Received XX Month, XXXX; revised XX Month, XXXX; accepted XX Month, XXXX; Date of publication XX Month, XXXX; date of current version XX Month, XXXX.

Digital Object Identifier 10.1109/XXXX.2022.1234567

Synthbuster: Towards Detection of Diffusion Model Generated Images

Quentin Bammey[1]

[1]Université Paris-Saclay, ENS Paris-Saclay, CNRS, Centre Borelli

Corresponding author: Quentin Bammey (email: quentin.bammey@ens-paris-saclay.fr).

This work has received funding by the European Union under the Horizon Europe vera.ai project, grant agreement number 101070093, and by the ANR under the APATE project, grant number ANR-22-CE39-0016.
Centre Borelli is also a member of Université Paris Cité, SSA and INSERM.

ABSTRACT Synthetically-generated images are getting increasingly popular. Diffusion models have advanced to the stage where even non-experts can generate photo-realistic images from a simple text prompt. They expand creative horizons but also open a Pandora's box of potential disinformation risks. In this context, the present corpus of synthetic image detection techniques, primarily focusing on older generative models like Generative Adversarial Networks, finds itself ill-equipped to deal with this emerging trend. Recognizing this challenge, we introduce a method specifically designed to detect synthetic images produced by diffusion models. Our approach capitalizes on the inherent frequency artefacts left behind during the diffusion process. Spectral analysis is used to highlight the artefacts in the Fourier transform of a residual image, which are used to distinguish real from fake images. The proposed method can detect diffusion-model-generated images even under mild JPEG compression, and generalizes relatively well to unknown models. By pioneering this novel approach, we aim to fortify forensic methodologies and ignite further research into the detection of AI-generated images.

INDEX TERMS Diffusion models, Image forensics, Media forensics, Multimedia forensics, Spectral analysis, Synthetic image detection

I. Introduction

HOW to assess the validity of an image as a proof to its content? Photographic images used to be considered the most reliable evidence possible, as they were difficult to realistically modify. With the proliferation of digital photography and the development of sophisticated image editing tools, this status of absolute proof is unfortunately long gone. It is increasingly easier to alter an image, not only to make it more aesthetically appealing, but also to change its semantic content and give it a different meaning than the truth.

In the fight against disinformation, the role of image forensics was thus to analyse whether an image was authentic or had been maliciously and locally altered to hide or distort the truth. However, a new source of disinformation has now appeared. Thanks to the advent of diffusion models [40], [48]–[50] and text-to-image joint embeddings, it is now possible and easy to generate images from scratch with nothing more than a text prompt describing the intended image, as seen in Figure 1. Although the resolution of generated images remains limited, these images have achieved a high level of photorealism, that can make them visually indistinguishable from real photographies.

This progress has enabled many innovations, for instance in the arts, to create movies or even in architecture. However, it also brings the risk of people pretending the synthetic images they created is in fact an actual photography representing a real scene, for instance to incriminate or ridicule someone or more globally spread disinformation.

A cardinal question thus arises: how can such images be distinguished from real ones? Until very recently, synthetic images were mainly generated using Generative Adversarial Networks (GANs) [23], [30]–[33]. The methods to detect synthetic images have thus also focused on this architecture, while the literature on detecting images synthetized by those newer diffusion-based methods is still lacking.

It has been noted [21], [25], [38], [55] that GAN-generated images feature frequency artefacts. This is also true, to

This work is licensed under a Creative Commons Attribution 4.0 License. For more information, see https://creativecommons.org/licenses/by/4.0/

This article has been accepted for publication in IEEE Open Journal of Signal Processing. This is the author's version which has not been fully edited and content may change prior to final publication. Citation information: DOI 10.1109/OJSP.2023.3337714

