# TOWARDS THE DETECTION OF DIFFUSION MODEL DEEPFAKES

**Anonymous authors**
Paper under double-blind review

## ABSTRACT

Diffusion models (DMs) have recently emerged as a promising method in image synthesis. They have surpassed generative adversarial networks (GANs) in both diversity and quality, and have achieved impressive results in text-to-image modeling. However, to date, only little attention has been paid to the *detection* of DM-generated images, which is critical to prevent adverse impacts on our society. While prior work has shown that GAN-generated images can be reliably detected using automated methods, it is unclear whether the same methods are effective against DMs. In this work, we address this challenge and take a first look at detecting DM-generated images. We approach the problem from two different angles: First, we evaluate the performance of state-of-the-art detectors on a variety of DMs. Second, we analyze DM-generated images in the frequency domain and study different factors that influence the spectral properties of these images. Most importantly, we demonstrate that GANs and DMs produce images with different characteristics, which requires adaptation of existing classifiers to ensure reliable detection. We believe this work provides the foundation and starting point for further research to detect DM deepfakes effectively.

## 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 allow 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 & 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 the public perception, to create sentiment for or against certain social groups, and thus influence public opinion. In their effect, deepfakes lead to an erosion of trust in institutions or individuals, support conspiracy theories, and promote a fundamental political camp formation. Despite the importance of this topic, there is only a limited amount of research on effective deepfake detection. Previous work on the detection of GAN-generated images (e.g., Wang et al. (2020), Gragnaniello et al. (2021), and Mandelli et al. (2022a)) showed promising results, but it remains unclear if any of these methods can be applied to DM-generated images.

In this paper, we present the first look at detection methods for DM-generated media. We tackle the problem from two different angles. On the one hand, we investigate whether DM-generated images can be effectively detected by existing methods that claim to be universal. We study ten models in total, five GANs and five DMs. We find that existing detection methods suffer from severe performance degradation when applied on DM-generated images, with the area under the receiver operating characteristic curve (AUROC) metric dropping by 15.2% on average compared to GANs. These results hint at a structural difference between synthetic images generated by GANs and DMs. We show that existing detection methods can be improved by fine-tuning, the detection is then almost perfect. However, our results also suggest that recognizing DM-generated images is a more difficult task than recognizing GAN images.

On the other hand, we analyze DM-generated images in the frequency domain and compare them to GAN-generated images. Although DMs do not exhibit strong frequency artifacts compared to GANs, their spectrum deviates from real images. We hypothesize that discrepancies in spectral properties are a possible reason for our identified differences. Therefore, we analyze the spectral properties of DM- and GAN-generated images in detail and find that high frequencies are systematically underestimated. Further analysis suggests that too little weight is given to these frequencies during training due to the choice of the training objective. We believe that our results provide the foundation for further research on the effective detection of deepfakes generated by DMs.

## 2 Related Work

**Universal Fake Image Detection** While in recent years a variety of successful methods to detect artificially generated images has been proposed (Verdoliva, 2020), generalization to unseen data remains a challenging task (Cozzolino et al., 2019). Constructing an effective classifier for a specific generator is considered straightforward, which is why more research effort is put into designing universal detectors (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., 2022a). This is especially important in the context of deceptive media, since new generative models emerge on a frequent basis and manually updating detectors is too slow to stop the propagation of harmful contents.

**Frequency-Based Deepfake Detection** Zhang et al. (2019) 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. These findings were later reproduced by Wang et al. (2020) and extended to the discrete cosine transform (DCT) by Frank et al. (2020). Another characteristic was discovered by 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 & Elgammal, 2022), while others suggest the discriminator is not able to provide an accurate training signal (Chen et al., 2021; Schwarz et al., 2021).

**Detection of DM-Generated Images** Despite the massive attention from the scientific community and beyond, DMs have not yet been studied from the perspective of image forensics. A very specific use case is considered by Mandelli et al. (2022b), where the authors evaluate methods for detecting western blot images synthesized by different models, including DDPM (Ho et al., 2020). More related to our analysis is the work of Wolter et al. (2022), which proposes to detect generated images based on their wavelet-packet representation, combining features from pixel- and frequency space. While the focus lies on GAN-generated images, they demonstrate that the images generated by ADM (Dhariwal & Nichol, 2021) can be detected using their approach. They also report that the classifier "appears to focus on the highest frequency packet", which is in accordance with our findings.

**Frequency Considerations in DMs** Both Kingma et al. (2021) and 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. Another interesting observation is made by Rissanen et al. (2022) who analyze the generative process of diffusion models in the frequency domain. They state that diffusion models have an inductive bias according to which, during the reverse process, higher frequencies are added to existing lower frequencies.

## 3 Background on DMs

DMs were first proposed by Sohl-Dickstein et al. (2015) and later advanced by Ho et al. (2020), who also pointed out the connections between DMs and score-based generative models (Song & Ermon, 2019; 2020; Song et al., 2022b). Since then, numerous modifications and improvements have been proposed, leading to higher perceptual quality (Nichol & Dhariwal, 2021; Dhariwal & Nichol, 2021; Choi et al., 2022; Rombach et al., 2022) and increased sampling speed (Song et al.,

2022a; Liu et al., 2022; Salimans & Ho, 2022; Xiao et al., 2022). In short, a DM models a data distribution by gradually disturbing a sample from this distribution and then learning to reverse this diffusion process. In the diffusion (or forward) process, a sample $\mathbf{x}_0$ (i.e., an image in most applications) is repeatedly corrupted by adding Gaussian noise in sequential steps $t = 1, \dots, T$:

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

where $\beta_t$ defines a noise schedule. With $\alpha_t \coloneqq 1 - \beta_t$ and $\bar{\alpha}_t \coloneqq \prod_{s=1}^{t}(1-\beta_t)$, we can directly sample a corrupted sample $x_t$ via

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

During the denoising (or reverse) process, we aim to sample from $q(\mathbf{x}_{t-1}|\mathbf{x}_t)$ to obtain a clean image given $\mathbf{x}_T$. 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

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

where $\mu_\theta$ and $\Sigma_\theta$ are given by the output of the model (or the latter is set to a constant as proposed by Ho et al. (2020)). Predicting the mean of the denoised sample $\mu_\theta(\mathbf{x}_t, t)$ is conceptually equivalent to predicting the noise via $\epsilon_\theta(\mathbf{x}_t, t)$ which should be removed. Predominantly, the latter approach is implemented (e.g., by Ho et al. (2020); Dhariwal & Nichol (2021)) such that training a DM boils down to minimizing a (weighted) mean-squared error $\|\epsilon - \epsilon_\theta(\mathbf{x}_t, t)\|^2$ between the true and predicted noise.

## 4 Generative Models Under Evaluation

In contrast to previous work, we evaluate all generative models trained on the same dataset. This avoids any biases due to the complexity of the dataset when estimating the accuracy of deep fake detection methods, e.g., a high detection accuracy for a model trained on a complex dataset that cannot be modeled as well as a simpler dataset. It also provides fair conditions for evaluating a fine-tuned detector's ability to generalize to unseen data. Besides, having only a single dataset is beneficial for frequency analysis. Although natural images have been shown to have similar power spectra across different domains (Burton & Moorhead, 1987; Field, 1987; 1999; Tolhurst et al., 1992; Torralba & Oliva, 2003), a model may be able to reproduce the frequencies of one dataset better than another.

We evaluate models trained on LSUN Bedroom (Yu et al., 2016) due to the availability of pre-trained models and/or samples for many state-of-the-art GANs and DMs. This dataset also has the advantage that no resizing operations (besides cropping)[1] are required, which are a potential source of error due to faulty implementations (Parmar et al., 2022) or a mismatch between pre-processing for real and generated images (Chai et al., 2020). To demonstrate that our findings are not dataset specific, we show results on additional data in Appendix D.

Table 1: **Models evaluated in this work.** Fréchet inception distances (FIDs) (Heusel et al., 2017) on LSUN Bedroom 256×256 are taken from the original publications and from Dhariwal & Nichol (2021) in the case of IDDPM.

| Model Class | Method | Publication | FID |
|---|---|---|---|
| GAN | ProGAN | Karras et al. (2018) | 8.34 |
| | StyleGAN | Karras et al. (2019) | 2.65 |
| | ProjectedGAN | Sauer et al. (2021) | 1.52 |
| | Diff-StyleGAN2 | Wang et al. (2022) | 3.65 |
| | Diff-ProjectedGAN | Wang et al. (2022) | 1.43 |
| DM | DDPM | Ho et al. (2020) | 6.36 |
| | IDDPM | Nichol & Dhariwal (2021) | 4.24 |
| | ADM | Dhariwal & Nichol (2021) | 1.90 |
| | PNDM | Liu et al. (2022) | 5.68 |
| | LDM | Rombach et al. (2022) | 2.95 |

An overview of the models under evaluation is shown in Table 1, and example images are given in Appendix A. We consider data from ten models in total, five GANs and five DMs. This includes the seminal models ProGAN and StyleGAN, as well as the more recent ProjectedGAN. Note that the recently proposed methods, Diff(usion)-StyleGAN2 and Diff(usion)-ProjectedGAN (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 original DDPM, its successors IDDPM, and ADM, the latter outperforming several GANs with

[1] All images have the smaller dimension equal to 256 pixels.

an FID of 1.90 on LSUN Bedroom. PNDM optimizes the sampling process by the factor of 20 using pseudo numerical methods, which can be applied to existing pre-trained DMs. Lastly, LDM uses an adversarially trained autoencoder that transforms an image from the pixel space to a latent space. Training the DM in this more suitable latent space significantly reduces the computational complexity and therefore enables training on higher resolutions. The success of this approach is underpinned by the recent publication of Stable Diffusion[2], a powerful and publicly available text-to-image model, which is based on LDM. An interesting detail is the fact that the autoencoder used to transform images is very similar to a VQGAN (Esser et al., 2021).

