Intriguing properties of synthetic images:
from generative adversarial networks to diffusion models

Riccardo Corvi[1]   Davide Cozzolino[1]   Giovanni Poggi[1]   Koki Nagano[2]   Luisa Verdoliva[1]

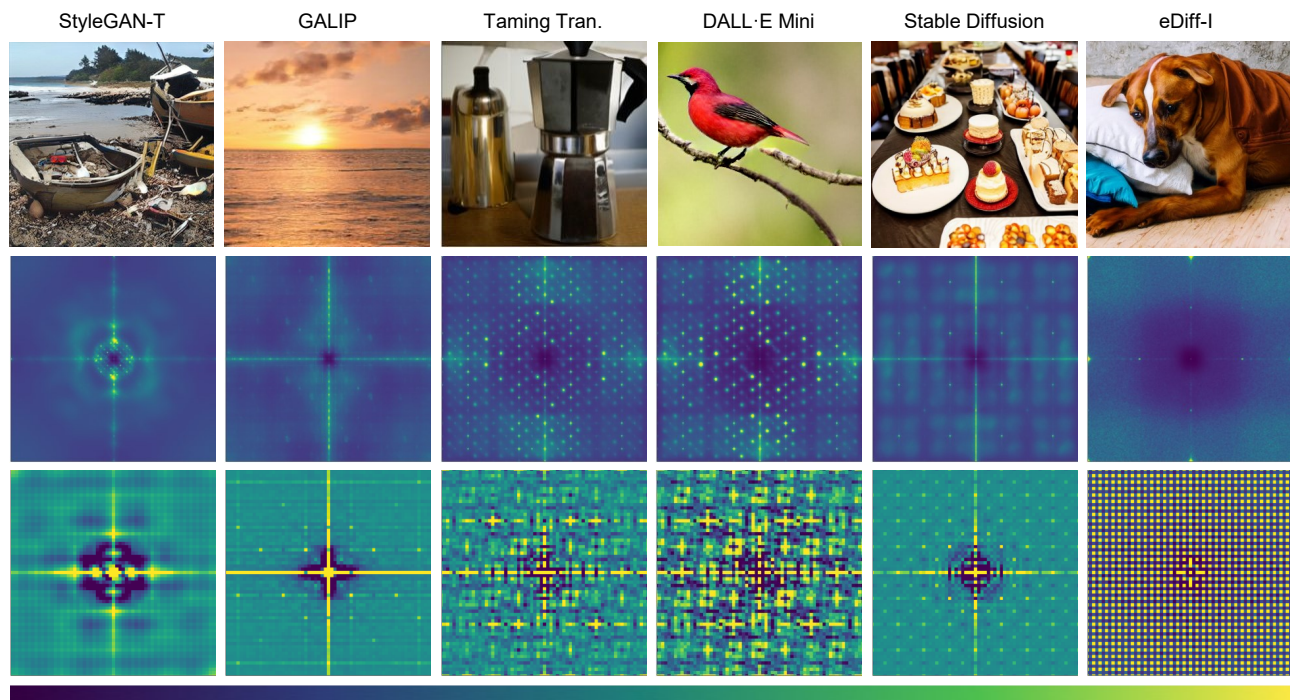[1]University Federico II of Naples   [2] NVIDIA

Figure 1. Examples of synthetic images (top) generated by state-of-the-art models: StyleGAN-T [44], GALIP [51], Taming Transformers [14], DALL· E Mini [9], Stable Diffusion [43], eDiff-I [2]. Each model leaves peculiar traces in the generated images, which can be regarded as a sort of artificial fingerprint and used for forensic analyses. They are typically visible in the frequency domain as spectral peaks in the power spectra (middle), or in the spatial domain as anomalous regular patterns in the autocorrelation (bottom). It is easy to observe that models based on the same architecture, like Taming Transformers and DALL· E Mini, give rise to similar artifacts.

Abstract

Detecting fake images is becoming a major goal of computer vision. This need is becoming more and more pressing with the continuous improvement of synthesis methods based on Generative Adversarial Networks (GAN), and even more with the appearance of powerful methods based on Diffusion Models (DM). Towards this end, it is important to gain insight into which image features better discriminate fake images from real ones. In this paper we report on our systematic study of a large number of image generators of different families, aimed at discovering the most forensically relevant characteristics of real and generated images. Our experiments provide a number of interesting observations and shed light on some intriguing properties of synthetic images: (1) not only the GAN models but also the DM and VQ-GAN (Vector Quantized Generative Adversarial Networks) models give rise to visible artifacts in the Fourier domain and exhibit anomalous regular patterns in the autocorrelation; (2) when the dataset used to train the model lacks sufficient variety, its biases can be transferred to the generated images; (3) synthetic and real images exhibit significant differences in the mid-high frequency signal content, observable in their radial and angular spectral power distributions.
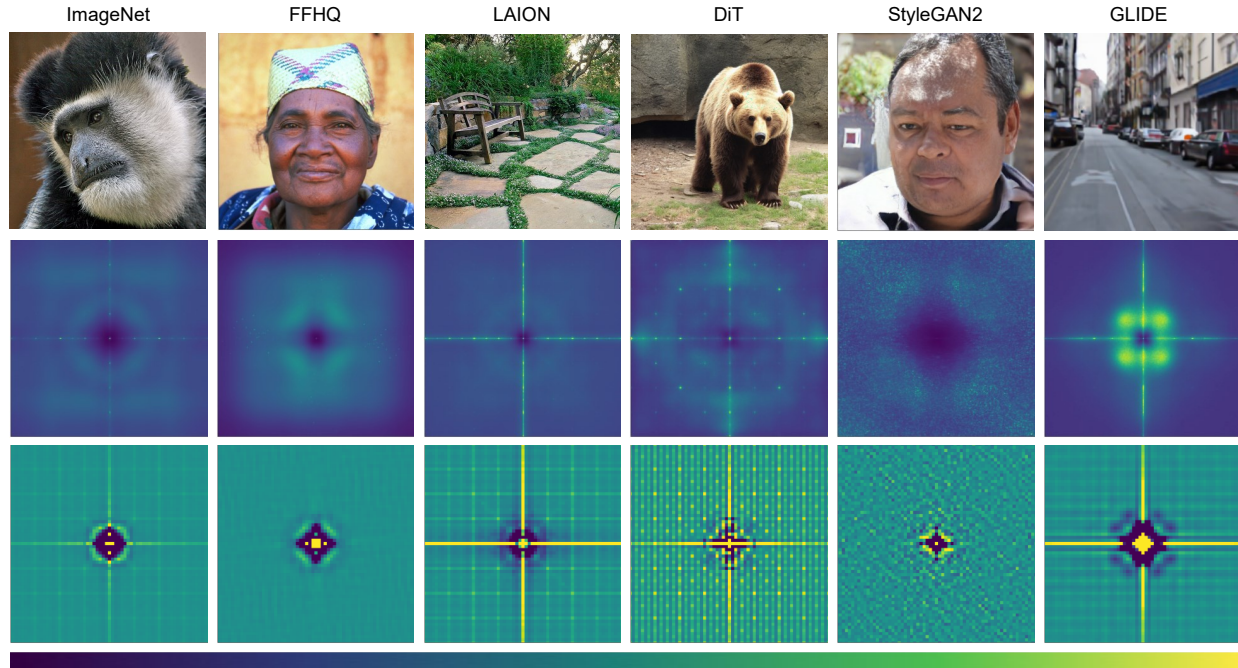
Figure 2. Examples of real and synthetic images drawn from popular datasets. Real datasets (left): Imagenet [10], FFHQ [27], LAION [45]. Synthetic datasets (right): DiT [38], StyleGAN2 [28], GLIDE [35]. For each dataset, we also show the average power spectra (middle) and autocorrelation (bottom) of noise residuals. It is possible to observe that real images embed some traces of JPEG artifacts (see also Section 4.2) while all generated images exhibit anomalous periodic patterns in the spatial domain.

