

BEYOND SPECTRAL PEAKS: INTERPRETING THE CUES BEHIND SYNTHETIC IMAGE DETECTION

Sara Mandelli¹, Diego Vila-Portela², David Vázquez-Padín², Paolo Bestagini¹, Fernando Pérez-González²

¹Dipartimento di Elettronica, Informazione e Bioingegneria, Politecnico di Milano - Milan, Italy

²atlanTTic Research Center, University of Vigo, E.E. de Telecomunicación - Vigo, Spain

ABSTRACT

Over the years, the forensics community has proposed several deep learning-based detectors to mitigate the risks of generative AI. Recently, frequency-domain artifacts (particularly periodic peaks in the magnitude spectrum), have received significant attention, as they have been often considered a strong indicator of synthetic image generation. However, state-of-the-art detectors are typically used as black-boxes, and it still remains unclear whether they truly rely on these peaks. This limits their interpretability and trust.

In this work, we conduct a systematic study to address this question. We propose a strategy to remove spectral peaks from images and analyze the impact of this operation on several detectors. In addition, we introduce a simple linear detector that relies exclusively on frequency peaks, providing a fully interpretable baseline free from the confounding influence of deep learning. Our findings reveal that most detectors are not fundamentally dependent on spectral peaks, challenging a widespread assumption in the field and paving the way for more transparent and reliable forensic tools.

Index Terms— Synthetic image detection, Frequency artifacts, Image forensics, Interpretability

1. INTRODUCTION

The advent of generative AI has fundamentally changed the way synthetic content is produced, making it possible for virtually anyone to generate high-quality media without advanced technical knowledge. Although such technologies hold promise for creative industries and data enhancement [1], they have also raised serious concerns about privacy, security, and the dissemination of misinformation [2]. The proliferation of deepfake generation techniques amplifies the potential for malicious use, ranging from identity theft and non-consensual pornography to political disinformation and fraud.

To counteract the spreading of malicious deepfakes, a wide range of forensic detectors has been proposed in recent years, almost all of which rely on deep learning [3, 4, 5, 6, 7, 8, 9]. However, since these models often operate as “black boxes”, one of the main challenges lies in the interpretability of their outcomes. In particular,

This work was supported by the FOSTERER project, funded by the Italian Ministry of Education, University, and Research within the PRIN 2022 program. This work was also partially supported by the European Union - Next Generation EU under the Italian National Recovery and Resilience Plan (NRRP), Mission 4, Component 2, Investment 1.3: CUP D43C22003080001, partnership on “Telecommunications of the Future” (PE00000001 - program “RESTART”); CUP D43C22003050001, partnership on “SEcurity and RIghts in the CyberSpace” (PE000000014 - program “FF4ALL-SERICS”). This work was also supported by Xunta de Galicia and the European Regional Development Fund under Project ED431C 2025/41.

it remains hard to understand which generative artifacts the detectors exploit in order to distinguish synthetic content from authentic data.

Interpreting the output of a forensic detector remains a highly complex task, as deep learning systems often lack transparency in their decision-making process. A critical concern is that a detector may inherit biases from its training dataset, leading it to base its predictions on spurious correlations rather than genuine generative artifacts. For example, differences in compression levels, semantic content, or other dataset-specific characteristics between real and synthetic images can unintentionally guide the detector’s classification [4]. This raises important questions regarding the reliability, generalization, and fairness of current deep learning-based detection methods. Moreover, even in the absence of dataset biases, when a detector genuinely captures characteristic traces of the generation process, explaining which specific cues are being leveraged remains highly challenging. It is also likely that different detectors rely on distinct types of generative traces, with some focusing on certain patterns and others exploiting entirely different signals.

Over the years, the forensic community has reported frequency domain artifacts as an important trace that allows to tell real and synthetic images apart [4, 7, 10]. In particular, significant attention has been devoted to energy peaks occurring in the magnitude spectrum, which are often considered a strong indicator of synthetic image generation. However, the vast majority of prior research has limited to describing these spectral artifacts without explicitly leveraging them as discriminative cues for detection. This limitation is particularly evident in deep learning-based detectors, where interpreting the origin of a specific prediction remains challenging.

