Did You Note My Palette? Unveiling Synthetic Images Through Color Statistics

Lea Uhlenbrock
Friedrich-Alexander University Erlangen-Nürnberg
Erlangen, Germany
lea.uhlenbrock@fau.de

Davide Cozzolino
Universita degli Studi di Napoli Federico II
Naples, Italy
davide.cozzolino@unina.it

Denise Moussa
Friedrich-Alexander University Erlangen-Nürnberg
Erlangen, Germany
denise.moussa@fau.de

Luisa Verdoliva
Universita degli Studi di Napoli Federico II
Naples, Italy
verdoliv@unina.it

Christian Riess
Friedrich-Alexander University Erlangen-Nürnberg
Erlangen, Germany
christian.riess@fau.de

ABSTRACT

High-quality artificially generated images are widely available now and increasingly realistic, posing challenges for image forensics in distinguishing them from real ones. Unfortunately, building a single detector that generalizes well to unseen generators is very difficult, creating the need for diverse cues. In this paper, we show that natural and synthetic images differ in their color statistics, possibly due to the widely used perceptual loss, which is more sensitive to brightness than to chroma differences. Consequently, color statistics offer valuable cues for forensic analysis and the development of robust detectors. Our experiments using simple hand-crafted color functions with a random forest achieve 90% accuracy across all tested Diffusion Models, even with limited training samples.

CCS CONCEPTS

• **Applied computing** → *Investigation techniques.*

KEYWORDS

Image Forensics, Color Spaces, Synthetic Image Detection, Diffusion Models

ACM Reference Format:
Lea Uhlenbrock, Davide Cozzolino, Denise Moussa, Luisa Verdoliva, and Christian Riess. 2024. Did You Note My Palette? Unveiling Synthetic Images Through Color Statistics. In *Proceedings of the 2024 ACM Workshop on Information Hiding and Multimedia Security (IH&MMSec '24), June 24–26, 2024, Baiona, Spain.* ACM, New York, NY, USA, 6 pages. https://doi.org/10.1145/3658664.3659652

Figure 1: Relational transformations of the color channels strongly enhance the visual detectability of synthetic image areas.

1 INTRODUCTION

The lines between real and generated content are blurred with the advent of advanced image synthesis techniques, facilitated by a rapid rise in the popularity of generative deep learning models. Especially recent Diffusion Models [17] are easily able to produce images that fool human perception. Popular examples of such generators include open-source software like Stable Diffusion [32], commercial products like Midjourney [18], or integrated tools like the recently introduced AI Firefly found in various Adobe Products [1]. Such images can serve various purposes, including illustration, humor, and even malicious use. This raises concerns not only regarding disinformation but also about the authenticity of legal evidence. With AI-generated images potentially entering courtrooms, there is a strong need for reliable and effective detection software. This raises the research interest in the detection of synthetic images.

A large line of research proposed numerous cues for detecting images from Generative Adversarial Networks (GANs) [4, 14, 24, 27, 37, 43], which preceded diffusion models. For example, GANs leave artificial fingerprints in the image frequencies that can be used for detection and model attribution [25]. It has been shown that some GAN-related findings can be transferred to diffusion-based images. For example, Diffusion Models also introduce distinct traces in the generated images similar to GAN fingerprints [8]. However, traces in diffusion-based images are often more subtle and the development of new detectors is an ongoing task.

Current detection strategies often focus on abstract traces learned by deep neural networks (DNNs) [3, 8, 15, 16, 23, 29, 35, 37, 38]. However, such black box systems lack transparency and interpretability. This poses challenges across applications where decisions must be carefully justified. Legal frameworks, such as the EU Artificial Intelligence Act, impose strong requirements for explainability

IH&MMSec '24, June 24–26, 2024, Baiona, Spain.

ACM ISBN 979-8-4007-0637-0/24/06…$15.00
https://doi.org/10.1145/3658664.3659652

in the deployment of detection tools [36], a criterion that many DNN-based tools struggle to meet. Further, to achieve good results, DNN-based approaches often need vast amounts of training data, which is additionally exacerbated by the rapid introduction of new image generators. Hence, it is desirable to research features for the discrimination of synthetic and real images that are both interpretable and well generalizable. If this can be achieved with relatively simple statistics, then it might become feasible to train classifiers with lower amounts of training data.

We hypothesize that color cues can contribute to fill this gap. As we outline in our related work section, there exist several reports that it is beneficial to perform the detection of synthetic images in alternative color spaces. However, there is no systematic study on color in synthetic image detection.

In this work, we offer several contributions towards a better understanding of the color properties of synthetic images.

