $\boldsymbol{W}\,\mathrm{diag}(\boldsymbol{L})\,\mathrm{diag}(\boldsymbol{E})\boldsymbol{R}$. See Figure 5 for an example. Hence additional color processing is needed to blend the donor object into the probe environment. Specifically, an attacker needs to apply a color transformation matrix $\boldsymbol{Q} \in \mathbb{R}^{3\times 3}$ satisfying the following property:

$$\begin{aligned}\boldsymbol{QWD} =& \boldsymbol{QW}\,\mathrm{diag}(\boldsymbol{L})(\boldsymbol{I} + \mathrm{diag}(\boldsymbol{E}))\boldsymbol{R}\\ =& \boldsymbol{W}\,\mathrm{diag}(\boldsymbol{L})(\boldsymbol{I} + \mathrm{diag}(\boldsymbol{E}'))\boldsymbol{R}.\end{aligned} \tag{7}$$

Solving for $\boldsymbol{Q}$ yields

$$\begin{aligned}\boldsymbol{Q} =& \boldsymbol{W}\,\mathrm{diag}(\boldsymbol{L})(\boldsymbol{I} + \mathrm{diag}(\boldsymbol{E}'))\\ & \times (\boldsymbol{I} + \mathrm{diag}(\boldsymbol{E}))^{-1}\,\mathrm{diag}(\boldsymbol{L})^{-1}\boldsymbol{W}^{-1}.\end{aligned} \tag{8}$$

In practice, the attacker is likely to find $\boldsymbol{Q}$ empirically by visual inspection, tweaking the color until $\boldsymbol{QWD}$ blends into the probe scene (i.e. not mathematically). But the achieved effect is the same.

To summarize, the "post-color transformed" spliced regions of the images have the color

$$\boldsymbol{Y} = \boldsymbol{MQX}. \tag{9}$$

Its corresponding chromaticity coordinate is

$$\begin{aligned}\boldsymbol{y} =& \frac{\boldsymbol{Y}}{\|\boldsymbol{Y}\|_1}\\ =& \gamma \underbrace{\frac{\boldsymbol{MQWL}}{\|\boldsymbol{MQWL}\|_1}}_{\boldsymbol{\ell}'} + (1-\gamma) \underbrace{\frac{\boldsymbol{MQWD}}{\|\boldsymbol{MQWD}\|_1}}_{\boldsymbol{d}'},\end{aligned} \tag{10}$$

$$\text{where} \qquad \gamma = \frac{\alpha\|\boldsymbol{MQWL}\|_1}{\|\alpha\boldsymbol{MQWL} + \boldsymbol{MQWD}\|_1}. \tag{11}$$

To the human eye, the spliced color $\boldsymbol{y}$ may appear more convincingly blended into the probe image than $\boldsymbol{x}$. However, the specular component $\boldsymbol{\ell}'$ has moved away from the white point color $\boldsymbol{w}$ reported by the EXIF—a fact we exploit in our proposed algorithm. See Figure 3(b) for illustration.

## 4. Proposed Methodology

### 4.1. Hypothesis and Test Statistics

We propose a method to detect image forgery by constructing a hypothesis test designed to detect a violation of the dichromatic model in (4). Hence we consider a hypothesis test of the following form:

$$\begin{cases} H_0 : \boldsymbol{\ell} = \boldsymbol{w}\\ H_1 : \boldsymbol{\ell} \neq \boldsymbol{w}, \end{cases} \tag{12}$$

where the null hypothesis $H_0$ is a proxy for the scenario that the image is authentic; and the alternative hypothesis $H_1$ indicates the presence of image manipulation as determined by the fact that the illuminant color and the white point do not coincide.

Let $\Lambda = \{\boldsymbol{x}_1, \dots, \boldsymbol{x}_n\}$ be a group of pixels belonging to the same microfaceted object surface. Denote by

$$\boldsymbol{p}(t) = \boldsymbol{\mu} + \boldsymbol{s}\cdot t \tag{13}$$

a parametric equation (about $t \in \mathbb{R}$) for a line that best fits $\Lambda$, where $\boldsymbol{\mu} \in [0,1]^3$ and $\boldsymbol{s} \in \mathbb{R}^3$ are intercept and

