# CMA: A Chromaticity Map Adapter for Robust Detection of Screen-Recapture Document Images

Changsheng Chen[1], Liangwei Lin[1], Yongqi Chen[1], Bin Li[1]*, Jishen Zeng[2], Jiwu Huang[3]
[1] Guangdong Key Laboratory of Intelligent Information Processing,
and Shenzhen Key Laboratory of Media Security, Shenzhen University, [2]Alibaba Group,
[3] Guangdong Laboratory of Machine Perception and Intelligent Computing,
Faculty of Engineering, Shenzhen MSU-BIT University
cschen@szu.edu.cn, 2210433032@email.szu.edu.cn

## Abstract

*The rebroadcasting of screen-recaptured document images introduces a significant risk to the confidential documents processed in government departments and commercial companies. However, detecting recaptured document images subjected to distortions from online social networks (OSNs) is challenging since the common forensics cues, such as moiré pattern, are weakened during transmission. In this work, we first devise a pixel-level distortion model of the screen-recaptured document image to identify the robust features of color artifacts. Then, we extract a chromaticity map from the recaptured image to highlight the presence of color artifacts even under low-quality samples. Based on the prior understanding, we design a chromaticity map adapter (CMA) to efficiently extract the chromaticity map, and feed it into the transformer backbone as multi-modal prompt tokens. To evaluate the performance of the proposed method, we collect a recaptured office document image dataset with over 10K diverse samples. Experimental results demonstrate that the proposed CMA method outperforms a SOTA approach (with RGB modality only), reducing the average EER from 26.82% to 16.78%. Robustness evaluation shows that our method achieves 0.8688 and 0.7554 AUCs under samples with JPEG compression (QF=70) and resolution as low as 534×503 pixels.*

## 1. Introduction

Document images, such as certificates, contracts, and identity documents, are gaining popularity in e-business and e-government applications, which brings both convenience and threat to our applications. Traditionally, an organization controls the distribution of hard-copy documents to guard against the leakage of confidential information. Many office documents with confidential information can be viewed on screen for a limited time but cannot be printed as hard copies [29]. However, illegal users could acquire the **genuine document image** shown on display with their smartphones and distribute the **recaptured document image** without being noticed [2, 31]. Worse still, these images are usually transmitted through online social networks (OSN), *e.g.*, WhatsApp, and WeChat, which introduces further distortions (compression, resizing, *etc.*) to the document images. Thus, there is a pressing need to develop a recaptured document detection scheme robust to various distortions.

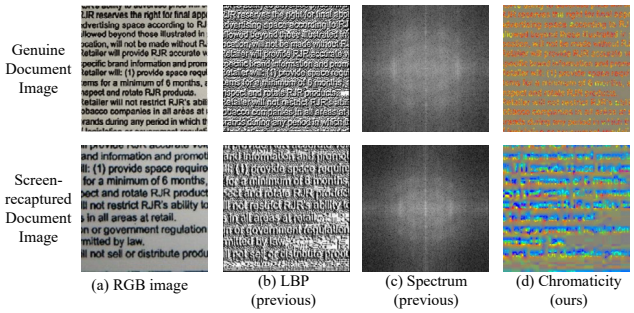


Figure 1. Illustrations of the low-quality genuine (top row) and recaptured (bottom row) samples and their transformed domain representations. (a) Image patches in RGB space. The samples are collected by Canon C3530 (4800×1200 DPI) printer, OnePlus 5T camera (resolution at 1280×960 pixels), Dell P2418D display (resolution at 2560×1440 pixels, size 23.8 inches), and subject to JPEG compression with a quality factor of 70. (b) LBP maps with a radius 1 containing 8 elements [40]. (c) Amplitude spectrum in the Fourier domain. (d) Chromaticity map extracted by Eq. (5) in our work. The color artifacts highlighted in (d) show clear differences between the genuine and recaptured samples.

Existing works on recaptured image detection focus on face and natural images. For instance, literature exploits physical traces of distortion, *e.g.*, specularity distribution [52], color saturation [7], edge blurriness [45] and moiré

*B. Li is the corresponding author.

texture [19, 57, 58] to identify the natural images that are recaptured from LCD screens. Many state-of-the-art (SOTA) face anti-spoofing (FAS) methods incorporate multi-modal forensic data, *e.g.*, depth-based cues [3] and remote photoplethysmography signals [37], to improve the robustness under challenging scenarios. There are significant limitations in the existing approaches. For these methods based on hand-crafted features [7, 45, 52, 57], the robustness is not satisfactory under documents with various contents and image qualities. For those multi-modal FAS approaches [3, 37], the additional modalities do not apply to document images [42, 55]. Specifically, there is no depth difference between genuine and recaptured document images, and no physiological signal is available in a document.

There are a few works on recaptured document detection. Recently, Chen *et al*. [9] proposed a DPAD scheme by a side-by-side forensic similarity comparison between the questioned document image and a genuine reference. Many existing DPAD methods only consider limited types of document contents [9, 28, 34]. To improve the generalization performance across different content, Benalcazar *et al*. [4] developed a synthetic document image generation scheme by overlaying moiré patterns extracted from recaptured images to genuine document images. However, as demonstrated in Appendix D, the performance of deep models trained by synthetic moiré data from [4] is not robust in the samples transmitted by OSN. A similar drawback is found in our recent work [10]. This is because the moiré patterns in the low-quality samples are removed during the blurring and down-sampling process. As shown in Fig. 1 (b) and (c), some common textural and spectral representations of document images are not distinctive under the OSN distortions.

In this work, we exploit a model of pixel-level screen-recapture distortion to address the limitations of existing methods, *i.e.*, generalization and robustness. According to our insight of the distortion model, the color artifacts are identified along the character edges. As shown in Fig. 1 (d), we aggregate the color artifacts of a document image into the chromaticity map. There are two merits for the chromaticity map. First, the color features in the chromaticity map are independent of the high-level semantic content of the document images, which improves the generalization performance. Second, the color artifacts are strong even in recaptured images without a moiré pattern, allowing robust detection of the low-quality recaptured samples.

Subsequently, we propose the chromaticity map adapter (CMA) to incorporate the extracted chromaticity map as multi-modal prompt tokens into the transformer backbones. Cross-domain experimental results show that the proposed CMA improves the performance of a SOTA approach (with RGB modality alone) by, on average 37.71 % reduction of EER or 10.04 percentage points (p.p.) improvement. Our CMA also shows 0.8688 AUC under a JPEG compression distortion with a quality factor of 70, which shows 0.1085 improvement in AUC over a generic ViT-based multi-modal transformer with RGB and chromaticity inputs.

The main contributions of this work are as follows.