All samples have dimensions 256×256 and are either directly downloaded or generated using code and pre-trained models provided by the original publications. More detailed descriptions are given in Appendix A. For each model, we collect 50k samples in total from which 39k are used for training, 1k for validation, and 10k for testing and frequency analysis, if not stated otherwise.

## 5 Detection of DM-Generated Images

We now investigate the effectiveness of state-of-the-art fake image detectors on DM-generated images. First, we analyze the performance of off-the-shelf pre-trained models which are highly capable of detecting images generated by GANs and then perform experiments with models fine-tuned on DM-generated images.

In this work, we evaluate three state-of-the-art detection methods by Wang et al. (2020), Gragnaniello et al. (2021), and Mandelli et al. (2022a). A description of these detectors is provided in Appendix B.1. All three claim to perform well on unseen data, but it is unclear whether this holds for DM-generated images as well.

The performance of the analyzed classifiers is estimated in terms of the widely used AUROC. As pointed out by Cozzolino et al. (2021), however, the AUROC tends to be overly optimistic since it captures merely the potential of a classifier, but the optimal threshold is usually unknown. Therefore, 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 positive 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 certain amount of false positives is tolerable. We consider FARs of 5% and 1%.

### 5.1 Can we use State-of-the-Art Detectors for DM-Generated Images?

Our first analysis aims to answer the question whether DM-generated images are sufficiently similar, with respect to detectable traces, to those generated by GANs.

Table 2: **Detection performance of pre-trained universal detectors.** For Wang et al. (2020) and Gragnaniello et al. (2021), we consider two different variants, respectively. The best score (determined by the highest Pd@1%) for each generator is highlighted in **bold**. We also report average scores per detector and model class in gray.

| AUROC / Pd@5% / Pd@1% | Wang et al. (2020) | | Gragnaniello et al. (2021) | | Mandelli et al. (2022a) |
|---|---|---|---|---|---|
| | 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 / 100.0** | **100.0 / 100.0 / 100.0** | 91.2 / 54.6 / 27.5 |
| StyleGAN | 98.7 / 93.7 / 81.4 | 99.0 / 95.5 / 84.4 | **100.0 / 100.0 / 100.0** | **100.0 / 100.0 / 100.0** | 89.6 / 43.6 / 14.7 |
| ProjectedGAN | 94.8 / 73.8 / 49.1 | 90.9 / 61.8 / 34.5 | **100.0 / 99.9 / 99.3** | 99.9 / 99.6 / 97.8 | 59.4 / 8.4 / 2.4 |
| Diff-StyleGAN2 | 99.9 / 99.6 / 97.9 | 100.0 / 99.9 / 99.3 | **100.0 / 100.0 / 100.0** | **100.0 / 100.0 / 100.0** | 100.0 / 100.0 / 99.9 |
| Diff-ProjectedGAN | 93.8 / 69.5 / 43.3 | 88.8 / 54.6 / 27.2 | **99.9 / 99.9 / 99.2** | 99.8 / 99.6 / 96.6 | 62.1 / 10.5 / 2.8 |
| Average | 97.4 / 87.3 / 74.3 | 95.7 / 82.4 / 69.1 | **100.0 / 100.0 / 99.7** | 99.9 / 99.8 / 98.9 | 80.4 / 43.4 / 29.5 |
| DDPM | 85.2 / 37.8 / 14.2 | 80.8 / 29.6 / 9.3 | **96.5 / 79.4 / 39.1** | 95.1 / 69.5 / 30.7 | 57.4 / 3.8 / 0.6 |
| IDDPM | 81.6 / 30.6 / 10.6 | 79.9 / 27.6 / 7.8 | **94.3 / 64.8 / 25.7** | 92.8 / 58.0 / 21.2 | 62.9 / 7.0 / 1.3 |
| ADM | 68.3 / 13.2 / 3.4 | 68.8 / 14.1 / 4.0 | **77.8 / 20.7 / 5.2** | 70.6 / 13.0 / 2.5 | 60.5 / 8.2 / 1.8 |
| PNDM | 79.0 / 27.5 / 9.2 | 75.5 / 22.6 / 6.3 | 91.6 / 52.0 / 16.6 | **91.5 / 53.9 / 22.2** | 71.6 / 15.4 / 4.0 |
| LDM | 78.7 / 24.7 / 7.4 | 77.7 / 24.3 / 6.9 | 96.7 / 79.9 / 42.1 | **97.0 / 81.8 / 48.9** | 54.8 / 7.7 / 2.1 |
| Average | 78.6 / 26.8 / 9.0 | 76.6 / 23.7 / 6.8 | **91.4 / 59.3 / 25.7** | 89.4 / 55.2 / 25.1 | 61.4 / 8.4 / 2.0 |

[2] https://github.com/CompVis/stable-diffusion

Using the dataset introduced in Section 4, we analyze the performance of the three state-of-the-art detectors. We use pre-trained models provided by the authors without any modifications and compute the evaluation metrics using 10k real and 10k generated images. The results are given in Table 2. For GAN-generated images, both classifiers provided by Gragnaniello et al. (2021) achieve almost perfect detection results across all five generators, even w.r.t. the challenging Pd@1% metric. For images generated by ProjectedGAN and Diff-ProjectedGAN, the variant trained on ProGAN images performs slightly better. The reduced scores for these two models could in part be explained by the high image quality indicated by low FIDs (see Table 1). In comparison, the performance of the classifier provided by Wang et al. (2020) is reduced (except for ProGAN, on which the model was trained), which confirms the findings of Gragnaniello et al. (2021). Similar to Gragnaniello et al. (2021), ProjectedGAN and Diff-ProjectedGAN achieve the lowest scores, with Pd@1% dropping below 50%. Except for Diff-StyleGAN2, Mandelli et al. (2022a) achieves significantly worse results than its contestants. A possible explanation could be that StyleGAN2 images make up a large portion of the training data for this classifier.

For images generated by DMs, we observe a sharp deterioration for all detectors, with AUROC dropping by 15.2% on average in comparison to GANs. The difference becomes even more severe when looking at the more realistic Pd@FAR metrics. While the best model by Gragnaniello et al. (2021) (trained on ProGAN images) achieves a Pd@5% of 100.0% and Pd@1% of 99.7% on GAN-generated images, the scores drop by 40.7% and 74.0% for DMs, respectively. Given a setting where a low FAR is required, we find that none of the detectors under evaluation are usable. Among the five DMs, ADM generated the images most difficult to detect, which corresponds to a low FID. However, images from LDM, which achieves the second lowest FID, can be detected relatively effectively. A potential reason for this could be the adversarial training procedure of LDM (see Section 4), which could result in properties similar to those of GAN-generated images.

Overall, we observe that pre-trained universal detectors suffer from a severe performance drop when applied to DM-generated images (compared to GANs). We conclude that there exists a structural difference between GAN- and DM-generated images that requires the adaptation of existing methods or the development of novel methods to ensure the reliable detection of synthesized images.

### 5.2 Can we Improve Detection by Fine-Tuning?

Given the findings presented above, the question remains whether DMs evade detection altogether, or whether the detection performance can be increased by fine-tuning.[3] We select Wang et al. (2020) as a baseline since the original training code is available[4], which guarantees using the same training procedure. Furthermore, we choose the configuration Blur+JPEG (0.5) as it yields slightly better scores on average, and use the validation set for adapting the learning rate and performing early stopping. Besides fine-tuning a model on data from each generator, we also consider three aggregated settings in which we fine-tuned on all images generated by GANs, DMs, and both, respectively.

The results are depicted in Figure 1. A model which is fine-tuned on images from a specific DM achieves near-perfect scores, with PD@1% ranging from 97.7% for ADM to 100.0% for PNDM and LDM. To a limited extent, models are also capable of generalizing to data from other DMs, indicating common, detectable characteristics. Interestingly, detectors fine-tuned on images from DMs are significantly more successful in detecting GAN-generated images than vice versa. The same holds for models fine-tuned using images from multiple GANs and DMs, respectively. Given these results, we hypothesize that the detection of DM-generated images poses a more difficult task than the detection of GAN-generated images. The analysis in the frequency domain, which we perform in Section 6, provides additional support for this hypothesis.

In addition, the results demonstrate that training a detector which is capable of correctly classifying both GAN- and DM-generated images is clearly possible (see rightmost column of Figure 1a). While this finding is promising, it is unclear whether a similar performance can be achieved in the model-agnostic setting addressed by Wang et al. (2020).

[3] In addition, we carry out the same evaluation for detectors trained from scratch, which leads to very similar results, as shown in Appendix B.2.