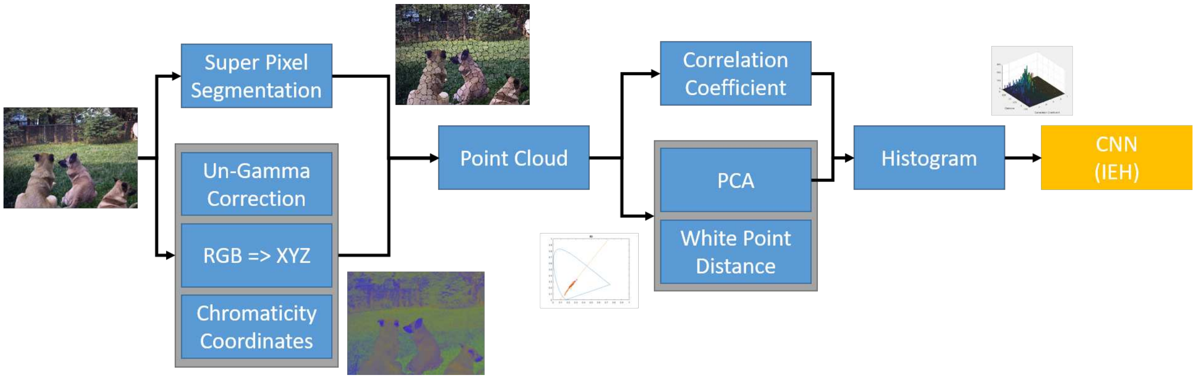


Figure 6. Block diagram of the proposed method.

slope of the line, respectively. From the dichromaticity analyzed in Section 2.2, we know that this line intersects the illuminant (specular) color:

$$\exists t \in \mathbb{R} \quad \ni \quad \boldsymbol{\ell} = \boldsymbol{p}(t). \tag{14}$$

The same line $\boldsymbol{p}(t)$ intersects the white point $\boldsymbol{w} \in [0,1]^3$ also if the image is authentic (the white point and illuminant color coincide), while the line $\boldsymbol{p}(t)$ moves away from $\boldsymbol{w}$ in manipulated images. Thus we consider the error between $\boldsymbol{w}$ and $\boldsymbol{p}(t)$ as test statistics:

$$\varepsilon = \min_{t \in \mathbb{R}} \|\boldsymbol{p}(t) - \boldsymbol{w}\|_2 = \left\| \boldsymbol{\mu} + \boldsymbol{s} \frac{\langle \boldsymbol{s}, \boldsymbol{w} - \boldsymbol{\mu} \rangle}{\langle \boldsymbol{s}, \boldsymbol{s} \rangle} - \boldsymbol{w} \right\|_2 . \tag{15}$$

Intuitively, $\varepsilon$ represents the perpendicular distance from the line to the white point $\boldsymbol{w}$.

In practice, we compute the line $\boldsymbol{p}(t)$ from $\Lambda$ via principal component analysis. Specifically, let $\boldsymbol{\mu} \in [0,1]^3$ and $\boldsymbol{\Sigma} \in \mathbb{R}^{3\times 3}$ be the mean and covariance matrix of these pixels, respectively, computed as:

$$\boldsymbol{\mu} = \frac{1}{n} \sum_{k=1}^{n} \boldsymbol{x}_k, \quad \boldsymbol{\Sigma} = \frac{1}{n} \sum_{k=1}^{n} (\boldsymbol{x}_k - \boldsymbol{\mu})(\boldsymbol{x}_k - \boldsymbol{\mu})^T. \tag{16}$$

The Euclidean distance between $\boldsymbol{x} \in \Lambda$ and $\boldsymbol{p}(t)$ in (13) is minimized when $\boldsymbol{s}$ is the eigenvector of $\boldsymbol{\Sigma}$ corresponding to the largest eigenvalue. We assess the quality of this line fitting by the magnitude of Pearson product-moment correlation coefficient,

$$\rho = \left| \frac{\boldsymbol{\Sigma}_{12}}{\sqrt{\boldsymbol{\Sigma}_{11} \boldsymbol{\Sigma}_{22}}} \right|, \tag{17}$$

where $\boldsymbol{\Sigma}_{ij}$ denotes the $(i,j)$th entry in matrix $\boldsymbol{\Sigma}$. Intuitively, $\rho$ represents the strength of the linear relationship among the pixels $\{\boldsymbol{x}_1, \ldots, \boldsymbol{x}_n\}$. In our work, we treat $\rho$ as an additional test statistics.