- We devise a pixel-level distortion model focusing on resampling operations in the screen-recaptured document images. This model reveals the underlying source and robustness properties of the color artifacts.
- We extract the chromaticity map to highlight the color artifacts in a document image and design a chromaticity map adapter (CMA) to efficiently input the forensic cues to a transformer backbone.
- We gather and share the Recaptured Office Document (ROD) dataset, with 4860 genuine images and 6027 recaptured images, which covers document images with various acquisition devices, contents, and image qualities.

## 2. Literature

### 2.1. Works on Forensic Color Cues

Color has been an important forensic cue. Riess and Angelopoulou [43] proposed the inverse-intensity chromaticity (IIC) space to estimate the illumination color. The illumination map based on IIC space is then employed in the manipulation detection of digital images. Carvalho *et al*. [15] exploited the fact that objects of similar material show similar color characteristics, which applies to the problem of face manipulation detection in a photo with two or more people. Hadwiger and Riess [24] learned a robust image splicing localization method for detecting image regions from different images based on the contrastive loss of color features. Li *et al*. [33] analyzed the residuals of color components in HSV and YCbCr color spaces to distinguish the real and deep network generated images.

However, existing methods on this topic are designed for different forensics tasks, which are different from the recaptured document detection task in our work. Compared with image tampering localization, the recaptured document detection task needs to identify the forensic discrepancies across different samples instead of finding the image regions with large intra-sample differences within an image. Without a counterpart for comparison, the latter task is more susceptible to variations in the questioned samples, *e.g.*, acquisition devices, contents, and image qualities.

As a side note, there have been many research efforts on color constancy which is an important step in estimating the illumination color. Color constancy is a fundamental low-level computer vision task that has been studied for decades [11, 23]. There is a growing interest in color constancy research based on deep learning [32, 39, 41, 53, 56]. However, these works do not aim at forensic applications.

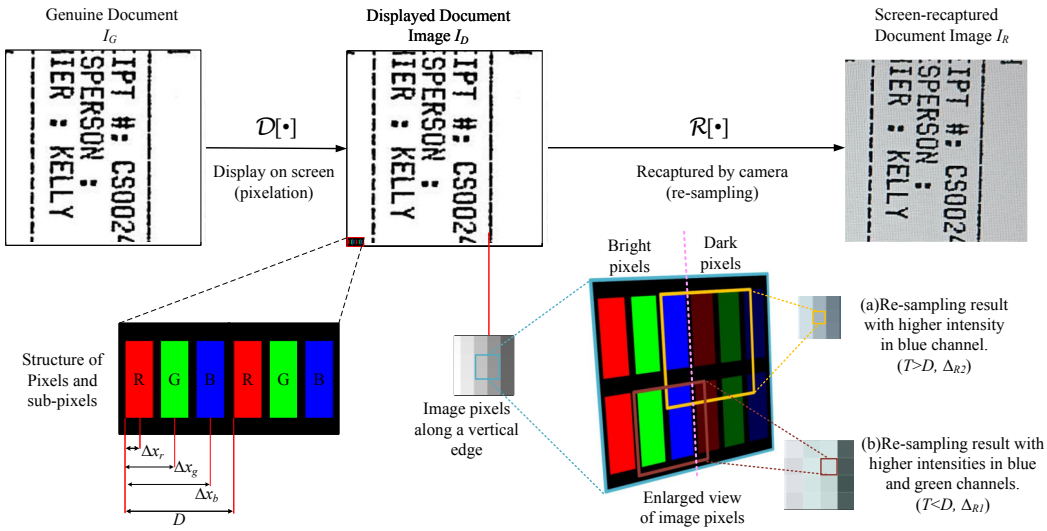


Figure 2. The block diagram illustrates the collection of screen-recaptured document images with emphasis on the re-sampling operation in pixels and sub-pixels. (a) The result of re-sampling covering bright blue sub-pixel along a dark edge. (b) The result of re-sampling covering bright blue and green sub-pixels along a dark edge.

### 2.2. Works on Adapter in Transformers

An adapter adds new modules between layers of a pre-trained network. Only the task-specific parameters are added and trained for a new task. Houlsby *et al*. [27] first introduced a bottleneck adapter structure within transformer blocks, freezing the original backbone to rapidly adapt the pre-trained model to downstream natural language processing tasks, achieving performance comparable to fine-tuning. In the domain of computer vision, Li *et al*. [35] proposed fine-tuning ViT [17] for object detection with minimal modifications. ViT-adapter [13] employed adapters to facilitate the standard ViT in handling various downstream tasks. Liu *et al*. [38] introduced the Explicit Visual Prompt (EVP) technique, which integrates explicit visual prompts into the adapter. Chen *et al*. [12] implemented adapters in the image segmentation model SAM. For the face anti-spoofing task, there are some novel adapter-based methods [5, 6, 14, 54]. Specifically, Cai *et al*. [6] proposed a novel S-Adapter to adapt pre-trained ViT models, extracting statistical information via token histograms to achieve better generalization performance in FAS.

Inspired by the above literature, we design an adapter that extracts discriminative forensic color cues for the task of recaptured document image detection.

## 3. Proposed Method

This section introduces the color artifacts with insights from the theoretical model and real samples. Then, we design the chromaticity map adapter based on prior knowledge from the color artifacts targeting the screen-recaptured document image detection task.

### 3.1. Distortion Model of the Color Artifacts

In this part, we present the proposed distortion model of the color artifact in screen-recaptured document images. Our model focuses on the signal discretization steps, involving pixelation (in display) and re-sampling (in imaging sensor) distortions, which are the main reasons for color artifacts. It reveals the origins of the color artifacts in screen-recaptured document images. To ensure the tractability of our model, we illustrate the distortion with the one-dimensional (1D) color artifacts in the horizontal direction.

Fig. 2 illustrates the displaying process $\mathcal{D}[\cdot]$. An image is rendered in pixels, each consisting of three sub-pixels in red, green, and blue. Considering an ideal spatial light modulation process [25], we can model the unit-intensity output from each sub-pixel as a rectangular function. That is

$$f_c(x, \Delta x_c) = \begin{cases} 1, & |x - \Delta x_c| \leq \frac{d}{2} \\ 0, & \text{otherwise} \end{cases} . \quad (1)$$

where $x$ is the horizontal coordinate in the physical domain (continuous), $c \in \{R, G, B\}$ represents the color channels, $d$ denotes the width of each sub-pixel, and $\Delta x_c$ represents the central coordinate of a sub-pixel in the $c$ channel within a display pixel.

