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

Supplementary Material

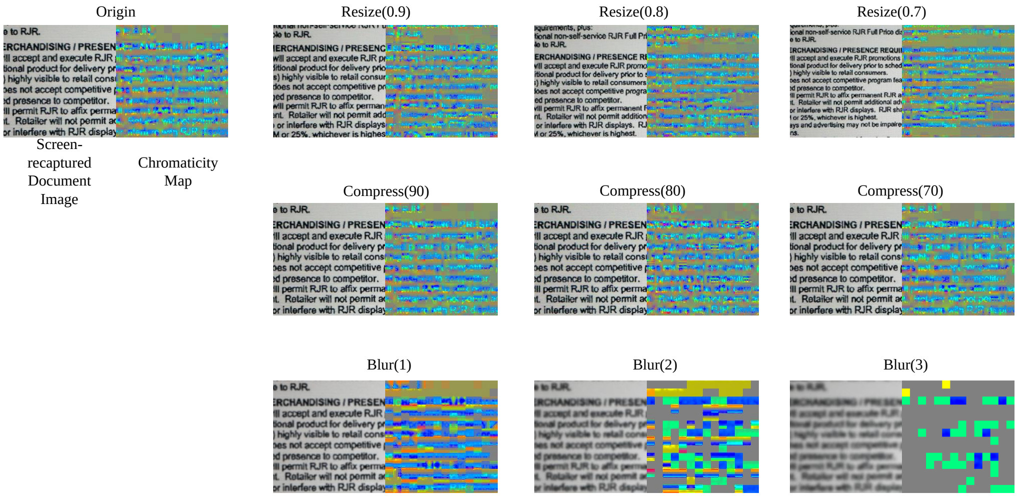


Figure B1. Examples of screen-recaptured images are shown in Appendix D. The original from ROD_M&F is collected by OnePlus 5T camera (resolution at 1280×960 pixels), Dell P2418D display (resolution at 2560×1440 pixels, size 23.8 inches). The image was rotated 90 degrees horizontally to fill the entire screen during re-sampling. Upon comparison, among three image distortions (image resizing, JPEG compression, and Gaussian blur), the chromaticity map exhibits more significant traces for recapturing.

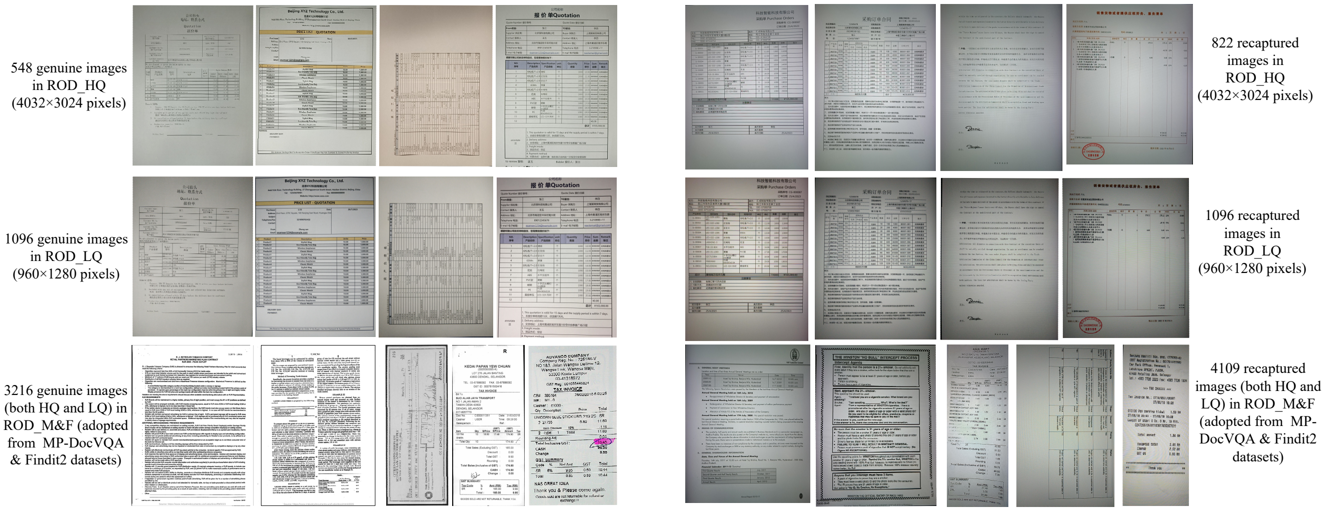
A. Our ROD Dataset

Tab. A1 reports all the devices used in our dataset. For the subsets ROD_HQ and ROD_LQ, we employ a printer/screen to print/display the genuine document, followed by imaging it with smartphones of higher and lower imaging quality, respectively. For ROD_M&F, there is no need for printers, as we use images randomly collected from the public datasets MP-DocVQA [46] and "Find it again! [47]" as genuine document images. We follow the rules in [9] during the collection of the ROD database to gather genuine and recaptured images. Specifically, our configurations are summarized as follows.