(1) We show that the popular perceptual loss for training of image generators prefers to optimize luminosity over chrominance. Hence, color, or more generally differences between intensity channels, can be interesting candidates for future-proof detection strategies. Based on our findings, we conclude that a key element to the detection of synthetic images is the relationship between color channels, as they are not generated correctly regarding natural image statistics.
(2) We show that a simple transformation of the color channels provides visually well-interpretable results that may enable an analyst to detect inpainted areas by visual inspection.
(3) To demonstrate the effectiveness of a lightweight detector based on color properties, we build a straightforward detector for synthetic images that uses simple relationships between color channels and a random forest classifier.

We test the features on seven different diffusion-based image generators, including the four most current generators and three predecessor versions. Our simple features perform well across all generators, and on average they outperform related work in a cross-dataset evaluation. It is furthermore encouraging to note that this detector operates well across different generator versions, which may indicate robustness along the evolution of generators. All in all, our experiments show that it is feasible to use handcrafted features and a lightweight classifier to achieve detection results across different generators with an average accuracy of 90% while maintaining a high degree of interpretability.

2 RELATED WORK

Extensive research has been conducted in the realm of synthetic image detection, with a signficant body of literature existing on detecting GAN images and a rapidly growing amount of research works concering diffusion-based images. Many works on synthetic image detection operate on general statistics that are calculated from the pixel domain [3, 8, 15, 16, 23, 29, 35, 37, 38]. These methods are among the most effective approaches in terms of detection accuracy. On the downside, they typically rely on black box systems that lack transparency and interpretability. On the other hand, there also exist several works that operate on directly interpretable, visible cues. Examples are the faulty generation of geometry like unusual amounts of fingers on a human hand, physically wrong shadows [12], or inconsistent scene lighting [11]. Early generative models sometimes introduced visible asymmetries like differences in the eye colors of a person [26], unnatural color tones [20], and broken structures or smudged blobs [41]. As generative AI advances, modern Diffusion Models like Midjourney 6 or DALL-E 3, for instance, create more sophisticated images with significantly fewer visible flaws compared to earlier models.

Arguably, methods focusing on simple image intensity statistics can offer a middle ground between abstract statistics and visual clues, thereby being easier to verify and interpret. For example, McCloskey and Albright observe that GAN-generated images are normalized during the generation process, which inherently bounds the range of color intensities, leading to the absence of under- or overexposed areas in these images [27]. Multiple works show that converting images into alternative color spaces like YCbCr, L*a*b*, and HSV can be beneficial for the detection performance of a classifier for distinguishing real and synthetic images. Zeng *et al.* [39] observe that generation artifacts are more visible in color spaces other than RGB and conduct a study of which colorspace works best, observing that chrominance components perform especially well. Li *et al.* [21], Mo *et al.* [28], and He *et al.* [16] also utilize chrominance components to detect synthetic images. Qiao *et al.* [31] investigate correlations of adjacent pixels of GAN images from various color channels in the color spaces RGB, HSV, and YCbCr. They select those channels where the correlation coefficients of real and synthetic images differs most. Chen *et al.* [6] report that the YCbCr colorspace is well suited to create detectors that are robust against some forms of postprocessing. Amin *et al.* [2] extract frequency properties of images from the color channels and use their correlation to detect synthetic images. Barni *et al.* [3] suggest enhancing the detection of GAN-generated images by incorporating cross-band co-occurrences describing the relationship between color channels along with spatial co-occurrences computed separately for each band.

These findings collectively suggest the effectiveness of color transformations and indicate that leveraging alternative color spaces can improve synthetic image detection. However, there is no further investigation into why intensity transforms may benefit synthetic image detection, nor into the space of possible transforms beyond pre-defined spaces such as HSV.

3 COLOR TRANSFORMS IN SYNTHETIC IMAGES

Synthetic and real images differ in their image formation, which is our starting point for understanding the origin of differences in color distributions. A real image taken by a camera is created by light falling through a lens onto the individual cells of the camera sensor. In contrast, a generative network creates a synthetic image by reproducing learned features from real images. This reproduction is an optimization task, where specific image statistics are approximated through an optimization loss. For contemporary generative networks, the primary goal is to create images that are visually pleasing. Conversely, features that contribute less to perceptual quality are typically also less constrained by the optimization. Such underconstrained statistics open an opportunity to obtain forensic traces to distinguish real from synthetic images.

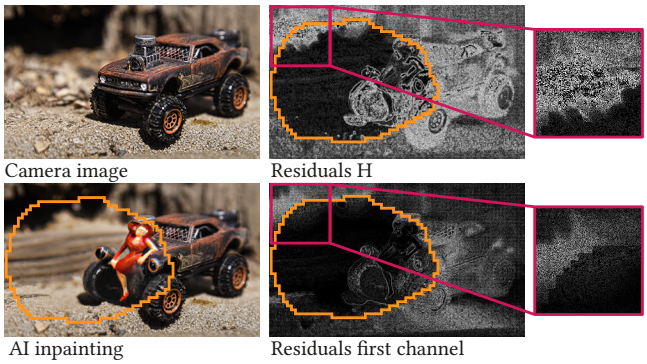

Figure 2: Left: Original image and below inpainting created in DALL-E 2 in RGB color space. Right: Using the residuals of the H component of HSV colorspace (top) in comparison with the log-scaled residuals from channels reordered by intensity (bottom).