Each color channel operates independently. Thus, the displayed content $I_D$ of the $n$-th pixel in a genuine image $I_G$ can be formulated as a concatenation of the modulated intensity of different color channels. That is

$$I_D(x, n) = \mathcal{D}\left[I_G(n)\right] = g\Big[\Big\|_c I_G^c(n) \cdot f_c(x, \Delta x_c)\Big] \quad (2)$$

where '$\|$' concatenates the results from different color channels, $I_G^c(n)$ represents the intensity of $c$ channel for the $n$-th

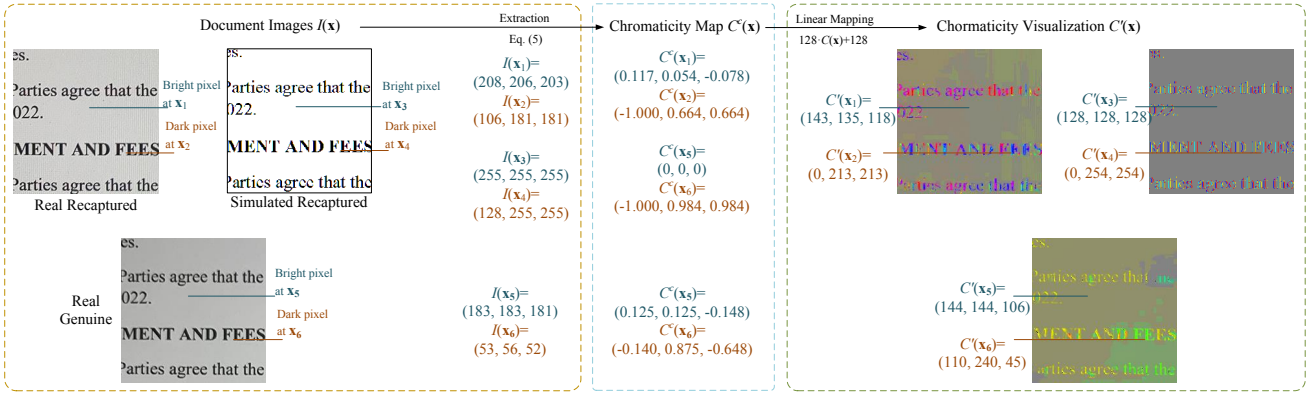


Figure 3. The process of chromaticity map extraction and visualization with a real recaptured image (top row), a simulated screen-recaptured image (top row) by theoretical distortion model in Sec. 3.1, and a real genuine image (bottom row). Obvious color artifacts can be seen in real and simulated recaptured images (top row), while no such artifact is observed in the bottom row.

pixel. Moreover, $g(\cdot) = (\cdot)^{-\gamma}$ represents the gamma correction of a display [20], and $\gamma$ is chose as 2.2 adapting to our experimental configuration.

Based on the single-pixel formulation in Eq. (2), we devise the 1D formula of the displayed content that consists of $N$ pixels by repeating Eq. (2) at the multiple widths of $D$, which is

$$I_D(x) = g\Big[\sum_{n=1}^{N} \Big|\Big|_c \, I_G^c(n) \cdot f_c\big(x - nD, \Delta x_c\big)\Big]. \quad (3)$$

In the recapturing process $\mathcal{R}[\cdot]$, the 1D pixels in the displayed document image are captured by the 1D imaging sensor. We focus on the blurring and re-sampling distortion since they are the primary source of the color artifacts. Following [8], the recaptured document image can be characterized as the result of both blurring and re-sampling distortion, which is

$$\begin{aligned} I_R(x) &= \mathcal{R}\big[I_D(x)\big] \\ &= I_D(x - \Delta_R) \otimes F_{lp}(x, \alpha) \cdot \textstyle\sum_k \delta(x - kT) \end{aligned} \quad (4)$$

where $\Delta_R \in [0, D)$ represents the positional offset between the imaging sensor and the displaying device during the recapturing process, $\delta(\cdot)$ denotes the Dirac delta function, $k$ is the 1D index of the sensor pixel in the camera, and $T$ denotes the re-sampling period, *i.e.*, the size of an imaging pixel. The blurring distortion introduced during the imaging process is also simulated by the low-pass filter $F_{lp}(\cdot)$ and its kernel diameter $\alpha = 1$ pixel in our simulation.

Our distortion model in Fig. 2 and Eq. (4) suggests that an image pixel is obtained by re-sampling a local region of the displayed content (*i.e.*, some sub-pixels with different offsets) covered by a blurring filter.

### 3.2. Color Artifacts in Recaptured Characters

Based on our model in Sec. 3.1, we elaborate on the color artifacts around characters in a recaptured image. According to Fig. 2 and Eq. (4), an imaging device recaptures the displayed pixels on a screen, where each pixel consists of horizontally arranged red, green, and blue sub-pixels[1]. Due to the misalignment between display and camera devices (specified by parameter $\Delta_R$) and the choice of sampling period (*i.e.*, parameter $T$), each imaging pixel seldom covers the exact region of one or multiple display pixels. More commonly, the alignment between display and camera pixels is not perfect. Each imaging pixel samples the intensity values from a fraction of one or more pixels in the horizontal direction. Consider a dark character on a white background, such imperfect re-sampled image pixels collect unbalanced emitting light from the red, green, and blue sub-pixels. As shown in Fig. 2 (a), the image pixels along the left side of a dark edge may consist of a higher intensity in the blue color channel. Similarly, those along the right side of a dark edge could cover the red sub-pixels. Thus, there are color artifacts along a dark edge in the recaptured document images, as illustrated by pixels $\mathbf{x}_2$ and $\mathbf{x}_4$ in Fig. 3.

*Generalization of the color artifacts*: First, the color artifacts can be observed in different types of documents. According to our analysis, the color artifacts are pronounced in edges with high contrast, such as dark characters on light backgrounds. Second, the color artifacts explained by Fig. 2 and Eq. (4) also apply to different channel variations, *e.g.*, RGB sub-pixel layouts, display, imaging resolutions, and perspective distortion. Different hardware specifications and recapturing setups lead to variations of parameters $\Delta_R$, $T$, and $\Delta x_c$ in our model, or even the need for a complicated two-dimensional distortion model. However, the camera pixels are commonly overlaid onto the display pixels without precise alignment. The re-sampling of display pixels with misalignment leads to unbalanced RGB intensities in some directions. Thus, color artifacts always appear

[1] We first consider the common stripe layout of RGB sub-pixels on LCD screens [25], while our observations apply to general layouts.

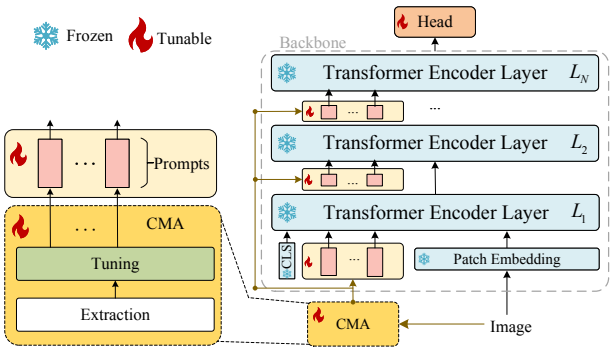


Figure 4. The architecture of the proposed CMA consists of Extraction and Tuning steps. The Extraction step extracts chromatic features according to Eq. (5), while the tuning phase maps the chromaticity patches into latent embedding.

in the recaptured document images.