[4] `https://github.com/peterwang512/CNNDetection`

Fine-tuned on

| Tested on | ProGAN | StyleGAN | ProjectedGAN | Diff-StyleGAN2 | Diff-ProjectedGAN | DDPM | IDDPM | ADM | PNDM | LDM | GAN | DM | All |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| ProGAN | 100.0 | 99.9 | 100.0 | 100.0 | 100.0 | 99.2 | 99.3 | 99.6 | 99.3 | 99.8 | 100.0 | 99.9 | 100.0 |
| StyleGAN | 98.5 | 100.0 | 98.0 | 99.9 | 98.9 | 95.5 | 98.2 | 98.1 | 96.8 | 97.8 | 100.0 | 99.9 | 100.0 |
| ProjectedGAN | 93.7 | 87.8 | 100.0 | 94.6 | 100.0 | 90.9 | 91.8 | 88.2 | 92.3 | 86.8 | 100.0 | 98.1 | 100.0 |
| Diff-StyleGAN2 | 99.4 | 99.8 | 99.3 | 100.0 | 99.5 | 97.6 | 96.9 | 84.6 | 99.1 | 99.3 | 100.0 | 98.9 | 100.0 |
| Diff-ProjectedGAN | 93.2 | 86.4 | 100.0 | 93.9 | 100.0 | 91.0 | 91.8 | 89.5 | 90.3 | 91.0 | 100.0 | 98.6 | 100.0 |
| DDPM | 82.4 | 77.5 | 81.8 | 88.0 | 82.6 | 100.0 | 99.9 | 99.7 | 99.1 | 98.2 | 92.4 | 100.0 | 100.0 |
| IDDPM | 80.5 | 77.9 | 79.7 | 82.7 | 81.7 | 99.8 | 100.0 | 99.9 | 97.9 | 97.9 | 89.6 | 100.0 | 100.0 |
| ADM | 65.7 | 61.3 | 65.2 | 60.3 | 67.9 | 93.3 | 97.4 | 99.9 | 85.9 | 93.0 | 69.5 | 100.0 | 100.0 |
| PNDM | 77.5 | 74.3 | 79.9 | 83.1 | 81.8 | 99.8 | 99.9 | 99.8 | 100.0 | 99.2 | 93.3 | 100.0 | 100.0 |
| LDM | 76.7 | 68.1 | 74.1 | 77.6 | 78.0 | 97.6 | 99.2 | 99.5 | 94.6 | 100.0 | 87.9 | 100.0 | 100.0 |

(a) AUROC

Fine-tuned on

| Tested on | ProGAN | StyleGAN | ProjectedGAN | Diff-StyleGAN2 | Diff-ProjectedGAN | DDPM | IDDPM | ADM | PNDM | LDM | GAN | DM | All |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| ProGAN | 100.0 | 98.9 | 99.4 | 100.0 | 99.8 | 95.9 | 96.5 | 97.6 | 94.8 | 98.0 | 100.0 | 99.4 | 100.0 |
| StyleGAN | 78.5 | 100.0 | 70.1 | 98.8 | 83.0 | 60.7 | 81.8 | 85.5 | 67.8 | 73.6 | 100.0 | 98.4 | 100.0 |
| ProjectedGAN | 40.2 | 20.0 | 100.0 | 55.2 | 100.0 | 43.8 | 51.7 | 48.8 | 40.2 | 37.6 | 100.0 | 86.2 | 100.0 |
| Diff-StyleGAN2 | 88.9 | 96.1 | 88.9 | 100.0 | 92.9 | 75.2 | 72.2 | 39.1 | 89.5 | 92.0 | 100.0 | 89.7 | 100.0 |
| Diff-ProjectedGAN | 39.8 | 17.9 | 99.5 | 49.2 | 100.0 | 42.0 | 50.0 | 47.4 | 31.0 | 45.1 | 100.0 | 88.9 | 100.0 |
| DDPM | 10.9 | 7.1 | 9.3 | 19.9 | 10.0 | 99.7 | 99.0 | 93.5 | 81.8 | 72.9 | 30.3 | 100.0 | 100.0 |
| IDDPM | 8.4 | 7.0 | 7.4 | 12.2 | 9.2 | 96.1 | 99.3 | 98.9 | 64.5 | 69.2 | 22.7 | 100.0 | 100.0 |
| ADM | 2.4 | 1.8 | 2.4 | 1.7 | 3.2 | 42.0 | 65.9 | 97.7 | 16.5 | 38.6 | 4.0 | 99.7 | 99.9 |
| PNDM | 6.6 | 6.3 | 7.4 | 14.4 | 10.9 | 96.7 | 98.7 | 96.6 | 100.0 | 89.1 | 39.4 | 100.0 | 100.0 |
| LDM | 5.7 | 2.9 | 4.4 | 7.8 | 6.6 | 68.4 | 86.2 | 92.7 | 40.8 | 100.0 | 20.6 | 100.0 | 100.0 |

(b) PD@1%.

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

## 6 Analysis of DM-generated images in the Frequency Domain

The visual quality of synthesized images has reached a level that makes them practically indistinguishable from real images for humans (Nightingale & Farid, 2022). However, in the case of GANs, the same does not hold when switching to the frequency domain (Zhang et al., 2019). In this section, we analyze the spectral properties of DM-generated images and compare them to those of GAN-generated images. We use three frequency transforms that have been used successfully in both traditional image forensics (Lyu, 2013) and deepfake detection: discrete Fourier transform (DFT), discrete cosine transform (DCT), and the reduced spectrum, which can be viewed as a 1D representation of the DFT. While DFT and DCT visualize frequency artifacts, the reduced spectrum can be used to identify spectrum discrepancies. The formal definitions of all transforms are provided in Appendix C.1. Furthermore, we evaluate the influence of different parameters during training and sampling of DMs and hypothesize on the origin of spectrum discrepancies.

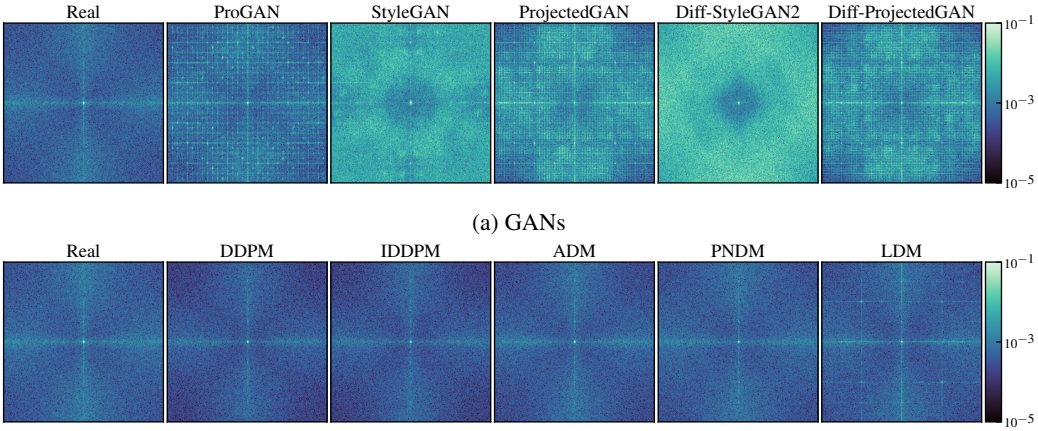
### 6.1 Spectral Properties of DM- vs GAN-generated Images

Figure 2 depicts the average absolute DFT spectrum of 10k images from each GAN and DM in our dataset.[5] 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. Taking the reduced spectra in Figure 3a into account we confirm the previously reported elevated high frequencies (see Section 2). For ProjectedGAN and Diff-ProjectedGAN, this characteristic is also visible. However, the spectral density does not exceed that of real images. Diff-StyleGAN2 is an exception as it shows a slight decrease towards the end of the spectrum.

In contrast, the DFT spectra of images generated by DMs (see Figure 2b), except LDM, are significantly more similar to the real spectrum with almost no visible artifacts. LDM, however, exhibits a thin grid across its spectrum, as well as the increased amount of high frequencies which is characteristic for GANs (see Figure 3b). As mentioned in Section 4, the architecture of LDM differs from the remaining DMs as the actual image is generated using an adversarially trained autoencoder, which could explain the discrepancies. This supports previous findings which suggest that the discriminator is responsible for spectrum deviations.

Although the reduced spectra of the remaining DMs do not exhibit the GAN-like rise towards the end of the spectrum, several DMs deviate stronger from the spectrum of real images than GANs. Especially for ADM, IDDPM, and DDPM, we observe an underestimation of the spectral density which becomes stronger towards higher frequencies (relative to the overall spectral density level). In Section 6.3 we discuss potential causes.

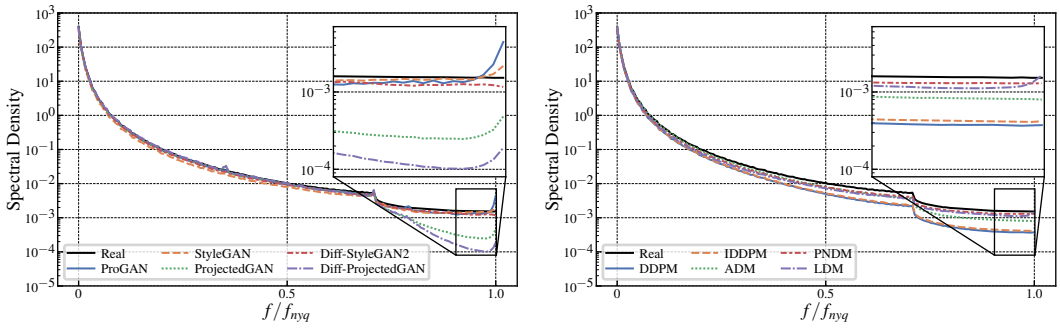
[5] We also report the DCT spectra of both GANs and DMs in Appendix C.2 which produce similar findings.