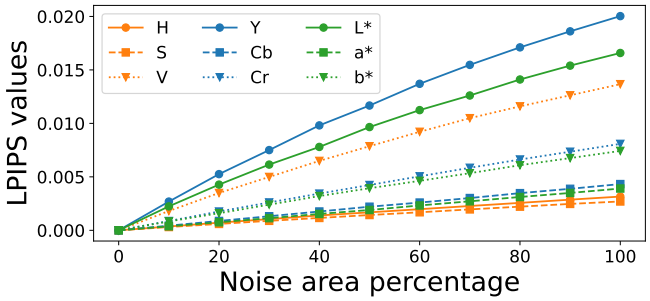

Figure 3: Impact of different color components on LPIPS loss [42]. Luminance components influence the loss more strongly than chrominance components.

3.1 Visible Noise Patterns in Color Transforms

Synthetic images are optimized for the RGB colorspace with no explicit regard to their representation in alternative colorspaces. This becomes apparent when examining natural images where a subregion is synthetically generated. Figure 2 shows such an example. A natural image (top left) is locally modified with the DALL-E 2 generator (bottom left). In the top right image of Fig. 2, the modified image is shown after conversion to HSV space and high-pass filtering (cf. Sec. 4) of its hue component H. The bottom right of Fig. 2 uses a custom transform instead, namely the largest intensity of every pixel's RGB triplet, and a logarithm on the residuals. In both transforms, the noise patterns of the generator are even visually identifiable, without further sophisticated computations.

One may hypothesize that there exists a whole space of inter-channel transforms of pixel colors that may increase visual interpretability and detectability. From this point of view, the custom transform in the bottom right of Fig. 2 is only an example. Nevertheless, we can further formalize this example transform: For each RGB pixel $p_{ij} = (r, g, b)$ at position (i, j) in image I, we extract the index of the color channel with maximum intensity with $\Uparrow_{ij} = \arg\max(p_{ij}) \in \{0, 1, 2\}$. We analogously extract per-pixel the index of the minimum intensity, namely $\Downarrow_{ij} = \arg\min(p_{ij}) \in \{0, 1, 2\}$. Finally, the remaining index per pixel is the median intensity, formally $\Updownarrow_{ij} \in \{0, 1, 2\} \setminus \{\Uparrow_{ij}, \Downarrow_{ij}\}$. To obtain then three new channels, where each only contains the upper, median or lower values per pixel, we use these per-pixel indices, formalized as $U_{ij} = I_{\Uparrow ij}$, $M_{ij} = I_{\Updownarrow ij}$ and $L_{ij} = I_{\Downarrow ij}$. Then, we high-pass filter the logarithm of U, M, and L to obtain residuals. Figure 2 shows that U channel residuals lead to even better visual results for an example DALL-E 2 image than H from HSV, which has been used in several previous color-related works [16, 21, 28, 31, 39].

3.2 Luminosity Bias in the Perceptual Loss

It is interesting to further investigate *why* color transforms are apparently beneficial for the detection of synthetic images. One important part of an image generator network is the conversion of the image from the latent to the pixel space. When examining the architecture of Stable Diffusion [33], one module of this conversion is the usage of a perceptual loss to control the image quality in pixel space, namely the Learned Perceptual Image Patch Similarity (LPIPS) [42]. LPIPS is a learned quality metric that is a popular choice also in many other generative networks [5, 10, 19]. It computes the perceptual distance between two images based on their features in latent space in a pre-trained standard network. Many deployed image generators are commercial, and their precise architecture is not disclosed, but we can assume that they use similar loss terms as all of them prioritize visual quality.

Interestingly, it turns out that LPIPS is much more sensitive to luminosity than to chromaticity, thereby encouraging a more accurate reproduction of luminosity and a potentially stronger deviation of chromaticity statistics from natural images.

We demonstrate this peculiar behavior of LPIPS in a small experiment. The idea is to calculate the LPIPS loss between a natural image and a copy of the image that is corrupted by Gaussian noise. The noise injection is performed in the channels of a transformed color space, namely HSV, L*a*b*, and YCbCr (noise magnitudes are controlled to be of equivalent level after transformation back to RGB). The results of the LPIPS loss calculation are shown in Fig. 3. The x-axis indicates the percentage of pixels affected by the noise injection. The y-axis indicates the LPIPS loss. It can be observed that the luminosity channels H (from HSV), Y (from YCbCr) and L (from L*a*b*) achieve approximately 3 times larger LPIPS losses compared to chroma channels. Hence, when used in neural network training, the LPIPS loss will punish errors in luminosity reproduction much more severely than errors in chromaticity reproduction. Conversely, statistical discrepancies between real and generated images are more likely to be found in the chromaticities.