In this work, we take a step toward shedding light on the actual interpretability of synthetic image detectors. A central question we address is whether deep learning-based detectors truly rely on spectral peaks introduced by synthetic generation, or if these artifacts play only a marginal role in the decision process. To investigate this, we conduct a systematic analysis of several state-of-the-art detectors and design experiments aimed at disentangling their reliance on such frequency-domain cues. Specifically, we design a strategy to remove periodic peaks from the frequency spectrum of images and we assess how this operation affects different detectors. In addition, we introduce a very simple detector that relies exclusively on frequency peaks, without any data-driven component. This experiment enables a clearer interpretation of the results, free from the possible confounding effects of deep learning’s black-box nature.

Our results suggest that, for most detectors, the presence of spectral peaks does not constitute a fundamental artifact for detection, challenging a common assumption in the field and opening the way to a deeper understanding of what features these models exploit.

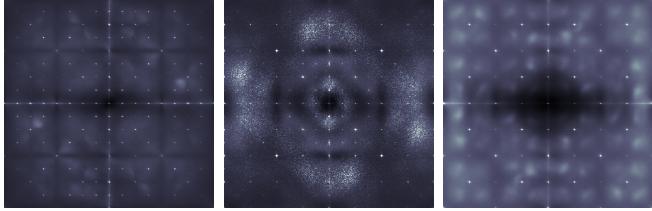


Fig. 1: Fourier transform analysis of synthetic images generated, respectively from left to right by Stable Diffusion (SD)3.5 [16], Flux 1.1Pro [17] and DALL-E 3 [18].

2. FREQUENCY-DOMAIN GENERATION ARTIFACTS

In the forensic community, it is widely recognized that synthetic image generation techniques introduce distinctive traces in their produced content [7, 10, 11]. All generated images exhibit such artifacts, whether produced from a text prompt or via the “img2img” modality (i.e., where a new image is synthesized from an existing one). Recent studies have further shown that even images simply passed through a generative model’s autoencoder (without any diffusion step) display artifacts similar to those found in fully synthetic generation [7]. These images have been referred to as “laundered”, since their semantic content is almost entirely preserved, with only minor imperceptible alterations, while still retaining synthetic-like traces in the frequency domain.

The generation artifacts often manifest as pronounced peaks in the Fourier spectrum of noise residuals extracted from synthetic images, typically appearing at components with periods of 4, 8 or 16 samples in both directions [11]. Together with these, artifacts may also appear as recurring structures like rings, ovals, or circular patterns in the Fourier domain [12]. Such artifacts are generally attributed to the upsampling operators employed during the decoding stage [13], although they may also arise from the characteristics of the training dataset used for a specific generator [10].

Artifact analysis is typically performed by subjecting the images to a high-pass filtering process, frequently implemented through a neural network. For instance, a very common choice is to adopt the DnCNN architecture [14] as a denoiser to reveal peaks and other spectral irregularities [7, 10]. To illustrate the frequency artifacts commonly observed in synthetic images, Fig. 1 reports the average power spectra computed from synthetic images of different generators included in the recently released Wild dataset [15]. Before averaging, all images have been processed with the DnCNN-based denoiser proposed in [10]. The spectra reveal a strong presence of peaks with different periodicity across all generators; moreover, each generator exhibits distinct spectral traces that could potentially allow for the unique characterization of its images.

Several recent studies have shown that synthetic and real images can be distinguished by analyzing their frequency spectra [4, 7, 10, 11, 19]. However, most prior research has limited to showing these spectral discrepancies rather than explicitly using them as discriminative detection traces. To our knowledge, only few studies have directly targeted frequency-domain mismatches between real and synthetic data [11, 20, 21, 22]. Moreover, all these works focused on artifacts from relatively old generators, with only SynthBuster [11] extending the investigation to diffusion models.

3. PROPOSED EXPERIMENTAL ANALYSIS

To take a step toward improving the interpretability of detectors, we focus on the Fourier spectrum of the tested images. Specifically, we

design two experiments: (i) peak removal from synthetic images; (ii) peak removal from laundered images.

