





Towards the Detection of Diffusion Model Deepfakes

Jonas Ricker¹ ^a, Simon Damm¹ ^b, Thorsten Holz² ^c and Asja Fischer¹ ^d

¹Ruhr University Bochum, Bochum, Germany

²CISPA Helmholtz Center for Information Security, Saarbrücken, Germany
fi

Keywords: Deepfake Detection, Diffusion Models, Generative Adversarial Networks, Frequency Analysis.


Abstract: In the course of the past few years, diffusion models (DMs) have reached an unprecedented level of visual quality. However, relatively little attention has been paid to the detection of DM-generated images, which is critical to prevent adverse impacts on our society. In contrast, generative adversarial networks (GANs), have been extensively studied from a forensic perspective. In this work, we therefore take the natural next step to evaluate whether previous methods can be used to detect images generated by DMs. Our experiments yield two key findings: (1) state-of-the-art GAN detectors are unable to reliably distinguish real from DM-generated images, but (2) re-training them on DM-generated images allows for almost perfect detection, which remarkably even generalizes to GANs. Together with a feature space analysis, our results lead to the hypothesis that DMs produce fewer detectable artifacts and are thus more difficult to detect compared to GANs. One possible reason for this is the absence of grid-like frequency artifacts in DM-generated images, which are a known weakness of GANs. However, we make the interesting observation that diffusion models tend to underestimate high frequencies, which we attribute to the learning objective.


1 INTRODUCTION


In the recent past, diffusion models (DMs) have shown a lot of promise as a method for synthesizing images. Such models provide better (or at least similar) performance compared to generative adversarial networks (GANs) and enable powerful text-to-image models such as DALL·E 2 (Ramesh et al., 2022), Imagen (Saharia et al., 2022), and Stable Diffusion (Rombach et al., 2022). Advances in image synthesis have resulted in very high-quality images being generated, and humans can hardly tell if a given picture is an actual or artificially generated image (so-called *deepfake*) (Nightingale and Farid, 2022). This progress has many implications in practice and poses a danger to our digital society: Deepfakes can be used for disinformation campaigns, as such images appear particularly credible due to their sensory comprehensibility. Disinformation aims to discredit opponents in public perception, to create sentiment for or against certain social groups, and thus influence public opinion. In effect, deepfakes lead to an erosion of trust


in institutions or individuals, support conspiracy theories, and promote a fundamental political camp formation. DM-based text-to-image models entail particular risks, since an adversary can specifically create images supporting their narrative, with very little technical knowledge required. A recent example of public deception featuring DM-generated images—although without malicious intent—is the depiction of Pope Francis in a puffer jacket (Huang, 2023). Despite the growing concern about deepfakes and the continuous improvement of DMs, there is only a limited amount of research on their detection.

In this paper, we conduct an extensive experimental study on the detectability of images generated by DMs. Since previous work on the detection of GAN-generated images (e.g., (Wang et al., 2020; Gragnaniello et al., 2021; Mandelli et al., 2022)) resulted in effective detection methods, we raise the question whether these can be applied to DM-generated images. Our analysis on five state-of-the-art GANs and five DMs demonstrates that existing detection methods suffer from severe performance degradation when applied to DM-generated images, with the AUROC dropping by 15.2% on average compared to GANs. However, we show that by re-training, the detection accuracy can be drastically improved, proving that

^a  <https://orcid.org/0000-0002-7186-3634>

^b  <https://orcid.org/0000-0002-4584-1765>

^c  <https://orcid.org/0000-0002-2783-1264>

^d  <https://orcid.org/0000-0002-1916-7033>

images generated by DMs *can* be detected. Remarkably, a detector trained on DM-generated images is capable of detecting images from GANs, while the opposite direction does not hold. Our analysis in feature space suggests that DM-generated images are harder to detect because they contain fewer generation artifacts, particularly in the frequency domain. However, we observe a previously overlooked mismatch towards higher frequencies. Further analysis suggests that this is caused by the training objective of DMs, which favors perceptual image quality instead of accurate reproduction of high-frequency details. We believe that our results provide the foundation for further research on the effective detection of deepfakes generated by DMs. Our code, data, and the extended version of this paper (with additional experiments) are available at <https://github.com/jonasricker/diffusion-model-deepfake-detection>.

2 RELATED WORK

Fake Image Detection. In the wake of the emergence of powerful image synthesis methods, the forensic analysis of deepfake images received increased attention, leading to a variety of detection methods (Verdoliva, 2020). Existing approaches can be broadly categorized into two groups. Methods in the first group exploit either semantic inconsistencies like irregular eye reflections (Hu et al., 2021) or known generation artifacts in the spatial (Nataraj et al., 2019; McCloskey and Albright, 2019) or frequency domain (Frank et al., 2020). The second group uses neural networks to learn a feature representation in which real images can be distinguished from generated ones. Wang et al. demonstrate that training a standard convolutional neural network (CNN) on real and fake images from a single GAN yields a classifier capable of detecting images generated by a variety of unknown GANs (Wang et al., 2020). Given the rapid evolution of generative models, developing detectors which generalize to new generators is crucial and therefore a major field of research (Xuan et al., 2019; Chai et al., 2020; Wang et al., 2020; Cozzolino et al., 2021; Gragnaniello et al., 2021; Girish et al., 2021; Mandelli et al., 2022; Jeong et al., 2022).

Since DMs have been proposed only recently, few works analyze their forensic properties. Farid performs an initial exploration of lighting (Farid, 2022a) and perspective (Farid, 2022b) inconsistencies in images generated by DALL-E 2 (Ramesh et al., 2022), showing that DMs often generate physically implausible scenes. A novel approach specifically targeted at

DMs is proposed in (Wang et al., 2023), who observe that DM-generated images can be more accurately reconstructed by a pre-trained DM than real images. The difference between the original and reconstructed image then serves as the input for a binary classifier. Another work (Sha et al., 2023) focuses on text-to-image models like Stable Diffusion (Rombach et al., 2022). They find that incorporating the prompt with which an image was generated (or a generated caption if the real prompt is not available) into the detector improves classification. In a work related to ours (Corvi et al., 2023b), it is shown that GAN detectors perform poorly on DM-generated images. Therefore, a pressing challenge is to develop *universal* detection methods that are effective against different kinds of generative models, mainly GANs and DMs. Ojha et al. make a first step in this direction (Ojha et al., 2023). Instead of training a classifier directly on real and fake images, which according to their hypothesis leads to poor generalization since the detector focuses on e.g., GAN-specific artifacts, they propose to use a pre-trained vision transformer (CLIP-ViT (Dosovitskiy et al., 2021; Radford et al., 2021)), extended with a final classification layer.

Frequency Artifacts in Generated Images. Zhang et al. were the first to demonstrate that the spectrum of GAN-generated images contains visible artifacts in the form of a periodic, grid-like pattern due to transposed convolution operations (Zhang et al., 2019). These findings were later reproduced (Wang et al., 2020) and extended to the discrete cosine transform (DCT) (Frank et al., 2020). Another characteristic was discovered in (Durall et al., 2020), who showed that GANs are unable to correctly reproduce the spectral distribution of the training data. In particular, generated images contain increased magnitudes at high frequencies. While several works attribute these spectral discrepancies to transposed convolutions (Zhang et al., 2019; Durall et al., 2020) or, more general, up-sampling operations (Frank et al., 2020; Chandrasegaran et al., 2021), no consensus on their origin has yet been reached. Some works explain them by the spectral bias of convolution layers due to linear dependencies (Dzanic et al., 2020; Khayatkhoei and Elgammal, 2022), while others suggest the discriminator is not able to provide an accurate training signal (Chen et al., 2021; Schwarz et al., 2021).