More generally, the transformations from RGB to the three alternative spaces HSV, YCbCr, and L*a*b* relate the R, G, and B color channels in quite different ways. YCbCr and L*a*b* calculate weighted sums of the RGB channels, while HSV reorders and normalizes the color channels by their intensity. Hence, one may hypothesize that there is a larger, unknown space of possible inter-channel functions that improve the detectability of synthetic images. We leave this question to future work.

Table 1: KL divergences between residual histograms of real and synthetic images for different color representations.

R	G	B	H	S	V	Y	Cb	Cr	$\frac{L}{U}$	$\frac{L}{M}$	$\frac{Md}{U}$
0.02	0.01	0.02	0.08	0.06	0.02	0.01	0.07	0.09	0.07	0.08	0.07

3.3 Distinctiveness of Color Spaces

The previous investigations suggest that some color channels in some color spaces provide more useful statistics for distinguishing real and synthetic images than others. To quantitatively demonstrate this, we take individual channels from RGB, HSV, YCbCr, and a (relatively arbitrarily chosen) custom color space consisting of pixel-wise fractions of the channel relationships U, M, and L (cf. Sec. 3.1), namely $\frac{Lo}{U}$, $\frac{L}{M}$, and $\frac{M}{U}$. For each channel of each color space, histograms of high-pass residuals are calculated (analogously to Sec. 4, details omitted here) and averaged over 1,000 real images from the COCO set or 1,000 generated images from Stable Diffusion.

Table 1 shows the Kullback-Leibler (KL) divergence between the residual histograms of real and synthetic images. Analogous to the behavior of the perceptual loss, the statistics over the plain RGB channels, as well as the luminosity channels V (from HSV) and Y (from YCbCr) exhibit by far the smallest KL divergences of 0.01 or 0.02, whereas the other channels, including the arbitrarily defined ratios, exhibit KL divergences between 0.06 and 0.09.

4 COLOR STATISTICS FOR SYNTHETIC IMAGE DETECTION

Color transforms are a relatively rich source of information. As such, it does not necessarily require a complex detection pipeline. We show this by feeding a few straightforward color statistics to a lightweight random forest classifier to detect synthetic images from state-of-the-art generative image models.

The feature extraction follows a very traditional pipeline. An image is transformed into one or more target color spaces, from which selected color channels are extracted. High-pass noise residuals are calculated from the logarithm of the color channel. In previous works, the high-pass filters for obtaining the noise residuals are oftentimes applied in horizontal or vertical direction. However, Corvi *et al.* [7] point out that the spectral energy of synthetic images compared to real images is not isotropic, and that they are more discriminative in diagonal direction. Following their observation, we use as a high-pass filter a diagonal variant of the 3×3 discrete Laplace filter, with coefficients $(1, 0, 1; 0, -4, 0; 1, 0, 1)$. In preliminary experiments, this filter performed better than a standard Laplace filter, subtraction of a Gaussian low-pass filtered image, non-local means, and the learning-based denoiser DnCNN [40]. The extracted residuals are grouped into co-occurrence matrices following the work by Fridrich and Kodovsky [13]. To this end, residuals are quantized by a factor of 2 and truncated to the value range $[-2; 2]$. Five of such quantized and truncated residuals in horizontal or vertical direction form a co-occurrence vector. The relative frequency of each vector forms a histogram, whereby mirrored and sign-flipped residuals are collected in one bin [13]. One such histogram describes the statistics of one transformed color channel of the image. We then use these histograms to train a classic Random Forest classifier with 200 trees to distinguish real and synthetic images.

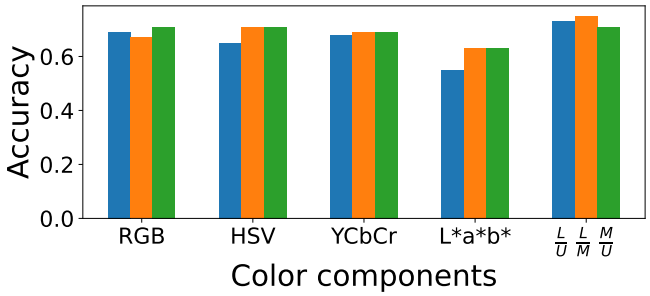

Figure 4: Accuracy when using different color components in our training setup. The use of color relationships specifically increases detection performance.