- Photographing smartphone: The captured images are saved in JPEG format with the highest quality coefficient. The camera is placed at a distance of 15-25 cm from the document/screen (to accommodate different document sizes) so that the area of interest covers most of the captured image in height.
- Environmental Light: Illuminated by an LED lamp and fluorescent lights in the office.
- Printer: Set to color mode, with the optimal print resolution of 4800×1200 dpi.
- Cropping: Utilizing an automatic cropping script, the document area can be accurately extracted from the images. With the support of the OpenCV library, the algorithm automatically detects the four corners based on the color difference in the background, and crops out the ROI. A final manual adjustment is made thereafter.

Printer/Screen Devices		Imaging Devices	
Printer	Canon C3530 (1200 DPI)	HQ Phones	Honor 50SE (4032×3024 pixels)
			OPPO K9x (4032×3024 pixels)
Screen	DELL-P2418D (123 ppi)	LQ Phones	Xiaomi MEE7 (1280×960 pixels)
	iPad-Mini5 (326 ppi)		OnePlus 5T (1280×960 pixels)
	ASUS-FX80ge (141 ppi)		iPhone 6 (1280×960 pixels)
			iPhone 8 (1280×960 pixels)

Table A1. The devices used for collecting ROD dataset categorized by print/screen and smartphone quality.


548 genuine images in ROD_HQ (4032×3024 pixels)
1096 genuine images in ROD_LQ (960×1280 pixels)
3216 genuine images (both HQ and LQ) in ROD_M&F (adopted from MP-DocVQA & Findit2 datasets)
822 recaptured images in ROD_HQ (4032×3024 pixels)
1096 recaptured images in ROD_LQ (960×1280 pixels)
4109 recaptured images (both HQ and LQ) in ROD_M&F (adopted from MP-DocVQA & Findit2 datasets)

(a) Genuine document images.

(b) Screen-recaptured document images.

Figure A2. Examples of document images in our dataset. Each row illustrates instances from ROD_HQ , ROD_LQ , and ROD_M&F. There are, in total, 10887 document images (4860 genuine and 6027 recaptured).

B. Chromaticity Map Visualization

As illustrated in Fig. B1, even when low-quality samples undergo various degrees of distortion, the chromaticity map still maintains good robustness, in contrast to spatial data. It highlights the presence of color artifacts. For operations like Resize and JPEG compression, there is a loss of high-frequency components during the DCT quantization steps, yet the chromaticity map still allows for a stable observation of the color artifact phenomena and regions. Regarding the Blur operation, the chromaticity map experiences a significant feature loss. This is challenging as the blurring distortion occurs because the Gaussian kernel covers multiple pixels in a local region. After applying the blurring kernel, the resulting pixel value at the center of the blurring kernel is a weighted sum of the 17×17 neighboring pixels [49]. The visualization effects of the chromaticity map against different distortions also align with the variation of experimental results in Tab. 2.

C. Cross-Dataset Evaluation

As shown in Tab. C1, methods relying on a single modality do not achieve satisfactory performance. Due to the absence of noticeable moiré patterns in low-quality screen recaptured samples and the blurriness of font edges due to the low quality of the images, the performance of LTC-PE [58] and CNN + ViT [19] is not satisfactory. Among different RGB-only approaches, MobileNetV2 [4] obtains the best AUC of 0.8210 and EER of 27.37%. Such inferior performances of the RGB-only approaches are due to the loss of forensic features, such as moiré pattern, in low-quality samples. Similarly, methods using only chromaticity data also showed poor results on ROD_LQ. MobileNetV2 achieves the highest AUC of 0.7232. Therefore, predicting image-level decisions based solely on RGB or chromaticity data is not reliable.

	ROD_HQ → ROD_LQ	
Method	AUC	EER
Single modality: RGB only		
LTC-PE [58]	0.6800	42.31%
CNN + ViT [19]	0.6102	49.33%
ResNet50	0.7691	35.52%
ResNeXt101	0.7738	34.17%
MobileNetV2 [4]	0.8210	27.37%
ViT	0.7669	35.64%
VPT [30]	0.7883	30.14%
CMA (w/o Ext.)	0.8039	33.40%
Single modality: Chromaticity Map only		
ResNet50	0.7104	41.40%
ResNeXt101	0.7019	38.98%
MobileNetV2	0.7232	37.19%
ViT	0.6883	34.79%
VPT	0.7200	37.07%
Multi-modalities: RGB + Chromaticity		
ResNet50	0.7389	29.36%
ResNeXt101	0.7562	27.85%
MobileNetV2	0.8122	24.55%
ViT	0.7714	34.63%
CMA (ours)	**0.8991**	**20.79%**

Table C1. Comparisons of different approaches under cross-dataset experiment. The best performance under each protocol is **boldfaced**.