1. Introduction

In recent years, deep learning-based generated images have gained immense popularity in industries such as entertainment, advertising, and design. These are high-quality images of impressive photorealism [29, 36]. Especially the recent text-image generators, which rely on text descriptions or preliminary sketches as a guide, can allow image creation across countless categories [2, 9, 23, 35, 40, 44, 51]. While this technology offers tremendous value to artists, game designers, and creative individuals, it also comes with risks. The potential security and ethical issues that can arise from the misuse of these images cannot be ignored.

Research on synthetic image detection has received a huge boost in recent years [31, 34], especially since the first realistic images generated by Generative Adversarial Networks [24, 60] appeared. Some works focus mainly on the inability of such generators to perfectly reproduce the high-level semantic features of natural images. In fact, they can produce visible artifacts, such as chromatic anomalies or lack of natural symmetries. For example, faces generated by GANs may have differently colored left and right eyes, unconvincing eye reflexes, or irregular pupil shapes [20, 33]. Similarly, some recent studies on images generated by diffusion models [15, 16] pointed out that the absence of explicit 3D modeling for objects and surfaces results in asym-

metries in shadows and reflections. In addition, a global inconsistency in lighting can be observed to some extent.

However, the generation methods keep improving over time and show fewer and fewer synthesis artifacts. Therefore, other approaches focus on low-level features that are related to the generative model architecture and can be exposed by removing the high-level semantic content of the captured scene [32, 56]. In fact, the complex processing pipeline needed to generate synthetic images inevitably leads to the introduction of digital artifacts that characterize the specific architecture and significantly differ from those typical of traditional acquisition devices. Some works have already analyzed the traces present in the synthetic images working in the Fourier domain. However, the analysis mainly focused on artifacts caused by GAN models, such as those due to the up-sampling steps used to achieve the desired spatial resolution [13, 59]. To date, little attention has been paid to more recent and promising architectures, such as those based on auto-encoders and diffusion models [7, 37, 41, 46].

In this work we carry out a systematic investigation of the traces left not only by established GAN-based generators but also by several recent VQ-based and DM-based generators. To this end, we consider second-order statistics both in the spatial domain and in the frequency domain.

For the spatial domain, we use the average autocorrelation function while its transform, the average power spectrum, is used for frequency-domain analyses together with some derived quantities: the radial and angular spectra. Examples of power spectra and (central crop) autocorrelations are shown in Fig. 1 and Fig. 2.

As a result of our analyses we contribute the following main observations:

- To date, no generator appears to be artifact-free. Most of the times, artifacts are clearly spotted as strong peaks in the power spectra of noise residuals, but some recent methods have succeeded in reducing them. Even in these cases, unnatural regular patterns are typically observed in the autocorrelations of noise residuals.

- Artifacts (*e.g.*, due to compression) present in the dataset used to train a given model can be transferred to generated images. This is especially evident when the training dataset has limited variety and strong biases.

- Synthetic and real images may be distinguished not only by the presence of artifacts but also because of mismatched statistical features of the high-level real or generated content. We highlight such consistent differences by studying the radial and angular spectra of the sources, derived from the average power spectra.

In the following of the paper, we account for related work, describe the generative models under investigation, report on experimental results, discuss their significance, and finally draw conclusions.

2. Related Work

In this section, we describe previous work that shows the evidence of peculiar invisible traces embedded in the synthetic images both in the spatial and in the frequency domain.

Artificial fingerprints. Early research on synthetic images has shown that, similar to real cameras that imprint each image with a device- or camera-specific signature [8,30], GAN architectures also embed a fingerprint in each generated image. Such subtle traces, that are originated by the processing steps used in the image synthesis process, betray the artificial nature of the images. Several detectors [1, 32, 56] rely on these traces to expose synthetic images and even to trace back the individual architecture that generated them. Differently from these works, we rely on the second order statistics of the images to better highlight the patterns created by the generation process.

Spectral discrepancies. It is well known that synthetic images generated by GANs present anomalies in the frequency domain and cannot perfectly reproduce the spectral distribution of natural images. More specifically, in [59] for the first time evidence of such traces was shown, which appear as peaks in the Fourier domain and can be used as discriminating features for the development of a forensic detector. These traces are caused by the up-sampling operation present in the decoder architecture of the generation process which causes a marked aliasing phenomenon [26]. Further studies [13] showed that real and synthetic images exhibit a different decay of the Fourier spectrum at the highest frequencies. The inability to correctly reproduce the spectral distributions of the real data used for training is often attributed to up-sampling methods [12, 17]. Some works also propose a modification to the generator loss in order to better fit the spectral decay of natural images [12, 22]. Interestingly, it has been recently shown that also diffusion models are affected by a frequency bias and are not able to perfectly reproduce high frequencies and image details [55].

Based on these findings, several methods for GAN image detection rely on the high frequency discrepancies between real and generated images. However, it is not easy to exploit such traces in the challenging conditions of real-world applications [5, 19]. In this work we show that also middle frequencies can play a significant role for discrimination. In particular, we observe that the generators possess an anisotropic behaviour and struggle to reproduce details along the diagonal directions.