4.1 Datasets

We use images from several state-of-the-art Diffusion Models. We use DALL-E 2 data from Corvi et al. [7]. Additionally, we generate data from DALL-E 3 [30], Midjourney 5, Midjourney 6, Stable Diffusion 1.5 and Stable Diffusion XL as well as Firefly, using the generation tool integrated in Adobe Photoshop. The captions of images from the COCO dataset[22] were used as prompts for generation. This helps to align the visual content of the generated diffusion-based images closer to real images, and hence to encourage the classifier to focus primarily on low-level statistics during detection. All images have the dimension of 1024×1024 pixels, except for Stable Diffusion, where images have a dimension of 512×512. During training, we perform a central crop of all samples to a size of 512×512 pixels to operate on identical image sizes.

4.2 Training and Evaluation Protocol

One training assumption is that there is an abundance of real images available, and also sufficiently many training images from at least one single synthetic image generator. The set of real images consists of 10,000 images from the COCO [22] and RAISE [9] datasets, with an additional 10,000 RAISE images downscaled to 512 pixels along the shorter axis to simulate lower quality. The set of generated images are 10,000 images from Stable Diffusion.

For testing, we use 200 images from LAION [34] as original images, and 200 images from each image generator, namely Stable Diffusion, DALL-E 2, DALL-E 3, Midjourney 5, Midjourney 6, FireFly, and Stable Diffusion XL (SDXL). By default, we evaluate on generated images from Stable Diffusion, except for the generalization experiment, where images from various generators are considered. Note that the real images for testing are always from a different source than the training images, hence each experiment exhibits a slight training-test mismatch.

The training data is randomly augmented with a probability of 10% by either JPEG compression, Blur, or Resize. For JPEG compression, we use a random quality factor of $75 + 5k$ with $0 \leq k \leq 4$, for blurring we use Gaussian kernel with $\sigma = l \cdot 0.5$ with $1 \leq l \leq 5$, and for resizing we use a random resizing factor $r \in [0.45, 0.75, 1.25, 1.55, 1.75]$.

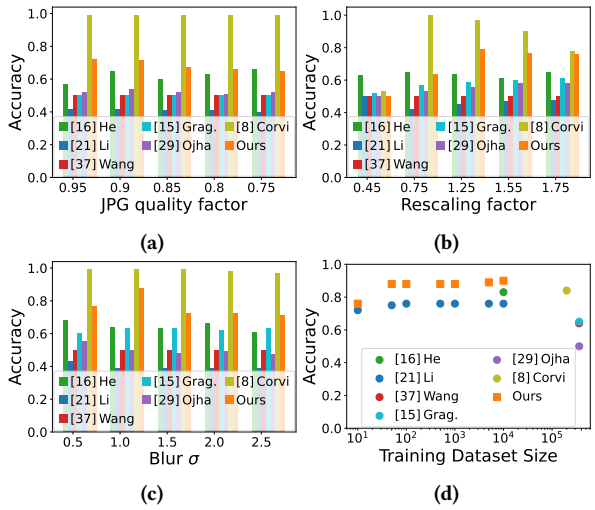

Figure 5: Detection accuracy under different post-processing attacks as well as with different sizes of training datasets.

The methods for comparison are four general learning-based methods and two color-based works. For general learning-based synthetic image detectors, we use the works by Wang *et al.* [37] and Gragnaniello *et al.* [15] based on ResNet50, Corvi *et al.* [8] trained on latent diffusion data and Ojha *et al.* [29] using CLIP for detection. These general image detectors are used in their pre-trained versions as they can be downloaded. The color-based work by Li *et al.* [21] uses a combination of color components and is trained with an ensemble of linear classifiers. The color-based work by He *et al.* [16] uses a shallow Convolutional Neural Network (CNN) to extract image features that are then classified by a Random Forest. Both color works are trained from scratch. Due to space constraints, we will reference those works by the first author names.

4.3 Evaluation Results

We first train the classifier separately on the color components, to get some insight in the performance of individual color channels. Training and testing is performed on LAION and Stable Diffusion images. A comparison can be seen in Fig. 4. Here, the fractions of re-ordered channels $\frac{L}{U}$, $\frac{L}{M}$, $\frac{M}{U}$ work best, but the differences to other color channels are not very pronounced. Hence, for all subsequent experiments, we form a combined feature vector that concatenates the feature vectors from multiple color channels in order to gain additional robustness. This combined feature vector consists of features from the three fractions of re-ordered channels $\frac{L}{U}$, $\frac{L}{M}$, $\frac{M}{U}$ and from the channels H from HSV and Cb and Cr from YCbCr based on the large KL divergences in Tab 1 in Sec. 3.3.