### 3.3. Our Chromaticity Map Adapter

In this section, we design a chromaticity map adapter (CMA) with layers for adapting Visual Prompt Tuning (VPT) [30] pre-trained on ImageNet [16] to the screen-recaptured document image detection task. CMA keeps the backbone frozen and contains tunable parameters to learn the chromatic features of color artifacts discussed in Sec. 3.1.

An adaptor adapts the input data to a new task by incorporating new layers. These layers extract task-specific information (or data modality) from the input and allow efficient multi-modal training for various tasks [12, 13, 27, 38]. The adapter outputs are then fed to the transformer backbone as prompt tokens. The vanilla Vision Transformer (ViT) [17] backbone is utilized due to its proven effectiveness [54]. As illustrated in the left part of Fig. 4, the CMA is divided into Extraction and Tuning.

In **Extraction**, we enhance the chromatic information by computing the chromaticity map. First, we divide an image $I \in \mathbb{R}^{H\times W\times 3}$ into small patches $I^p \in \mathbb{R}^{224\times 224\times 3}$ to fit the input size of our network. Then, we devise our chromaticity map according to the analysis in Sec. 3.2 where we identify that the color artifacts arise from the imbalance of RGB color components. Inspired by the IIC space [44], we enlarge the disparity in RGB color components by computing the chromaticity map (*a.k.a.*, the normalized RGB values) $C(\mathbf{x})$ of the input image (genuine or recaptured) as

$$C^c(\mathbf{x}) = \mathcal{N}\left[\frac{I^c(\mathbf{x})}{\sum_c I^c(\mathbf{x})}\right], c \in \{R, G, B\} \tag{5}$$

where $\mathbf{x}$ is the 2D coordinate of image $I^c(\mathbf{x})$ in color channel $c$, and $\mathcal{N}[\cdot]$ denotes the $Z$-score normalization [21].

To facilitate our understanding, we visualize the chromaticity map with a linear mapping to yield $C'(\mathbf{x})$ as demonstrated in Fig. 3. A large value of a color channel in $C'(\mathbf{x})$ indicates the dominating of such color component in the document image $I(\mathbf{x})$. Chromaticity value $C^c(\mathbf{x})$ is large when the denominator is small but the nominator is large. As illustrated by the visualization of the chromaticity map in Fig. 3, such a scenario occurs along the edges of a dark character shown on a bright background in a recaptured document image, *i.e.*, pixels $\mathbf{x}_2$ and $\mathbf{x}_4$. However, the chromaticity values are small in a genuine document image (see pixels $\mathbf{x}_5$ and $\mathbf{x}_6$). This is because the sampling rate of a camera (∼300 dpi on an A4-size paper with a high-resolution camera) is much smaller than that of the resolution of a printing device (as high as 1200 dpi).

In **Tuning**, adaptation is conducted across all layers efficiently and effectively by considering the features derived from the chromaticity map. The chromaticity map extracted by Eq. (5) enhances the color artifacts mentioned in Sec. 3.2 and provides distinctive forensic clues for detecting recaptured documents. Traditional methods concatenate the data from different modalities (*i.e.*, RGB data and chromaticity map) with a fully connected (FC) classifier. However, such FC-based multi-modal methods tend to over-fit the training samples [22], and neglect the multi-modal features from cross-modal tokens [54]. Motivated by the enhanced high-frequency components in [38], we learn explicit prompts from chromatic features with trainable linear layers. Specifically, the linear layer $\mathtt{L}_{cm}$ projects the chromaticity patch $C^p$ into a $m$-dimension feature $F_{cm} \in \mathbb{R}^m$, that is

$$F_{cm} = \mathtt{L}_{cm}(C^p) \tag{6}$$

where $m$ is the product of the ViT-B16 embedding size and the number of prompts (*i.e.*, 10 in our implementation).

## 4. Dataset

*1) ROD dataset* : We have built a large-scale and practical document image database named the ROD dataset, which comprises 10,887 genuine and screen-recaptured document images containing business information in offices. Considering the disparities in imaging quality and document content, we partitioned it into three subsets, *i.e.*, ROD_HQ, ROD_LQ, and ROD_M&F. The database collection process involves collecting genuine documents and recapturing them from document images. We summarize the ROD dataset's three subsets in Tab. 1.

*Collecting genuine documents:* The genuine documents should be carefully chosen since they serve as the source images in our dataset. Firstly, the content of the chosen documents should be of practical significance. As illustrated in Fig. A2, the documents include Chinese and English contracts, sales orders, electronic invoices, etc. We

| Subsets \ Types | Genuine (# imgs, Config.) | Recaptured (# imgs, Config.) |
|---|---|---|
| HQ (4032 × 3024) | 548 images, by 2 pairs of P-Cs | 822 images, by 6 pairs of D-Cs |
| LQ (1280 × 960) | 1096 images, by 4 pairs of P-Cs | 1096 images, by 12 pairs of D-Cs |
| M&F (4032 × 3024 & 1280 × 960) | 3216 images, from MP-DocVQA [46] and "Find it again! [47]" | 4109 images, by 18 pairs of D-Cs |

Table 1. Summary of different subsets in our ROD dataset. P-C and D-C are the abbreviations for print-capture and display-capture devices, respectively.

obtained 137 genuine documents for subsets ROD_HQ and ROD_LQ. Secondly, we adopt 3,216 samples with rich textual content from the public datasets MP-DocVQA [46] and "Find it again! [47]" as the genuine document images in ROD_M&F. The images in ROD_M&F encounter common distortions for document images in practice, such as binarization, *etc*.

*Recapturing from document images:* We accounted for the diverse range of real-world display devices and imaging configurations in our data collection. We took into account the variations in imaging configurations found in practical settings. Our dataset encompasses 24 different print/display-imaging device combinations, including 1 printing device, 3 screen devices, and 6 imaging devices, resulting in 4860 genuine images and 6027 screen-recaptured images of varying quality. These devices cover a diverse range of hardware parameters. During the printing or displaying of document images, we resize the document content to cover the A4 size paper or the whole screen, respectively. Regarding imaging devices, there are 2 smartphones with high imaging resolution (at 4032×3024 pixels) and 4 smartphones with lower resolution (at 1280×960 pixels, achieved by a third-party App, Open Camera [1]). The process of collecting recaptured document images follows the rules outlined in [9], except that we adjusted the capturing distance to accommodate different document sizes. More details of the ROD dataset are presented in Appendix A.