3. Generative models

This section aims at describing the generators of synthetic images analyzed in this work and their most important properties. A list of the models is reported in Tab. 1.

Generative Adversarial Networks. GANs adopt a training strategy based on a min-max game between two networks: a generator, and a discriminator [18]. The former tries to create realistic samples, able to fool the discriminator, while this latter tries to correctly distinguish real from generated samples. In 2019, Brock et al. proposed BigGAN [3] a class-conditional image synthesizer. Karras et al. proposed a series of GAN architectures for unconditional image generation, called StyleGAN [26–28], where the latent code controls the modulation of the convolutional kernels of the generator. StyleGAN-T [44] uses a StyleGAN-like architecture to generate text-guided images relying on CLIP (Contrastive Language-Image Pre-Training) [39] a powerful model that matches an image with its textual description. Recently, Tao et al. proposed GALIP (Generative Adversarial CLIP) [51] a GAN for text-to-image synthesis where both the discriminator and generator architectures are based on CLIP.

Vector Quantized Generative Networks. The Vector Quantized Variational Autoencoders (VQ-VAE) proposed

	Method	Backbone	Dataset	Conditioning
GAN	[3] BigGAN	ResNet-style	ImageNet	Class
	[28] StyleGAN2	StyleGAN like	FFHQ	Uncond.
	[26] StyleGAN3	StyleGAN like	FFHQ	Uncond.
	[44] StyleGAN-T	StyleGAN like	CC12M, CC, YFCC100m, Redcaps, LAION-aesthetic-6+	Text
	[51] GALIP	Transf. based	CUB, MS-COCO, CC3M, CC12M	Text
VQ	[14] Taming Transf.	VQGAN + Transf.	ImageNet, FFHQ	Class, Uncond.
	[9] DALL·E Mini	VQGAN + Transf.	CC, YFCC100M, CC12M	Text
DM	[49] Score-SDE	U-Net	FFHQ	Uncond.
	[11] ADM	U-Net	ImageNet, LSUN	Class, Uncond.
	[35] GLIDE	U-Net	CC, YFCC100M, Wikipedia text-image pairs	Text
	[40] DALL·E 2	U-Net	Unknown	Text
	[2] Ediff-I	Mul. U-Nets	Unknown	Text
LDM	[42] Latent Diff.	VAE + U-Net	ImageNet, LSUN	Class, Uncond.
	[43] Stable Diff.	VAE + U-Net	LAION	Text
	[38] DiT	VAE + ViT	ImageNet	Class

Table 1. Main properties of the generative models analyzed in this work.

in [54] are auto-encoders with an encoder network to compress data into a low-dimensional latent space and a decoder network to reconstruct the original data from the latent space. In particular, the VQ-VAEs quantize the latent space using a discrete codebook, which allows the model to learn latent vectors with a more meaningful and structured prior. In [6], the prior is derived from an autoregressive model based on a transformer network, while in [14], the authors propose VQ-GAN, a variant of the original VQ-VAE that induces an adversarial loss. Both Taming Transformers [14] and DALL·E Mini [9], considered in this work, are VQ-GANs with a transformer network as the autoregressive model of the prior.

Diffusion Models. To generate synthetic samples mimicking real ones, diffusion models rely on the inversion of an additive noise process. The model takes as input a noisy image composed of white noise and image content, and generates progressively less noisy versions of it until reaching the desired noiseless output. Recently, DMs [21, 25, 47, 48] have gained popularity for the task of image synthesis due to their favorable properties, such as stable training and better mode coverage compared to previous methods [18]. Ho et al. [21] first demonstrated the capability of diffusion models to generate high-quality images. Later, Dhariwal et al. [11] showed that DMs can produce images of better quality than those generated by GANs. The use of U-Net architecture or its variations is prevalent among diffusion models [2, 35, 40, 42, 43, 49]. In particular, [35, 40] extend the previous work [11] to image generation from text based on CLIP. The recent work Ediff-I [2] adopts multiple U-Net models specialized for different synthesis stages.

Latent Diffusion Models. These models, proposed in [42], combine variational autoencoders (VAEs) with diffusion models. a VAE is used to project images into a latent space and back. In this low-dimensional space, a diffusion model can generate the desired latent vector with limited computational cost. This combination allows for the generation of high-dimensional images with good quality and with a reduced computational burden compared to other diffusion models. A noteworthy model of this class is Stable Diffusion [43], which is part of an open-source project and is trained on the 5.85 billion images of the LAION dataset [45]. A recent work, Diffusion Transformers (DiTs) [38], proposes to replace the commonly-used U-Net architecture with a transformer architecture.

4. Forensic analysis

A crucial step in developing increasingly effective detectors is understanding which image features can help distinguish fake images from real ones. Interestingly, a few carefully designed simple experiments provide us already with several insightful observations. This section serves as a guided journey through some of these experimental results, with the aim of conveying the importance of these findings. Our analysis is carried out by relying on second-order statistics observed both in the spatial domain, the image autocorrelation function, and in the frequency domain, the image power spectrum[1].

For all sources of interest we consider a set of $I = 1000$ images, either real or generated under controlled conditions.

---

[1]In the literature, "power" is sometimes replaced by "energy", but it is only a matter of theoretical image modeling.

For a single image $x_i(m,n)$ the autocorrelation function is

$$R_{x_i}(\Delta m, \Delta n) = \langle x_i(m,n)x_i(m+\Delta m, n+\Delta n)\rangle \quad (1)$$

where $\langle \cdot \rangle$ indicates spatial average. The average autocorrelation of the source should be obtained through statistical averaging, which we approximate by averaging on all I images

$$R_x(\Delta m, \Delta n) = \frac{1}{I}\sum_{i=1}^{I} R_{x_i}(\Delta m, \Delta n) \quad (2)$$

This function summarizes the statistical correlation between pixels separated by a $(\Delta m, \Delta n)$ displacement.