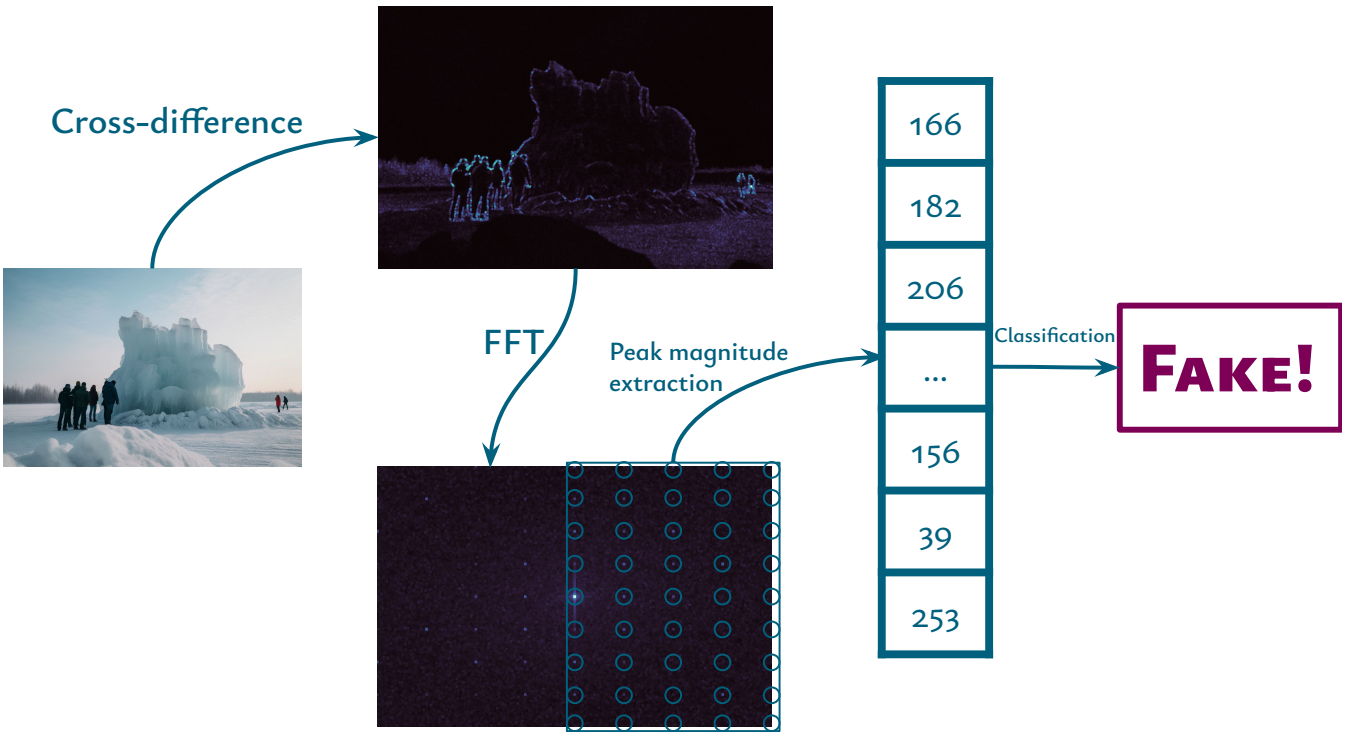


FIGURE 1: The proposed method detects synthetic images generated by diffusion models in the spectral domain. It computes a high-pass residual of a suspect image, and analyses suspected peaks in the Fourier transform of the image to detect whether an image is synthetic or authentic.

some extent, of DM-generated images [14], [15]. Can these artefacts be used to identify synthetic images? Such an enterprise is challenging. These artefacts are subtle and not immediately visible, they must be revealed with suitable filters. While previous work [14], [29] reveal these artefacts, they could only do so by aggregating a large number of images together. To identify whether an image is synthetic, those artefacts must be extracted from a single image, which is a much more challenging undertaking. Furthermore, the frequency artefacts of DM-generated images lie at the same frequency spots as the artefacts caused by a common JPEG compression. It is thus crucial to be able to distinguish the artefacts that come from frequency-based methods from those coming from JPEG compression, lest natural but JPEG-compressed images be mistakenly detected as synthetic.

In this paper, we propose a method based on spectral analysis to detect synthetic images generated by diffusion models. We set up a simple method to highlight and analyse the frequency artefacts in images, distinguishing DM-generated images from authentic ones. Experiments show that the proposed method can reliably detect artefacts even under mild JPEG compression, and distinguish the artefacts caused by compression than those caused by diffusion processes. The method adapts well to unseen architectures, a gap that is yet to be overcome by existing models.

Our main contribution is four-fold:

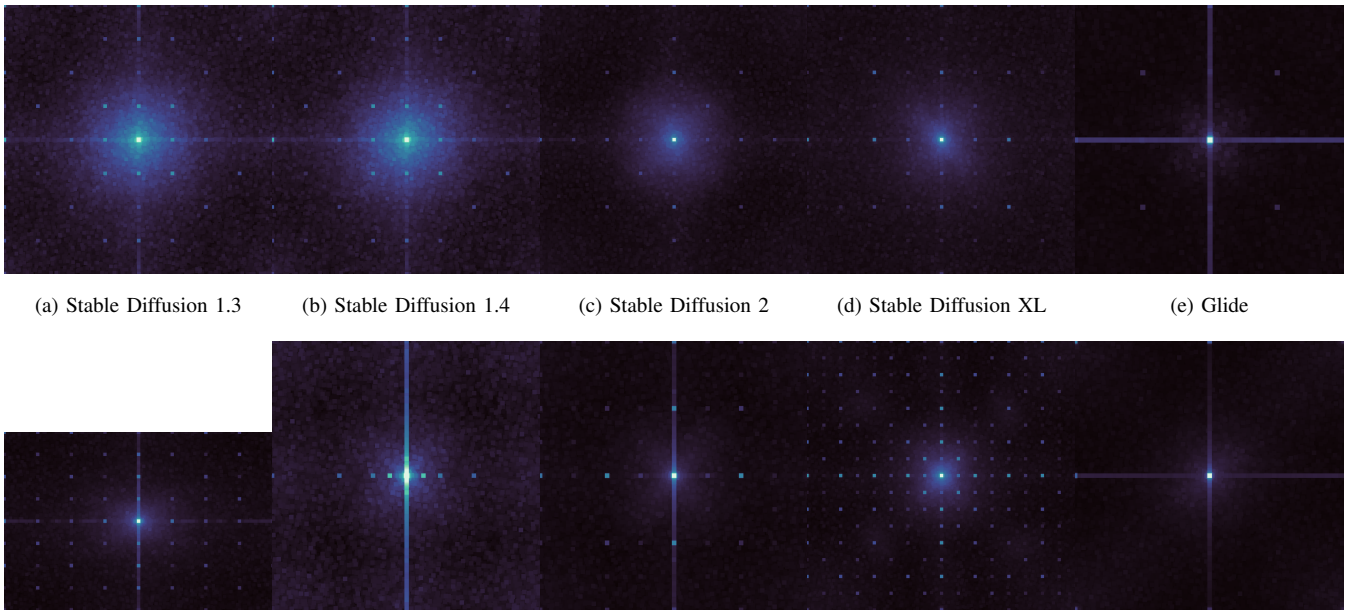
- We show that the cross-difference, a simple high-pass filter, can outperform the state of the art to highlight frequency artefacts in images, to a point they can be detected on individual images,
- Based on this, we introduce a spectral method to detect AI-generated images from diffusion models,
- We design a database of synthetic images to compare the existing methods. The dataset includes the most recent available generation methods to date.
- We study the ability of the proposed method and of the state of the art to distinguish real from fake images against JPEG compression and on unseen models.