To summarize, we make inference on the hypothesis test in (12) via the test statistics $(\varepsilon, \rho)$. The error $\varepsilon$ is expected to be small if the image is authentic because the line $\boldsymbol{p}(t)$ intersects the white point $\boldsymbol{w}$. The value of $\varepsilon$ may become large if the image has been manipulated, or if $\rho$ is small (i.e. our confidence in a line is low).

### 4.2. White Point-Illuminant Consistency Algorithm

A high-level schematic of the proposed White Point-Illuminant Consistency (WPIC) algorithm for detecting manipulated images is shown in Figure 6. The detailed description of the blocks are provided below.

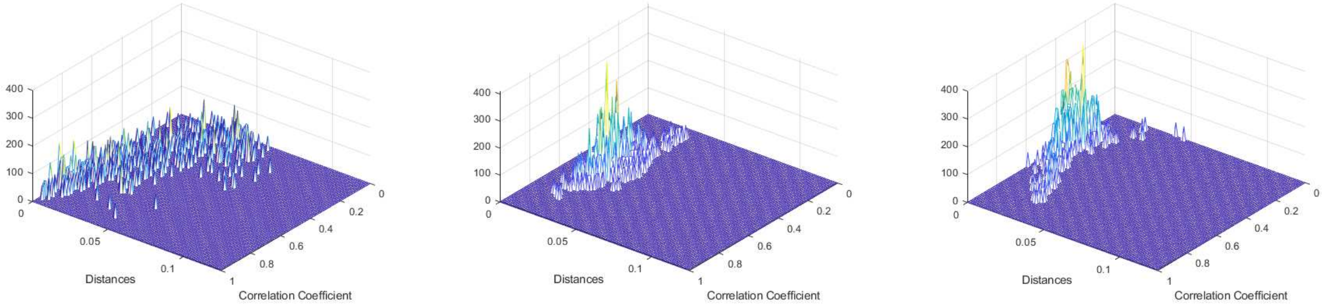
#### Preprocessing

Assuming that a color image is encoded in sRGB space (true for most modern digital cameras), we begin by estimating the XYZ coordinate $\boldsymbol{X}$ from a given color image. Specifically, we apply inverse gamma correction to each sRGB pixel value to recover the corresponding linear sRGB coordinates. We then convert the linear sRGB values to XYZ coordinates via a matrix transformation. Finally we compute the chromaticity coordinates via $\boldsymbol{x} = \boldsymbol{X}/\|\boldsymbol{X}\|_1$. The recovered chromaticity coordinate $\boldsymbol{x}$ has a specular/diffuse decomposition as described by the dichromatic model in (4).

#### Superpixel and Illuminant Error Histogram

Let $k \in \{1, \ldots, K\}$ be the index of superpixels, and denote by $(\varepsilon_k, \rho_k)$ the test statistics (as described in Section 4.1) computed from the pixels within the $k$th superpixel. Drawing on prior by work in [4, 3], we propose the notion of illuminant error histogram (IEH), or a two dimensional histogram of the test statistics:

$$\begin{aligned} H(i,j) =& \sum_{k=1}^{K} \delta\left(i - \left\lfloor \frac{\varepsilon_k}{\Delta\varepsilon} \right\rfloor\right) \delta\left(j - \left\lfloor \frac{\rho_k}{\Delta\rho} \right\rfloor\right) \\ =& \# \left\{ k \in \{1, \ldots, K\} \,\middle|\, i = \left\lfloor \frac{\varepsilon_k}{\Delta\varepsilon} \right\rfloor, j - \left\lfloor \frac{\rho_k}{\Delta\rho} \right\rfloor \right\}, \end{aligned}$$



(a) authentic image (b) color manipulated image (c) spliced image

Figure 7. Examples of Illuminant Error Histogram (IEH).

where $\delta(\cdot)$ is a Kronecker delta function, $\lfloor\cdot\rfloor$ denotes floor function, and $\Delta\varepsilon$ and $\Delta\rho$ are the histogram bin sizes. We interpret $H$ as a global feature for the image.