*2) DLC2021* [18]: The Document Liveness Detection Dataset (DLC2021) dataset comprises 1,424 video clips of various identity documents, captured at resolutions of 1080×1920 and 2160×3840 pixels, and frame rates of 30 fps and 60 fps, illustrating original and screen-recaptured documents. The latter exhibit moiré patterns, sourced from displays of office desktops and laptops. To enhance image quality and mitigate video compression artifacts, the FFmpeg [**?** ] library was used to extract six intra-coded frames (I-frames) from each clip. The refined dataset contains 1,740 genuine and 2,400 screen-recaptured samples, serving as a testing set in our experiment.

# 5. Experiment

## 5.1. Experimental Protocols

Our preliminary results show that all models achieve AUCs of 1.0 in the intra-dataset experiment. Therefore, we focus on two practical and challenging cross-domain experimental protocols.

*1) Cross-Dataset Experiment:* The training and testing sets contain the same document types but are collected by acquisition devices of high and low imaging quality, respectively. Specifically, experimental protocol ROD_HQ→ROD_LQ is carried out. The model trained by ROD_HQ learns the chromatic features from high-quality devices and is evaluated in low-quality samples from ROD_LQ.

*2) In-the-Wild Experiment:* In this experiment, two protocols (ROD_HQ→ROD_M&F and ROD_HQ→DLC2021) are executed to test the model under various conditions. These protocols involve training and testing sets that include different types of documents, and the testing set featuring samples with varying distortions (e.g., binarization, blurring, compression).

## 5.2. Experimental Results

### 5.2.1 Cross-Dataset Results

In this protocol, genuine and screen-recaptured document images have identical content but differ in imaging quality. We utilize ROD_HQ as the training set and ROD_LQ for testing. Within the training set, 20% of samples are reserved for validation.

In the experiment, we consider SOTA approaches for detecting screen-recaptured natural images [51] and ID document images [4]. These methods [4, 51] propose data augmentation methods by overlaying the moiré patterns extracted from screen-recaptured images onto some genuine samples to produce synthesized recaptured samples. Benalcazar *et al*. utilizes a generic CNN-based backbone, MobileNetV2 in [4]. To allow an extensive comparison, we incorporate the moiré patterns synthesis strategy [4, 51] into more CNN-based models (ResNet50 [26], ResNeXt101 [50]) and Transformer-based models (ViT-B16 [17] and VPT [30]). To further contextualize our proposal with recent advancements in the field, we extend our comparisons to include LTC-PE [58] and CNN + ViT [19]. The settings for these models are consistent with those presented in [4]. All methods process patches and implement majority voting to aggregate the patch-level decisions into image-level ones. Performance is evaluated using the Area Under the ROC Curve (AUC) and Equal Error Rate (EER) metrics.

For the single-modality approaches, the networks receive input patches of size 224×224 pixels. They are trained using an Adam optimizer with a batch size of 32. The training

| Methods / Distortions | RGB only | | | | RGB+Chromaticity | | | |
|---|---|---|---|---|---|---|---|---|
| | ViT | | VPT | | ViT | | CMA (ours) | |
| | AUC | EER | AUC | EER | AUC | EER | AUC | EER |
| Origin | 0.7669 | 35.64% | 0.7883 | 30.14% | 0.7714 | 34.63% | **0.8991** | **20.79%** |
| Resize (0.9) | 0.7351 | 37.99% | 0.7438 | 35.61% | 0.7391 | 37.39% | **0.8410** | **27.31%** |
| Resize (0.8) | 0.7163 | 39.26% | 0.7232 | 37.30% | 0.7289 | 39.07% | **0.7900** | **33.24%** |
| Resize (0.7) | 0.6954 | 41.01% | 0.6908 | 41.54% | 0.7080 | 40.86% | **0.7554** | **36.52%** |
| Compress (90) | 0.7531 | 35.71% | 0.7680 | 32.83% | 0.7605 | 35.06% | **0.8888** | **22.07%** |
| Compress (80) | 0.7489 | 37.07% | 0.7592 | 35.68% | 0.7784 | 33.90% | **0.8731** | **23.51%** |
| Compress (70) | 0.7402 | 38.13% | 0.7536 | 36.79% | 0.7603 | 35.96% | **0.8688** | **24.62%** |
| Blur (1) | 0.6582 | 40.69% | 0.6356 | 39.35% | 0.6990 | 43.09% | **0.7450** | **29.85%** |
| Blur (2) | 0.5973 | 43.17% | 0.6017 | 41.84% | 0.6487 | 46.32% | **0.6861** | **34.28%** |
| Blur (3) | 0.5509 | 45.40% | 0.5680 | 43.92% | 0.6230 | 47.78% | **0.6282** | **39.69%** |

Table 2. Comparisons of different approaches under ROD_HQ→ROD_LQ protocol with different post-processing, *i.e.*, resizing, JPEG compression, and Gaussian blur. The best performance under each case is **boldfaced**. "Origin" denotes the experimental protocol without distortions. The numbers in parentheses denote the resizing ratio, JPEG quality factor, and variances of Gaussian blur kernels, respectively.

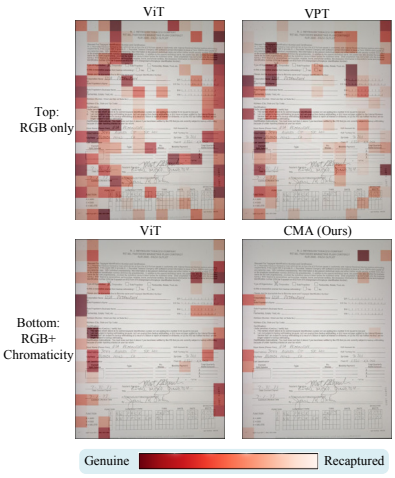


Figure 5. The patch-level responses of different approaches for a sample in ROD_M&F. Under the ideal case, responses should contain no red patch.

spans 20 epochs with a learning rate of $1 \times 10^{-4}$. For the multi-modal CNNs, we process the RGB and chromaticity data, respectively, with the same CNN backbones and employ the cross-modal focal (CMF) loss [22] in merging information from different modalities. For the multi-modal ViT-based backbone, we divide the chromaticity map into patches according to [54]. These patches are then flattened and fed to the backbone with a trainable linear projection.

**ROD_HQ → ROD_LQ:** As shown in Tab. C1, methods relying on a single modality do not achieve satisfactory performance. Due to missing moiré patterns and blurriness of font edges in low-quality samples, many important forensic features are lost. Therefore, the performance of RGB-modality methods relying on texture, such as LTC-PE and CNN + ViT is compromised. Methods using only chromaticity data also underperformed on ROD_LQ, indicating reliance solely on RGB or chromaticity data for image-level decisions is unreliable. In contrast, by integrating RGB and chromaticity data, our multi-modal CMA model achieved superior performance with an AUC of 0.8991 and an EER of 20.79%. More details are shown in Appendix C.