II. Related works

A. Synthetic Image Generation

Recently, the domain of image generation has undergone profound transformations, predominantly fueled by the triad of Variational Autoencoders (VAEs), Generative Adversarial Networks (GANs), and Diffusion Models (DMs). These advancements have revolutionized image synthesis, paving the way towards crafting photorealistic synthetic images.

While GANs [28] have deeply influenced the landscape of image generation [9], [32], they have recently been surpassed by diffusion models [51]. These models conceptualize data distribution as a diffusion process, iteratively distorting the image using a simplistic prior and gradually converting it back into the target distribution. Notably, the Ablated Diffusion Model (ADM) [20] has exceeded the capabilities



(a) Stable Diffusion 1.3 (b) Stable Diffusion 1.4 (c) Stable Diffusion 2 (d) Stable Diffusion XL (e) Glide

(f) Midjourney v5 (g) DALL·E 2 (h) DALL·E 3 (i) Adobe Firefly (j) Raise (real images)

FIGURE 2: FFT of the averaged cross-difference of the models and of real images, computed on the proposed database. For a given model, we compute the cross-difference of each of the 1000 images, then average the computed cross-difference as well as the colour channels. We then display the magnitude of the Fourier transform of the averaged result. For better visual legibility at display size, the magnitude is augmented by a morphological dilation, which increases the size of the peaks in the images. For models where images vary in size, only those of the most frequent size are used. We can see that most diffusion models feature traces at periods of 2, 4, and 8. Firefly even features a 16-periodic artifact component, possibly due to a higher number of upsampling steps. Glide images feature fewer, but more visible artefacts, possibly due to the fact that it performs only one small super-resolution step, which is less than the other models. Curiously, DALL·E 2 images only feature artefacts on the horizontal axis of the Fourier transform, hinting at a strongly different treatment of both axes in the weights of the model.

of GANs and VAEs in image generation, marking a vital inflection point in the evolution of diffusion models.

In parallel with diffusion models, Transformer Models [52] have witnessed expanding applications in computer vision, primarily fueled by the advent of CLIP [47], a model adept at embedding images and text into a shared space. Capitalizing on this capability, latent diffusion models [49] such as Stable Diffusion (SD) [50] and DALL·E [48] have extended diffusion models to synthesize images from text prompts in a latent feature space, resulting in a leap forward in the realm of image generation capabilities, both in terms of variety and photorealism.

Nonetheless, the swift advances in image generation have birthed societal apprehensions, primarily the threat of deepfakes, posing substantial security risks. The necessity to devise robust methods for synthetic image detection and potential misuse mitigation cannot be overstated.

B. Synthetic image detection

The central thrust of this paper lies in the authentication of synthetic images, an area where existing literature remains sparse. AutoGAN [55] utilizes a classifier in the spectral domain to identify synthetic images by their frequency artefacts. PatchForensics [10] investigates the unique properties of fake images, particularly face images, that render them detectable and discerns what generalizes across varying model architectures, datasets, and training alterations. McCloskey and Albright [39] take advantage of the fact that the intensity values of synthetic images are rarely saturated, while Wang et al. [53] and Gragnaniello et al. [29] train CNNs to differentiate real and GAN-generated images. However, these studies largely predate the prevalence of diffusion models and text-to-image techniques, hence they are primarily trained and evaluated on GAN-generated images sampled from specific classes. Two methods have already been proposed to detect DM-generated images. Corvi et al. [15] retrains the existing architecture of Gragnaniello et al. [29] on DM-generated images, while Ojha et al. [46] train a network to distinguish real and fake images in the latent domain of a CLIP-trained architecture [22]. However, neither methods achieve good generalizability against methods unseen during training.

III. Proposed method

This article has been accepted for publication in IEEE Open Journal of Signal Processing. This is the author's version which has not been fully edited and content may change prior to final publication. Citation information: DOI 10.1109/OJSP.2023.3337714

FIGURE 3: Examples of the generated images in the database. The images are generated with different diffusion models using a text prompt, that loosely based on a natural image from Raise-1k [17]. As the goal of the database is to evaluate methods that can distinguish natural photographies from synthetic images, attention is paid in the prompts to generate images with photorealistic styles and textures rather than artistic styles.

AS seen in Figure 1, the frequency artefacts from DM-generated images lie at very specific frequencies, corresponding to components of periods 2, 4, and 8. We propose to use a cross-difference filter on the image to highlight the frequency artefacts in an image, and extract the magnitude of the points corresponding to components of periods 0, 2, 4, or 8, in both directions. A simple classifier is then trained to distinguish real from generated images.