Consider examples of IEH in Figure 7(a), computed from an authentic image. If the dichromatic model in (4) holds, and if the mixing parameter $\beta$ is indeed stochastic due to microfacets, then we expect small $\varepsilon_k$ and high $\rho_k$ values—a fact verified empirically by IEH in Figure 7(a). In some superpixels, $\rho_k$ is small to indicate that the line in (13) does not fit the pixels in $\Lambda$ well—this may have occurred because the superpixel grouping of pixels span across multiple objects in the scene, or due to the heterogeneity of the texture regions. As such, our classifier must tolerate large $\varepsilon$ value when $\rho$ is small (due to increased uncertainty).

Compare this to the IEH in Figure 7(b), computed from a color transformed image. Because the colors have been manipulated globally (e.g. color balance), IEH is entirely shifted away from the $\varepsilon = 0$ axis (even with large $\rho_k$ value), where the offset roughly represents the deviation of the new illuminant color relative to $\boldsymbol{\ell}$. On the other hand, splicing merges multiple illuminations together, resulting in "high $\varepsilon$, high $\rho$" test statistics. See Figure 7(c). We conclude that IEH is sensitive to image manipulations, providing opportunities to detect edited images.

### Classification

We designed a convolutional neural network designed to detect the presence of image manipulation based on the IEH, modeled after the histogram-based CNN in [4, 3]. The configuration of CNN is shown in Figure 8. The input into the CNN is a 101×101 IEH, normalized by the maximum value. In the first stage, there are two Conv-ReLU layers of 3×3×64, followed by a MaxPool by stride 2. Next it passes through the second stage with two 3×3×128 Conv-ReLu layers, followed by a MaxPool by stride 2. In the third stage, there are four 3×3×256 Conv-ReLu layers, followed by a MaxPool by stride 2. The fourth level consists of three 3×3×512 Conv-Relu layers, followed by a MaxPool by stride 2. In the final stage, data is combined by a fully connected layer, whose output is used to make a final determination on whether an image is manipulated or not by thresholding the output score values yielded by the previous layer.

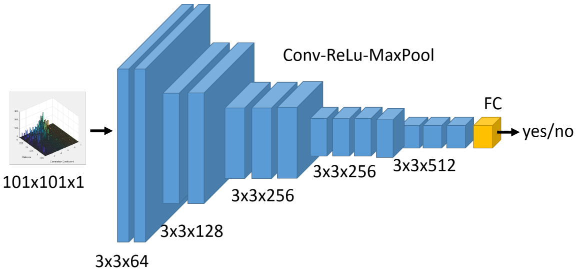


Figure 8. CNN layers.

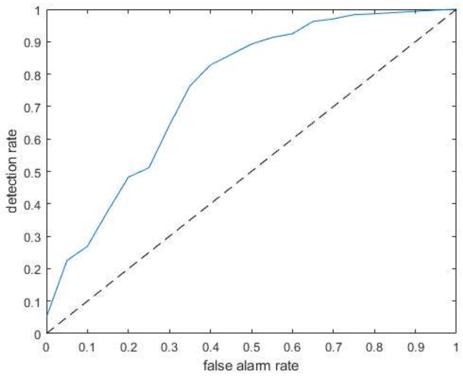


Figure 9. ROC curve for DSO-1 dataset[5].

## 5. Experimental Verification

### 5.1. Experiment Setup

The DSO-1 dataset was developed to test algorithms that detect image splicing [5]. DSO-1 consists of 200

Table 1. Area under curve (AUC) computed for DSO-1 [5].

| Category | Method | AUC |
|---|---|---|
| Supervised | Carvalho [5] | 86.3% |
| | Carvalho [1] | 97.2% |
| Unsupervised | Carvalho[5] | 63.0% |
| | Gholap [7] | 55.5% |
| | Wu[17] | 57.0% |
| | Proposed | 76.0% |

images—100 authentic and 100 manipulated images, which are created by splicing an additional person into a picture that already had at least one other person. Images are comprised of indoor and outdoor scenes, and the image resolution is 2048×1536 pixels.

Due to the limited size of the DSO-1 dataset, we augment the training data with the image forensics dataset from NIST Media Forensics Challenge (MFC18) [12]. This dataset is comprised of many manipulation types, including splicing, color balancing, among others. In order to compare our performance to the prior art, testing was conducted exclusively on DSO-1. NIST MFC18 dataset was used only for training purposes.