(a) GANs

(b) DMs

Figure 2: **Mean of DFT spectrum from 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.




(a) GANs

(b) DMs

Figure 3: **Mean reduced spectrum from real and generated images.** The higher end of the spectrum where GAN-characteristic discrepancies occur is magnified.

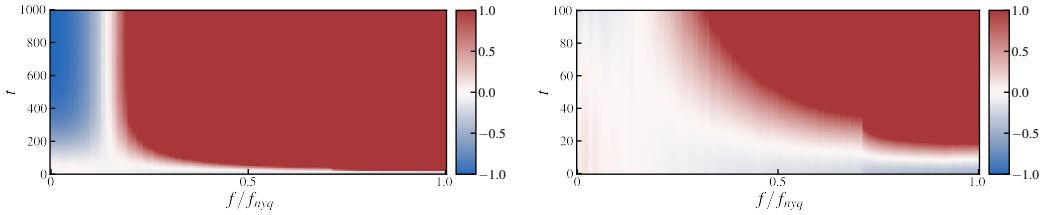
### 6.2 Are DM-generated images easier to detect in the frequency domain?

So far we showed that although DMs do not exhibit strong frequency artifacts, their spectrum deviates from that of real images. A natural question is therefore whether these discrepancies can be used to detect DM-generated images more effectively. Following Frank et al. (2020), we perform a simple logistic regression on each dataset with different transforms: Pixel (no transform), DFT (and taking the absolute value), and DCT. We use 20k samples for training, 2k for validation, and 20k for testing, each set equally split between real and fake images. To reduce the number of features, we transform all images to grayscale and take a center crop with dimensions 64x64. Additionally, all features are independently standardized to have zero mean and unit variance. We apply $L_2$ regularization and identify the optimal regularization weight by performing a grid search over the range $\{10^k \mid k \in \{-4, -3, \dots, 4\}\}$.

The accuracy of all GANs and DMs in our dataset is given in Table 3. We also report the results for log-scaled DFT and DCT coefficients, as this leads to significant improvements for some generators. For both GANs and DMs, using information from the frequency domain increases classification accuracy. On average, the performance gain of the best transform compared to no transform is 5.72% and 10.6%, respectively. Although the gain for DMs is more than double that for GANs, the overall maximum accuracy is significantly lower (90.9% for GANs and 63.1% for DMs on average). Therefore, we do not conclude that DMs exhibit stronger discriminative features in the frequency domain compared to GANs.

Table 3: **Accuracy of logistic regression on pixels and different transforms.** Next to each transform column we report the gain compared to the accuracy on pixels.

| | Pixel | DFT | | log(DFT) | | DCT | | log(DCT) | |
|---|---|---|---|---|---|---|---|---|---|
| ProGAN | 64.8 | 74.7 | +9.9 | 72.6 | +7.8 | 65.4 | +0.6 | **74.9** | +10.1 |
| StyleGAN | 91.1 | 87.4 | -3.7 | 86.2 | -4.9 | **92.6** | +1.5 | 86.4 | -4.7 |
| ProjectedGAN | 90.0 | 90.8 | +0.8 | 90.3 | +0.3 | 91.0 | +1.0 | **95.3** | +5.3 |
| Diff-StyleGAN2 | 92.4 | 80.3 | -12.0 | 80.5 | -11.9 | **93.8** | +1.4 | 87.6 | -4.7 |
| Diff-ProjectedGAN | 87.4 | 93.9 | +6.5 | 93.1 | +5.7 | 88.1 | +0.6 | **97.7** | +10.3 |
| DDPM | 51.7 | 64.2 | +12.6 | 64.2 | +12.5 | 52.4 | +0.7 | **64.3** | +12.6 |
| IDDPM | 51.6 | **62.1** | +10.6 | 61.7 | +10.1 | 51.7 | +0.1 | 61.7 | +10.1 |
| ADM | 50.1 | **54.7** | +4.6 | 52.3 | +2.3 | 50.1 | +0.0 | 53.8 | +3.7 |
| PNDM | 52.5 | 57.0 | +4.5 | 58.2 | +5.7 | 51.1 | -1.4 | **61.4** | +8.9 |
| LDM | 56.6 | 63.7 | +7.1 | 66.1 | +9.5 | 58.5 | +1.9 | **73.0** | +16.3 |




(a) $0 \leq t \leq 1000$ (b) $0 \leq t \leq 100$

Figure 4: **Evolution of spectral density error throughout the denoising process.** The error is computed relatively 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.

### 6.3 Investigating the Source of Spectrum Discrepancies

As explained in Section 2, several recent works attempted to identify the source of spectrum discrepancies in GAN-generated images. In this section, we take an effort to launch a similar line of work for DMs. For the experiments, we used code and model from ADM (Dhariwal & Nichol, 2021) pre-trained on LSUN Bedroom. Spectra are generated using 512 grayscale images.

**Spectrum Evolution During the Denoising Process** Our first analysis targets the denoising process. We generate samples at different time steps $t$ and compare their average reduced spectrum to that of 50k real images. The results are shown in Figure 4. 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 $t = 1000$, the image is pure Gaussian noise, which causes the strong spectrum deviations. Around $t = 300$, the error starts to decrease, but interestingly it appears to not continuously improve until $t = 0$. With regard to high frequencies, we observe an optimum at approximately $t = 10$ followed by an increasing underestimation. In Appendix C.3, we additionally provide the spectral density error between the denoising process and the diffusion process at the corresponding step, which further visualizes this behavior. However, at $t = 10$, the images are easier to detect and also visually noisier (see Appendix C.3 Figures 11 and 12). Thus, stopping the denoising process early is not an effective means to make synthetic images less detectable.

We hypothesize that the underestimation towards higher frequencies is due to a training objective that is sub-optimal to capture high frequency content. As noted by Nichol & Dhariwal (2021), the objective proposed by Ho et al. (2020), $L_{\text{simple}} = \mathbb{E}_{t,x_0,\epsilon}[\|\epsilon - \epsilon_\theta(\mathbf{x}_t, t)\|^2]$, does not lead to good log-likelihood values. These depend on high-frequency details synthesized at the challenging denoising tasks near $t = 0$ (Kingma et al., 2021). In contrast, the classical variational lower bound $L_{\text{vlb}}$ forces the model to focus on the denoising steps towards $t = 0$. However, as $L_{\text{vlb}}$ is extremely difficult to optimize (see e.g., Nichol & Dhariwal (2021)) the modified objective $L_{\text{hybrid}} = L_{\text{simple}} + \lambda L_{\text{vlb}}$ (with $\lambda = 0.001$) is proposed by (Nichol & Dhariwal, 2021) and

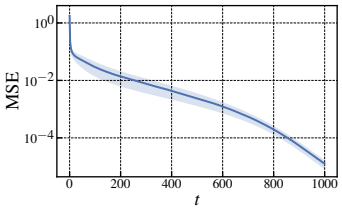
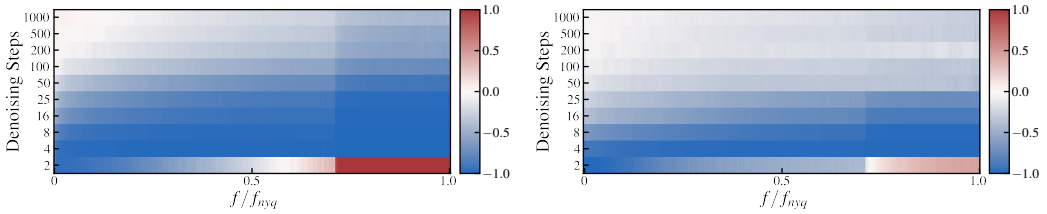



Figure 5: **Mean and standard deviation of the mean-squared error at different steps.**




(a) Without DDIM (b) With DDIM

Figure 6: **Evolution of spectral density error 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.

also employed by (Dhariwal & Nichol, 2021). And indeed, when looking at the mean-squared error at different $t$ (see Figure 5), we observe a steep increase near $t = 0$, suggesting the model struggles with accurately removing the small perturbations which correspond to high frequencies. We conclude that the training objective used in state-of-the-art DMs enables impressive perceptual image quality (also regarding FID), but struggles to accurately model the high-frequency content of real images – which in turn can be exploited for detection. We elaborate in Appendix C.5.

**Effect of Number of Sampling Steps** In a second step, we analyze how varying the number of sampling steps affects the spectral properties of generated images. While DMs are usually trained using $T = 1000$ steps, sampling speed can be increased by reducing the number of steps (Nichol & Dhariwal, 2021). For ADM on LSUN Bedroom, however, images were sampled using 1000 steps since this led to much better results (Dhariwal & Nichol, 2021). The authors also evaluate DDIM (Song et al., 2022a), an alternative sampling method which allows for sampling with fewer steps.

Using the previous experimental settings, we analyze the reduced spectra of images sampled using different numbers of steps (Figure 6). With only a few steps, the spectral density is too low across the entire spectrum, resulting in blurry images with little contrast (see Appendix C.4). Performing more steps leads to lower errors, with DDIM improving faster than normal sampling. This finding is coherent with previous results, as DDIM achieves better samples at fewer sampling steps (Song et al., 2022a). In Appendix C.4 we also analyze the detectability with different numbers of sampling steps. For both sampling methods, images become easier to detect with fewer sampling steps.