In contrast, whether images generated by DMs exhibit grid-like frequency patterns appears to strongly depend on the specific model (Sha et al., 2023; Corvi et al., 2023a; Ojha et al., 2023). Another interesting observation is made by Rissanen et al. who analyze the generative process of diffusion models in the frequency domain (Rissanen et al., 2023). They

state that diffusion models have an inductive bias according to which, during the reverse process, higher frequencies are added to existing lower frequencies. Other works (Kingma et al., 2021; Song et al., 2022b) experiment with adding Fourier features to improve learning of high-frequency content, the former reporting it leads to much better likelihoods.

3 BACKGROUND ON DMs

DMs are a class of probabilistic generative models, originally inspired by nonequilibrium thermodynamics (Sohl-Dickstein et al., 2015). The most common formulations build either on DDPM (Ho et al., 2020) or the score-based modeling perspective (Song and Ermon, 2019; Song and Ermon, 2020; Song et al., 2022b). Numerous modifications and improvements have been proposed, leading to higher perceptual quality (Nichol and Dhariwal, 2021; Dhariwal and Nichol, 2021; Choi et al., 2022; Rombach et al., 2022) and increased sampling speed (Song et al., 2022a; Liu et al., 2022; Salimans and Ho, 2022; Xiao et al., 2022). In short, DMs model a data distribution by gradually disturbing a sample from this distribution and then learning to reverse this diffusion process. To be more precise, we briefly review the forward and backward process for the seminal work in (Ho et al., 2020). In the diffusion (or forward) process for DDPMs, a sample \mathbf{x}_0 (an image in most applications) is repeatedly corrupted by Gaussian noise in sequential steps $t = 1, \dots, T$ in dependence of a monotonically increasing noise schedule $\{\beta_t\}_{t=1}^T$:

$$q(\mathbf{x}_t|\mathbf{x}_{t-1}) = \mathcal{N}(\sqrt{1 - \beta_t}\mathbf{x}_{t-1}, \beta_t\mathbf{I}) . \quad (1)$$

With $\alpha_t := 1 - \beta_t$ and $\bar{\alpha}_t := \prod_{s=1}^t \alpha_s$, we can directly sample from the forward process at arbitrary times:

$$q(\mathbf{x}_t|\mathbf{x}_0) = \mathcal{N}(\sqrt{\bar{\alpha}_t}\mathbf{x}_0, (1 - \bar{\alpha}_t)\mathbf{I}) . \quad (2)$$

The noise schedule is typically designed to satisfy $q(\mathbf{x}_T|\mathbf{x}_0) \approx \mathcal{N}(\mathbf{0}, \mathbf{I})$. During the denoising (or reverse) process, we aim to iteratively sample from $q(\mathbf{x}_{t-1}|\mathbf{x}_t)$ to ultimately obtain a clean image from $\mathbf{x}_T \sim \mathcal{N}(\mathbf{0}, \mathbf{I})$. However, since $q(\mathbf{x}_{t-1}|\mathbf{x}_t)$ is intractable as it depends on the entire underlying data distribution, it is approximated by a deep neural network. More formally, $q(\mathbf{x}_{t-1}|\mathbf{x}_t)$ is approximated by

$$p_\theta(\mathbf{x}_{t-1}|\mathbf{x}_t) = \mathcal{N}(\mu_\theta(\mathbf{x}_t, t), \Sigma_\theta(\mathbf{x}_t, t)) , \quad (3)$$

where mean μ_θ and covariance Σ_θ are given by the output of the model (or the latter is set to a constant as proposed in (Ho et al., 2020)). Predicting the mean of the denoised sample $\mu_\theta(\mathbf{x}_t, t)$ is conceptually equivalent to predicting the noise that should be removed,

denoted by $\varepsilon_\theta(\mathbf{x}_t, t)$. Predominantly, the latter approach is implemented (e.g., (Ho et al., 2020; Dhariwal and Nichol, 2021)) such that training a DM boils down to minimizing a (weighted) mean squared error (MSE) $\|\varepsilon - \varepsilon_\theta(\mathbf{x}_t, t)\|^2$ between the true and predicted noise. Note that this objective can be interpreted as a weighted ELBO with data augmentation (Kingma and Gao, 2023). For a recent overview on DMs see (Yang et al., 2023).

4 DATASET

To ensure technical correctness, we decide to analyze a set of generative models for which pre-trained checkpoints and/or samples of *the same* dataset, namely LSUN Bedroom (Yu et al., 2016) (256×256), are available. Otherwise, both the detectability of generated samples and their spectral properties might suffer from biases, making them difficult to compare. An overview of the dataset is given in Table 1, and we provide details and example images in the appendix.

All samples are either directly downloaded or generated using code and pre-trained models provided by the original publications. We consider data from ten models in total, five GANs and five DMs. This includes the seminal models ProGAN (Karras et al., 2018) and StyleGAN (Karras et al., 2019), as well as the more recent ProjectedGAN (Sauer et al., 2021). Note that Diff(usion)-StyleGAN2 and Diff(usion)-ProjectedGAN (Wang et al., 2022a) (the current state of the art on LSUN Bedroom) use a forward diffusion process to optimize GAN training, but this does not change the GAN model architecture. From the class of DMs, we consider the

Table 1: Models evaluated in this work. Fréchet inception distances (FIDs) on LSUN Bedroom are taken from the original publications and from (Dhariwal and Nichol, 2021) in the case of IDDPM. The lower the FID, the higher the image quality.

Model Class	Method	FID
GAN	ProGAN	8.34
	StyleGAN	2.65
	ProjectedGAN	1.52
	Diff-StyleGAN2	3.65
	Diff-ProjectedGAN	1.43
DM	DDPM	6.36
	IDDPM	4.24
	ADM	1.90
	PNDM	5.68
	LDM	2.95

original DDPM (Ho et al., 2020), its successor ID-DPM (Nichol and Dhariwal, 2021), and ADM (Dhariwal and Nichol, 2021), the latter outperforming several GANs with an FID (Heusel et al., 2017) of 1.90 on LSUN Bedroom. PNDM (Liu et al., 2022) speeds up the sampling process by a factor of 20 using pseudo numerical methods, which can be applied to existing pre-trained DMs. Lastly, LDM (Rombach et al., 2022) uses an adversarially trained autoencoder that transforms an image from the pixel space to a latent space (and back). Training the DM in this more suitable latent space reduces the computational complexity and therefore enables training on higher resolutions. The success of this approach is underpinned by the groundbreaking results of Stable Diffusion, a powerful and publicly available text-to-image model based on LDM.

5 DETECTION ANALYSIS

In this section we analyze how well state-of-the-art fake image detectors can distinguish DM-generated from real images. At first, we apply pre-trained detectors known to be effective against GANs, followed by a study on the generalization abilities of re-trained detectors. Based on our findings, we conduct an in-depth feature space analysis to gain a better understanding on how fake images are detected.

Detection Methods. We evaluate three state-of-the-art CNN-based detectors: Wang2020 (Wang et al., 2020), Gragnaniello2021 (Gragnaniello et al., 2021), and Mandelli2022 (Mandelli et al., 2022). They are supposed to perform well on images from unseen generative models, but it is unclear whether this holds for DM-generated images as well.