**Ablation Study:** An ablation study of our method can be performed by comparing the performance of multi-modality ViT and our CMA (both with and without the chromaticity map extraction step). The difference between these methods lies in the way that chromaticity maps are input to the backbone. The chromaticity maps are fed to the ViT backbone as with linear embedding layers, while they are extracted by an adapter by our CMA method. The CMA (w/o Ext.) method removes the chromaticity extraction layers in our CMA, and it degrades to a single modality approach with RGB data only. As shown in Tab. C1, both the multi-modality ViT and CMA (w/o Ext.) suffer from significant performance loss compared to the proposed CMA method. On the one hand, the limited multi-modal fusion capability in the ViT model [54] leads to poor performance of 34.63% EER (an increment of 13.84 p.p.) with the direct input of multi-modalities data. On the other hand, the missing chromaticity map in CMA (w/o Ext.) leads to a significant degradation of generalization performance (a reduction of 0.0952 in AUC and an increment of 12.61 p.p. in EER) compared to our CMA. Thus, our CMA is more robust under low-quality samples.

**Robustness Experiment:** Transformer-based models pretrained on ROD_HQ were assessed under three image distortions (image resizing, JPEG compression, and Gaussian blur) which reflect common distortions under Online Social Network (OSN) transmissions. As shown in Tab. 2, our findings indicate that the proposed CMA model demonstrated superior robustness across all tested distortions, particularly under JPEG compression, where it shows small performance degradation even under a JPEG quality factor of 70. More details are shown in Appendix D.

### 5.2.2 In-the-Wild Results

In this protocol, all methods and training strategies remain consistent with those in Sec. 5.2.1.

**ROD_HQ → ROD_M&F:** Genuine and recaptured document images differ in content, and recaptured samples cover a range of imaging qualities. ROD_HQ is used as the training set and ROD_M&F for testing. It is noted that there are equal amounts of low and high-quality samples in the ROD_M&F dataset.

The models with only RGB data receive poor performance in the ROD_M&F testing set, while methods using only chromaticity data maintain stable performance

| Method | ROD_HQ → ROD_M&F | | ROD_HQ → DLC2021 | |
|---|---|---|---|---|
| | AUC | EER | AUC | EER |
| Single modality: RGB only | | | | |
| LTC-PE [58] | 0.6116 | 46.14% | 0.6272 | 42.87% |
| CNN + ViT [19] | 0.7406 | 33.60% | 0.7836 | 29.08% |
| ResNet50 | 0.6634 | 38.18% | 0.6770 | 39.41% |
| ResNeXt101 | 0.7585 | 36.32% | 0.6880 | 39.73% |
| MobileNetV2 [4] | 0.8094 | 26.27% | 0.7235 | 33.96% |
| ViT | 0.7546 | 33.27% | 0.7356 | 32.23% |
| VPT [30] | 0.7967 | 28.45% | 0.7547 | 30.89% |
| CMA (w/o Ext.) | 0.8282 | 27.39% | 0.7765 | 27.04% |
| Single modality: Chromaticity Map only | | | | |
| ResNet50 | 0.7281 | 40.91% | 0.5596 | 44.64% |
| ResNeXt101 | 0.8035 | 32.19% | 0.5796 | 46.20% |
| MobileNetV2 | 0.7274 | 32.88% | 0.5768 | 44.04% |
| ViT | 0.7304 | 29.26% | 0.6221 | 41.50% |
| VPT | 0.7499 | 32.20% | 0.6805 | 36.24% |
| Multi-modalities: RGB + Chromaticity | | | | |
| ResNet50 | 0.6892 | 40.65% | 0.7051 | 35.49% |
| ResNeXt101 | 0.7767 | 29.85% | 0.7114 | 35.93% |
| MobileNetV2 | 0.8051 | 27.40% | 0.7400 | 33.48% |
| ViT | 0.8094 | 28.22% | 0.7651 | 33.22% |
| CMA (ours) | **0.9475** | **12.77%** | **0.8489** | **25.82%** |

**All networks**: Trained by the moiré augmentation strategy in [4].
**Single modality methods**:
LTC-PE, CNN+ViT: recent approaches from [58], [19], respectively;
ViT: the Vanilla Vision Transformer [17]; VPT: Visual Prompt Tuning with ViT [30];
CMA (w/o Ext.): Our CMA method without chromaticity map extraction step.
**Multi-modalities methods**:
CNNs: fusion by the CMF loss [22];
ViT: ViT backbone with RGB and chromaticity input tokens;
CMA: ViT with RGB tokens and chromaticity prompts processed by our adapter.

Table 3. Comparisons of different approaches under in-the-wild experiment. The best performance under each protocol is **boldfaced**. By efficiently incorporating a chromaticity map, our CMA achieves the best performance under both protocols.

compared to the performance under the ROD_LQ testing set. This is because the chromaticity map is insensitive to changes in document content, as demonstrated by examples shown in Appendix B. For multi-modality cases, the proposed model exhibits SOTA performance under this protocol with an AUC of 0.9475 and an EER of 12.77%.

Notably, there are significant performance improvements for the proposed CMA (0.0484 in AUC and 8.02 p.p. in EER) under the ROD_M&F compared to that under the ROD_LQ. This is because the document images in ROD_M&F are with black text and white background, as illustrated in Fig. A2 in the supplementary material. Some of the document images adopted from MP-DocVQA [46] have been post-processed by a binarization algorithm for quality enhancement. Such content with high contrast facilitates the observation of the color artifacts as demonstrated in Fig. 3. Thus, our CMA achieves better performance in ROD_M&F than that in ROD_LQ though the former is more challenging for the other methods.

The above findings are further supported by network responses in Fig. 5. By comparing the patch-level responses from different methods, we confirm that both the robust chromaticity map and the efficient chromaticity adapter contribute to the improvement of generalization performance in our CMA approach. The t-SNE visualization demonstrates similar results as shown in Appendix E. The latent spaces of our CMA method show a clear separation of the genuine and screen-recaptured samples.

**ROD_HQ → ROD_DLC2021:** In this protocol, ROD_HQ is utilized as the training set, while the testing samples are derived from the public dataset DLC2021 [18]. It is noted that this protocol is very challenging due to serious degradation (e.g., low resolution, slow auto-focus, and compression) in the video recording mode employed in gathering the DLC2021 dataset.

As illustrated in Tab. 3, methods relying on single-modal data experience a performance decline, because samples in DLC2021 contain numerous distortions. In particular, blurring distortion makes chromaticity maps less effective than before. Consistent with the conclusions in Appendix D, our method shows a performance drop when facing blur attacks. Despite this, CMA outperforms the other methods with an AUC of 0.8489 and an EER of 25.82%.