## 7 Conclusion

In this work we take a much-needed first step towards detection of DM-generated images. We find that existing detectors, which claim to be universal, fail to effectively recognize images synthesized by different DMs. This hints at fundamental differences between images generated by DMs and GANs, one of which we identify as different spectral properties. While DMs produce little to no frequency artifacts, there appears to be a systematic underestimation of high frequencies. Our hypothesis is that during training, too little weight is attached to these frequencies due to the choice of the training objective. However, this is an appropriate choice because correctly reproducing high frequencies is less important for the perceived quality of generated images than matching lower frequencies. Whether spectral discrepancies can be reduced (e.g., by adding spectral regularization (Durall et al., 2020)) without impairing image quality should be subject of future work. Most importantly, more efforts are needed to develop novel detection methods, both universal and DM-specific. For tackling this task, a sufficiently large amount of DM-generated images is necessary, which poses a challenge due to the slow generation speed compared to GANs. We therefore urge researchers presenting novel generation methods to publish sample images alongside pre-trained models, to enable the development and evaluation of detection methods without the need for immense computational resources. This is especially relevant due to the emergence of powerful text-to-image models based on DMs like DALL-E 2 (Ramesh et al., 2022), Imagen (Saharia et al., 2022), and Stable Diffusion (Rombach et al., 2022). We hope that our work can spark further research in this direction and emphasizes the fact that advances in realistic media synthesis should always be accompanied by considerations regarding deepfake detection methods.

### ETHICS STATEMENT

While generative models have many beneficial applications, deepfakes pose a serious threat to society by undermining trust in media and digital communication (e.g., via fake profiles in social media that are used in misinformation campaigns). This work explores the challenges of deepfake detection for the novel promising class of diffusion models. We believe that the urge for reliable deepfake detection – ideally agnostic to the underlying generative model – is amplified by the ever easier use of generative models (e.g., via text-to-image synthesis). By highlighting the (current) lack of "universal" detectors, we strongly encourage the community to address this shortcoming and our results can serve as a starting point for future work in this direction. While insights from such research could potentially be used to exacerbate or eventually elude detection, we strongly believe further insights in this area are much needed in order to sustain identifiability of deepfakes.

### REPRODUCIBILITY STATEMENT

To ensure reproducibility, we provide instructions to recreate the dataset in Appendix A and, for the additional datasets, in Appendix D. All datasets considered in this work are publicly available. We train/fine-tune the detector proposed by Wang et al. (2020) using the settings provided by the authors and describe training and validation data in the text (Section 5.2). All tables and figures can be reproduced using the source code provided in the supplementary materials.

## REFERENCES

G. J. Burton and Ian R. Moorhead. Color and spatial structure in natural scenes. *Applied Optics*, 26 (1):157–170, January 1987.

Lucy Chai, David Bau, Ser-Nam Lim, and Phillip Isola. What makes fake images detectable? Understanding properties that generalize. In *Computer Vision – ECCV 2020*, Lecture Notes in Computer Science, pp. 103–120, 2020.

Keshigeyan Chandrasegaran, Ngoc-Trung Tran, and Ngai-Man Cheung. A closer look at fourier spectrum discrepancies for CNN-generated images detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 7200–7209, 2021.

Yuanqi Chen, Ge Li, Cece Jin, Shan Liu, and Thomas Li. SSD-GAN: Measuring the realness in the spatial and spectral domains. *Proceedings of the AAAI Conference on Artificial Intelligence*, 35 (2):1105–1112, May 2021.

Jooyoung Choi, Jungbeom Lee, Chaehun Shin, Sungwon Kim, Hyunwoo Kim, and Sungroh Yoon. Perception prioritized training of diffusion models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 11472–11481, 2022.

Davide Cozzolino, Justus Thies, Andreas Rössler, Christian Riess, Matthias Nießner, and Luisa Verdoliva. ForensicTransfer: Weakly-supervised domain adaptation for forgery detection. *arXiv:1812.02510 [cs]*, November 2019.

Davide Cozzolino, Diego Gragnaniello, Giovanni Poggi, and Luisa Verdoliva. Towards universal GAN image detection. In *2021 International Conference on Visual Communications and Image Processing (VCIP)*, pp. 1–5, December 2021.

Prafulla Dhariwal and Alexander Nichol. Diffusion models beat GANs on image synthesis. In *Advances in Neural Information Processing Systems*, volume 34, pp. 8780–8794, 2021.

Ricard Durall, Margret Keuper, and Janis Keuper. Watch your up-convolution: CNN based generative deep neural networks are failing to reproduce spectral distributions. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2020.

Tarik Dzanic, Karan Shah, and Freddie Witherden. Fourier spectrum discrepancies in deep network generated images. In *Advances in Neural Information Processing Systems*, volume 33, pp. 3022–3032, 2020.

Patrick Esser, Robin Rombach, and Björn Ommer. Taming transformers for high-resolution image synthesis. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 12873–12883, 2021.

D. J. Field. Wavelets, vision and the statistics of natural scenes. *Philosophical Transactions: Mathematical, Physical and Engineering Sciences*, 357(1760):2527–2542, 1999.

David J. Field. Relations between the statistics of natural images and the response properties of cortical cells. *JOSA A*, 4(12):2379–2394, December 1987.

Joel Frank, Thorsten Eisenhofer, Lea Schönherr, Asja Fischer, Dorothea Kolossa, and Thorsten Holz. Leveraging frequency analysis for deep fake image recognition. In *Proceedings of the 37th International Conference on Machine Learning*, volume 119 of *Proceedings of Machine Learning Research*, pp. 3247–3258, July 2020.

Krzysztof J. Geras and Charles Sutton. Scheduled denoising autoencoders. In *3rd International Conference on Learning Representations, ICLR 2015, San Diego, CA, USA, May 7-9, 2015, Conference Track Proceedings*, 2015.

Krzysztof J. Geras and Charles Sutton. Composite denoising autoencoders. In *Machine Learning and Knowledge Discovery in Databases*, Lecture Notes in Computer Science, pp. 681–696, 2016.

Sharath Girish, Saksham Suri, Sai Saketh Rambhatla, and Abhinav Shrivastava. Towards discovery and attribution of open-world GAN generated images. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 14094–14103, 2021.

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

Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 770–778, 2016.

Martin Heusel, Hubert Ramsauer, Thomas Unterthiner, Bernhard Nessler, and Sepp Hochreiter. GANs trained by a two time-scale update rule converge to a local nash equilibrium. In *Advances in Neural Information Processing Systems 30: Annual Conference on Neural Information Processing Systems 2017, December 4-9, 2017, Long Beach, CA, USA*, pp. 6626–6637, 2017.

Jonathan Ho, Ajay Jain, and Pieter Abbeel. Denoising diffusion probabilistic models. In *Advances in Neural Information Processing Systems*, volume 33, pp. 6840–6851, 2020.

Tero Karras, Timo Aila, Samuli Laine, and Jaakko Lehtinen. Progressive growing of GANs for improved quality, stability, and variation. In *International Conference on Learning Representations*, 2018.

Tero Karras, Samuli Laine, and Timo Aila. A style-based generator architecture for generative adversarial networks. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 4401–4410, June 2019.

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

Mahyar Khayatkhoei and Ahmed Elgammal. Spatial frequency bias in convolutional generative adversarial networks. *Proceedings of the AAAI Conference on Artificial Intelligence*, 36(7):7152–7159, June 2022.

Diederik Kingma, Tim Salimans, Ben Poole, and Jonathan Ho. Variational diffusion models. In *Advances in Neural Information Processing Systems*, volume 34, pp. 21696–21707, 2021.

Luping Liu, Yi Ren, Zhijie Lin, and Zhou Zhao. Pseudo numerical methods for diffusion models on manifolds. In *International Conference on Learning Representations*, February 2022.

Siwei Lyu. Natural image statistics in digital image forensics. In *Digital Image Forensics: There Is More to a Picture than Meets the Eye*, pp. 239–256. 2013.

Sara Mandelli, Nicolò Bonettini, Paolo Bestagini, and Stefano Tubaro. Detecting GAN-generated images by orthogonal training of multiple CNNs. *arXiv:2203.02246 [cs]*, March 2022a.

Sara Mandelli, Davide Cozzolino, Edoardo D. Cannas, João P. Cardenuto, Daniel Moreira, Paolo Bestagini, Walter J. Scheirer, Anderson Rocha, Luisa Verdoliva, Stefano Tubaro, and Edward J. Delp. Forensic analysis of synthetically generated western blot images. *IEEE Access*, 10:59919–59932, 2022b.

Francesco Marra, Diego Gragnaniello, Luisa Verdoliva, and Giovanni Poggi. Do GANs leave artificial fingerprints? In *2019 IEEE Conference on Multimedia Information Processing and Retrieval (MIPR)*, pp. 506–511, March 2019.

Alexander Quinn Nichol and Prafulla Dhariwal. Improved denoising diffusion probabilistic models. In *Proceedings of the 38th International Conference on Machine Learning*, pp. 8162–8171, July 2021.

Sophie J. Nightingale and Hany Farid. AI-synthesized faces are indistinguishable from real faces and more trustworthy. *Proceedings of the National Academy of Sciences*, 119(8), February 2022.

Gaurav Parmar, Richard Zhang, and Jun-Yan Zhu. On aliased resizing and surprising subtleties in GAN evaluation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 11410–11420, 2022.

Aditya Ramesh, Prafulla Dhariwal, Alex Nichol, Casey Chu, and Mark Chen. Hierarchical text-conditional image generation with CLIP latents, April 2022.