In a nutshell, we apply a binary mask operator to the entire spectrum of the images (phase included) for removing the energy of the spectral components around the peaks. Our goal is to assess how detectors’ performance changes when these spectral modifications are applied. If detectors relied primarily on spectral peaks, we would expect their scores to vary significantly after peak removal. Conversely, if the spectral energy at peak positions were not a key factor for detection, the scores should remain largely unaffected.

Peaks removal from synthetic images. As shown in Section 2, the Fourier spectrum of each generator exhibits peaks at different positions. To suppress these peaks, we adopt a straightforward masking strategy in the Fourier domain, designed to remove frequency components lying on a $P \times P$ grid, where P denotes the periodicity of the peaks to be eliminated.

Formally, we define two sets of normalized spatial frequencies along two dimensions, $\mathcal{F}_x = \{n/P, n \in \mathbb{N}\}$ and $\mathcal{F}_y = \{m/P, m \in \mathbb{N}\}$. Then, we define a binary mask $\mathbf{M}(f_x, f_y)$ as

$$\mathbf{M}(f_x, f_y) = \begin{cases} 0, & \text{if } (f_x, f_y) \in \mathcal{F}_x \times \mathcal{F}_y \setminus \{(0, 0)\} \\ 1, & \text{elsewhere} \end{cases}. \quad (1)$$

To delete the peaks, we apply the mask to the full spectrum of each input image in a coefficient-wise fashion. However, we experimentally verified that simply deleting the periodic pattern at the exact peak positions is insufficient to fully suppress the spectral energy associated with the peaks. To address this, we apply a dilation operator with a disk-shaped structuring element, thereby slightly enlarging the “holes” in their surrounding area. Note that we preserve the peak at frequency $(0, 0)$ to avoid altering the low-pass behaviour of the image, i.e., its mean value and immediate surroundings. After applying the mask, we go back to the pixel domain, adjusting the output dynamics to match that of the input image, and we quantize it to 8-bit¹.

Fig. 2 illustrates the effect of this procedure by reporting the average power spectra of 1000 synthetic images generated with Midjourney [23], DALL-E 3 [18] and SDXL [24], before and after peak removal with periodicity $P = 8$. To show these examples, we omit the denoising step to directly highlight the plain spectral magnitudes before and after the proposed modification.

Peaks removal from laundered images. Alongside synthetic images, we also evaluate detectors on laundered images, i.e., real images passed through an encoding-decoding chain to erase their original traces and simulate synthetic generation [5, 7]. As a matter of fact, image laundering poses a concrete threat in forensic contexts, as it provides an effective way to conceal user traces and disguise content as if it were synthetically generated. Recently, it has been shown that even the most advanced detectors may be highly vulnerable to this manipulation, often misclassifying laundered images as synthetic [7]. As potential consequence, sensitive or harmful material may be overlooked or not appropriately flagged, increasing the risk of its dissemination online.

To experiment on these images, we apply the same peak-removal pipeline described previously for fully synthetic content. Fig. 3 illustrates an example of average spectral magnitudes computed from laundered images produced using the autoencoder of SD3.5 [16]. As it can be inspected, the laundered spectrum contains less visible artifacts than that of fully synthetic images; nonetheless, the frequency peaks are still there (kindly refer to the close-up shown in Fig. 3).

¹The peak removal code can be found at <https://github.com/polimi-ispl/beyond-spectral-peaks>

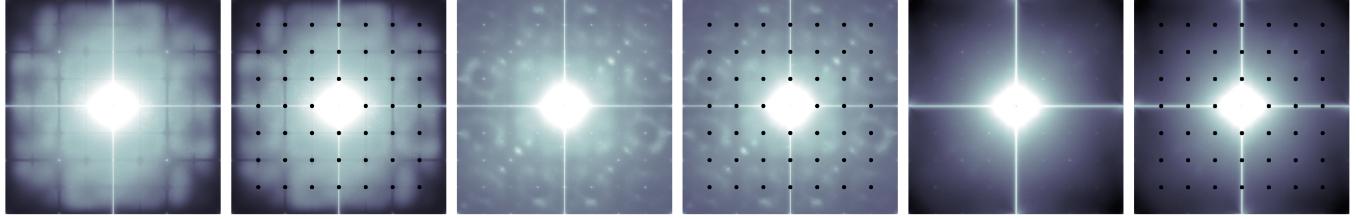


Fig. 2: Average Fourier spectra (magnitude, in logarithmic scale) of synthetic images generated with Midjourney (first column), DALL-E 3 (third) and SDXL (fifth) before and after peak removal with periodicity $P = 8$. Best viewed in electronic format.