For frequency domain analyses, we start from the Fourier transform of the $M \times N$ image

$$X_i(k,l) = \mathcal{F}[x_i(m,n)] =$$
$$= \sum_{m=1}^{M}\sum_{n=1}^{N} x_i(m,n)e^{-j2\pi(\frac{k}{M}m + \frac{l}{N}n)} \quad (3)$$

and obtain the source power spectrum, again, by averaging all individual power spectra

$$S_x(k,l) = \frac{1}{I}\sum_{i=1}^{I} S_{x_i}(k,l) = \frac{1}{I}\sum_{i=1}^{I} |X_i(k,l)|^2 \quad (4)$$

The power spectrum accounts for the fraction of the total image power concentrated at a given (horizontal,vertical) frequency pair $(\frac{k}{M}, \frac{l}{N})$. Note that these two functions are themselves related by a frequency transform

$$S_x(k,l) = \mathcal{F}[R_x(\Delta m, \Delta n)] \quad (5)$$

hence redundant, in theory. However, some phenomena are more visible in the spatial domain and some in frequency, so we study both of them.

Often, phenomena of interest are only visible in the noise residuals, obtained by removing the high-level semantic content from the image by means of a denoising filter.

$$r_i(m,n) = x_i(m,n) - \mathcal{D}(x_i(m,n); \sigma) \quad (6)$$

In our experiments, we use always the denoiser by Zhang *et al.* [58] with noise parameter $\sigma=1$.

All autocorrelations and spectra, associated to images or residuals, are normalized to the average power per pixel computed as the mean value of $S_x(k,l)$. Note that we show only the central 65×65 crop of the autocorrelation function.

4.1. Architecture related traces

It is well known that many GAN-based generators leave clear traces of their processing pipeline in the images. These can be exposed by extracting and averaging the noise
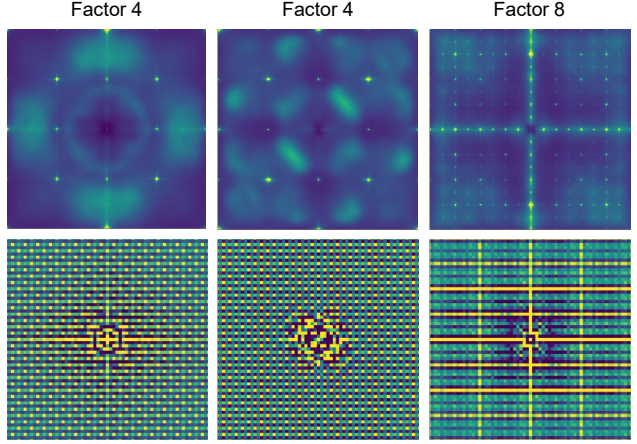

Figure 3. Power spectra (top) and autocorrelation functions (bottom) of noise residuals of three slightly different latent diffusion models. The upsampling factor between latent space and pixel space is 4 for the first two architectures and 8 for the third one. This single parameter appears to be responsible for the positioning of the peaks in the power spectra and the periodicity of the regular patterns in the autocorrelations.

residuals, to compute a model artificial fingerprint, and analyzing the results in the spatial and/or spectral domain. Images generated by diffusion models appear to show artifacts of a similar nature and, arguably, a similar origin. Fig.3 shows the power spectra (top) and autocorrelation functions (bottom) of the artificial fingerprints of three slightly different latent diffusion architectures. The first two models are characterized by an upsampling factor of 4 between latent space and pixel space and, not by chance, exhibit clear peaks in the spectrum at frequencies multiples of 1/4, and strong regular patterns with periodicity 4 in the autocorrelation function. The minor differences between the two are mostly due to the different training set used, ImageNet vs. CelebHQ. On the contrary, the third model, trained on LSUN churches images, has an upsampling factor of 8. As expected, peaks in the power spectrum are now denser, at frequencies multiple of 1/8, while they are sparser in the autocorrelation function.

While these artifacts are clearly present in current generation architectures, and can be exploited for forensic needs, these analyses refer to images observed "in the lab", as they leave the generator, with no further processing. This is not a realistic condition, just think of all the standard processing steps, especially compression and resizing, applied to images as soon as they are uploaded on a social network. These post-processing steps may significantly modify and hide the artifacts highlighted before and should be always taken into account when designing a forensic detector.

Fig.4 shows some interesting examples. We consider images generated by Stable Diffusion and show, on the top-
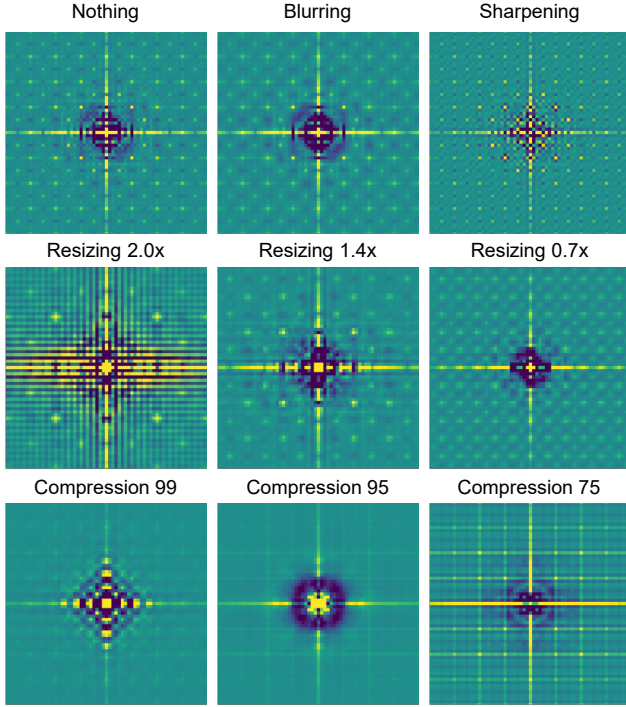
Nothing	Blurring	Sharpening

Resizing 2.0x	Resizing 1.4x	Resizing 0.7x

Compression 99	Compression 95	Compression 75

Figure 4. Effect of post-processing operations on the autocorrelation of noise residuals. We consider Stable Diffusion images and apply to them different operations to see how the autocorrelation changes. In some cases the original generation traces (top-left) are completely hidden by post-processing artifacts.

ADM	Latent Diff.	ImageNet

ADM	Latent Diff.	LSUN bedrooms

Figure 5. Autocorrelation of noise residuals for real and synthetic images. Top: two diffusion models (ADM and Latent) and the dataset used for their training (ImageNet). Bottom: two diffusion models (ADM and Latent) and the dataset used for their training (LSUN Bedroom). In the first case the JPEG traces (8×8 grid) are barely visible in the training dataset and they are not present in the generated images. Instead, in the second case they are clearly visible both in the training dataset and in the generated images.

left corner, the autocorrelation function observed in the absence of any post-processing. A clear regular pattern is visible which suggests the artificial nature of these images and may even be used to trace back the individual generator used. In the other slots, we show the autocorrelation function after various types of post-processing, blurring, sharpening, resizing at three different scales, JPEG compression at three different quality factors. The autocorrelation function changes more or less markedly depending on the nature and intensity of the post-processing, but in most cases it can be hardly associated with the original one. In the presence of strong JPEG compression, for example, the compression artifacts dominate the scene and hide completely the generation artifacts.