Severi Rissanen, Markus Heinonen, and Arno Solin. Generative modelling with inverse heat dissipation, July 2022.

Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. High-resolution image synthesis with latent diffusion models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 10684–10695, 2022.

Olga Russakovsky, Jia Deng, Hao Su, Jonathan Krause, Sanjeev Satheesh, Sean Ma, Zhiheng Huang, Andrej Karpathy, Aditya Khosla, Michael Bernstein, Alexander C. Berg, and Li Fei-Fei. ImageNet large scale visual recognition challenge. *International Journal of Computer Vision*, 115 (3):211–252, December 2015.

Chitwan Saharia, William Chan, Saurabh Saxena, Lala Li, Jay Whang, Emily Denton, Seyed Kamyar Seyed Ghasemipour, Burcu Karagol Ayan, S. Sara Mahdavi, Rapha Gontijo Lopes, Tim Salimans, Jonathan Ho, David J. Fleet, and Mohammad Norouzi. Photorealistic text-to-image diffusion models with deep language understanding, May 2022.

Tim Salimans and Jonathan Ho. Progressive distillation for fast sampling of diffusion models. In *International Conference on Learning Representations*, March 2022.

Axel Sauer, Kashyap Chitta, Jens Müller, and Andreas Geiger. Projected GANs converge faster. In *Advances in Neural Information Processing Systems*, volume 34, pp. 17480–17492, 2021.

Katja Schwarz, Yiyi Liao, and Andreas Geiger. On the frequency bias of generative models. In *Advances in Neural Information Processing Systems*, volume 34, pp. 18126–18136, 2021.

Jascha Sohl-Dickstein, Eric Weiss, Niru Maheswaranathan, and Surya Ganguli. Deep unsupervised learning using nonequilibrium thermodynamics. In *Proceedings of the 32nd International Conference on Machine Learning*, pp. 2256–2265, June 2015.

Jiaming Song, Chenlin Meng, and Stefano Ermon. Denoising diffusion implicit models. In *International Conference on Learning Representations*, February 2022a.

Yang Song and Stefano Ermon. Generative modeling by estimating gradients of the data distribution. In *Advances in Neural Information Processing Systems*, volume 32, 2019.

Yang Song and Stefano Ermon. Improved techniques for training score-based generative models. In *Advances in Neural Information Processing Systems*, volume 33, pp. 12438–12448, 2020.

Yang Song, Jascha Sohl-Dickstein, Diederik P. Kingma, Abhishek Kumar, Stefano Ermon, and Ben Poole. Score-based generative modeling through stochastic differential equations. In *International Conference on Learning Representations*, February 2022b.

Mingxing Tan and Quoc Le. EfficientNet: Rethinking model scaling for convolutional neural networks. In *Proceedings of the 36th International Conference on Machine Learning*, pp. 6105–6114, May 2019.

D. J. Tolhurst, Y. Tadmor, and Tang Chao. Amplitude spectra of natural images. *Ophthalmic and Physiological Optics*, 12(2):229–232, 1992.

Antonio Torralba and Aude Oliva. Statistics of natural image categories. *Network: Computation in Neural Systems*, 14(3):391–412, January 2003.

Luisa Verdoliva. Media forensics and DeepFakes: An overview. *IEEE Journal of Selected Topics in Signal Processing*, 14(5):910–932, 2020.

Pascal Vincent, Hugo Larochelle, Isabelle Lajoie, Yoshua Bengio, and Pierre-Antoine Manzagol. Stacked denoising autoencoders: Learning useful representations in a deep network with a local denoising criterion. *Journal of Machine Learning Research*, 11(110):3371–3408, 2010.

Gregory K. Wallace. The JPEG still picture compression standard. *Communications of the ACM*, 34(4):30–44, April 1991.

Sheng-Yu Wang, Oliver Wang, Richard Zhang, Andrew Owens, and Alexei A. Efros. CNN-generated images are surprisingly easy to spot... for now. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2020.

Zhendong Wang, Huangjie Zheng, Pengcheng He, Weizhu Chen, and Mingyuan Zhou. Diffusion-GAN: Training gans with diffusion, June 2022.

Moritz Wolter, Felix Blanke, Raoul Heese, and Jochen Garcke. Wavelet-packets for deepfake image analysis and detection. *Machine Learning*, August 2022.

Zhisheng Xiao, Karsten Kreis, and Arash Vahdat. Tackling the generative learning trilemma with denoising diffusion GANs. In *International Conference on Learning Representations*, May 2022.

Xinsheng Xuan, Bo Peng, Wei Wang, and Jing Dong. On the generalization of GAN image forensics. In *Biometric Recognition*, Lecture Notes in Computer Science, pp. 134–141, 2019.

Fisher Yu, Ari Seff, Yinda Zhang, Shuran Song, Thomas Funkhouser, and Jianxiong Xiao. LSUN: Construction of a large-scale image dataset using deep learning with humans in the loop, June 2016.

Xu Zhang, Svebor Karaman, and Shih-Fu Chang. Detecting and simulating artifacts in GAN fake images. In *2019 IEEE International Workshop on Information Forensics and Security (WIFS)*, pp. 1–6, December 2019.

## A Details on the Dataset

**LSUN Bedroom (Yu et al., 2016)** We download and extract the lmbd database files using the official repository[6]. The images are center-cropped to 256x256 pixels.

**ProGAN (Karras et al., 2018)** We download the first 10k samples from the non-curated collection provided by the authors.[7]

[6] https://github.com/fyu/lsun

[7] https://github.com/tkarras/progressive_growing_of_gans

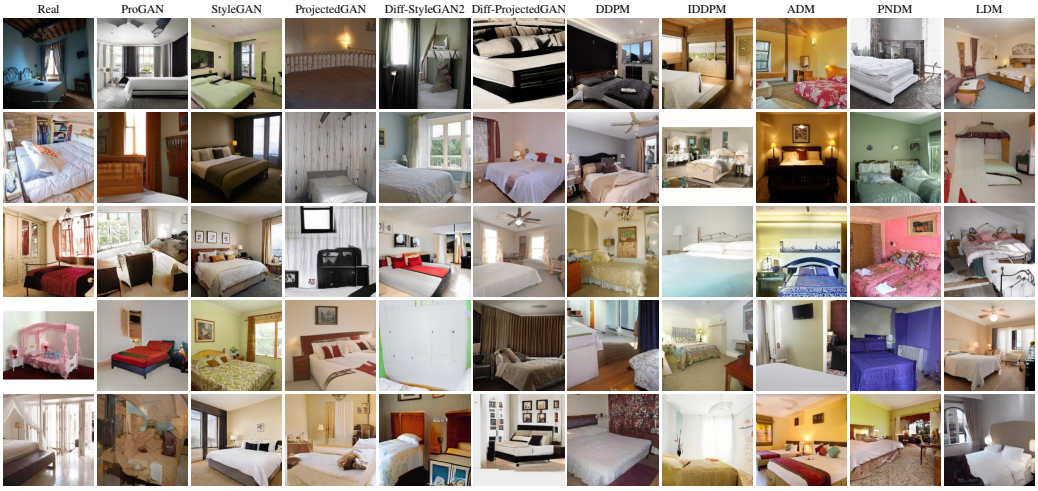



Figure 7: **Non-curated examples for real LSUN Bedroom, GAN-generated, and DM-generated images.**

**StyleGAN (Karras et al., 2019)** We download the first 10k samples generated with $\psi = 0.5$ from the non-curated collection provided by the authors.[8]

**ProjectedGAN (Sauer et al., 2021)** We sample 10k images using code and pre-trained models provided by the authors using the default configuration (`--trunc=1.0`).[9]

**Diff-StyleGAN2 (Wang et al., 2022) and Diff-ProjectedGAN (Wang et al., 2022)** We sample 10k images using code and pre-trained models provided by the authors using the default configuration.[10]

**DDPM (Ho et al., 2020), IDDPM (Nichol & Dhariwal, 2021), and ADM (Dhariwal & Nichol, 2021)** We download the samples provided by the authors of ADM[11] and extract the first 10k samples for each generator. For ADM on LSUN, we select the models trained with dropout.

**PNDM (Liu et al., 2022)** We sample 10k images using code and pre-trained model provided by the authors.[12] 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 10k images using code and pre-trained models provided by the authors using settings from the corresponding table in the repository.[13] 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.

[8] https://github.com/NVlabs/stylegan
[9] https://github.com/autonomousvision/projected_gan
[10] https://github.com/Zhendong-Wang/Diffusion-GAN
[11] https://github.com/openai/guided-diffusion
[12] https://github.com/luping-liu/PNDM
[13] https://github.com/CompVis/latent-diffusion

# B DETECTION

## B.1 DESCRIPTIONS OF DETECTORS

**Wang et al. (2020)** In this influential work, the authors demonstrate that a standard deep convolutional neural network (CNN) trained on data from a single generator performs surprisingly well on unseen images. They train a ResNet-50 (He et al., 2016) on 720k images from 20 LSUN (Yu et al., 2016) categories (not including Bedroom and Church), equally divided into real images and images generated by ProGAN (Karras et al., 2018). The trained binary classifier is able to distinguish real from generated images from a variety of generative models and datasets. The authors further show that extensive data augmentation in the form of random flipping, blurring, and JPEG compression generally improves generalization. Moreover, the authors provide two pre-trained model configurations, Blur+JPEG (0.1) and Blur+JPEG (0.5), where the value in parentheses denotes the probability of blurring and JPEG compression, respectively. They achieve an average precision of 92.6% and 90.8%, respectively. The work suggests that CNN-generated images contain common artifacts (or fingerprints) which make them distinguishable from real images.