## 6. Conclusion

In this work, we have demonstrated the efficacy of a forensic feature from the sub-pixel sampling color artifacts. The color artifacts are aggregated in our chromaticity map, which is fed to a ViT backbone with the proposed CMA approach. The generalization and robustness of our CMA approach have been successfully demonstrated under recaptured document samples with different acquisition devices, contents, and image qualities. Robustness evaluation confirms that the proposed method shows superior performances compared to the RGB-only approaches and a multi-modal ViT backbone without using our CMA.

The chromaticity map studied in this work is promising in different forensics applications. For example, the color artifacts presented in a recaptured document image should be in-homogeneous under the splicing attack. This is because the re-sampling factors (between different pairs of display and imaging devices) vary across different document images. Moreover, the chromaticity map, highlighting the disparity in different color channels, may also be employed as a new data modality in identifying generated document images by deep networks or locating regions tampered by photo editing tools [33]. This is due to the fact that the synthetic document data is not generated by the re-sampling operation elaborated in Sec. 3.1.

**Acknowledgement** This work was supported in part by NSFC under Grant 62072313, 62371301, and U23B2022; and in part by the CCF-Alibaba Innovative Research Fund for Young Scholars.

## References


[1] Open camera. https://opencamera.org.uk/. [Accessed: 2023-10-17]. 6

[2] Shruti Agarwal, Wei Fan, and Hany Farid. A diverse large-scale dataset for evaluating rebroadcast attacks. In *ICASSP*, pages 1997–2001, 2018. 1

[3] Yousef Atoum, Yaojie Liu, Amin Jourabloo, and Xiaoming Liu. Face anti-spoofing using patch and depth-based cnns. In *IJCB*, pages 319–328, 2017. 2

[4] Daniel Benalcazar, Juan E Tapia, Sebastian Gonzalez, and Christoph Busch. Synthetic ID card image generation for improving presentation attack detection. *TIFS*, 18:1814–1824, 2023. 2, 6, 8, 3

[5] Rizhao Cai, Yawen Cui, Zhi Li, Zitong Yu, Haoliang Li, Yongjian Hu, and Alex Kot. Rehearsal-free domain continual face anti-spoofing: Generalize more and forget less. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 8037–8048, 2023. 3

[6] Rizhao Cai, Zitong Yu, Chenqi Kong, Haoliang Li, Changsheng Chen, Yongjian Hu, and Alex Kot. S-adapter: Generalizing vision transformer for face anti-spoofing with statistical tokens. *arXiv preprint arXiv:2309.04038*, 2023. 3

[7] Hong Cao and Alex C. Kot. Identification of recaptured photographs on LCD screens. In *ICASSP*, pages 1790–1793, 2010. 1, 2

[8] Changsheng Chen, Mulin Li, Anselmo Ferreira, Jiwu Huang, and Rizhao Cai. A copy-proof scheme based on the spectral and spatial barcoding channel Models. *TIFS*, 15:1056–1071, 2019. 4

[9] Changsheng Chen, Shuzheng Zhang, Fengbo Lan, and Jiwu Huang. Domain-agnostic document authentication against practical recapturing attacks. *TIFS*, 17:2890–2905, 2022. 2, 6, 1

[10] Changsheng Chen, Bokang Li, Rizhao Cai, Jishen Zeng, and Jiwu Huang. Distortion model-based spectral augmentation for generalized recaptured document detection. *TIFS*, 19: 1283–1298, 2023. 2

[11] Haoyu Chen, Zhihua Wang, Yang Yang, Qilin Sun, and Kede Ma. Learning a deep color difference metric for photographic images. In *CVPR*, pages 22242–22251, 2023. 2

[12] Tianrun Chen, Lanyun Zhu, Chaotao Deng, Runlong Cao, Yan Wang, Shangzhan Zhang, Zejian Li, Lingyun Sun, Ying Zang, and Papa Mao. Sam-adapter: Adapting segment anything in underperformed scenes. In *CVPR*, pages 3367–3375, 2023. 3, 5

[13] Zhe Chen, Yuchen Duan, Wenhai Wang, Junjun He, Tong Lu, Jifeng Dai, and Yu Qiao. Vision transformer adapter for dense predictions. *arXiv preprint arXiv:2205.08534*, 2022. 3, 5

[14] Yawen Cui, Zitong Yu, Rizhao Cai, Xun Wang, Alex C Kot, and Li Liu. Generalized few-shot continual learning with contrastive mixture of adapters. *arXiv preprint arXiv:2302.05936*, 2023. 3

[15] Tiago José De Carvalho, Christian Riess, Elli Angelopoulou, Helio Pedrini, and Anderson de Rezende Rocha. Exposing digital image forgeries by illumination color classification. *TIFS*, 8(7):1182–1194, 2013. 2

[16] Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. Imagenet: A large-scale hierarchical image database. In *CVPR*, pages 248–255, 2009. 5

[17] Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, Jakob Uszkoreit, and Neil Houlsby. An image is worth 16x16 words: Transformers for image recognition at scale. *ICLR*, 2021. 3, 5, 6, 8

[18] D. Polevoy et al. Document liveness challenge dataset (dlc-2021). *J. Imaging*, 2022. 6, 8

[19] G. Li et al. Recaptured screen image identification based on vision transformer. *J. VCIP*, 2023. 2, 6, 8

[20] Hany Farid. Blind inverse gamma correction. *TIP*, 10(10): 1428–1433, 2001. 4

[21] Nanyi Fei, Yizhao Gao, Zhiwu Lu, and Tao Xiang. Z-score normalization, hubness, and few-shot learning. In *ICCV*, pages 142–151, 2021. 5

[22] Anjith George and Sebastien Marcel. Cross modal focal loss for rgbd face anti-spoofing. In *CVPR*, pages 7882–7891, 2021. 5, 7, 8

[23] Julian Jorge Andrade Guerreiro, Mitsuru Nakazawa, and Björn Stenger. Pct-net: Full resolution image harmonization using pixel-wise color transformations. In *CVPR*, pages 5917–5926, 2023. 2

[24] Benjamin Christopher Hadwiger and Christian Riess. Deep metric color embeddings for splicing localization in severely degraded images. *TIFS*, 17:2614–2627, 2022. 2

[25] Rolf R Hainich and Oliver Bimber. *Displays: fundamentals & applications*. CRC press, 2016. 3, 4

[26] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *CVPR*, pages 770–778, 2016. 6

[27] Neil Houlsby, Andrei Giurgiu, Stanislaw Jastrzebski, Bruna Morrone, Quentin De Laroussilhe, Andrea Gesmundo, Mona Attariyan, and Sylvain Gelly. Parameter-efficient transfer learning for nlp. In *ICML*, pages 2790–2799, 2019. 3, 5