The generalization capability of the combined feature vector is shown in a cross-generator experiment. As before, the Stable Diffusion dataset is used for training. All methods are then tested on the images from Stable Diffusion, but in particular also on images from DALL-E 2, DALL-E 3, Midjourney 5, Midjourney 6, Firefly, and SDXL. Table 2 shows the results of this experiment. The reported metrics are accuracy and Area under the Curve (AUC). In the first column, the Stable Diffusion results indicate in-distribution accuracy for the methods that were explicitly retrained on this data (Li *et al.*, He *et al.*, and ours), the other columns indicate the generalization performance. The last column reports the average accuracy and AUC per method. While most methods achieve good scores on a subset of the generators, it is challenging to provide consistent scores across all generators. The proposed color statistics perform competitively and yield good generalization accuracies of 74% at worst, and 90% on average across all evaluated methods.

Another aspect of robustness is the resilience to postprocessing operations. For images from social media, one may commonly expect lossy JPEG compression and rescaling. It is also customary to investigate the impact of blurring as a potential postprocessing operation. We study the impact of these operations on the color-based methods. The training protocol is identical to the previous experiments, only the testing data is subject to various degrees of JPEG compression, resizing, and blurring. The results are shown in Fig. 5 (a) to (c). Though Corvi *et al.* performs better across all postprocessing operations,the proposed color statistics perform comparably well and better than most related works on these tasks.

One remarkable benefit of the proposed color statistic is its ability to operate on a limited amount of training data. This may be beneficial when adapting the method to novel generators. To demonstrate this, the classifier is re-trained with dataset sizes of $\{10, 50, 100, 500, 1000, 5000, 10000\}$ images while leaving the other experimental settings fixed. We also retrained the method by Li *et al.* on the same dataset sizes, since this method is also relatively light-weight. For the remaining methods, we report the amount of training data in relation to accuracy. The results are shown in Fig. 5 (d). The proposed color statistics (orange) achieves good performance already for 50 training images. As such, it provides a better training data/accuracy tradeoff than related works.

5 DISCUSSION AND CONCLUSIONS

We demonstrate that image generators focus their training on reproducing luminosity, which opens an angle for forensic analysis of the color statistics of generated images. Earlier works reported cursory insights that, e.g., HSV is a good color space for DeepFake classification. We show that the class of possible color statistics is potentially much larger, and provide empirical evidence by using next to standard color transforms also the (rather arbitrary) fraction of re-ordered pixel intensities as an inter-channel feature.

Our proof-of-concept experiments with a lightweight classifier show that these features exhibit excellent generalization capabilities, while requiring only few training images. Moreover, suitable inter-channel relationships may even provide visual cues, where the change in noise pattern is even perceptually visible, which may be useful for providing interpretable forensic evidence. In future work, it will be interesting to extend the exploration of the space of inter-channel functions, and to apply more complex classifiers to further probe the performance limits of this forensic cue.

ACKNOWLEDGMENTS

This work was supported by Deutsche Forschungsgemeinschaft (DFG, German Research Foundation) as part of the Research and Training Group 2475 "Cybercrime and Forensic Computing" (grant number 393541319/GRK2475/2-2024).

Table 2: Generalization capability of different classifiers in terms of accuracy. All classifiers are tested on Stable Diffusion, DALL-E 2, DALL-E 3, Midjourney 5, Midjourney 6, Firefly, and SDXL.

Acc./AUC%	Stable Diff.	DALL-E 2	DALL-E 3	Midjourney 5	Midjourney 6	FireFly	SDXL	Avg. Acc. / Avg. AUC
[37] Wang	0.50 / 0.51	0.50 / 0.78	0.50 / 0.32	0.50 / 0.85	0.50 / 0.60	0.51 / 0.79	0.50 / 0.54	0.50 / 0.63
[15] Grag.	0.58 / 0.87	0.59 / 0.93	0.49 / 0.66	0.70 / 0.97	0.53 / 0.84	0.98 / 0.99	0.67 / 0.92	0.65 / 0.88
[29] Ojha	0.58 / 0.74	0.78 / 0.92	0.49 / 0.51	0.64 / 0.80	0.50 / 0.50	0.82 / 0.93	0.64 / 0.80	0.64 / 0.74
[8] Corvi	0.99 / 1.00	0.49 / 0.45	0.98 / 0.99	0.91 / 0.95	0.99 / 1.00	0.53 / 0.63	0.99 / 1.00	0.84 / 0.86
[21] Li	0.92 / 0.90	0.99 / 0.98	0.49 / 0.49	0.50 / 0.52	0.96 / 0.94	0.49 / 0.49	0.99 / 0.97	0.76 / 0.75
[16] He	0.72 / 0.80	0.80 / 0.86	0.93 / 0.87	0.50 / 0.50	0.77 / 0.85	0.78 / 0.85	0.97 / 0.92	0.78 / 0.81
Ours	0.98 / 1.00	0.96 / 0.99	0.74 / 0.93	0.91 / 0.98	0.94 / 0.98	0.85 / 0.97	0.96 / 0.99	0.91 / 0.98

REFERENCES

[1] Adobe. 2024. *Firefly.* https://www.adobe.com/sensei/generative-ai/firefly.html.