**Gragnaniello et al. (2021)** Building upon the architecture and dataset from Wang et al. (2020), the authors of this work experiment with different variations to further improve the detection performance in real-world scenarios. The most promising variant, no-down, removes downsampling from the first layer, increasing the average accuracy from 80.71% to 94.42%, at the cost of more trainable parameters. They also train a model with the same architecture on images generated by StyleGAN2 (Karras et al., 2020) instead of ProGAN, which further improves accuracy to 98.48%.

**Mandelli et al. (2022a)** Unlike the other two methods, this work uses an ensemble of five orthogonal CNNs to detect fake images not seen during training. All CNNs are based on the EfficientNet-B4 model (Tan & Le, 2019) but trained on different datasets. Dataset orthogonality refers to images having different content, different processing (e.g., JPEG compression), or being generated by different GANs. They argue that by having different datasets, each CNN learns to detect different characteristics of real and generated images, improving the overall performance and generalization ability. For each CNN, a score is computed for $\approx 200$ random patches extracted from the image. These scores are combined using a novel patch aggregation strategy which assumes an image is fake if at least one patch is classified as being fake. Finally, the output score is computed by averaging the individual scores from all five CNNs. It should be noted that, besides several GANs, the training dataset also includes samples generated by Song et al. (2022b), a score-based model.

## B.2 DETECTION RESULTS FOR CLASSIFIERS TRAINED FROM SCRATCH

We repeat the experiment described in Section 5.2 but train each classifier from scratch instead of fine-tuning it, the results are shown in Figure 8. Note that here "from scratch" means that we start from a ResNet-50 model pre-trained on ImageNet, which is exactly how the authors trained their models (Wang et al., 2020). For most constellations of training and test data, the results are similar compared to fine-tuning, with a tendency towards worse scores for the models trained from scratch.

Trained on

| Tested on | ProGAN | StyleGAN | ProjectedGAN | Diff-StyleGAN2 | Diff-ProjectedGAN | DDPM | IDDPM | ADM | PNDM | LDM | GAN | DM | All |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| ProGAN | 100.0 | 92.5 | 98.7 | 99.9 | 98.0 | 98.0 | 98.5 | 98.5 | 98.0 | 97.5 | 100.0 | 99.7 | 100.0 |
| StyleGAN | 94.2 | 100.0 | 91.6 | 99.7 | 92.8 | 89.3 | 97.5 | 98.2 | 94.4 | 92.2 | 100.0 | 99.9 | 100.0 |
| ProjectedGAN | 86.9 | 66.2 | 100.0 | 90.0 | 100.0 | 89.0 | 88.5 | 82.6 | 88.8 | 75.3 | 100.0 | 97.9 | 100.0 |
| Diff-StyleGAN2 | 96.1 | 94.1 | 94.8 | 100.0 | 93.3 | 93.5 | 93.1 | 83.4 | 97.1 | 91.3 | 100.0 | 99.0 | 100.0 |
| Diff-ProjectedGAN | 86.2 | 66.5 | 99.9 | 88.6 | 100.0 | 87.6 | 88.0 | 82.6 | 86.5 | 79.9 | 100.0 | 98.4 | 100.0 |
| DDPM | 83.7 | 63.0 | 78.8 | 87.2 | 79.1 | 100.0 | 99.9 | 99.6 | 99.1 | 97.9 | 93.1 | 100.0 | 100.0 |
| IDDPM | 81.6 | 67.3 | 77.7 | 81.1 | 78.2 | 99.9 | 100.0 | 100.0 | 97.9 | 97.2 | 90.0 | 100.0 | 100.0 |
| ADM | 67.8 | 56.5 | 64.5 | 56.6 | 65.0 | 93.2 | 98.0 | 99.9 | 87.0 | 91.0 | 70.5 | 100.0 | 100.0 |
| PNDM | 81.3 | 62.7 | 78.0 | 84.3 | 74.1 | 99.6 | 99.9 | 99.8 | 100.0 | 97.8 | 95.1 | 100.0 | 100.0 |
| LDM | 75.2 | 54.7 | 71.6 | 73.0 | 69.0 | 97.0 | 99.1 | 99.4 | 93.9 | 100.0 | 88.3 | 100.0 | 100.0 |

(a) AUROC

Trained on

| Tested on | ProGAN | StyleGAN | ProjectedGAN | Diff-StyleGAN2 | Diff-ProjectedGAN | DDPM | IDDPM | ADM | PNDM | LDM | GAN | DM | All |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| ProGAN | 100.0 | 42.2 | 78.8 | 98.8 | 76.4 | 86.8 | 92.4 | 93.7 | 82.0 | 75.7 | 100.0 | 98.9 | 100.0 |
| StyleGAN | 45.1 | 100.0 | 34.4 | 95.0 | 42.6 | 27.7 | 75.5 | 84.9 | 48.0 | 37.0 | 100.0 | 98.8 | 100.0 |
| ProjectedGAN | 24.1 | 2.8 | 100.0 | 35.6 | 99.5 | 34.2 | 42.1 | 38.5 | 27.8 | 15.2 | 100.0 | 85.4 | 100.0 |
| Diff-StyleGAN2 | 58.3 | 42.8 | 50.7 | 100.0 | 48.1 | 43.8 | 43.8 | 25.9 | 66.6 | 44.6 | 100.0 | 90.6 | 100.0 |
| Diff-ProjectedGAN | 23.7 | 3.0 | 98.2 | 28.0 | 100.0 | 29.9 | 37.8 | 32.9 | 20.2 | 17.3 | 100.0 | 87.6 | 100.0 |
| DDPM | 13.0 | 2.1 | 8.3 | 18.8 | 9.0 | 99.8 | 99.1 | 92.6 | 80.2 | 68.7 | 33.9 | 100.0 | 100.0 |
| IDDPM | 10.6 | 2.5 | 7.7 | 10.3 | 8.1 | 96.3 | 99.7 | 99.3 | 62.4 | 61.6 | 23.9 | 100.0 | 100.0 |
| ADM | 3.2 | 0.9 | 2.5 | 1.3 | 3.1 | 37.5 | 69.7 | 98.2 | 16.7 | 30.1 | 4.2 | 100.0 | 99.9 |
| PNDM | 11.0 | 2.3 | 8.4 | 16.1 | 7.0 | 91.6 | 97.7 | 96.6 | 100.0 | 75.2 | 48.9 | 100.0 | 100.0 |
| LDM | 5.8 | 0.7 | 4.3 | 4.9 | 4.0 | 62.7 | 85.3 | 91.5 | 36.0 | 100.0 | 20.8 | 100.0 | 100.0 |

(b) PD@1%.

Figure 8: **Detection performance for detectors trained from scratch.**

# C FREQUENCY ANALYSIS

## C.1 FREQUENCY TRANSFORMS

**Discrete Fourier Transform (DFT)** The DFT maps a discrete signal to the frequency domain by expressing it as a sum of periodic basis functions. Given a grayscale image $I$ with height $H$ and width $W$, the two-dimensional DFT (with normalization term omitted) is defined as

$$I_{\text{DFT}}[k,l] = \sum_{x=0}^{H-1} \sum_{y=0}^{W-1} I[x,y] \ \exp^{-2\pi i \frac{x \cdot k}{H}} \ \exp^{-2\pi i \frac{y \cdot l}{W}}, \tag{4}$$

with $k = 0, \dots, H-1$ and $l = 0, \dots, W-1$. For visualization, the zero-frequency component is shifted to the center of the spectrum. Therefore, coefficients towards the edges of the spectrum correspond to higher frequencies.

**Discrete Cosine Transform (DCT)** The DCT is closely related to the DFT, however it uses real-valued cosine functions as basis functions. It is used in the JPEG compression standard due to its high degree of energy compaction (Wallace, 1991), which ensures that a large portion of a signal's energy can be represented using only a few DCT coefficients. The type-II DCT, which the term DCT usually refers to, is given as

$$I_{\text{DCT}}[k,l] = \sum_{x=0}^{H-1} \sum_{y=0}^{W-1} I[x,y] \cos\left[\frac{\pi}{H}\left(x + \frac{1}{2}\right)k_x\right] \cos\left[\frac{\pi}{W}\left(y + \frac{1}{2}\right)k_y\right], \tag{5}$$

again omitting the normalization factor. In the resulting spectrum, the low frequencies are located in the upper left corner, with frequencies increasing along both spatial dimensions.

**Reduced Spectrum** While the DFT provides a useful visual representation of an image's spectrum, it is less suitable for comparing different spectra quantitatively. Previous works therefore use the reduced spectrum[14], a one-dimensional representation of the Fourier spectrum (Durall et al., 2020; Dzanic et al., 2020; Schwarz et al., 2021; Chandrasegaran et al., 2021). It is obtained by azimuthally averaging over the spectrum in normalized polar coordinates $r \in [0,1]$, $\theta \in [0, 2\pi)$ according to

$$\tilde{S}(r) = \frac{1}{2\pi} \int_0^{2\pi} S(r,\theta) d\theta \quad \text{with} \quad r = \sqrt{\frac{k^2 + l^2}{\frac{1}{4}(H^2 + W^2)}} \quad \text{and} \quad \theta = \text{atan2}(k,l), \tag{6}$$

with $S[k,l] = |I_{DFT}[k,l]|^2$ being the squared magnitudes of the Fourier coefficients. The maximum frequency is given by the Nyquist frequency $f_{\text{nyq}} = \sqrt{k^2 + l^2} = H/\sqrt{2}$ for a square image with $H = W$.