A. High-pass residual to reveal the artefacts

The cross-difference filter [11] has been introduced and used to reveal periodic artefacts coming from JPEG compression [11], [43] and image demosaicing [4]. The cross-difference is an high-pass filter defined as the absolute difference between the two diagonals of a 2×2 block on an image. Let I be a 2-dimensional image, the cross-difference at location (x, y) is defined as

$$C_{x,y} = |I_{x,y} + I_{x+1,y+1} - I_{x,y+1} - I_{x+1,y}| . \quad (1)$$

We propose to use the cross-difference to dampen the low frequencies and highlight the frequency artefacts from DM-generated images, shown in Figure 2 For each colour channel of the image to analyse, the cross-difference filter defined in Equation 1 is used to extract a simple fingerprint of the image. As the cross-difference acts as an high-pass filter, the high-frequency artefacts we expect to find in synthetic images are much more prominent on the cross-difference than on the original image.

The Fast Fourier Transform (FFT) of the cross-difference is then computed. To avoid any bias linked to the image size, the FFT is normalized by the size. Peaks representative of DM-generated images occur on components of period 0, 2, 4, and 8, in both directions. We extract the magnitude of the 45 peaks from each of the three colour channels, leading to 135 extracted magnitudes.

B. Analysis of the extracted peaks

Using only the 135 potential magnitude peaks as features, we then train a classifier to distinguish real from fake images. We use a histogram-based gradient boosting tree classifier (HBGB) [34], [35]. This variant of the traditional Gradient Boosting Trees leverages the concept of gradient boosting with an histogram-based approach to accelerate the tree-growing process. The algorithm maintains the robustness of gradient boosting, while the histogram-based technique enhances its scalability, making it suitable for large-scale datasets. It is able to handle the dimensionality of the data, as 135 features are to consider, and can maintain a high accuracy. Although it was historically shadowed by neural network, this much simpler model is sufficient as we only use a relatively small number of features. Trained with both real and synthetic images from different diffusion models, our classifier learns to distinguish authentic and generated images using only the magnitude of these peaks.

The model is trained on both natural images and diffusion-model generated ones, to detect whether the analysed features correspond to a natural or synthetic images. Different training schemes are presented and discussed in the Experiments section.

C. Robustness to JPEG compression

The proposed method analyses FFT peaks corresponding to periods 0, 2, 4, and 8, in both directions. However, JPEG compression also leaves strong artefacts in these periods [2], [7], [8], [44]. To train the network to only detect artefacts coming from diffusion methods, we apply JPEG compression to the training images.

One model is trained for each JPEG quality factor, as well as without compression. At inference, the JPEG potential quantization table of the tested image is estimated using a quantization table estimator [45], and the appropriate model is selected, a strategy that has already proved its efficiency in the forensic literature [16].

IV. Database

TRAINING and evaluating the proposed method requires sets of real and fake images. Real images are plentiful. In particular, the Raise dataset [17] and the Dresden dataset [27] contain 8156 and 1488 uncompressed photographs. Using uncompressed images is particularly useful, as we can then apply various post-processing such as JPEG compression on a clean image.

On the other hand, as diffusion models are quite recent, the available data on such images is scarce. To the best of our knowledge, the only such published database is proposed by Corvi et al. [15], consisting of 1000 images generated with different GAN and diffusion models, including DALL·E 2 [48], Glide [41], and Latent Diffusion [50].

To address this scarcity, we propose our own dataset of DM-generated images. This enables us to provide a way to evaluate methods on current diffusion models, such as Stable Diffusion [50] 1.3, 1.4, 2, and XL, Midjourney [40], Adobe Firefly [24],DALL·E [48] 2 and 3, for which no publicly-available datasets are available yet. This newly-constructed dataset is also useful to train and test models on independently-generated data, ensuring a fair evaluation.

While the synthetic images are generated from a text prompt, we use an existing database of real image as guideline for the generated image, the Raise-1k dataset, which is a varied subset of the full Raise [17] dataset. This dataset contains one thousand high-quality, uncompressed photographies of diverse categories: indoor, outdoor, landscape, nature, people, objects, and buildings. While using an existing dataset of natural images is not strictly needed, it provides several advantages:

1) Being of the same categories, the natural images themselves provide a fair comparison point for the methods to check both their ability to detect fake images and to avoid false positives,
2) As already established in the literature [21], it is crucial to evaluate synthetic image detection methods on

	Glide	SD1.3	SD1.4	SD2	SD XL	Midjourney	DALL·E 2	DALL·E 3	Firefly	Overall
Proposed, generic	0.915	0.933	0.943	0.866	0.915	**0.867**	0.932	0.920	**0.913**	0.872
Proposed, specific	**0.944**	**0.969**	**0.971**	*0.860*	**0.956**	0.852	**0.972**	**0.932**	0.728	**0.953**
Proposed, generalization	*0.827*	0.833	0.830	0.702	0.592	*0.793*	*0.478*	*0.845*	*0.700*	*0.527*
UFD [46]	0.101	0.243	0.218	0.344	0.215	0.000	0.424	0.000	0.617	0.143
Wang et al. [53]	0.052	0.000	0.000	0.031	0.000	0.000	0.000	0.000	0.390	0.004
Corvi et al. [15]	0.000	*0.923*	*0.933*	**0.889**	0.730	0.701	0.000	0.000	0.122	0.250
Grag. et al [29]	0.000	0.048	0.039	0.000	0.000	0.022	0.263	0.000	0.447	0.000
PatchFor [10]	0.016	0.193	0.184	0.357	0.185	0.207	0.114	0.327	0.113	0.121
Mandelli et al. [37]	0.612	0.734	0.745	0.614	0.507	0.542	0.372	0.490	0.758	0.449

TABLE 1: Comparative results of the state of the art and the proposed method on Glide (GD) [41], stable diffusion (SD) [50] 1.3 and 1.4, Midjourney(MJ) v5 [40], and DALL·E 2(DE) [48]. The proposed method is displayed when the diffusion model of the studied images was encountered during training (generic), when the detection method is trained specifically on this model (specific), and when the method is **not** trained on the image diffusion model (generalization). The Matthew's Correlation Coefficient MCC score is displayed, by setting the optimal threshold per dataset for each method. The MCC is widely regarded as the most representative single metric on detection scores [12]. The **best**, second and *third-best* results are highlighted. As can be seen the method consistently yields good detection scores across all models, and displays good generalization ability.

varied image classes. Using an already-diverse dataset as a guideline ensures the generated images are varied.