Performance Metrics. The performance of the analyzed classifiers is estimated in terms of the widely used area under the receiver operating characteristic curve (AUROC). However, the AUROC is overly optimistic as it captures merely the potential of a classifier, but the optimal threshold is usually unknown (Cozzolino et al., 2021). Thus, we adopt the use of the probability of detection at a fixed false alarm rate (Pd@FAR) as an additional metric, which is given as the true positive rate at a fixed false alarm rate. Intuitively, this corresponds to picking the y-coordinate of the ROC curve given an x-coordinate. This metric is a valid choice for realistic scenarios such as large-scale content filtering on social media, where only a small amount of false positives is tolerable. We consider a fixed false alarm rate of 1 %.

Evaluating Pre-Trained Detectors. At first, we test the performance of the pre-trained detectors based on 20000 samples, equally divided into real and generated images. While Wang2020 and Gragnaniello2021 are trained on images from a single GAN (ProGAN or StyleGAN2), Mandelli2022 is trained on images from a diverse set of generative models. The results in the upper half of Table 2 show that all GAN-generated images can be effectively distinguished from real images, with Gragnaniello2021 yielding the best results. For DM-generated, however, the performance of all detectors significantly drops, on average by 15.2 % AUROC compared to GANs. Although the average AUROC of 91.4 % achieved by the best-performing model Gragnaniello2021 (ProGAN variant) appears promising, we argue that in a realistic setting with 1 % tolerable false positives, detecting only 25.7 % of all fake images is unacceptable.

To verify that our findings are not limited to our dataset, we extend our evaluation to images from DMs trained on other datasets, variations of ADM, and popular text-to-image models. We provide details on these additional datasets in the appendix. The results, given in the lower half of Table 2, support the finding that detectors perform significantly worse on DM-generated images. Images from PNDM and LDM trained on LSUN Church are detected better, which we attribute to a dataset-specific bias.

Generalization of Re-Trained Detectors. Given the findings presented above, the question arises whether DMs evade detection in principle, or whether the detection performance can be increased by re-training a detector. We select the architecture from Wang2020 since the original training code is available and training is relatively efficient. Furthermore, we choose the configuration Blur+JPEG (0.5) as it yields slightly better scores on average. For each of the ten generators, we train a detector according to the authors' instructions, using 78000 samples for training and 2000 samples for validation (equally divided into real (LSUN Bedroom) and fake). We also consider three aggregated settings in which we train on all images generated by GANs, DMs, and both, respectively.

We report AUROC and Pd@1%FAR for each detector evaluated on all datasets in Figure 1, based on 20000 held-out test samples (10000 real and 10000 fake per generator). All detectors achieve near-perfect scores when evaluated on the dataset they were trained on (represented by the values in the diagonal). While this is unsurprising for GANs, it shows that DMs *do* exhibit detectable features that a detector can learn. Regarding generalization, it appears that detectors trained on images from a single DM perform better on images from unseen DMs compared to

Table 2: Detection performance of pre-trained universal detectors. For Wang2020 and Gragnaniello2021, we consider two different variants, respectively. In the upper half, we report the performance of models trained on LSUN Bedroom, while results on additional datasets are given in the second half. The best score (determined by the highest Pd@1%) for each generator is highlighted in bold. We report average scores in gray.

AUROC / Pd@1%	Wang2020		Gragnaniello2021		Mandelli2022	
	Blur+JPEG (0.5)	Blur+JPEG (0.1)	ProGAN	StyleGAN2		
ProGAN	100.0 / 100.0	100.0 / 100.0	100.0 / 100.0	100.0 / 100.0	91.2 / 27.5	
StyleGAN	98.7 / 81.4	99.0 / 84.4	100.0 / 100.0	100.0 / 100.0	89.6 / 14.7	
ProjectedGAN	94.8 / 49.1	90.9 / 34.5	100.0 / 99.3	99.9 / 97.8	59.4 / 2.4	
Diff-StyleGAN2	99.9 / 97.9	100.0 / 99.3	100.0 / 100.0	100.0 / 100.0	100.0 / 99.9	
Diff-ProjectedGAN	93.8 / 43.3	88.8 / 27.2	99.9 / 99.2	99.8 / 96.6	62.1 / 2.8	
Average	97.4 / 74.3	95.7 / 69.1	100.0 / 99.7	99.9 / 98.9	80.4 / 29.5	
DDPM	85.2 / 14.2	80.8 / 9.3	96.5 / 39.1	95.1 / 30.7	57.4 / 0.6	
IDDPM	81.6 / 10.6	79.9 / 7.8	94.3 / 25.7	92.8 / 21.2	62.9 / 1.3	
ADM	68.3 / 3.4	68.8 / 4.0	77.8 / 5.2	70.6 / 2.5	60.5 / 1.8	
PNDM	79.0 / 9.2	75.5 / 6.3	91.6 / 16.6	91.5 / 22.2	71.6 / 4.0	
LDM	78.7 / 7.4	77.7 / 6.9	96.7 / 42.1	97.0 / 48.9	54.8 / 2.1	
Average	78.6 / 9.0	76.6 / 6.8	91.4 / 25.7	89.4 / 25.1	61.4 / 2.0	
ADM (LSUN Cat)	58.4 / 2.5	58.1 / 3.3	60.2 / 4.2	51.7 / 1.8	55.6 / 1.3	
ADM (LSUN Horse)	55.5 / 1.5	53.4 / 2.2	56.1 / 2.7	50.2 / 1.4	44.2 / 0.5	
ADM (ImageNet)	69.1 / 4.1	71.7 / 4.5	72.1 / 3.5	83.9 / 16.6	60.1 / 0.9	
ADM-G-U (ImageNet)	67.2 / 3.7	62.3 / 1.2	66.8 / 1.6	78.9 / 10.2	60.0 / 1.0	
PNDM (LSUN Church)	76.9 / 10.2	77.6 / 12.0	90.9 / 24.5	99.3 / 85.8	56.4 / 1.9	
LDM (LSUN Church)	86.3 / 19.8	82.2 / 14.2	98.8 / 75.5	99.5 / 90.2	58.9 / 1.3	
LDM (FFHQ)	69.4 / 3.6	71.0 / 3.6	91.1 / 25.4	67.2 / 2.1	63.0 / 0.6	
ADM' (FFHQ)	77.7 / 8.7	81.4 / 8.8	87.7 / 17.8	89.0 / 17.2	69.8 / 2.0	
P2 (FFHQ)	79.5 / 8.9	83.2 / 9.2	89.2 / 11.5	91.1 / 18.9	72.5 / 2.7	
Stable Diffusion v1-1	42.4 / 1.5	51.4 / 2.0	73.2 / 4.0	75.2 / 13.6	76.1 / 4.2	
Stable Diffusion v1-5	43.7 / 1.4	52.6 / 2.1	72.9 / 2.8	79.8 / 18.3	75.3 / 4.1	
Stable Diffusion v2-1	46.1 / 1.4	47.3 / 1.1	62.8 / 1.1	55.1 / 1.1	37.0 / 0.5	
Midjourney v5	52.7 / 3.0	57.1 / 3.0	69.9 / 3.3	67.1 / 3.3	18.3 / 0.3	