Fig. 3: From left to right: average Fourier spectrum (magnitude, in logarithmic scale) of laundered versions of real images through SD3.5; close-up of one quadrant; close-up of the peak-removed spectrum with periodicity $P = 16$. Best viewed in electronic format.

4. EXPERIMENTAL SETUP

Dataset. Synthetic images have been selected from the recently released Wild dataset [15], which was built using some of the most popular commercial and open-source generators currently available. Specifically, we consider its “closed-set” sub-dataset, containing 1000 text-to-image samples for each of 10 generators, with average resolution of 1024×1024 pixels. All images are provided in uncompressed format, with the exception of those produced by Adobe Firefly [25] and Freepik [26], which are JPEG-compressed files.

Real images have been selected from the Raise dataset [27]; specifically, we randomly selected 1000 images and cropped them to a standard size of 1024×1024 pixels from the top-left corner. Laundered versions of these images were then synthesized following the procedure described in [7], using recent open-source generators, namely SDXL [24], SD3.5 [16] and Flux.1 [28].

Importantly, all images used in our analysis are kept uncompressed. We do this on purpose, as it is well known that JPEG compression introduces a characteristic 8×8 grid pattern in the frequency domain, and this can obscure synthetic generation traces [11]. Similar distortions could arise from resizing or rotation, as these operations typically leave interpolation artifacts in the spectrum. Since our goal is to investigate whether the removal of frequency peaks specific to synthetic generation affects forensic detectors, we restrict our analysis to uncompressed images not undergone any post-processing. For this reason, we also omit synthetic images of Adobe Firefly and Freepik from our analysis.

Synthetic image detectors. We evaluate several state-of-the-art detectors spanning from CNNs to transformer-based architectures, trained under different paradigms and on diverse datasets. Specifically, we consider the detectors proposed in [3, 4, 5, 6, 7, 8, 9], all of them being publicly available, with both testing code and pretrained weights. For consistency, we run all experiments using the official implementations released by the original authors.

We initially considered also the SynthBuster detector [11] which, as discussed in Section 2, relies on feeding a machine learning model with information about frequency peaks. However, it experimentally reported excessively high false alarms on original content, and, for this reason, we excluded it from our evaluation.

Nonetheless, inspired by the SynthBuster strategy, we also evaluate an extremely simple linear detector that focuses solely on the frequency peaks, without relying on any data-driven approach. This allows us to provide a clearer interpretation of the results, without the confounding influence of the black-box nature of deep learning solutions. Our developed detector extracts a high-pass residue from each image by applying a Laplacian of Gaussian Kernel \mathbf{H} [29]:

$$[\mathbf{H}]_{n_1, n_2} = -\frac{1}{\pi\sigma^4} \left[1 - \frac{n_1^2 + n_2^2}{2\sigma^2} \right] e^{-\frac{n_1^2 + n_2^2}{2\sigma^2}}, \quad (2)$$

with $\sigma = 0.7$ and $n_1, n_2 \in [-5, 5]$. Then, the detector computes the magnitude spectrum of the residual and averages the frequency contributions lying on a 8×8 or on a 16×16 grid, obtaining a real score that can be used for classifying the images.