For methods incorporating both RGB and chromaticity, the proposed model demonstrated superior performance with an AUC of 0.8991 and an EER of 20.79%. By ex-

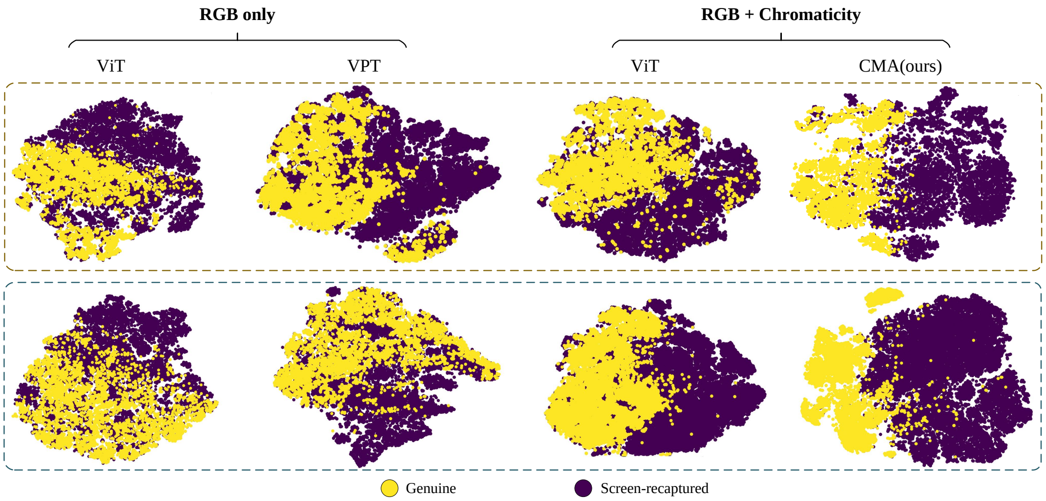


Figure E1. The t-SNE visualizations of the latent space for the proposed scheme trained by ROD_HQ. Top row: visualization for ROD_LQ as a testing set. Bottom row: visualization for ROD_M&F as a testing set.

tracting distinctive chromaticity prompt tokens with our CMA, our CMA achieves good generalization performance. However, the generalization performance of other models is limited. For example, the multi-modal MobileNetV2 only presents a slight improvement (2.82 p.p. in EER) over the RGB-only counterpart. The unstable performance of multi-modal CNN-based models is due to the CMF loss employed in the fusion step. It determines the fusion weights for each channel based solely on their prediction scores without establishing the correspondence between the features of the two modalities [36].

D. Robustness Evaluation

In a practical forensics scenario, the questioned samples often contain various distortions that are introduced during the transmission over OSNs. Given that transformer backbones are popular choices for multi-modal learning tasks [54], the performances of the transformer-based models are studied in this experiment. Transformers pre-trained by ROD_HQ (without additional distortions) are evaluated in ROD_LQ under three image distortions (image resizing, JPEG compression, and Gaussian blur). The distortion parameters are determined by considerations of the typical OSN applications. Note that some testing samples in this experiment are of low quality. For example, the samples are in a resolution of 534×503 pixels after a resizing ratio of 0.7.

Tab. 2 shows that the proposed CMA achieved the best robustness across different distortions. For example, under JPEG compression with a quality factor of 70, the EER of our CMA is 24.62%, which increases by 3.83 p.p. or 18.42% compared to that under no compression. However, the EERs of other approaches are over 35.96% under the same setting. For the RGB-only approaches, such as the ViT trained by strategy in [4], it works well under samples with moiré patterns. The performance degradation is mainly due to the loss of high-frequency components during the DCT quantization steps in JPEG compression. Similar conclusions can be achieved for the resizing distortion.

The blurring distortion is challenging since the Gaussian kernel covers multiple pixels in a local region. After applying the blurring kernel, the resulting pixel value at the center of the blurring kernel is a weighted sum of the 17×17 neighboring pixels [49]. Therefore, the pixel-level color artifacts are disturbed after such distortion.

E. t-SNE Visualization

Fig. E1 shows t-SNE [48] visualizations of embeddings of [CLS] after the last transformer layer and before the classification head. The visualization illustrates the separation of sampling within a two-dimensional space. Experimental setup refers to [30]. To plot the t-SNE graph, we employed the ROD_HQ dataset in accordance with the experimental protocols outlined in Sec. 5.1 to train our deep model. The integration of the chromaticity map modality into our model enhances the aggregation of genuine document images under two experimental protocols when compared to using the RGB modality alone. Additionally, the adaptor facilitates a more effective combination of chromaticity features. These outcomes are in line with the analyses presented in Sec. 5.2.