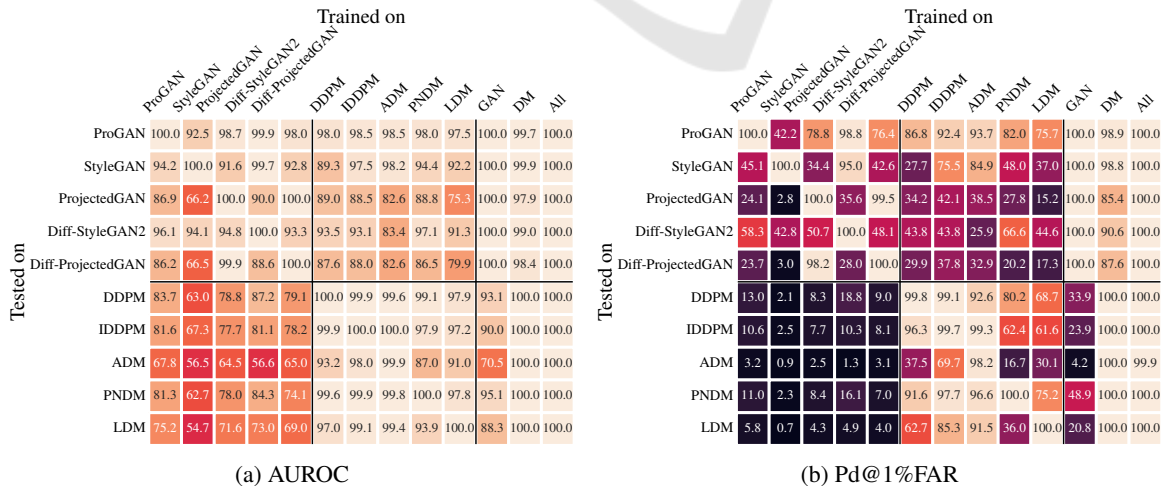


Figure 1: Detection performance for re-trained detectors. The columns *GAN*, *DM*, and *All* correspond to models trained on samples from all GANs, all DMs, and both, respectively.

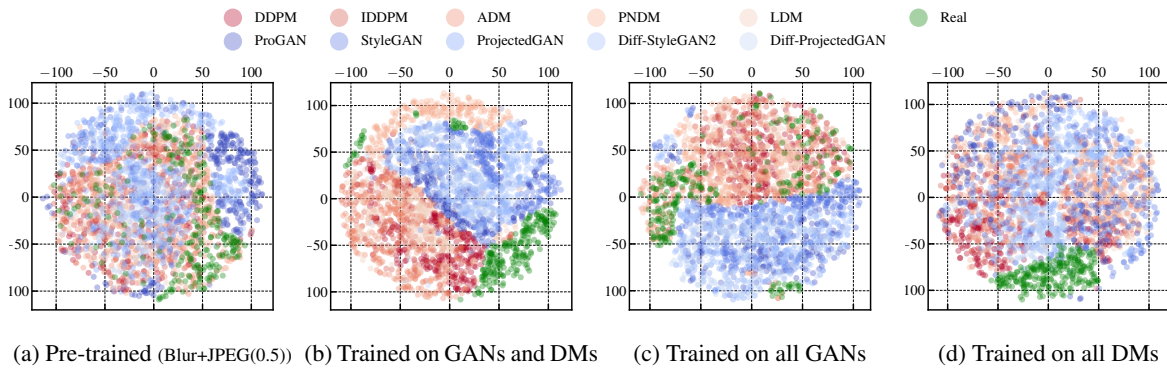


Figure 2: Feature space visualization for the detector Wang2020 via t-SNE of real and generated images in two dimensions. The features correspond to the representation prior to the last fully-connected layer of the given detector.

detectors trained on images from a single GAN. For instance, the detector trained solely on images from ADM achieves a $\text{Pd}@1\%FAR$ greater than 90% for all other DMs. These findings suggest that images generated by DMs not only contain detectable features, but that these are similar across different architectures and training procedures.

Surprisingly, detectors trained on images from DMs are significantly more successful in detecting GAN-generated images than vice versa. This becomes most apparent when analyzing the detectors that are trained on all GANs and DMs, respectively. While the detector trained on images from all GANs achieves an average $\text{Pd}@1\%FAR$ of 26.34% on DM-generated images, the detector trained on images from all DMs on average detects 94.26% of all GAN-generated samples.

Analysis of the Learned Feature Spaces. We conduct a more in-depth analysis of the learned feature spaces to better understand this behavior. We utilize t-SNE (van der Maaten and Hinton, 2008) to visualize the extracted features prior to the last fully-connected layer in Figure 2. For the pre-trained Wang2020 we observe a relatively clear separation between real and GAN-generated images, while there exists a greater overlap between real and DM-generated images (Figure 2a). These results match the classification results from Table 2. Looking at the detector which is trained on DM-generated images only (Figure 2d), the feature representations for GAN- and DM-generated images appear to be similar. In contrast, the detectors trained using GAN-generated images or both (Figures 2c and 2b) seem to learn distinct feature representations for GAN- and DM-generated images.

Based on these results, we argue that the hypothesis, according to which a detector trained on one family of generative models cannot generalize to a different family (Ojha et al., 2023), only holds true “in one direction”. Given the feature space visualizations, de-

ectors trained on GAN-generated images appear to focus mostly on GAN-specific artifacts, which may be more prominent and easier to learn. In contrast, a detector trained exclusively on DM-generated images learns a feature representation in which images generated by GANs and DMs are mapped to similar embeddings. As a consequence, this detector *can* generalize to GAN-generated images, since it is not “distracted” by family-specific patterns, but learns to detect artifacts which are present in both GAN- and DM-generated images.

This also implies that DM-generated images contain fewer family-specific artifacts. This becomes apparent when analyzing them in the frequency domain, which we demonstrate in the following section.

6 FREQUENCY ANALYSIS

For detecting GAN-generated images, exploiting artifacts in the frequency domain has proven to be highly effective (Frank et al., 2020). Since DMs contain related building blocks as GANs (especially up-sampling operations in the underlying U-Net (Ronneberger et al., 2015)), it seems reasonable to suspect that DM-generated exhibit similar artifacts. In this section, we analyze the spectral properties of DM-generated images and compare them to those of GAN-generated images. We investigate potential reasons for the identified frequency characteristics by analyzing the denoising process.

Transforms. We use two frequency transforms that have been applied successfully in both traditional image forensics (Lyu, 2008) and deepfake detection: discrete Fourier transform (DFT) and the reduced spectrum (Durall et al., 2020; Dzanic et al., 2020; Schwarz et al., 2021), which is as a 1D representation of the DFT. While DFT visualizes frequency ar-

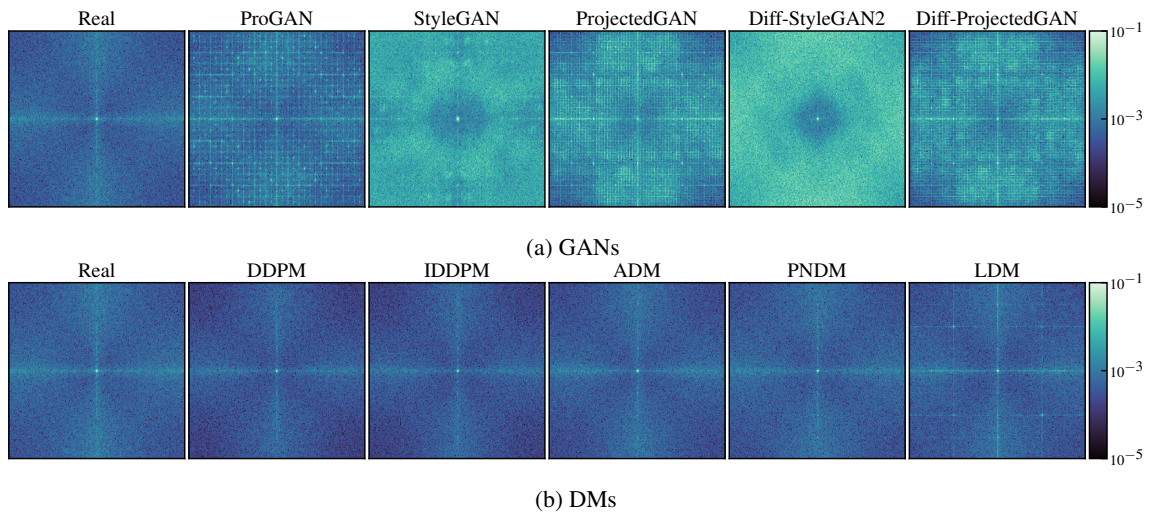


Figure 3: Mean DFT spectrum of real and generated images. To increase visibility, the color bar is limited to $[10^{-5}, 10^{-1}]$, with values lying outside this interval being clipped.

tifacts, the reduced spectrum can be used to identify spectrum discrepancies.