Evaluation metrics. We recall that each deep learning-based detector outputs a score s that is thresholded at 0 to distinguish between real ($s \leq 0$) and synthetic ($s > 0$) images. To quantify the effects of peaks removal, we evaluate the percentage of images that are classified as synthetic. Since all images we are working with are either fully synthetic or laundered, we consider all of them as part of the “positive” set, exploiting the True Positive Rate at fixed threshold ($\text{TPR}_{@th}$) as metrics for evaluating the detectors before and after peaks removal. This threshold is equal to 0 for all deep learning-based detectors. For the linear detectors (i.e., that with grid 8 and the one with grid 16), we calibrated them such to obtain a 5% of false alarms over an internal dataset of real images. The resulting threshold has been applied for evaluating them over the testing data.

5. RESULTS

Peaks removal from synthetic images. We start evaluating detectors on the “untouched” synthetic data (i.e., without peaks removal), reporting results in Table 1. All detectors achieve satisfactory performance across all generators, except for detectors [5, 8] which return an average $\text{TPR}_{@th}$ below 70%. Thus, we focus our further analysis on the remaining detectors. Notably, the linear detector achieves remarkably high performance on several generation techniques. While it would likely be easily fooled by simple JPEG compression (as noted earlier), it is striking that averaging contributions at specific frequencies alone produces such strong detection results.

Table 2 reports the relative differences between the $\text{TPR}_{@th}$ obtained under the peaks removal scenario and that of the standard case. To avoid confusion over deep learning-based detectors, we exclude datasets where the initial $\text{TPR}_{@th}$ (i.e., without peaks removal) was below 70%, thereby retaining only scenarios in which these detectors already demonstrated good detection performance.

Interestingly, among the deep learning-based detectors, only the one proposed in [3] appear to be strongly affected by the peaks removal operation, showing an average drop larger than 45% in $\text{TPR}_{@th}$. Moreover, this effect is not consistent across datasets, as

Table 1: TPR_{@th} achieved over uncompressed images of the Wild dataset (without peaks removal). In bold, we highlight average results greater than 70%.

Detector	DALL-E 3	Flux.1	Flux 1.1Pro	Leonardo AI	Midjourney	SD3.5	SDXL	Starry AI	Average
[3]	1.000	0.727	0.932	1.000	0.841	1.000	0.269	1.000	0.846
[5]	0.675	0.131	0.108	0.040	0.046	0.337	0.071	0.257	0.208
[7]	1.000	1.000	1.000	1.000	0.984	1.000	1.000	1.000	0.998
[4]	1.000	0.738	0.975	0.915	0.632	0.997	0.866	0.998	0.890
[8]	0.935	0.755	0.593	0.539	0.395	0.650	0.236	0.878	0.623
[6]	0.963	0.965	0.965	0.986	0.920	0.950	0.699	0.932	0.922
[9]	0.997	0.883	0.993	0.871	0.638	0.947	0.946	0.219	0.812
Linear-8	0.985	0.644	1.000	1.000	1.000	0.594	0.536	0.304	0.859
Linear-16	0.938	0.516	1.000	1.000	0.997	0.974	0.334	0.233	0.750

Table 2: Relative difference of the TPR_{@th} achieved with peaks removal at periodicity 8 (i.e., R_m8) and periodicity 16 (R_m16) with respect to standard conditions, over the Wild synthetic dataset. We do not report results for deep learning-based detectors in which the original TPR_{@th} was below 70%. In bold, we highlight results with absolute value greater than 30%.

Detector	DALL-E 3	Flux.1	Flux 1.1Pro	Leonardo AI	Midjourney	SD3.5	SDXL	Starry AI	Average
	R _m 8/R _m 16								
[3]	-0.00/-0.10	-0.98/-0.92	-0.54/-0.63	+0.00/+0.00	-0.86/-0.98	-0.80/-0.89	--/-	-0.00/-0.05	-0.45/-0.51
[7]	-0.00/-0.01	-0.00/-0.00	+0.00/-0.00	+0.00/+0.00	-0.09/-0.47	-0.00/-0.02	+0.00/-0.01	+0.00/+0.00	-0.01/-0.06
[4]	+0.00/+0.00	-0.46/-0.83	-0.05/-0.19	-0.01/-0.08	--/-	-0.01/-0.14	-0.24/-0.49	-0.01/-0.05	-0.11/-0.26
[6]	+0.01/-0.04	-0.04/-0.14	-0.01/-0.03	+0.00/-0.02	-0.02/-0.16	-0.02/-0.10	--/-	-0.09/-0.15	-0.02/-0.09
[9]	+0.00/-0.00	+0.06/-0.02	+0.00/+0.01	+0.07/+0.06	--/-	+0.02/-0.01	+0.00/-0.02	--/-	+0.02/+0.00
Linear-8	-0.95/-0.94	-0.99/-0.99	-0.19/-0.17	-0.52/-0.51	-0.99/-0.99	-1.00/-1.00	-1.00/-1.00	-1.00/-1.00	-0.83/-0.83
Linear-16	-0.90/-0.99	-0.97/-1.00	+0.00/-0.75	-0.58/-0.99	-0.94/-1.00	-0.150/-1.00	-1.00/-1.00	-0.98/-1.00	-0.69/-0.97