[28] Zhaoxu Hu, Changsheng Chen, Wai Ho Mow, and Jiwu Huang. Document recapture detection based on a unified distortion model of halftone cells. *TIFS*, 17:2800–2815, 2022. 2

[29] ISO/IEC 27001:2022. Information security management systems. Standard, International Organization for Standardization, 2022. 1

[30] Menglin Jia, Luming Tang, Bor-Chun Chen, Claire Cardie, Serge Belongie, Bharath Hariharan, and Ser-Nam Lim. Visual prompt tuning. In *ECCV*, pages 709–727, 2022. 5, 6, 8, 2, 3

[31] Korean JoongAng Daily. Trade secrets may have been stolen from samsung electronics. https://koreajoongangdaily.joins.com/2022/03/23/business/tech/Samsung-Electronics-foundry/20220323175613170.html, 2022. [Accessed: 2023-10-26]. 1

[32] Hoang M Le, Brian Price, Scott Cohen, and Michael S Brown. Gamutmlp: A lightweight mlp for color loss recovery. In *CVPR*, pages 18268–18277, 2023. 2

[33] Haodong Li, Bin Li, Shunquan Tan, and Jiwu Huang. Identification of deep network generated images using disparities in color components. *SIGPRO*, 174:107616, 2020. 2, 8

[34] Jiaxing Li, Chenqi Kong, Shiqi Wang, and Haoliang Li. Two-branch multi-scale deep neural network for generalized document recapture attack detection. In *ICASSP*, pages 1–5, 2023. 2

[35] Yanghao Li, Hanzi Mao, Ross Girshick, and Kaiming He. Exploring plain vision transformer backbones for object detection. In *ECCV*, pages 280–296, 2022. 3

[36] Ajian Liu, Zichang Tan, Zitong Yu, Chenxu Zhao, Jun Wan, Yanyan Liang Zhen Lei, Du Zhang, Stan Z Li, and Guodong Guo. Fm-vit: Flexible modal vision transformers for face anti-spoofing. *TIFS*, 2023. 3

[37] Siqi Liu, Pong C Yuen, Shengping Zhang, and Guoying Zhao. 3d mask face anti-spoofing with remote photoplethysmography. In *ECCV*, pages 85–100, 2016. 2

[38] Weihuang Liu, Xi Shen, Chi-Man Pun, and Xiaodong Cun. Explicit visual prompting for low-level structure segmentations. In *CVPR*, pages 19434–19445, 2023. 3, 5

[39] Yi-Chen Lo, Chia-Che Chang, Hsuan-Chao Chiu, Yu-Hao Huang, Chia-Ping Chen, Yu-Lin Chang, and Kevin Jou. Clcc: Contrastive learning for color constancy. In *CVPR*, pages 8053–8063, 2021. 2

[40] Timo Ojala, Matti Pietikainen, and Topi Maenpaa. Multiresolution gray-scale and rotation invariant texture classification with local binary patterns. *TPAMI*, 24(7):971–987, 2002. 1

[41] Jueqin Qiu, Haisong Xu, and Zhengnan Ye. Color constancy by reweighting image feature maps. *TIP*, 29:5711–5721, 2020. 2

[42] Chenfan Qu, Chongyu Liu, Yuliang Liu, Xinhong Chen, Dezhi Peng, Fengjun Guo, and Lianwen Jin. Towards robust tampered text detection in document image: new dataset and new solution. In *CVPR*, pages 5937–5946, 2023. 2

[43] Christian Riess and Elli Angelopoulou. Scene illumination as an indicator of image manipulation. In *IH*, pages 66–80, 2010. 2

[44] Robby T Tan, Ko Nishino, and Katsushi Ikeuchi. Color constancy through inverse-intensity chromaticity space. *JOSA A*, 21(3):321–334, 2004. 5

[45] Thirapiroon Thongkamwitoon, Hani Muammar, and Pier-Luigi Dragotti. An image recapture detection algorithm based on learning dictionaries of edge rofiles. *TIFS*, 10(5): 953–968, 2015. 1, 2

[46] Rubèn Tito, Dimosthenis Karatzas, and Ernest Valveny. Hierarchical multimodal transformers for multipage docvqa. *PR*, 144:109834, 2023. 6, 8, 1

[47] Beatriz Martínez Tornés, Théo Taburet, Emanuela Boros, Kais Rouis, Antoine Doucet, Petra Gomez-Krämer, Nicolas Sidere, and Vincent Poulain d'Andecy. Receipt dataset for document forgery detection. In *ICDAR*, pages 454–469, 2023. 6, 1

[48] Laurens Van der Maaten and Geoffrey Hinton. Visualizing data using t-sne. *JMLR*, 9(11), 2008. 3

[49] Stefan Van der Walt, Johannes L Schönberger, Juan Nunez-Iglesias, François Boulogne, Joshua D Warner, Neil Yager, Emmanuelle Gouillart, and Tony Yu. Scikit-image: image processing in python. *PeerJ*, 2:e453, 2014. 2, 3

[50] Saining Xie, Ross Girshick, Piotr Dollár, Zhuowen Tu, and Kaiming He. Aggregated residual transformations for deep neural networks. In *CVPR*, pages 1492–1500, 2017. 6

[51] Cong Yang, Zhenyu Yang, Yan Ke, Tao Chen, Marcin Grzegorzek, and John See. Doing more with moiré pattern detection in digital photos. *TIP*, 32:694–708, 2023. 6

[52] Hang Yu, Tian-Tsong Ng, and Qibin Sun. Recaptured photo detection using specularity distribution. In *ICIP*, pages 3140–3143, 2008. 1, 2

[53] Huanglin Yu, Ke Chen, Kaiqi Wang, Yanlin Qian, Zhaoxiang Zhang, and Kui Jia. Cascading convolutional color constancy. In *AAAI*, pages 12725–12732, 2020. 2

[54] Zitong Yu, Rizhao Cai, Yawen Cui, Xin Liu, Yongjian Hu, and Alex Kot. Rethinking vision transformer and masked autoencoder in multimodal face anti-spoofing. *arXiv preprint arXiv:2302.05744*, 2023. 3, 5, 7

[55] Ling Zhang, Yinghao He, Qing Zhang, Zheng Liu, Xiaolong Zhang, and Chunxia Xiao. Document image shadow removal guided by color-aware background. In *CVPR*, pages 1818–1827, 2023. 2

[56] Zhifeng Zhang, Xuejing Kang, and Anlong Ming. Domain adversarial learning for color constancy. In *IJCAI*, pages 1693–1699, 2022. 2

[57] Nan Zhu and Zhiqin Liu. Recaptured image forensics based on local ternary count of high order prediction error. *SPIC*, 104:116662, 2022. 2

[58] N. Zhu and Z. Liu. Recaptured image forensics based on local ternary count of high order prediction error. *SPIC*, 2022. 2, 6, 8