### 5.2. CNN Training Procedure

In order to overcome the difficulties of relatively small dataset, we developed a multi-step training procedure. As a first step to CNN training, we take authentic images and perform "random" white balance—adjusting white point to a color temperature that is not D65 (i.e. to simulate the behavior of $\boldsymbol{Q}$ matrix in (10)). We then take convex combinations of two IEHs—first is an IEH corresponding to color transformed images (proxy for donor image), and the second is an IEH of another image without color transformation (proxy for probe). These "synthetic" spliced IEHs were used to train the CNN initially. During the second stage of CNN training, we used the MFC18 dataset. We randomly selected 8000 manipulated and 8000 authentic images, which was subsequently used to update the CNN coefficients via the transfer learning. In our final step of training, we carried out a five-fold cross-validation to fine-tune the CNN using DSO-1 dataset.

### 5.3. Results

The receiver operator characteristics (ROC) curve for the DSO-1 dataset is shown in Figure 9—it is averaged over the testing images of the cross-validated results. As tabulated in Table 1, the area under the curve (AUC) is 76.0%, which is signifiantly better than chance (which has a 50% AUC).

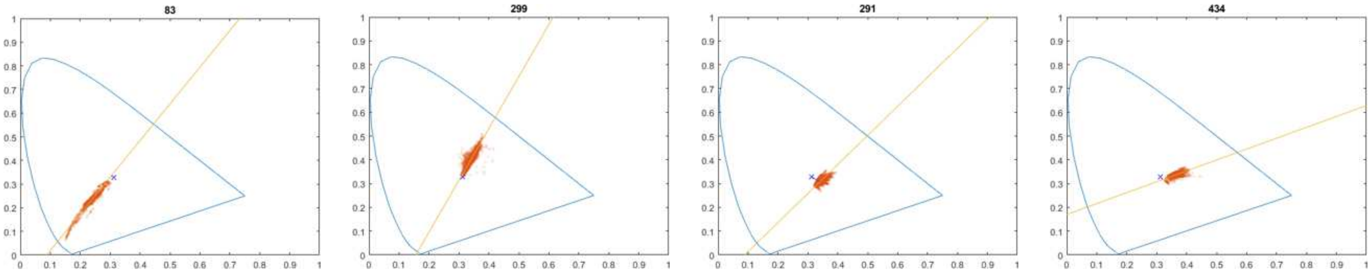
We compared our results to Carvalho2013 method in [5], Carvalho2016 method in [1], Gholap2008 method in [7], and Wu2011 method in [17]. The AUC values for the DSO-1 dataset were reported in [5, 1], which we reproduced in Table 1. The best performing methods were "supervised"—i.e., a human operator manually segemented faces in the DSO-1 images. The best performance among the unsupervised state-of-the-art methods was also Carvalho2013 method in [5] with automatic face recognition replacing the human operator, achieving 63.0% AUC. Despite the "automatic" approach, this method assumes *a priori* knowledge that the DSO-1 dataset's splicing included human faces. That is, the Carvalho2013 metho can not be applied to images *without* faces, whereas our method is more general since it does not restrict the semantic content of the image. Despite this, our proposed method achieved a higher AUC score, and is compatible with fully-automated forensic analysis.

### 5.4. Discussions

In the context of the color-based image forensics techniques, the proposed WPIC algorithm shares with the state-of-the-art methods in [2, 5, 1, 7, 17] the goal to detect inconsistencies in illumination. We highlight two major differences. First, the previous methods leveraged various white balance techniques for spatially local illuminant color estimation. By and large, these methods give "black box" treatment to white balance techniques, where the novelty focused on higher level details of how the estimated illuminant colors were handled. By contrast, WPIC is developed at a lower level, explicitly modifying the white balance techniques of [16, 9] in order to repurpose the theory for image forensics. Second, the previous methods relied on the "across patch" variation among estimated local illuminant. Our WPIC work instead focused on the "within patch" variations of the chromaticity coordinates for the evidences of image forgery. In our future research, we plan to explore the joint optimization of "across patch" and "within patch" variations.

## 6. Conclusion

In this paper, we proposed a novel methodology for detecting image manipulations by localizing violations of dichromatic property of Fresnel reflection model. Specifically, we convert the color images into chromaticity coordinate, and measure the shift of illumination color away from the white point that may have occurred due to image spicing or color manipulations. This shift is then classificed using CNN. The experimental results showed that our method is approaching the performance of the state-of-the-art supervised image forensics technique, and surpasses the performance of the state-of-the-art unsupervised methods.