Table 3: Relative difference of the TPR_{@th} achieved with peaks removal at periodicity 8 and 16 with respect to standard conditions, over laundered images. In bold, we highlight results with absolute value greater than 30%.

Detector	SDXL	SD3.5	Flux1	Average
	R _m 8/R _m 16			
[3]	-0.18/-0.37	-0.21/-0.32	-0.49/-0.54	-0.29/-0.41
[5]	+1.27/+1.69	+1.41/+1.90	+1.42/+1.83	+1.36/+1.81
[7]	+0.00/+0.00	-0.14/-0.32	-0.15/-0.23	-0.10/-0.18
[4]	-0.02/-0.04	+0.32/+0.37	+0.59/+0.96	+0.30/+0.43
[8]	+0.02/-0.14	+0.02/+0.11	+0.00/+0.10	+0.01/+0.02
[6]	+0.02/-0.14	+0.02/+0.11	+0.00/+0.10	+0.01/+0.02
[9]	+0.43/+0.25	+0.18/+0.11	+0.42/+0.14	+0.34/+0.17
Linear-8	-0.61/-0.65	-0.77/-0.85	-0.44/-0.47	-0.61/-0.66
Linear-16	-0.61/-0.79	-0.77/-0.93	-0.46/-0.75	-0.61/-0.82

this detector is not affected on DALL-E 3, Leonardo AI, and Starry AI images. A similar dataset-dependent behaviour is observed for other detectors, such as [4, 7], while some detectors like [6, 9] are not affected at all by the peaks removal procedure.

As expected, the linear detectors are strongly impaired by peaks removal, with two notable exceptions: Flux 1.1Pro and SD3.5 generators. For these images, performance consistently drops only with the linear-16 detector under the “R_m16” scenario. We hypothesize these generators carry significant energy contributions at periodicity 16, thus removing peaks with grid 8 is insufficient to degrade performance. This is consistent with the spectra reported in Fig. 1, which clearly show the 16-step periodicity of both generators.

Peaks removal from laundered images. For brevity’s sake, we omit the results obtained on real and laundered images without peak removal, though full details are available in the paper repository¹. Table 3 reports the relative difference in TPR_{@th} between the peak removal scenario and the standard case. Interestingly, detector [3] shows a drop comparable to that observed on synthetic data, while detectors [4, 5, 9] even show an increase in TPR_{@th}. As expected, the linear detectors are far more interpretable, as both exhibit a performance drop consistent with the results previously observed.

Results’ discussion. Overall, the results suggest that deep learning-based detectors do not exhibit a behaviour directly tied to the presence or absence of spectral peaks. The only detector showing an average behaviour that may be linked to these peaks is the one proposed in [3]. Indeed, its performance resembles that of linear de-

tectors which, by design, must experience a performance drop when the spectral energy at the peak positions is reduced to zero.

By contrast, most of the other detectors appear largely unaffected by the presence of peaks. For instance, the detectors proposed in [6, 7, 8] do not seem to rely on this information for detection. While it would be premature to conclude that frequency peaks are irrelevant for detection, this study represents a first step toward a clearer interpretability of results, questioning the assumption (often implicit in prior work) that such artifacts are necessarily exploited by deep learning algorithms, which are inherently black-box models.