4.2. Training dataset bias

In our analysis, we singled out also a further problem that undermines the reliability of high-frequency fingerprints. This is exemplified by the images of Fig. 5 relative to the DM models, where noise residuals exhibit an obvious regular pattern that may suggest their artificial origin. However, this is the very same pattern observed in the residuals of real images JPEG compressed at quality factor Q=75 (see
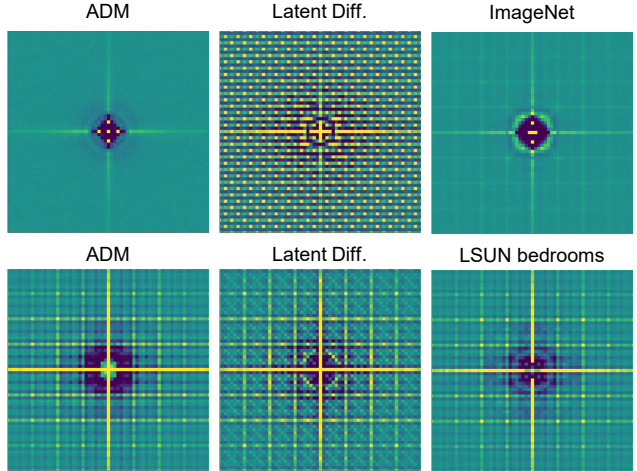
Fig. 4). Analyzing in more detail the generation pipeline, we discovered that the dataset used to train the model was composed entirely by images JPEG compressed at Q=75. The model learned to include the compression artifacts in the generated images, that is, it transferred dataset-related artifacts in the generated image with the potential to trick fingerprint-based detectors. This example rings also a bell about the experimental protocol. In fact, most of the times we do not know the details of the dataset used to train a model, but such details may well impact the characteristics of the generated images. A fully reliable analysis must keep this aspect under control.

Interestingly, the ability of GANs models to transfer the same fingerprints from training data to the generated images has been exploited to develop an approach to protect synthetic images by embedding the same artificial fingerprint as proposed in [57].

4.3. Fidelity to natural images

Natural images are known to exhibit strong statistical regularities in the frequency domain. In particular, several studies [4, 52] observed that the average power spectrum decades with a $1/f^\alpha$ law, with $\alpha \simeq 2$. Moreover, it has been also recognized [50, 53] that the power spectrum is far from isotropic, with power concentrated especially at the horizontal and vertical orientations and much less at the others. These may be precious pieces of information to-
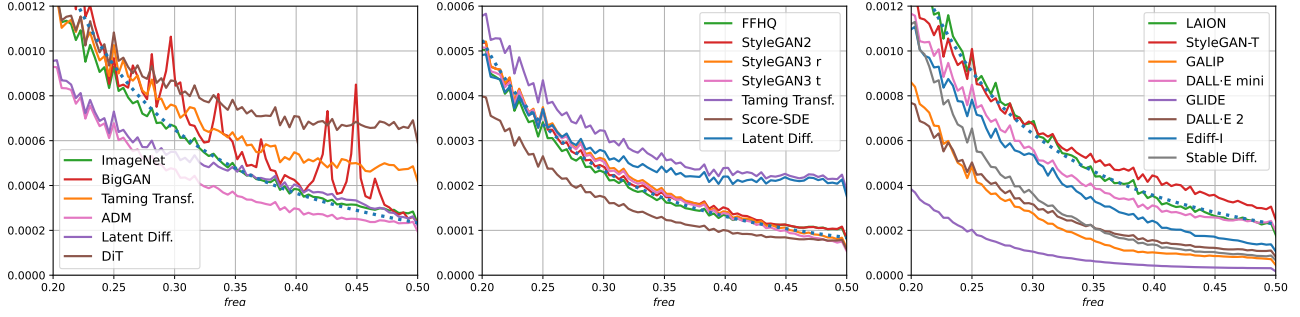
Figure 6. Radial spectrum power density. Synthetic images are compared with the real images used for training the correspondent model. Real images (green) fit very well the expected theoretical curve (dotted).

wards the detection of synthetic images. Indeed, considering their good photorealism and high semantic quality, we expect their spectra to be very similar to those of real images, at low frequencies. However, it is certainly possible that significant deviations arise at mid-high frequencies, not because of generation artifacts but due to the inability to correctly replicate the statistics of natural images. This conjecture is indirectly supported also by the literature where there is wide agreement on the superior performance of dual-stream detectors, which exploit both high-frequency artifacts and original RGB images. With this approach, the detector has the opportunity to analyze not only the highest frequencies, with their subtle generation artifacts, but also the intermediate frequency, which carry information on the signal itself, with beneficial effects.

Analysis of radial spectra. To test this conjecture, we now analyze the radial spectra of real and synthetic images. The radial spectrum of the i-th image is computed by averaging the values of its Fourier transform (normalized by its standard deviation) at distance ρ from the origin, irrespective of the angular direction

$$X_i(\rho) = \langle X_i(k,l) \rangle_{A(\rho)} \qquad (7)$$

where $A(\rho)$ is the annulus between radii $\rho - \Delta\rho$ and $\rho + \Delta\rho$, and ρ goes from 0 to 0.5 in 128 discrete steps. Then, as usual, an average on $I = 1000$ images is taken to characterize the whole source. In our analyses we will discard low and very high frequencies and focus on the range $\rho \in [0.2 - 0.5]$.

Our goal here is to test whether the radial spectra of synthetic images follow the same "correct" law as those of real images. However, to avoid dataset-related biases, for a given real source (*e.g.*, imagenet), we study only the images that are generated using the very same source as training set. With this care, statistical difference can be attributed only to the generation engine. In Fig. 6, for a few widespread real sources, we compare radial spectra of real and corresponding generated images. The former (green) decay with a nearly perfect power law (dashed black). On

the contrary, all radial spectra of synthetic sources deviate more or less markedly and in one or more frequency ranges from that law. Generally, but not always, real images seem to possess a richer content at the intermediate frequencies. Only for the FFHQ dataset, the radial spectra of some synthetic sources approximate pretty well that of the real source. However, the FFHQ dataset is composed by images that have been interpolated, that is low-pass filtered, a process that inevitably modifies the original spectral content and makes it indistinguishable from that of some synthetic images. In hindsight, this further supports our conjecture.