Analysis of Frequency Artifacts. Figure 3 depicts the absolute DFT spectrum averaged over 10000 images from each GAN and DM trained on LSUN Bedroom. Before applying the DFT, images are transformed to grayscale and, following previous works (Marra et al., 2019; Wang et al., 2020), high-pass filtered by subtracting a median-filtered version of the image. For all GANs we observe significant artifacts, predominantly in the form of a regular grid, corresponding to previous findings (Zhang et al., 2019; Frank et al., 2020). In contrast, the DFT spectra of images generated by DMs (see Figure 3b), are significantly more similar to the real spectrum with almost no visible artifacts. LDM is an exception: while being less pronounced than for GANs, generated images exhibit a clearly visible grid across their spectrum. As mentioned in Section 4, the architecture of LDM differs from the remaining DMs as the final image is generated using an adversarially trained autoencoder, which could explain the discrepancies. This observation supports previous findings which suggest that the discriminator is responsible for spectrum deviations (Chen et al., 2021; Schwarz et al., 2021).

We conclude that “traditional” DMs, which generate images by gradual denoising, do *not* produce the frequency artifacts known from GANs. Regarding our results in Section 5, this could explain why detectors trained on GAN images do not generalize to DMs, while training on DM-generated images leads to better generalization.

Analysis of Spectrum Discrepancies. In a second experiment we analyze how well GANs and DMs are

able to reproduce the spectral distribution of real images. We visualize the reduced spectra for all generators in Figure 4, again averaged over 10000 images. Except for Diff-StyleGAN2, all GANs contain the previously reported elevated high frequencies. Among the DMs, these can only be observed for LDM. This strengthens the hypothesis that it is the autoencoder which causes GAN-like frequency characteristics. However, we observe that all DMs have a tendency to underestimate the spectral density towards the higher end of the frequency spectrum. This is particularly noticeable for DDPM, IDDPM, and ADM.

Source of Spectrum Underestimation. Based on these findings, we conduct an additional experiment to identify the source of this spectrum underestimation. Since DMs generate images via gradual denoising, we analyze how the spectrum evolves during this denoising process. For this experiment, we use code and model from ADM (Dhariwal and Nichol, 2021) trained on LSUN Bedroom. We generate samples at different time steps t and compare the reduced spectrum (averaged over 512 images) to that of 50000 real images. The results are shown in Figure 5.

We adopt the figure type from (Schwarz et al., 2021) and depict the relative spectral density error $\tilde{S}_{\text{err}} = \tilde{S}_{\text{fake}}/\tilde{S}_{\text{real}} - 1$, with the colorbar clipped at -1 and 1. At $t = T = 1000$, the image is pure Gaussian noise, which naturally causes strong spectrum deviations. Around $t = 300$, the error starts to decrease, but interestingly it appears that the optimum is not reached at $t = 0$, but at $t \approx 10$. It should be noted that while at this step the frequency spectrum is closest to that of real images, they still contain visible noise.

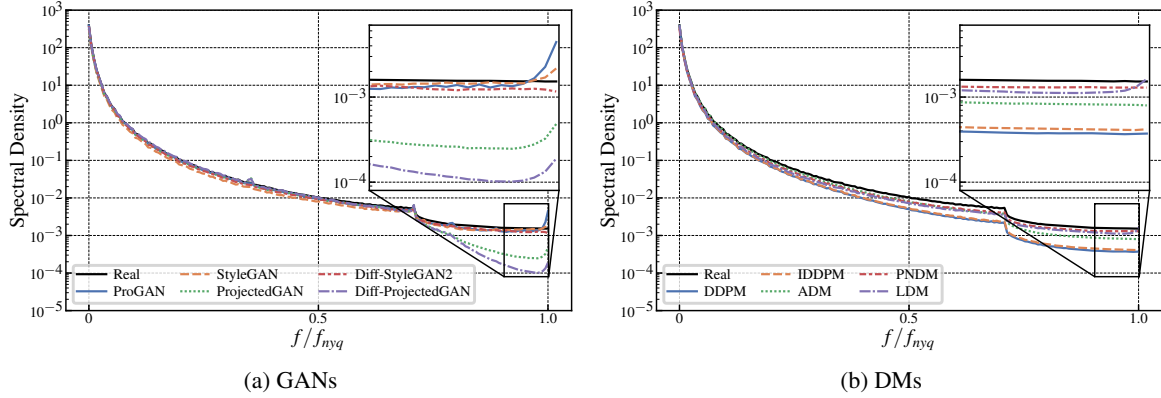


Figure 4: Mean reduced spectrum of real and generated images. The part of the spectrum where GAN-characteristic discrepancies occur is magnified.

During the final denoising steps, \tilde{S}_{err} becomes *negative*, predominantly for higher frequencies, which corresponds to our observations in Figure 4b.

We hypothesize that this underestimation towards higher frequencies stems from the learning objective used to train DMs. Recalling Section 3, DMs are trained to minimize the MSE between the true and predicted noise at different time steps. The weighting of the MSE therefore controls the relative importance of each step. While the semantic content of an image is generated early during the denoising process, high-frequency details are synthesized near $t = 0$ (Kingma et al., 2021). Theoretically, using the variational lower bound L_{vlb} as the training objective would yield the highest log-likelihood. However, training DMs

with L_{vlb} is difficult (Ho et al., 2020; Nichol and Dhariwal, 2021), which is why in practice modified objectives are used. The loss proposed in (Ho et al., 2020), $L_{\text{simple}} = \mathbb{E}_{t, \mathbf{x}_0, \varepsilon} [\|\varepsilon - \varepsilon_\theta(\mathbf{x}_t, t)\|^2]$, for example, considers each denoising step as equally important. Compared to L_{vlb} , the steps near $t = 0$ are significantly down-weighted, trading off a higher perceptual image quality for higher log-likelihood values. The MSE of ADM over t shown in Figure 6 demonstrates that the final denoising steps are the most difficult (which is already plain to see as the signal-to-noise ratio increases for $t \rightarrow 0$, i.e., the to-be-predicted noise makes up ever smaller fractions of \mathbf{x}_t). The hybrid training objective $L_{\text{hybrid}} = L_{\text{simple}} + \lambda L_{\text{vlb}}$ (Nichol and Dhariwal, 2021), used in IDDPM and ADM, incorporates L_{vlb} (with $\lambda = 0.001$) and already improves upon DDPM in modeling the high-frequency details of an image, but still does not match it accurately.