One possible explanation for the apparent independence of these detectors from spectral peak energy is that they have been trained to be robust against data compression, which, as previously discussed, is known to introduce a periodicity in the frequency domain. Extensive data augmentation during training may encourage detectors to disregard such artifacts, as these could be easily masked by simple processing operations.

In any case, we find it striking that a simple linear detector can achieve extremely high accuracy. While its performance would inevitably drop under post-processing, we believe it is worth exploring future research on hybrid detectors that can combine the representational power of deep learning with the interpretability of model-based methods.

6. CONCLUSIONS

This work provides new insights into the role of frequency-domain artifacts, particularly spectral peaks, in synthetic image detection. Through systematic removal of these peaks and evaluation across multiple detectors, we find that most state-of-the-art deep learning detectors do not seem to significantly rely on them, challenging a common assumption in the field (frequently implicit in previous studies) that forensic detectors inevitably rely on such artifacts.

At the same time, the impressive performance of a simple linear peak-based detector highlights the potential of interpretable, model-based approaches, paving the way for possible hybrid strategies that can combine the transparency of linear methods with the representational power of deep learning.

7. REFERENCES

- [1] Tao Yang, Rongyuan Wu, Peiran Ren, Xuansong Xie, and Lei Zhang, “Pixel-aware stable diffusion for realistic image super-resolution and personalized stylization,” in *European conference on computer vision*. Springer, 2024, pp. 74–91.
- [2] Irene Amerini, Mauro Barni, Sebastiano Battiatto, Paolo Bestagini, Giulia Boato, Vittoria Bruni, Roberto Caldelli, Francesco De Natale, Rocco De Nicola, Luca Guarnera, et al., “Deepfake media forensics: Status and future challenges,” *Journal of Imaging*, vol. 11, no. 3, pp. 73, 2025.
- [3] Riccardo Corvi, Davide Cozzolino, Giada Zingarini, Giovanni Poggi, Koki Nagano, and Luisa Verdoliva, “On the detection of synthetic images generated by diffusion models,” in *ICASSP 2023-2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2023, pp. 1–5.
- [4] Fabrizio Guillaro, Giada Zingarini, Ben Usman, Avneesh Sud, Davide Cozzolino, and Luisa Verdoliva, “A bias-free training paradigm for more general ai-generated image detection,” in *Proceedings of the Computer Vision and Pattern Recognition Conference*, 2025, pp. 18685–18694.
- [5] Davide Cozzolino, Giovanni Poggi, Riccardo Corvi, Matthias Nießner, and Luisa Verdoliva, “Raising the bar of ai-generated image detection with clip,” in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2024, pp. 4356–4366.
- [6] Zeyang Sha, Zheng Li, Ning Yu, and Yang Zhang, “De-fake: Detection and attribution of fake images generated by text-to-image generation models,” in *Proceedings of the 2023 ACM SIGSAC conference on computer and communications security*, 2023, pp. 3418–3432.
- [7] Sara Mandelli, Paolo Bestagini, and Stefano Tubaro, “When synthetic traces hide real content: Analysis of stable diffusion image laundering,” in *2024 IEEE International Workshop on Information Forensics and Security (WIFS)*. IEEE, 2024, pp. 1–6.
- [8] Lorenzo Baraldi, Federico Cocchi, Marcella Cornia, Lorenzo Baraldi, Alessandro Nicolosi, and Rita Cucchiara, “Contrasting deepfakes diffusion via contrastive learning and global-local similarities,” in *European Conference on Computer Vision*. Springer, 2024, pp. 199–216.
- [9] Christos Koutlis and Symeon Papadopoulos, “Leveraging representations from intermediate encoder-blocks for synthetic image detection,” in *European conference on computer vision*. 2024, vol. 15130, pp. 394–411, Springer.