[2] M.A. Amin, Y. Hu, H. She, J. Li, Y. Guan, and M.Z. Amin. 2023. Exposing Deepfake Frames through Spectral Analysis of Color Channels in Frequency Domain. In *International Workshop on Biometrics and Forensics*. IEEE, 1–6.

[3] M. Barni, K. Kallas, E. Nowroozi, and B. Tondi. 2020. CNN Detection of GAN-Generated Face Images Based on Cross-Band Co-Occurrences Analysis. In *IEEE International Workshop on Information Forensics and Security)*. IEEE, 1–6.

[4] N. Bonettini, P. Bestagini, S. Milani, and S. Tubaro. 2021. On the Use of Benford's Law to Detect GAN-Generated Images. In *International Conference on Pattern Recognition*. IEEE, 5495–5502.

[5] J. Bruna, P. Sprechmann, and Y. LeCun. 2016. Super-Resolution with Deep Convolutional Sufficient Statistics.

[6] B. Chen, X. Liu, Y. Zheng, G. Zhao, and Y.-Q. Shi. 2021. A Robust GAN-Generated Face Detection Method Based on Dual-Color Spaces and an Improved Xception. *IEEE Transactions on Circuits and Systems for Video Technology* 32, 6 (2021), 3527–3538.

[7] R. Corvi, D. Cozzolino, G. Poggi, K. Nagano, and L. Verdoliva. 2023. Intriguing Properties of Synthetic Images: From Generative Adversarial Networks to Diffusion Models. In *IEEE Conference on Computer Vision and Pattern Recognition*. 973–982.

[8] R. Corvi, D. Cozzolino, G. Zingarini, G. Poggi, K. Nagano, and L. Verdoliva. 2023. On the Detection of Synthetic Images Generated by Diffusion Models. In *International Conference on Acoustics, Speech and Signal Processing*. IEEE, 1–5.

[9] D.-T. Dang-Nguyen, C. Pasquini, V. Conotter, and G. Boato. 2015. RAISE: A Raw Images Dataset for Digital Image Forensics. In *ACM Multimedia Systems Conference*. 219–224.

[10] A. Dosovitskiy and T. Brox. 2016. Generating Images with Perceptual Similarity Metrics Based on Deep Networks. In *Advances in Neural Information Processing Systems*, Vol. 29.

[11] H. Farid. 2022. *Lighting (In)consistency of Paint by Text.* Technical Report 2207.13744. arXiv.

[12] H. Farid. 2022. *Perspective (In)consistency of Paint by Text.* Technical Report 2206.14617. arXiv.

[13] J. Fridrich and J. Kodovsky. 2012. Rich Models for Steganalysis of Digital Images. *IEEE Transactions on Information Forensics and Security* 7, 3 (2012), 868–882.

[14] M. Goebel, L. Nataraj, T. Nanjundaswamy, T. Manhar Mohammed, S. Chandrasekaran, and B.S. Manjunath. 2020. Detection, Attribution and Localization of GAN Generated Images.

[15] D. Gragnaniello, D. Cozzolino, F. Marra, G. Poggi, and L. Verdoliva. 2021. Are GAN Generated Images Easy to Detect? A Critical Analysis of the State-of-the-Art. In *IEEE International Conference on Multimedia and Expo*. IEEE, 1–6.

[16] P. He, H. Li, and H. Wang. 2019. Detection of Fake Images via the Ensemble of Deep Representations from Multi Color Spaces. In *IEEE International Conference on Image Processing*. IEEE, 2299–2303.

[17] J. Ho, A. Jain, and P. Abbeel. 2020. Denoising Diffusion Probabilistic Models. In *Advances in Neural Information Processing Systems*, Vol. 33. 6840–6851.

[18] Midjourney Inc. 2024. *Midjourney.* https://www.midjourney.com/home.

[19] J. Johnson, A. Alahi, and L. Fei-Fei. 2016. Perceptual Losses for Real-Time Style Transfer and Super-Resolution. In *European Conference on Computer Vision*. Springer, 694–711.

[20] E.-G. Lee, I. Lee, and S.-B. Yoo. 2023. ClueCatcher: Catching Domain-Wise Independent Clues for Deepfake Detection. *Mathematics* 11, 18 (2023), 3952.

[21] H. Li, B. Li, S. Tan, and J. Huang. 2020. Identification of Deep Network Generated Images Using Disparities in Color Components. *Signal Processing* 174 (2020), 107616.

[22] T.-Y. Lin, M. Maire, S. Belongie, J. Hays, P. Perona, D. Ramanan, P. Dollár, and C.L. Zitnick. 2014. Microsoft COCO: Common Objects in Context. In *European Conference on Computer Vision*. Springer, 740–755.