In summary, we conclude that the denoising steps near $t = 0$, which govern the high-frequency content of generated images, are the most difficult to model. By down-weighting the importance of these steps (relatively to the L_{vlb}), DMs achieve remarkable perceptual image quality (or benchmark metrics such as FID), but seem to fall short of accurately matching the high-frequency distribution of real data.

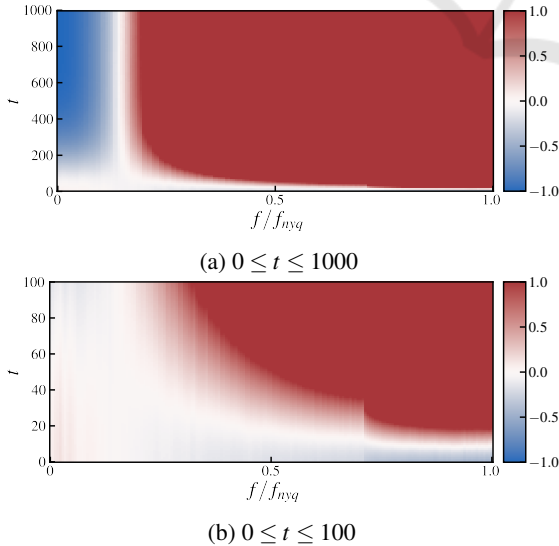


Figure 5: Spectral density error \tilde{S}_{err} throughout the denoising process. The error is computed relative to the spectrum of real images. We display the error for (a) all sampling steps and (b) a close-up of the last 100 steps. The colorbar is clipped at -1 and 1.

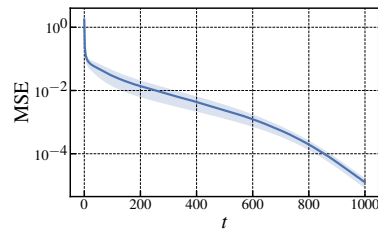


Figure 6: Mean and standard deviation of the MSE for ADM on LSUN Bedroom after training. The denoising steps towards $t = 0$, accounting for high frequencies, have a higher error.

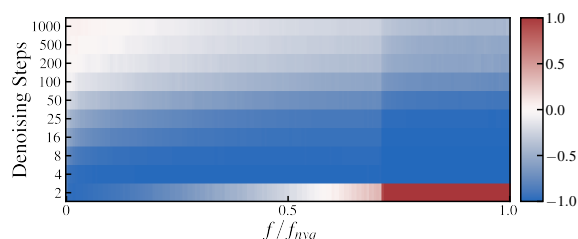


Figure 7: Spectral density error \tilde{S}_{err} for different numbers of denoising steps. The error is computed relatively to the spectrum of real images. The colorbar is clipped at -1 and 1. Note that the y-axis is not scaled linearly.

Effect of the Number of Sampling Steps. Lastly, we analyze how the number of sampling steps during the denoising process affects the frequency spectrum. Previous work reported that increasing the number of steps leads to an improved log-likelihood, corresponding to better reproduction of higher frequencies (Nichol and Dhariwal, 2021). Our results in Figure 7 confirm these findings, increasing the number of denoising steps reduces the underestimation.

7 CONCLUSION

Deepfakes pose a severe risk for society, and diffusion models have the potential to raise disinformation campaigns to a new level. Despite the urgency of the problem, research about detecting DM-generated images is still in its infancy. In this work, we provide a much-needed step towards the detection of DM deepfakes. Instead of starting from the ground up, we build on previous achievements in the forensic analysis of GANs. We show that, after re-training, current state-of-the-art detection methods can successfully distinguish real from DM-generated images. Further analysis suggests that DMs produce fewer detectable artifacts than GANs, explaining why detectors trained on DM-generated images generalize to GANs, but not vice versa. While artifacts in the frequency domain have been shown to be a characteristic feature of GAN-generated images, we find that DMs predominantly do not have this weakness. However, we observe a systematic underestimation of the spectral density, which we attribute to the loss function of DMs. Whether this mismatch can be exploited for novel detection methods should be part of future research. We hope that our work can foster the forensic analysis of images generated by DMs and spark further research towards the effective detection of deepfakes.

ACKNOWLEDGEMENTS

Funded by the Deutsche Forschungsgemeinschaft (DFG, German Research Foundation) under Germany's Excellence Strategy - EXC 2092 CASA - 390781972.

REFERENCES

- Chai, L., Bau, D., Lim, S.-N., and Isola, P. (2020). What makes fake images detectable? Understanding properties that generalize. In *European Conference on Computer Vision (ECCV)*.
- Chandrasegaran, K., Tran, N.-T., and Cheung, N.-M. (2021). A closer look at Fourier spectrum discrepancies for CNN-generated images detection. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*.
- Chen, Y., Li, G., Jin, C., Liu, S., and Li, T. (2021). SSD-GAN: Measuring the realness in the spatial and spectral domains. In *AAAI Conference on Artificial Intelligence (AAAI)*.
- Choi, J., Lee, J., Shin, C., Kim, S., Kim, H., and Yoon, S. (2022). Perception prioritized training of diffusion models. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*.
- Corvi, R., Cozzolino, D., Poggi, G., Nagano, K., and Verdoliva, L. (2023a). Intriguing properties of synthetic images: From generative adversarial networks to diffusion models. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR) Workshops*.
- Corvi, R., Cozzolino, D., Zingarini, G., Poggi, G., Nagano, K., and Verdoliva, L. (2023b). On the detection of synthetic images generated by diffusion models. In *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*.
- Cozzolino, D., Gragnaniello, D., Poggi, G., and Verdoliva, L. (2021). Towards universal GAN image detection. In *International Conference on Visual Communications and Image Processing (VCIP)*.
- Dhariwal, P. and Nichol, A. (2021). Diffusion models beat GANs on image synthesis. In *Advances in Neural Information Processing Systems (NeurIPS)*.
- Dosovitskiy, A., Beyer, L., Kolesnikov, A., Weissenborn, D., Zhai, X., Unterthiner, T., Dehghani, M., Minderer, M., Heigold, G., Gelly, S., Uszkoreit, J., and Houlsby, N. (2021). An image is worth 16x16 words: Transformers for image recognition at scale. *International Conference on Learning Representations (ICLR)*.
- Durall, R., Keuper, M., and Keuper, J. (2020). 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)*.
- Dzanic, T., Shah, K., and Witherden, F. (2020). Fourier spectrum discrepancies in deep network generated images. In *Advances in Neural Information Processing Systems (NeurIPS)*.