- [10] Riccardo Corvi, D. Cozzolino, G. Poggi, Koki Nagano, and L. Verdoliva, “Intriguing properties of synthetic images: from generative adversarial networks to diffusion models,” *2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*, pp. 973–982, 2023.
- [11] Quentin Bammey, “Synthbuster: Towards detection of diffusion model generated images,” *IEEE Open Journal of Signal Processing*, vol. 5, pp. 1–9, 2023.
- [12] Davide Alessandro Cocomini, Roberto Caldelli, Claudio Gennaro, Giuseppe Fiameni, Giuseppe Amato, and Fabrizio Falchi, “Deepfake detection without deepfakes: Generalization via synthetic frequency patterns injection,” *arXiv preprint arXiv:2403.13479*, 2024.
- [13] Ricard Durall, Margret Keuper, and Janis Keuper, “Watch your up-convolution: Cnn based generative deep neural networks are failing to reproduce spectral distributions,” in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2020, pp. 7890–7899.
- [14] Kai Zhang, Wangmeng Zuo, Yunjin Chen, Deyu Meng, and Lei Zhang, “Beyond a gaussian denoiser: Residual learning of deep cnn for image denoising,” *IEEE transactions on image processing*, vol. 26, no. 7, pp. 3142–3155, 2017.
- [15] Pietro Bongini, Sara Mandelli, Andrea Montibeller, Mirko Casu, Orazio Pontorno, Claudio Vittorio Ragaglia, Luca Zanchetta, Mattia Aquilina, Taiba Majid Wani, Luca Guarnera, et al., “Wild: a new in-the-wild image linkage dataset for synthetic image attribution,” *arXiv preprint arXiv:2504.19595*, 2025.
- [16] Stability AI, “Stable diffusion 3.5-large,” <https://huggingface.co/stabilityai/stable-diffusion-3.5-large>.
- [17] Black Forest Labs, “FLUX 1.1 [pro]: Advanced Text-to-Image Generation Model,” 2024.
- [18] OpenAI, “Improving image generation with better captions,” <https://cdn.openai.com/papers/dall-e-3.pdf>, 2024.
- [19] Edoardo Daniele Cannas, Sara Mandelli, Nataša Popović, Ayman Alkhateeb, Alessandro Gnutti, Paolo Bestagini, and Stefano Tubaro, “Is jpeg ai going to change image forensics?,” *arXiv preprint arXiv:2412.03261*, 2024.
- [20] Xu Zhang, Svebor Karaman, and Shih-Fu Chang, “Detecting and simulating artifacts in gan fake images,” in *2019 IEEE international workshop on information forensics and security (WIFS)*. IEEE, 2019, pp. 1–6.
- [21] Tarik Dzanic, Karan Shah, and Freddie Witherden, “Fourier spectrum discrepancies in deep network generated images,” *Advances in neural information processing systems*, vol. 33, pp. 3022–3032, 2020.
- [22] Joel Frank, Thorsten Eisenhofer, Lea Schönher, Asja Fischer, Dorothea Kolossa, and Thorsten Holz, “Leveraging frequency analysis for deep fake image recognition,” in *International conference on machine learning*. PMLR, 2020, pp. 3247–3258.
- [23] MidJourney, “MidJourney: An AI-powered image generation tool,” 2024.
- [24] Dustin Podell, Zion English, Kyle Lacey, Andreas Blattmann, Tim Dockhorn, Jonas Müller, Joe Penna, and Robin Rombach, “Sdxl: Improving latent diffusion models for high-resolution image synthesis,” *arXiv preprint arXiv:2307.01952*, 2023.
- [25] Adobe, “Adobe firefly,” <https://firefly.adobe.com/>, 2023.
- [26] Freepik, “Freepik AI Image Generator,” <https://docs.freepik.com/api-reference/mystic/post-mystic>, 2024.
- [27] Duc-Tien Dang-Nguyen, Cecilia Pasquini, Valentina Conotter, and Giulia Boato, “Raise: A raw images dataset for digital image forensics,” in *Proceedings of the 6th ACM multimedia systems conference*, 2015, pp. 219–224.
- [28] Black Forest Labs, “Flux,” <https://github.com/black-forest-labs/flux>, 2024.

- [29] David Marr and Ellen Hildreth, “Theory of edge detection,” *Proceedings of the Royal Society of London. Series B. Biological Sciences*, vol. 207, no. 1167, pp. 187–217, 1980.