Analysis of angular spectra. A similar analysis is carried out for the angular spectra, where now

$$X_i(\theta) = \langle X_i^{hp}(k,l) \rangle_{C(\theta)} \qquad (8)$$

$C(\theta)$ is a double cone between angles $\theta - \Delta\theta$ and $\theta + \Delta\theta$, and θ goes from 0 to π in 16 discrete steps. Note that the non-informative low-pass content is filtered beforehand with cut-off frequency 0.1. The resulting plots are reported in Fig. 7(top) for the same real and synthetic sources as before. Mostly, the very same considerations already done for radial spectra apply. The only exceptions are observed again for the FFHQ dataset and for the same reasons as before. In this case, there is no theoretical reference pattern from the literature. We observe again significant deviations from the angular spectra of the real sources (green). Such deviations can be better appreciated in the plots of Fig. 7(bottom), showing the square root of the Fisher's discriminant ratios

$$F_s(\theta) = \frac{\mu_s(\theta) - \mu_0(\theta)}{\sqrt{\sigma_s^2(\theta) + \sigma_0^2(\theta)}} \qquad (9)$$

where $\mu_s(\theta) = X_s(\theta)$ and $\mu_0(\theta) = X_0(\theta)$ are the mean angular spectra of synthetic (s) and real (0) images, respectively, and $\sigma_s^2(\theta), \sigma_0^2(\theta)$ the corresponding variances. In this case the real source is represented by the green unit circle. Finally, it is worth noting that the StyleGAN2 and StyleGAN3 models better fit the spectral distribution of real images. For StyleGAN3, this is justified by the development of GAN architectures that were able to avoid aliasing [26].
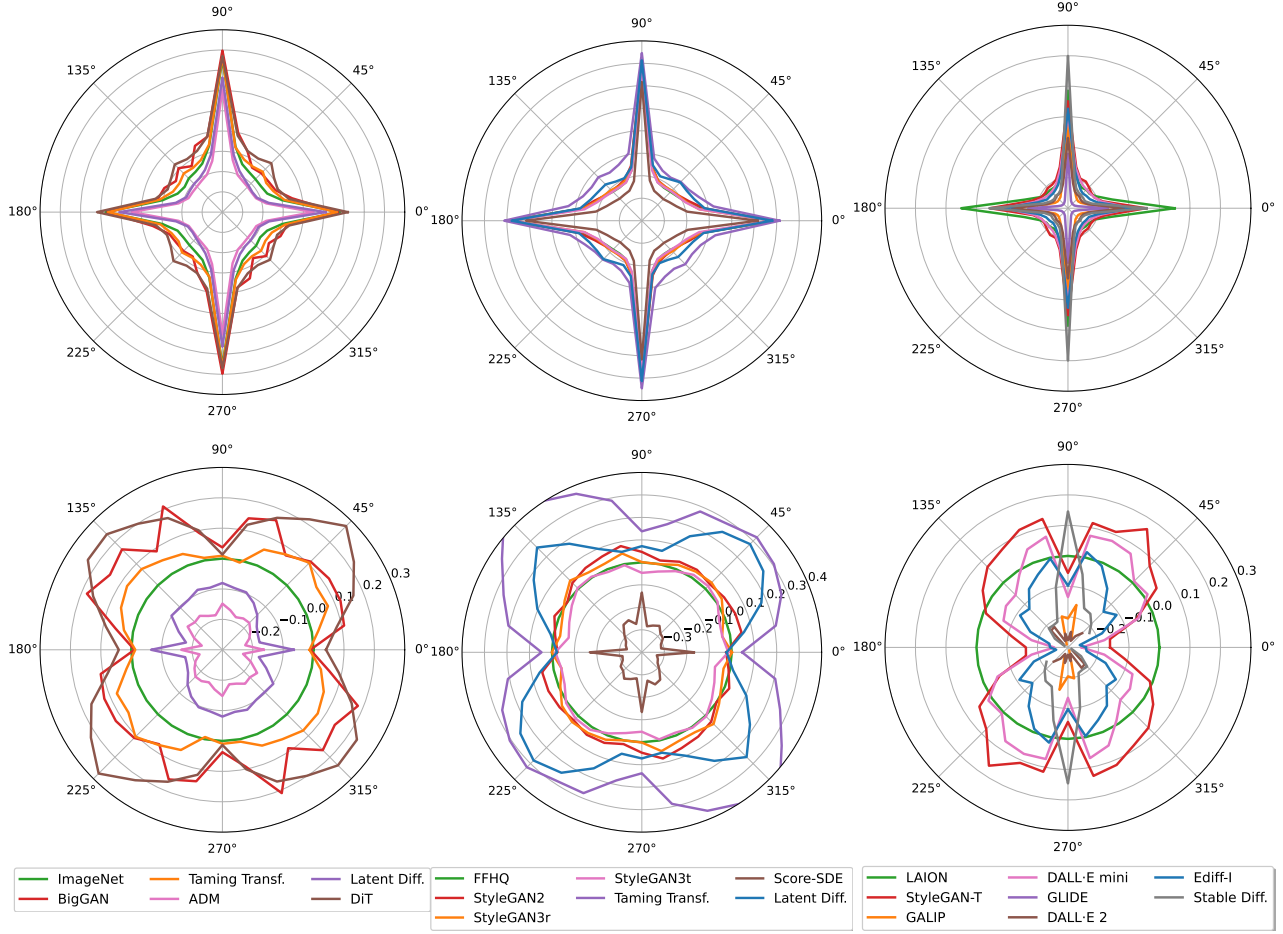
Figure 7. Angular spectrum power density (top) and the square-root of Fisher's discriminant ratio applied at each direction (bottom).

5. Conclusions

Generators of synthetic images are becoming more and more sophisticated. Casual viewers are already unable to tell real images from fake ones and even forensic tools struggle to cope with the rapid progress of new models. In this context, brute-force detectors, even trained on huge amounts of data, may not be up to the task, especially when asked to generalize to new generators. Therefore, it is important to single out and possibly exploit all features that distinguish real from fake data. In this work, we analyzed a large number of modern and popular generators looking for such features both in the spatial domain and in the frequency domain. In both domains we relied on second-order statistics, autocorrelation functions and power spectra.