(a) (b) (c) (d)

Figure 10. Scatter plot of chormaticity coordinates from (a,b) authentic and (c,d) manipulated image patches. The white point is indicated by the ×. The shift of the scatter plot away from white point is evident in manipulated images.

## 7. Acknowledgements

We thank the authors of [5] for providing access to the DSO-1 dataset. This material is based upon work supported by the United States Air Force and the Defense Advanced Research Projects Agency under Contract No. FA8750-16-C-0190. Any opinions, findings and conclusions or recommendations expressed in this material are those of the authors and do not necessarily reflect the views of the United States Air Force or the Defense Advanced Research Projects Agency.

## References

[1] T. Carvalho, F. A. Faria, H. Pedrini, R. d. S. Torres, and A. Rocha. Illuminant-based transformed spaces for image forensics. *IEEE transactions on information forensics and security*, 11(4):720–733, 2016. 1, 7

[2] T. Carvalho, H. Farid, and E. R. Kee. Exposing photo manipulation from user-guided 3d lighting analysis. In *Media Watermarking, Security, and Forensics 2015*, volume 9409, page 940902. International Society for Optics and Photonics, 2015. 1, 7

[3] C. Chen, S. McCloskey, and J. Yu. Image splicing detection via camera response function analysis. In *Proceedings of the IEEE conference on computer vision & pattern recognition*, pages 5087–5096, 2017. 5, 6

[4] C. Chen, S. McCloskey, and J. Yu. Focus manipulation detection via photometric histogram analysis. In *Proceedings of the IEEE Conference on Computer Vision & Pattern Recognition*, pages 1674–1682, 2018. 5, 6

[5] T. J. De Carvalho, C. Riess, E. Angelopoulou, H. Pedrini, and A. de Rezende Rocha. Exposing digital image forgeries by illumination color classification. *IEEE Transactions on Information Forensics and Security*, 8(7):1182–1194, 2013. 1, 6, 7, 8

[6] D. A. Forsyth. A novel algorithm for color constancy. *International Journal of Computer Vision*, 5(1):5–35, 1990. 1

[7] S. Gholap and P. Bora. Illuminant colour based image forensics. In *TENCON 2008-2008 IEEE Region 10 Conference*, pages 1–5. IEEE, 2008. 1, 7

[8] X. D. He, K. E. Torrance, F. X. Sillion, and D. P. Greenberg. A comprehensive physical model for light reflection. *SIGGRAPH Comput. Graph.*, 25(4):175–186, July 1991. 2

[9] H.-C. Lee. Method for computing the scene-illuminant chromaticity from specular highlights. *JOSA A*, 3(10):1694–1699, 1986. 1, 3, 7

[10] L. T. Maloney and B. A. Wandell. Color constancy: a method for recovering surface spectral reflectance. *JOSA A*, 3(1):29–33, 1986. 1

[11] J. C. Maxwell. *The Scientific Papers of James Clerk Maxwell...*, volume 2. University Press, 1890. 2

[12] N. I. of Standards and Technology. Media Forensics Challenge. https://www.nist.gov/itl/iad/mig/media-forensics-challenge-2019-0, 2019. [Online; accessed 3-May-2019]. 7

[13] T. Pomari, G. Ruppert, E. Rezende, A. Rocha, and T. Carvalho. Image splicing detection through illumination inconsistencies and deep learning. In *2018 25th IEEE International Conference on Image Processing (ICIP)*, pages 3788–3792. IEEE, 2018. 1

[14] S. Rusinkiewicz. A survey of brdf representation for computer graphics. *http://www-graphics. stanford. edu/cs348c/surveypaper. html*, 1997. 1, 3

[15] E. Silva, T. Carvalho, A. Ferreira, and A. Rocha. Going deeper into copy-move forgery detection: Exploring image telltales via multi-scale analysis and voting processes. *Journal of Visual Communication and Image Representation*, 29:16–32, 2015. 1

[16] S. Tominaga and B. A. Wandell. Standard surface-reflectance model and illuminant estimation. *JOSA A*, 6(4):576–584, 1989. 1, 3, 7

[17] X. Wu and Z. Fang. Image splicing detection using illuminant color inconsistency. In *2011 Third International Conference on Multimedia Information Networking and Security*, pages 600–603. IEEE, 2011. 1, 7