Note that the original images are not used as image prompts to try to recreate a similar image or modify it. The original images are only used as a guideline to create the new text prompt of the presentation, to ensure the resulting image broadly belongs to the same category as the original one.

For each of the 1000 images, descriptions of the images are generated using Midjourney descriptor [40] and CLIP Interrogator [13]. These descriptions are used as a basis to manually write a text prompt to generate a photo-realistic image loosely based on the original image. The objective is not to recreate a perfectly similar image, but rather to obtain an image from the same category, so as to keep the variety of the images.

The parameters that are used to guide the methods are selected randomly, within reasonable bounds.

V. Experiments

We now have a database of synthetic and authentic images tailored to evaluating methods, as the synthetic images are matched with real images from the Raise [17] database. We train our model on a separate fake images dataset [1] and on real images from the Dresden database [27], guaranteeing a fair evaluation on a challenging case where fake, but also real images from the training [27] and testing [17] datasets are wildly different.

For evaluation, we compare our results to the state of the art on the proposed database, naturally combined with the raise-1k [17] real images on which the dataset is based. Three scenarii are initially considered for training, to show potential results on the method depending on whether the tested synthetic image is generated by a model seen during training (generic), if the diffusion model is exactly known (specific) or in the worst case where the diffusion model is entirely unseen during training (generalization):

1) **Generic** training: the proposed method can be trained generically on images coming from all known diffusion models in the augmented Corvi et al. database. This is the most realistic case, as fake images for disinformation are usually created with fake images from existing, publicly-available diffusion model, but it is rarely known specifically which model was used. The generic-trained method constitutes the final method proposed in this paper, whereas the specific and generalization scenarii should be viewed as experiments to test the strengths and limits of the method.
2) **Specific** training: the method can be trained specifically on the diffusion model used for the images. While this approach can be seen as unrealistic, it enables us to know the limits of the method in an ideal case where it the exact model used to generate an image is known.
3) **Generalization** training: Reversely, the method can be trained on all known diffusion models, except the one used to generate the image, to assess whether the method can generalize to unknown models.

Results of this experiment are reported in Figure 4 and in Table 1. Under the generic training, the proposed method yields consistently good results across diffusion models and beats the state of the art on all, even against stable diffusion images which are already well-detected by Corvi et al. [15]. Knowing the specific model used is shown to slightly enhance the results, although this is only significant against Midjourney [40] and DALL·E 2 images. The model also shows great generalization ability, although the results

	Glide	SD1.3	SD1.4	SD2	SD XL	Midjourney	DALL·E 2	DALL·E 3	Firefly	Overall
Uncompressed	0.915	0.933	0.943	0.866	0.915	0.867	0.932	0.920	0.913	0.872
JPEG $Q = 95$	0.761	0.913	0.910	0.807	0.680	0.769	0.749	0.879	0.729	0.708
JPEG $Q = 90$	0.699	0.903	0.894	0.804	0.684	0.753	0.749	0.848	0.534	0.609
JPEG $Q = 80$	0.580	0.913	0.904	0.823	0.696	0.716	0.653	0.853	0.502	0.583
JPEG $Q = 70$	0.579	0.915	0.904	0.806	0.687	0.707	0.610	0.848	0.524	0.598

TABLE 2: Study of the robustness of the proposed method (generic training) against JPEG compression on the different models. The model is trained at different quality factors. At inference, the image JPEG quantization matrix is estimated to select the appropriate model. The Matthew's Correlation Coefficient MCC score is displayed, by setting the optimal threshold per dataset. DM artefacts and JPEG compression artefacts lie at the very same frequencies, rendering synthetic images detection difficult against JPEG compression. Despite that, the model still shows robustness against JPEG compression, even at a $Q = 70$ quality factor.

	Glide	SD1.3	SD1.4	SD2	SD XL	Midjourney	DALL·E 2	DALL·E 3	Firefly	Overall
Proposed (with Cross-difference)	0.915	0.933	0.943	0.866	0.915	0.867	0.932	0.920	0.913	0.872
Ablated (with DnCNN)	0.153	0.863	0.851	0.735	0.802	0.631	0.519	0.620	0.412	

TABLE 3: Ablation of the proposed method (generic training), with the cross-difference and the DnCNN denoiser proposed in existing works [14], [21], [38], which used DnCNN [54] to reveal frequency artefacts on synthetic images, but had to aggregate the results over a large number of images. We instead use a cross-difference filter, which can reveal artefacts on single images and yields much better results with our method.

are expectedly worse than when the model has been seen during training. Generalization results are significantly worse on Midjourney images, and even more so on DALL·E 2 images, suggesting these models architectures are dissimilar to the other known ones. We also note that, surprisingly and seemingly inexplicably, generalization results against Glide images are better than results when the model belongs to the training set.