- Farid, H. (2022a). Lighting (in)consistency of paint by text. *arXiv preprint*.
- Farid, H. (2022b). Perspective (in)consistency of paint by text. *arXiv preprint*.
- Frank, J., Eisenhofer, T., Schönherr, L., Fischer, A., Kolossa, D., and Holz, T. (2020). Leveraging frequency analysis for deep fake image recognition. In *International Conference on Machine Learning (ICML)*.
- Girish, S., Suri, S., Rambhatla, S. S., and Shrivastava, A. (2021). Towards discovery and attribution of open-world GAN generated images. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*.
- Gagnaniello, D., Cozzolino, D., Marra, F., Poggi, G., and Verdoliva, L. (2021). Are GAN generated images easy to detect? A critical analysis of the state-of-the-art. In *IEEE International Conference on Multimedia and Expo (ICME)*.
- Heusel, M., Ramsauer, H., Unterthiner, T., Nessler, B., and Hochreiter, S. (2017). GANs trained by a two time-scale update rule converge to a local nash equilibrium. In *Advances in Neural Information Processing Systems (NeurIPS)*.
- Ho, J., Jain, A., and Abbeel, P. (2020). Denoising diffusion probabilistic models. In *Advances in Neural Information Processing Systems (NeurIPS)*.
- Hu, S., Li, Y., and Lyu, S. (2021). Exposing GAN-Generated faces using inconsistent corneal specular highlights. In *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*.
- Huang, K. (2023). Why Pope Francis is the star of A.I.-generated photos. *The New York Times*.
- Jeong, Y., Kim, D., Ro, Y., Kim, P., and Choi, J. (2022). FingerprintNet: Synthesized fingerprints for generated image detection. In *European Conference on Computer Vision (ECCV)*.
- Karras, T., Aila, T., Laine, S., and Lehtinen, J. (2018). Progressive growing of GANs for improved quality, stability, and variation. In *International Conference on Learning Representations (ICLR)*.
- Karras, T., Laine, S., and Aila, T. (2019). A style-based generator architecture for generative adversarial networks. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*.
- Khayatkhoei, M. and Elgammal, A. (2022). Spatial frequency bias in convolutional generative adversarial networks. *AAAI Conference on Artificial Intelligence (AAAI)*.
- Kingma, D., Salimans, T., Poole, B., and Ho, J. (2021). Variational diffusion models. In *Advances in Neural Information Processing Systems (NeurIPS)*.
- Kingma, D. P. and Gao, R. (2023). Understanding diffusion objectives as the ELBO with simple data augmentation. In *Advances in Neural Information Processing Systems (NeurIPS)*.
- Liu, L., Ren, Y., Lin, Z., and Zhao, Z. (2022). Pseudo numerical methods for diffusion models on manifolds. In *International Conference on Learning Representations (ICLR)*.
- Lyu, S. (2008). *Natural Image Statistics in Digital Image Forensics*. PhD thesis, Dartmouth College.
- Mandelli, S., Bonettini, N., Bestagini, P., and Tubaro, S. (2022). Detecting GAN-generated images by orthogonal training of multiple CNNs. In *IEEE International Conference on Image Processing (ICIP)*.
- Marra, F., Gagnaniello, D., Verdoliva, L., and Poggi, G. (2019). Do GANs leave artificial fingerprints? In *IEEE Conference on Multimedia Information Processing and Retrieval (MIPR)*.
- McCloskey, S. and Albright, M. (2019). Detecting GAN-generated imagery using saturation cues. In *IEEE International Conference on Image Processing (ICIP)*.
- Nataraj, L., Mohammed, T. M., Manjunath, B. S., Chandrasekaran, S., Flenner, A., Bappy, J. H., and Roy-Chowdhury, A. K. (2019). Detecting GAN generated fake images using co-occurrence matrices. *Electronic Imaging*.
- Nichol, A. Q. and Dhariwal, P. (2021). Improved denoising diffusion probabilistic models. In *International Conference on Machine Learning (ICML)*.
- Nightingale, S. J. and Farid, H. (2022). AI-synthesized faces are indistinguishable from real faces and more trustworthy. *Proceedings of the National Academy of Sciences*.
- Ojha, U., Li, Y., and Lee, Y. J. (2023). Towards universal fake image detectors that generalize across generative models. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*.
- Radford, A., Kim, J. W., Hallacy, C., Ramesh, A., Goh, G., Agarwal, S., Sastry, G., Askell, A., Mishkin, P., Clark, J., Krueger, G., and Sutskever, I. (2021). Learning transferable visual models from natural language supervision. In *International Conference on Machine Learning (ICML)*.
- Ramesh, A., Dhariwal, P., Nichol, A., Chu, C., and Chen, M. (2022). Hierarchical text-conditional image generation with CLIP latents. *arXiv preprint*.
- Rissanen, S., Heinonen, M., and Solin, A. (2023). Generative modelling with inverse heat dissipation. In *International Conference on Learning Representations (ICLR)*.
- Rombach, R., Blattmann, A., Lorenz, D., Esser, P., and Ommer, B. (2022). High-resolution image synthesis with latent diffusion models. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*.
- Ronneberger, O., Fischer, P., and Brox, T. (2015). U-Net: Convolutional networks for biomedical image segmentation. In *International Conference on Medical Image Computing and Computer Assisted Intervention*.
- Russakovsky, O., Deng, J., Su, H., Krause, J., Satheesh, S., Ma, S., Huang, Z., Karpathy, A., Khosla, A., Bernstein, M., Berg, A. C., and Fei-Fei, L. (2015). ImageNet large scale visual recognition challenge. *International Journal of Computer Vision (IJCV)*.
- Saharia, C., Chan, W., Saxena, S., Li, L., Whang, J., Denton, E., Ghasemipour, S. K. S., Ayan, B. K., Mahdavi, S. S., Lopes, R. G., Salimans, T., Ho, J., Fleet, D. J., and Norouzi, M. (2022). Photorealistic text-to-image

- diffusion models with deep language understanding. *arXiv preprint*.
- Salimans, T. and Ho, J. (2022). Progressive distillation for fast sampling of diffusion models. In *International Conference on Learning Representations (ICLR)*.
- Sauer, A., Chitta, K., Müller, J., and Geiger, A. (2021). Projected GANs converge faster. In *Advances in Neural Information Processing Systems (NeurIPS)*.
- Schwarz, K., Liao, Y., and Geiger, A. (2021). On the frequency bias of generative models. In *Advances in Neural Information Processing Systems (NeurIPS)*.
- Sha, Z., Li, Z., Yu, N., and Zhang, Y. (2023). DE-FAKE: Detection and attribution of fake images generated by text-to-image diffusion models. *ACM SIGSAC Conference on Computer and Communications Security (CCS)*.
- Sohl-Dickstein, J., Weiss, E., Maheswaranathan, N., and Ganguli, S. (2015). Deep unsupervised learning using nonequilibrium thermodynamics. In *International Conference on Machine Learning (ICML)*.
- Song, J., Meng, C., and Ermon, S. (2022a). Denoising diffusion implicit models. In *International Conference on Learning Representations (ICLR)*.
- Song, Y. and Ermon, S. (2019). Generative modeling by estimating gradients of the data distribution. In *Advances in Neural Information Processing Systems (NeurIPS)*.
- Song, Y. and Ermon, S. (2020). Improved techniques for training score-based generative models. In *Advances in Neural Information Processing Systems (NeurIPS)*.
- Song, Y., Sohl-Dickstein, J., Kingma, D. P., Kumar, A., Ermon, S., and Poole, B. (2022b). Score-based generative modeling through stochastic differential equations. In *International Conference on Learning Representations (ICLR)*.
- van der Maaten, L. and Hinton, G. (2008). Visualizing data using t-SNE. *Journal of Machine Learning Research (JMLR)*.
- Verdoliva, L. (2020). Media forensics and DeepFakes: An overview. *IEEE Journal of Selected Topics in Signal Processing*.
- Wang, S.-Y., Wang, O., Zhang, R., Owens, A., and Efros, A. A. (2020). CNN-generated images are surprisingly easy to spot... for now. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*.
- Wang, Z., Bao, J., Zhou, W., Wang, W., Hu, H., Chen, H., and Li, H. (2023). DIRE for diffusion-generated image detection. *IEEE International Conference on Computer Vision (ICCV)*.
- Wang, Z., Zheng, H., He, P., Chen, W., and Zhou, M. (2022a). Diffusion-GAN: Training GANs with diffusion. *arXiv preprint*.
- Wang, Z. J., Montoya, E., Munechika, D., Yang, H., Hoover, B., and Chau, D. H. (2022b). DiffusionDB: A large-scale prompt gallery dataset for text-to-image generative models. *arXiv preprint*.
- Xiao, Z., Kreis, K., and Vahdat, A. (2022). Tackling the generative learning trilemma with denoising diffusion GANs. In *International Conference on Learning Representations (ICLR)*.
- Xuan, X., Peng, B., Wang, W., and Dong, J. (2019). On the generalization of GAN image forensics. In *Biometric Recognition (CCBR)*.
- Yang, L., Zhang, Z., Song, Y., Hong, S., Xu, R., Zhao, Y., Zhang, W., Cui, B., and Yang, M.-H. (2023). Diffusion models: A comprehensive survey of methods and applications. *ACM Computing Surveys*.
- Yu, F., Seff, A., Zhang, Y., Song, S., Funkhouser, T., and Xiao, J. (2016). LSUN: Construction of a large-scale image dataset using deep learning with humans in the loop. *arXiv preprint*.
- Zhang, X., Karaman, S., and Chang, S.-F. (2019). Detecting and simulating artifacts in GAN fake images. In *IEEE International Workshop on Information Forensics and Security (WIFS)*.