This investigation provides some interesting results. On one hand, even the most sophisticated architectures keep generating artifacts that can be exploited for detection. These may be more visible in the spatial or in the spectral domain and both should be always explored. In addition, generators do not seem able to mimic the spectral distribution of real images at the mid-high frequencies and radial and angular spectra may help expose such discrepancies. Finally, we noted that a strongly biased training set may teach the model to transfer the very same biases in the generated images, a possible problem for both generators and detectors. In future work we will pursue the design of forensic detectors based also on these traces.

References

[1] V. Asnani, X. Yin, T. Hassner, and X. Liu. Reverse engineering of generative models: Inferring model hyperparameters from generated images. *arXiv preprint arXiv/2106.07873*, 2021. 3

[2] Y. Balaji, S. Nah, X. Huang, A. Vahdat, J. Song, K. Kreis, M. Aittala, T. Aila, S. Laine, B. Catanzaro, T. Karras, and M.Y. Liu. eDiff-I: Text-to-Image Diffusion Models with an Ensemble of Expert Denoisers. *arXiv preprint arXiv:2211.01324*, 2022. 1, 2, 4

[3] A. Brock, J. Donahue, and K. Simonyan. Large Scale GAN Training for High Fidelity Natural Image Synthesis. In *International Conference on Learning Representations (ICLR)*, 2018. 3, 4

[4] G.J. Burton and I.R. Moorhead. Color and spatial structure in natural scenes. *Applied Optics*, 26:157–170, 1987. 6

[5] K. Chandrasegaran, N.-T. Tran, and N.-M. Cheung. A Closer Look at Fourier Spectrum Discrepancies for CNN-generated Images Detection. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2021. 3

[6] M. Chen, A. Radford, R. Child, J. Wu, H. Jun, D. Luan, and I. Sutskever. Generative pretraining from pixels. In *International Conference on Machine Learning (ICML)*, 2020. 4

[7] R. Corvi, D. Cozzolino, G. Zingarini, G. Poggi, K. Nagano, and L. Verdoliva. On the detection of synthetic images generated by diffusion models. In *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2023. 2

[8] D. Cozzolino and L. Verdoliva. Noiseprint: A CNN-based camera model fingerprint. *IEEE Transactions on Information Forensics and Security*, 15(1):14–27, 2020. 3

[9] B. Dayma, S. Patil, P. Cuenca, K. Saifullah, T. Abraham, P. Lê Khàc, L. Melas, and R. Ghosh. DALL-E Mini. https://github.com/borisdayma/dalle-mini, 2021. 1, 2, 4

[10] J. Deng, W. Dong, R. Socher, L.-J. Li, K. Li, and L. Fei-Fei. ImageNet: A large-scale hierarchical image database. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 248–255, 2009. 2

[11] P. Dhariwal and A. Nichol. Diffusion models beat GANs on image synthesis. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2021. 4

[12] R. Durall, M. Keuper, and J. Keuper. Watch your up-convolution: CNN based Generative Deep Neural Networks are failing to reproduce spectral distributions. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2020. 3

[13] T. Dzanic, K. Shah, and F. D. Witherden. Fourier spectrum discrepancies in deep network generated images. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2020. 2, 3

[14] P. Esser, R. Rombach, and B. Ommer. Taming transformers for high-resolution image synthesis. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 12873–12883, 2021. 1, 4

[15] H. Farid. Lighting (in)consistency of paint by text. *arXiv preprint arXiv:2207.13744v2*, 2022. 2

[16] H. Farid. Perspective (in)consistency of paint by text. *arXiv preprint arXiv:2206.14617v1*, 2022. 2

[17] J. Frank, T. Eisenhofer, L. Schönherr, A. Fischer, D. Kolossa, and T. Holz. Leveraging Frequency Analysis for Deep Fake Image Recognition. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2020. 3

[18] I. Goodfellow, J. Pouget-Abadie, M. Mirza, B. Xu, D. Warde-Farley, S. Ozair, A. Courville, and Y. Bengio. Generative adversarial nets. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2014. 3, 4

[19] 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 *IEEE International Conference on Multimedia and Expo (ICME)*, 2021. 3

[20] H. Guo, S. Hu, X. Wang, M.C. Chang, and S. Lyu. Eyes tell all: Irregular pupil shapes reveal GAN-generated faces. In *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2022. 2

[21] J. Ho, A. Jain, and P. Abbeel. Denoising diffusion probabilistic models. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2020. 4

[22] L. Jiang, B. Dai, W. Wu, and C. C. Loy. Focal frequency loss for image reconstruction and synthesis. In *IEEE International Conference on Computer Vision (ICCV)*, 2021. 3

[23] M. Kang, J.-Y. Zhu, R. Zhang, J. Park, E. Shechtman, S. Paris, and T. Park. Scaling up GANs for Text-to-Image Synthesis. *arXiv preprint arXiv:2303.05511*, 2023. 2

[24] T. Karras, T. Aila, S. Laine, and J. Lehtinen. Progressive Growing of GANs for Improved Quality, Stability, and Variation. In *International Conference on Learning Representations (ICLR)*, 2018. 2

[25] T. Karras, M. Aittala, T. Aila, and S. Laine. Elucidating the design space of diffusion-based generative models. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2022. 4

[26] T. Karras, M. Aittala, S. Laine, E. Härkönen, J. Hellsten, J. Lehtinen, and T. Aila. Alias-free generative adversarial networks. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2021. 3, 4, 7

[27] T. Karras, S. Laine, and T. Aila. A style-based generator architecture for generative adversarial networks. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2019. 2, 3

[28] T. Karras, S. Laine, M. Aittala, J. Hellsten, J. Lehtinen, and T. Aila. Analyzing and improving the image quality of StyleGAN. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2020. 2, 3, 4

[29] F. Lago, C. Pasquini, R. Boehme, H. Dumont, V. Goffaux, and G. Boato. More real than real: A study on human visual perception of synthetic faces. *IEEE Signal Processing Magazine*, 39(1):109–116, 20021. 2

[30] J. Lukàš, J. Fridrich, and M. Goljan. Digital camera identification from sensor pattern noise. *IEEE Transactions on Information Forensics and Security*, 1(2):205–214, 2006. 3

[31] F. Marra, D. Gragnaniello, D. Cozzolino, and L. Verdoliva. Detection of GAN-generated fake images over social networks. In *IEEE Conference on Multimedia Information Processing and Retrieval (MIPR)*, 2018. 2