A. Robustness to JPEG compression

It was stated earlier that DM artefacts and JPEG compression artefacts lie at the very same frequency, potentially rendering their distinction difficult. To assess this, we test the proposed model on images at different JPEG compression levels, as seen in Table 2.

The test images, both real and synthetic, are JPEG-compressed at the mentioned quality factor. As can be seen, the model is very robust even against mild JPEG compression, and can still distinguish real from fake images even at JPEG quality 70, albeit with reduced performance.

B. Ablation study

Frequency artefacts on synthetic images were previously highlighted using denoising with DnCNN [14], [15], [21], [54]. However, this was only be performed by aggregating numerous images to reveal the artefacts, rather than on a single image. We propose the use of a cross-difference filter, that can highlight the frequency artefacts on single images. Table 3 shows that this filter indeed improves performance over using DnCNN denoising.

VI. Discussion and limitations

DESPITE its simplicity, the proposed method is indeed able to detect synthetic images better than the existing state of the art. It shows some generalization ability, as well as robustness to JPEG compression. Despite that, those two points remain an important challenge. Indeed, while the proposed method performs better than the existing ones both against JPEG images and on unseen architectures, false positives are still impossible to avoid in these complex situations. Yet, simple Bayesian reasoning shows that even a small number of false positives can be sufficient to drown true detection from false alarms, due to the high proportion of authentic images in the wild. In addition, wrongly accusing someone of fraud can have disastrous consequences. As a consequence, current synthetic image detection methods, including the proposed one, should still be considered a research artefact, and not be used as proof that an image is actually forged. For practical usability, setting an automatic threshold would be crucial, for instance with *a contrario* analysis [18], [19], a promising approach in forensics [2]–[6], [26], [36], [42], [44]

Finally, we note that the proposed method is trained on **diffusion-model-generated**, **photorealistic** images. It is not trained to work on GAN images, for which numerous tools already exists. Given that the frequency artefacts are usually stronger on GAN images than on DM images, it would be easy to adapt the proposed method to GANs should the need arise. We also note that our method has only been tested on photo-realistic images; it remains untested, and thus not suited for, digital art examination. Indeed, not only

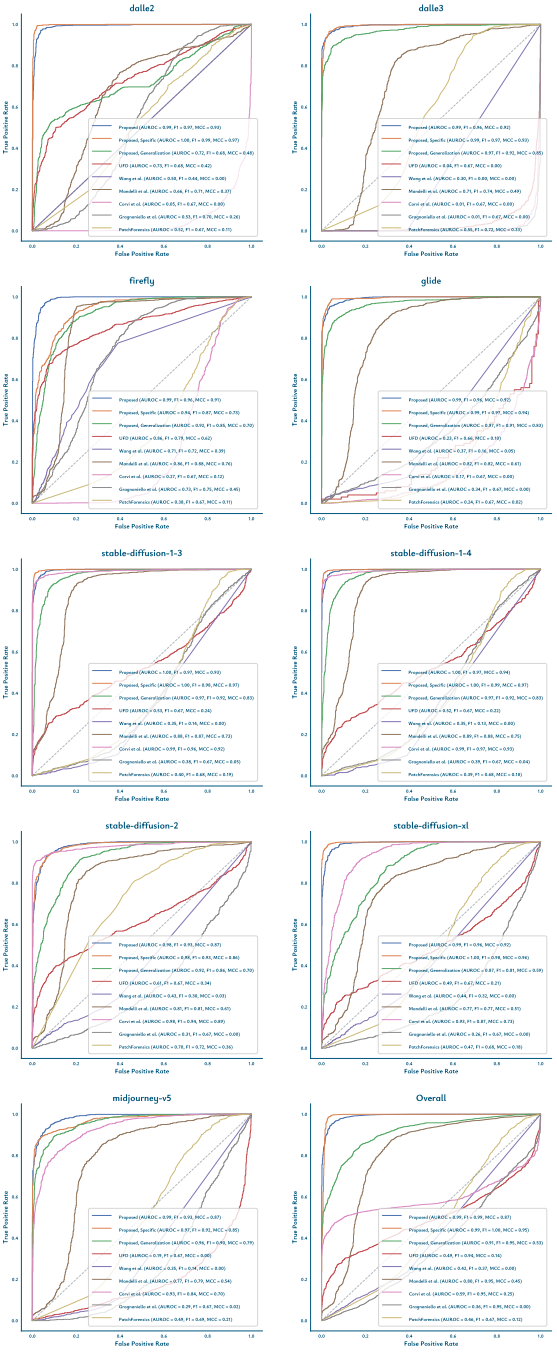


FIGURE 4: Comparative ROC curves of the proposed method with the existing state of the art, on the different detection models. The F1 and MCC scores are computed using the standard thresholds (0 or 0.5 depending on the method). We can see that the proposed method consistently get excellent results, even when not trained on the specific model to be tested. It thus shows some decent generalization ability, except on the DALL·E 2 and Firefly models, which seem to yield slightly different artefacts.

is the method not trained on such images, it is likely they would be more challenging, as digital art usually present flatter textures than natural photographies, and thus fewer opportunities for frequency artefacts to be revealed.

VII. Conclusion

IN this paper, we have trained a simple method to detect synthetic images generated by diffusion models. The method reveals the frequency artefacts using an high pass filter, then distinguishes real and fake images using the presence of these artefacts with a simple classifier on the FFT magnitude peaks.

This method performs well even in difficult situations such as JPEG compression and unseen models. Still, the risk of false positives and their consequences should before all draw future work into preventing and controlling the risk of false alarms.