[23] B. Liu, F. Yang, X. Bi, B. Xiao, W. Li, and X. Gao. 2022. Detecting Generated Images by Real Images. In *European Conference on Computer Vision*. Springer, 95–110.

[24] S. Mandelli, N. Bonettini, P. Bestagini, and S. Tubaro. 2022. Detecting GAN-Generated Images by Orthogonal Training of Multiple CNNs. In *IEEE International Conference on Image Processing*. IEEE, 3091–3095.

[25] F. Marra, D. Gragnaniello, L. Verdoliva, and G. Poggi. 2018. Do GANs Leave Artificial Fingerprints?. In *IEEE Conference on Multimedia Information Processing and Retrieval (MIPR)*. 506–511.

[26] F. Matern, C. Riess, and M. Stamminger. 2019. Exploiting Visual Artifacts to Expose Deepfakes and Face Manipulations. In *IEEE Winter Applications of Computer Vision Workshops*. IEEE, 83–92.

[27] S. McCloskey and M. Albright. 2019. Detecting GAN-Generated Imagery Using Saturation Cues. In *EEE International Conference on Image Processing*. IEEE, 4584–4588.

[28] S. Mo, P. Lu, and X. Liu. 2022. AI-Generated Face Image Identification with Different Color Space Channel Combinations. *Sensors* 22, 21 (2022), 8228.

[29] U. Ojha, Y. Li, and Y.J. Lee. 2023. Towards Universal Fake Image Detectors that Generalize Across Generative Models. In *IEEE Conference on Computer Vision and Pattern Recognition*. 24480–24489.

[30] OpenAI. 2024. *DALL-E 3.* https://openai.com/dall-e-3.

[31] T. Qiao, Y. Chen, X. Zhou, R. Shi, H. Shao, K. Shen, and X. Luo. 2023. CSC-Net: Cross-Color Spatial Co-occurrence Matrix Network for Detecting Synthesized Fake Images. *IEEE Transactions on Cognitive and Developmental Systems* (2023).

[32] R. Rombach, A. Blattmann, D. Lorenz, P. Esser, and B. Ommer. 2022. High-Resolution Image Synthesis with Latent Diffusion Models. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 10684–10695.

[33] R. Rombach, A. Blattmann, D. Lorenz, P. Esser, and B. Ommer. 2024. *Stable Diffusion.* https://github.com/CompVis/stable-diffusion.

[34] C. Schuhmann, R. Vencu, R. Beaumont, R. Kaczmarczyk, C. Mullis, A. Katta, T. Coombes, J. Jitsev, and A. Komatsuzaki. 2021. Laion-400M: Open Dataset of CLIP-Filtered 400 Million Image-Text Pairs.

[35] C. Tan, Y. Zhao, S. Wei, G. Gu, and Y. Wei. 2023. Learning on Gradients: Generalized Artifacts Representation for GAN-Generated Images Detection. In *IEEE Conference on Computer Vision and Pattern Recognition*. 12105–12114.

[36] European Union. 2021. *Artificial Intelligence Act, Article 10, paragraph 3.* https://artificialintelligenceact.eu/the-act/.

[37] S.-Y. Wang, O. Wang, R. Zhang, A. Owens, and A.A. Efros. 2020. CNN-Generated Images Are Surprisingly Easy to Spot... for Now. In *IEEE Conference on Computer Vision and Pattern Recognition*. 8695–8704.

[38] Z. Wang, J. Bao, W. Zhou, W. Wang, H. Hu, H. Chen, and H. Li. 2023. DIRE for Diffusion-Generated Image Detection.

[39] K. Zeng, X. Yu, B. Liu, Y. Guan, and Y. Hu. 2023. Detecting Deepfakes in Alternative Color Spaces to Withstand Unseen Corruptions. In *International Workshop on Biometrics and Forensics*. IEEE, 1–6.

[40] K. Zhang, W. Zuo, Y. Chen, D. Meng, and L. Zhang. 2017. Beyond a Gaussian Denoiser: Residual Learning of Deep CNN for Image Denoising. *IEEE Transactions on Image Processing* 26, 7 (2017), 3142–3155.

[41] L. Zhang, Y. Zhou, C. Barnes, S. Amirghodsi, Z. Lin, E. Shechtman, and J. Shi. 2022. Perceptual Artifacts Localization for Inpainting. In *European Conference on Computer Vision*. Springer, 146–164.

[42] R. Zhang, P. Isola, A. A Efros, E. Shechtman, and O. Wang. 2018. The Unreasonable Effectiveness of Deep Features as a Perceptual Metric. In *IEEE Conference on Computer Vision and Pattern Recognition*. 586–595.

[43] Xu Zhang, Svebor Karaman, and Shih-Fu Chang. 2019. Detecting and Simulating Artifacts in GAN Fake Images. In *IEEE International Workshop on Information Forensics and Security*.