[32] F. Marra, D. Gragnaniello, L. Verdoliva, and G. Poggi. Do GANs Leave Artificial Fingerprints? In *IEEE Conference on Multimedia Information Processing and Retrieval (MIPR)*, 2019. 2, 3

[33] F. Matern, C. Riess, , and M. Stamminger. Exploiting visual artifacts to expose deepfakes and face manipulations. In *IEEE Winter Conference on Applications of Computer Vision Workshops (WACVW)*, 2019. 2

[34] S. McCloskey and M. Albright. Detecting GAN-Generated Imagery using Saturation Cues. In *IEEE International Conference on Image Processing (ICIP)*, 2019. 2

[35] A. Q. Nichol, P. Dhariwal, A. Ramesh, P. Shyam, P. Mishkin, B. Mcgrew, I. Sutskever, and M. Chen. GLIDE: Towards Photorealistic Image Generation and Editing with Text-Guided Diffusion Models. In *International Conference on Machine Learning (ICML)*, 2022. 2, 4

[36] S. J. Nightingale and H. Farid. AI-synthesized faces are indistinguishable from real faces and more trustworthy. *Proceedings of the National Academy of Sciences*, 119(8):e2120481119, 2022. 2

[37] U. Ojha, Y. Li, and Y. Jae Lee. Towards universal fake image detectors that generalize across generative models. *arXiv preprint arXiv:2302.10174v1*, 2022. 2

[38] W. Peebles and S. Xie. Scalable diffusion models with transformers. *arXiv preprint arXiv:2212.09748*, 2022. 2, 4

[39] A. Radford, J. W. Kim, C. Hallacy, A. Ramesh, G. Goh, S. Agarwal, G. Sastry, A. Askell, P. Mishkin, J. Clark, et al. Learning transferable visual models from natural language supervision. In *International Conference on Machine Learning (ICML)*, 2021. 3

[40] A. Ramesh, P. Dhariwal, A. Nichol, C. Chu, and M. Chen. Hierarchical text-conditional image generation with clip latents. *arXiv preprint arXiv:2204.06125v1*, 2022. 2, 4

[41] J. Ricker, S. Damm, T. Holz, and A. Fischer. Towards the detection of diffusion model deepfakes. *arXiv preprint arXiv:2210.14571v1*, 2022. 2

[42] R. Rombach, A. Blattmann, D. Lorenz, P. Esser, and B. Ommer. High-resolution image synthesis with latent diffusion models. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 10684–10695, 2022. 4

[43] R. Rombach, A. Blattmann, D. Lorenz, P. Esser, and B. Ommer. Stable diffusion. https://github.com/CompVis/stable-diffusion, 2022. 1, 4

[44] A. Sauer, T. Karras, S. Laine, A. Geiger, and T. Aila. StyleGAN-T: Unlocking the Power of GANs for Fast Large-Scale Text-to-Image Synthesis. *arXiv preprint arXiv:2301.09515*, 2023. 1, 2, 3, 4

[45] C. Schuhmann, R. Kaczmarczyk, A. Komatsuzaki, A. Katta, R. Vencu, R. Beaumont, J. Jitsev, T. Coombes, and C. Mullis. Laion-400m: Open dataset of clip-filtered 400 million image-text pairs. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2021. 2, 4

[46] Z. Sha, Z. Li, N. Yu, and Y. Zhang. DE-FAKE: Detection and Attribution of Fake Images Generated by Text-to-Image Diffusion Models. *arXiv preprint arXiv:2210.06998*, 2022. 2

[47] J. Sohl-Dickstein, E. Weiss, N. Maheswaranathan, and S. Ganguli. Deep unsupervised learning using nonequilibrium thermodynamics. In *International Conference on Machine Learning (ICML)*, pages 2256–2265. PMLR, 2015. 4

[48] Y. Song and S. Ermon. Generative modeling by estimating gradients of the data distribution. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2019. 4

[49] Y. Song, J. Sohl-Dickstein, D. P. Kingma, A. Kumar, S. Ermon, and B. Poole. Score-based generative modeling through stochastic differential equations. In *International Conference on Learning Representations (ICLR)*, 2021. 4

[50] E. Switkes, M.J. Mayer, and J.A. Sloan. Spatial frequency analysis of the visual environment: anisotropy and the carpentered environment hypothesis. *Vision Research*, 18:1393–1399, 1978. 6

[51] M. Tao, B.-K. Bao, H. Tang, and C. Xu. GALIP: Generative Adversarial CLIPs for Text-to-Image Synthesis. *arXiv preprint arXiv:2301.12959*, 2023. 1, 2, 3, 4

[52] D.J. Tolhurst, Y. Tadmor, and C. Tang. The amplitude spectra of natural images. *Ophthalmic Physiol. Opt.*, 12:229–232, 1992. 6

[53] A. Torralba and A. Oliva. Statistics of natural image categories. *Network: Computation in Neural Systems*, 14:391–412, 2003. 6

[54] A. Van Den Oord, O. Vinyals, and K. Kavukcuoglu. Neural discrete representation learning. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2017. 4

[55] X. Yang, D. Zhou, J. Feng, and X. Wang. Diffusion probabilistic model made slim. *arXiv preprint arXiv:2211.17106*, 2022. 3

[56] N. Yu, L. Davis, and M. Fritz. Attributing Fake Images to GANs: Learning and Analyzing GAN Fingerprints. In *IEEE International Conference on Computer Vision (ICCV)*, 2019. 2, 3

[57] N. Yu, V. Skripniuk, S. Abdelnabi, and M. Fritz. Artificial Fingerprinting for Generative Models: Rooting Deepfake Attribution in Training Data. In *IEEE International Conference on Computer Vision (ICCV)*, 2021. 6

[58] K. Zhang, W. Zuo, Y. Chen, D. Meng, and L. Zhang. Beyond a gaussian denoiser: Residual learning of deep cnn for image denoising. *IEEE Transactions on Image Processing*, 26(7):3142–3155, 2017. 5

[59] X. Zhang, S. Karaman, and S.-F. Chang. Detecting and Simulating Artifacts in GAN Fake Images. In *IEEE international workshop on information forensics and security (WIFS)*, pages 1–6, 2019. 2, 3

[60] J.-Y. Zhu, T. Park, P. Isola, and A. Efros. Unpaired image-toimage translation using cycle-consistent adversarial networks. In *IEEE International Conference on Computer Vision (ICCV)*, 2017. 2