REFERENCES

[1] Quentin Bammey. Positional learning for reliable ai-generated images detection. https://github.com/qbammey/polardiffshield.
[2] Quentin Bammey. A contrario mosaic analysis for image forensics. In *Advanced Concepts for Intelligent Vision Systems (ACIVS)*. Springer, Aug. 2023.
[3] Quentin Bammey. Jade owl: Jpeg 2000 forensics by wavelet offset consistency analysis. In *8th International Conference on Image, Vision and Computing (ICIVC)*. IEEE, 2023.
[4] Quentin Bammey, Rafael Grompone von Gioi, and Jean-Michel Morel. Reliable demosaicing detection for image forensics. In *2019 27th European Signal Processing Conference (EUSIPCO)*, pages 1–5, 2019.
[5] Quentin Bammey, Tina Nikoukhah, Marina Gardella, Rafael Grompone von Gioi, Miguel Colom, and Jean-Michel Morel. Non-semantic evaluation of image forensics tools: Methodology and database. In *WACV*, 2022.
[6] Quentin Bammey, Rafael Grompone von Gioi, and Jean-Michel Morel. Automatic detection of demosaicing image artifacts and its use in tampering detection. In *2018 IEEE Conference on Multimedia Information Processing and Retrieval (MIPR)*, pages 424–429. IEEE, 2018.
[7] Quentin Bammey, Rafael Grompone von Gioi, and Jean-Michel Morel. An adaptive neural network for unsupervised mosaic consistency analysis in image forensics. In *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 14182–14192, 2020.
[8] Quentin Bammey, Rafael Grompone von Gioi, and Jean-Michel Morel. Forgery detection by internal positional learning of demosaicing traces. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV)*, pages 328–338, January 2022.
[9] Andrew Brock, Jeff Donahue, and Karen Simonyan. Large scale gan training for high fidelity natural image synthesis, 2019.
[10] Lucy Chai, David Bau, Ser-Nam Lim, and Phillip Isola. What makes fake images detectable? understanding properties that generalize. In *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part XXVI 16*, pages 103–120. Springer, 2020.
[11] Yi-Lei Chen and Chiou-Ting Hsu. Image tampering detection by blocking periodicity analysis in jpeg compressed images. In *2008 IEEE 10th Workshop on Multimedia Signal Processing*, pages 803–808. IEEE, 2008.
[12] Davide Chicco and Giuseppe Jurman. The advantages of the matthews correlation coefficient (mcc) over f1 score and accuracy in binary classification evaluation. *BMC genomics*, 2020.
[13] clip-interrogator, 2022.
[14] Riccardo Corvi, Davide Cozzolino, Giovanni Poggi, Koki Nagano, and Luisa Verdoliva. Intriguing properties of synthetic images: from generative adversarial networks to diffusion models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 973–982, 2023.
[15] 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), pages 1–5. IEEE, 2023.

[16] Davide Cozzolino and Luisa Verdoliva. Noiseprint: A cnn-based camera model fingerprint. *IEEE Transactions on Information Forensics and Security*, 15:144–159, 2019.

[17] 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*, pages 219–224, 2015.

[18] A Desolneux, L Moisan, and JM Morel. From gestalt theory to image analysis. interdisciplinary applied mathematics, vol. 35, 2007.

[19] Agnes Desolneux, Lionel Moisan, and Jean-Michel Morel. Meaningful alignments. *IJCV*, 2000.

[20] Prafulla Dhariwal and Alexander Nichol. Diffusion models beat gans on image synthesis. *Advances in neural information processing systems*, 34:8780–8794, 2021.

[21] Pantelis Dogoulis, Giorgos Kordopatis-Zilos, Ioannis Kompatsiaris, and Symeon Papadopoulos. Improving synthetically generated image detection in cross-concept settings. In *Proceedings of the 2nd ACM International Workshop on Multimedia AI against Disinformation*, MAD '23, page 28–35, New York, NY, USA, 2023. Association for Computing Machinery.

[22] Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, et al. An image is worth 16x16 words: Transformers for image recognition at scale. *arXiv preprint arXiv:2010.11929*, 2020.

[23] Patrick Esser, Robin Rombach, and Bjorn Ommer. Taming transformers for high-resolution image synthesis. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 12873–12883, June 2021.

[24] Adobe Firefly. https://firefly.adobe.com/.

[25] Joel Frank, Thorsten Eisenhofer, Lea Schönherr, Asja Fischer, Dorothea Kolossa, and Thorsten Holz. Leveraging frequency analysis for deep fake image recognition. In *International conference on machine learning*, pages 3247–3258. PMLR, 2020.

[26] Marina Gardella, Pablo Musé, Jean-Michel Morel, and Miguel Colom. Noisesniffer: a fully automatic image forgery detector based on noise analysis. In *IWBF*. IEEE, 2021.

[27] Thomas Gloe and Rainer Böhme. The'dresden image database'for benchmarking digital image forensics. In *Proceedings of the 2010 ACM symposium on applied computing*, pages 1584–1590, 2010.

[28] Ian Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. Generative adversarial nets. In Z. Ghahramani, M. Welling, C. Cortes, N. Lawrence, and K.Q. Weinberger, editors, *Advances in Neural Information Processing Systems*, volume 27. Curran Associates, Inc., 2014.

[29] D. Gragnaniello, D. Cozzolino, F. Marra, G. Poggi, and L. Verdoliva. Are gan generated images easy to detect? a critical analysis of the state-of-the-art. In *2021 IEEE International Conference on Multimedia and Expo (ICME)*, pages 1–6, 2021.