APPENDIX

Details on LSUN Bedroom Dataset

LSUN Bedroom (Yu et al., 2016). We download and extract the lmbd database files using the official repository¹. The images are center-cropped to 256×256 pixels.

ProGAN (Karras et al., 2018). We download the first 10000 samples from the non-curated collection provided by the authors.²

StyleGAN (Karras et al., 2019). We download the first 10000 samples generated with $\psi = 0.5$ from the non-curated collection provided by the authors.³

ProjectedGAN (Sauer et al., 2021). We sample 10000 images using code and pre-trained models provided by the authors using the default configuration (`--trunc=1.0`).⁴

Diff-StyleGAN2 and Diff-ProjectedGAN (Wang et al., 2022a). We sample 10000 images using code and pre-trained models provided by the authors using the default configuration.⁵

DDPM (Ho et al., 2020), IDDP (Nichol and Dhariwal, 2021), and ADM (Dhariwal and Nichol, 2021). We download the samples provided by the authors of ADM⁶ and extract the first 10000 samples for each generator. For ADM on LSUN, we select the models trained with dropout.

¹<https://github.com/fyu/lsun>

²https://github.com/tkarras/progressive_growing_of_gans

³<https://github.com/NVLabs/stylegan>

⁴https://github.com/autonomousvision/projected_gan

⁵<https://github.com/Zhendong-Wang/Diffusion-GAN>

⁶<https://github.com/openai/guided-diffusion>

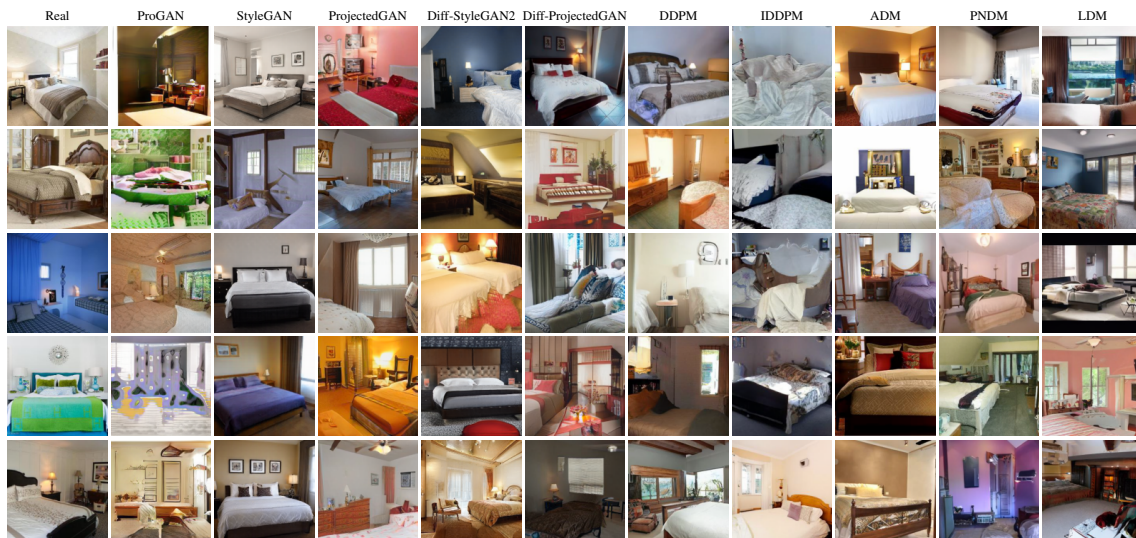


Figure 8: Non-curated example images for real LSUN Bedroom, GAN-generated, and DM-generated images.

PNDM (Liu et al., 2022). We sample 10000 images using code and pre-trained model provided by the authors.⁷ We specify `--method F-PNDM` and `--sample_speed 20` for LSUN Bedroom and `--sample_speed 10` for LSUN Church, as these are the settings leading to the lowest FID according to Tables 5 and 6 in the original publication.

LDM (Rombach et al., 2022). We sample 10000 images using code and pre-trained models provided by the authors using settings from the corresponding table in the repository.⁸ For LSUN Church there is an inconsistency between the repository and the paper, we choose 200 DDIM steps (`-c 200`) as reported in the paper.

Details on Additional Datasets

Here we provide details on the additional datasets analyzed in Table 2. Note that ADM-G-U refers to the two-stage up-sampling stack in which images are generated at a resolution of 64×64 and subsequently up-sampled to 256×256 pixels using a second model (Dhariwal and Nichol, 2021). The generated images are obtained according to the instructions given in the previous section.

Due to the relevance of facial images in the context of deepfakes, we also include two DMs not yet considered, P2 and ADM' (Choi et al., 2022), trained on FFHQ (Karras et al., 2019). ADM' is a smaller version of ADM with 93 million instead of more than 500 million parameters.⁹ P2 is similar to ADM'

⁷<https://github.com/luping-liu/PNDM>

⁸<https://github.com/CompVis/latent-diffusion>

⁹<https://github.com/jychoi118/P2-weighting#>

but features a modified weighting scheme which improves performance by assigning higher weights to diffusion steps where perceptually rich contents are learned (Choi et al., 2022). We download checkpoints for both models from the official repository and sample images according to the authors' instructions.

Real images from LSUN (Yu et al., 2016), ImageNet (Russakovsky et al., 2015), and FFHQ (Karras et al., 2019) are downloaded from their official sources. Images from LSUN Cat/Horse, FFHQ, and ImageNet are resized and cropped to 256×256 pixels by applying the same pre-processing that was used when preparing the training data for the model they are compared against. For all datasets we collect 10000 real and 10000 generated images.

Images from Stable Diffusion¹⁰ are generated using the *diffusers* library¹¹ with default settings. For each version, we generate 10000 images using prompts from DiffusionDB (Wang et al., 2022b). Since Midjourney¹² is proprietary, we collect 300 images created using the `"-v 5"` flag from the official Discord server. As real images, we take a subset of 10000 images from LAION-Aesthetics V2¹³ with aesthetics scores greater than 6.5. For the detection experiments, we use the entire images, for computing frequency spectra we take center crops of size 256×256 .

training-your-models

¹⁰<https://stability.ai/blog/stable-diffusion-public-release>

¹¹<https://huggingface.co/docs/diffusers/index>

¹²<https://www.midjourney.com>

¹³<https://laion.ai/blog/laion-aesthetics/>