[30] Tero Karras, Timo Aila, Samuli Laine, and Jaakko Lehtinen. Progressive growing of gans for improved quality, stability, and variation, 2018.

[31] Tero Karras, Miika Aittala, Samuli Laine, Erik Härkönen, Janne Hellsten, Jaakko Lehtinen, and Timo Aila. Alias-free generative adversarial networks. In M. Ranzato, A. Beygelzimer, Y. Dauphin, P.S. Liang, and J. Wortman Vaughan, editors, *Advances in Neural Information Processing Systems*, volume 34, pages 852–863. Curran Associates, Inc., 2021.

[32] Tero Karras, Samuli Laine, and Timo Aila. A style-based generator architecture for generative adversarial networks. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2019.

[33] Tero Karras, Samuli Laine, Miika Aittala, Janne Hellsten, Jaakko Lehtinen, and Timo Aila. Analyzing and improving the image quality of stylegan. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2020.

[34] Guolin Ke, Qi Meng, Thomas Finley, Taifeng Wang, Wei Chen, Weidong Ma, Qiwei Ye, and Tie-Yan Liu. Lightgbm: A highly efficient gradient boosting decision tree. *Advances in neural information processing systems*, 30, 2017.

[35] Oliver Kramer and Oliver Kramer. Scikit-learn. *Machine learning for evolution strategies*, pages 45–53, 2016.

[36] Yanhao Li, Marina Gardella, Quentin Bammey, Tina Nikoukhah, Jean-Michel Morel, Miguel Colom, and Rafael Grompone von Gioi. A contrario detection of h. 264 video double compression. In *2023 IEEE International Conference on Image Processing (ICIP)*, pages 1765–1769. IEEE, 2023.

[37] Sara Mandelli, Nicolò Bonettini, Paolo Bestagini, and Stefano Tubaro. Detecting gan-generated images by orthogonal training of multiple cnns. In *2022 IEEE International Conference on Image Processing (ICIP)*, pages 3091–3095, 2022.

[38] Francesco Marra, Diego Gragnaniello, Luisa Verdoliva, and Giovanni Poggi. Do gans leave artificial fingerprints? In *2019 IEEE conference on multimedia information processing and retrieval (MIPR)*, pages 506–511. IEEE, 2019.

[39] Scott McCloskey and Michael Albright. Detecting gan-generated imagery using saturation cues. In *2019 IEEE international conference on image processing (ICIP)*, pages 4584–4588. IEEE, 2019.

[40] Midjourney. Midjourney v5, 2023.

[41] Alex Nichol, Prafulla Dhariwal, Aditya Ramesh, Pranav Shyam, Pamela Mishkin, Bob McGrew, Ilya Sutskever, and Mark Chen. Glide: Towards photorealistic image generation and editing with text-guided diffusion models, 2022.

[42] Tina Nikoukhah, Jérémy Anger, Thibaud Ehret, Miguel Colom, Jean-Michel Morel, and Rafael Grompone von Gioi. JPEG grid detection based on the number of DCT zeros and its application to automatic and localized forgery detection. In *CVPRW*, 2019.

[43] Tina Nikoukhah, Miguel Colom, Jean-Michel Morel, and Rafael Grompone von Gioi. Local JPEG Grid Detector via Blocking Artifacts, a Forgery Detection Tool. *Image Processing On Line*, 10:24–42, 2020.

[44] Tina Nikoukhah, Miguel Colom, Jean-Michel Morel, and Rafael Grompone von Gioi. Local JPEG Grid Detector via Blocking Artifacts, a Forgery Detection Tool. *IPOL*, 10, 2020.

[45] Tina Nikoukhah, Miguel Colom, Jean-Michel Morel, and Rafael Grompone von Gioi. A Reliable JPEG Quantization Table Estimator. *Image Processing On Line*, 12:173–197, 2022. https://doi.org/10.5201/ipol.2022.399.

[46] Utkarsh Ojha, Yuheng Li, and Yong Jae Lee. Towards universal fake image detectors that generalize across generative models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 24480–24489, June 2023.

[47] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, Gretchen Krueger, and Ilya Sutskever. Learning transferable visual models from natural language supervision, 2021.

[48] Aditya Ramesh, Prafulla Dhariwal, Alex Nichol, Casey Chu, and Mark Chen. Hierarchical text-conditional image generation with clip latents, 2022.

[49] Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. High-resolution image synthesis with latent diffusion models, 2021.

[50] Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. Stable diffusion, 2022.

[51] Jascha Sohl-Dickstein, Eric Weiss, Niru Maheswaranathan, and Surya Ganguli. Deep unsupervised learning using nonequilibrium thermodynamics. In Francis Bach and David Blei, editors, *Proceedings of the 32nd International Conference on Machine Learning*, volume 37 of *Proceedings of Machine Learning Research*, pages 2256–2265, Lille, France, 07–09 Jul 2015. PMLR.

[52] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Ł ukasz Kaiser, and Illia Polosukhin. Attention is all you need. In I. Guyon, U. Von Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, and R. Garnett, editors, *Advances in Neural Information Processing Systems*, volume 30. Curran Associates, Inc., 2017.

[53] Sheng-Yu Wang, Oliver Wang, Richard Zhang, Andrew Owens, and Alexei A. Efros. Cnn-generated images are surprisingly easy to spot... for now. In *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 8692–8701, 2020.

[54] 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*, 26(7):3142–3155, 2017.

[55] 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)*, pages 1–6, 2019.