## C.2 DCT SPECTRA

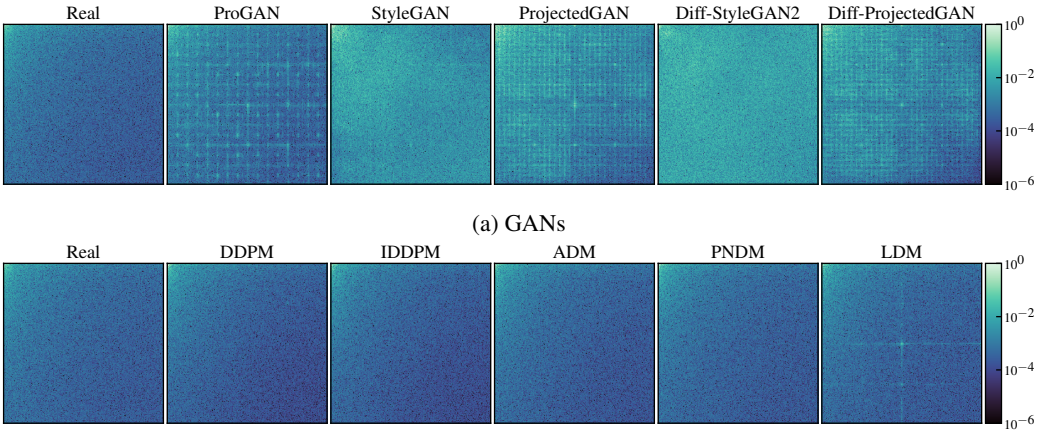
Figure 9 depicts the DCT spectra of both GANs and DMs. Similar to DFT, images generated by DMs exhibit fewer artifacts, with the exception of LDM.

## C.3 SPECTRUM EVOLUTION DURING THE DENOISING PROCESS

Analogous to Figure 4, we show the spectrum error evolution during the denoising process in Figure 10. Here, however, the error is computed relative to the spectrum of noised images at the corresponding step during the diffusion process. Similar to the denoising process, the spectra of the diffusion process are averaged over 512 samples. While the relative error is close to zero for a long time during the denoising process, the model fails to accurately reproduce higher frequencies towards $t = 0$. More precisely, too much high-frequency components are removed, which explains why the sweet spot does not seem to be at $t = 0$ but around $t = 10$ (see Figure 4).

[14]The definition of the reduced spectrum slightly differs between different existing works, we decide to follow that of Schwarz et al. (2021).

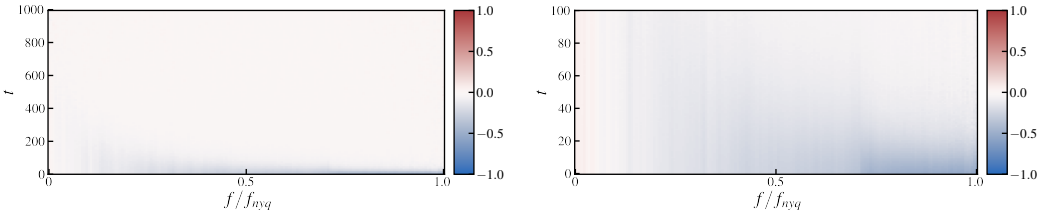
One could think that, by stopping the denoising process early, DM-generated images might be harder to detect. To test this hypothesis, we perform a logistic regression at every $0 \leq t \leq 100$ to distinguish real from increasingly denoised generated images. The model is trained using 512 real and 512 generated samples, from which 20% are used for testing. We select the optimal regularization weight by performing a 5-fold cross-validation over $\{10^k \mid k \in \{-4, -3, \ldots, 4\}$. Similar to Section 6.2, we compare the performance on pixels and different transforms. The results in Figure 11 do not indicate that around $t = 10$ fake images are less detectable. In Figure 12 it becomes apparent that at this $t$ the images are noticeably noisier, which probably explains the increased accuracy.




(a) GANs

(b) DMs

Figure 9: **Mean of DCT spectrum from real and generated images.** To increase visibility, the color bar is limited to $[10^{-6}, 10^0]$, with values lying outside this interval being clipped.




(a) $0 \leq t \leq 1000$

(b) $0 \leq t \leq 100$

Figure 10: **Evolution of spectral density error throughout the denoising process.** The error is computed relatively to the corresponding step in the diffusion process. The colorbar is clipped at -1 and 1. If the model was able to perfectly denoise, the difference would be 0 at all $t$.

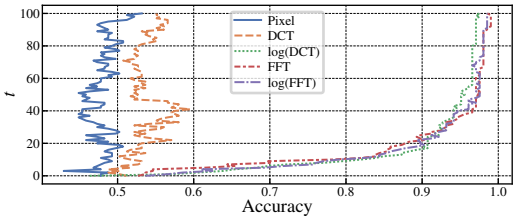



Figure 11: **Accuracy of logistic regression during the denoising process.** Note that only the last 100 steps of the denoising process are depicted.

### C.4 EFFECT OF NUMBER OF SAMPLING STEPS

We repeat the logistic regression experiment from the previous section for samples generated with different numbers of timesteps, the results are shown in Figure 13. We also provide example images in Figure 14. As expected, images become harder to detect with more sampling steps and therefore higher image quality. For DDIM, the accuracy decreases more quickly, which is likely because it was found to generate images of higher quality with fewer sampling steps (Song et al., 2022a).

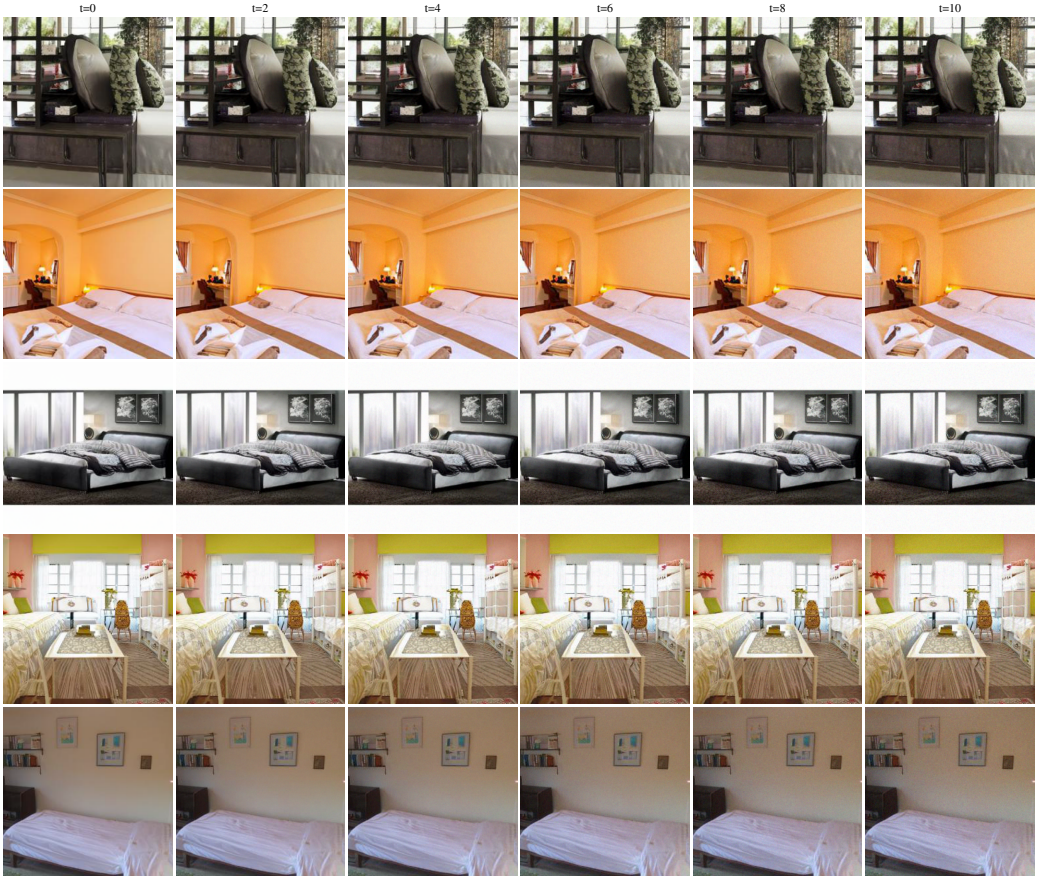
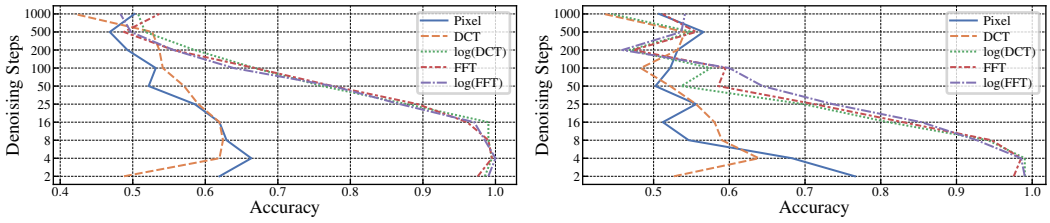


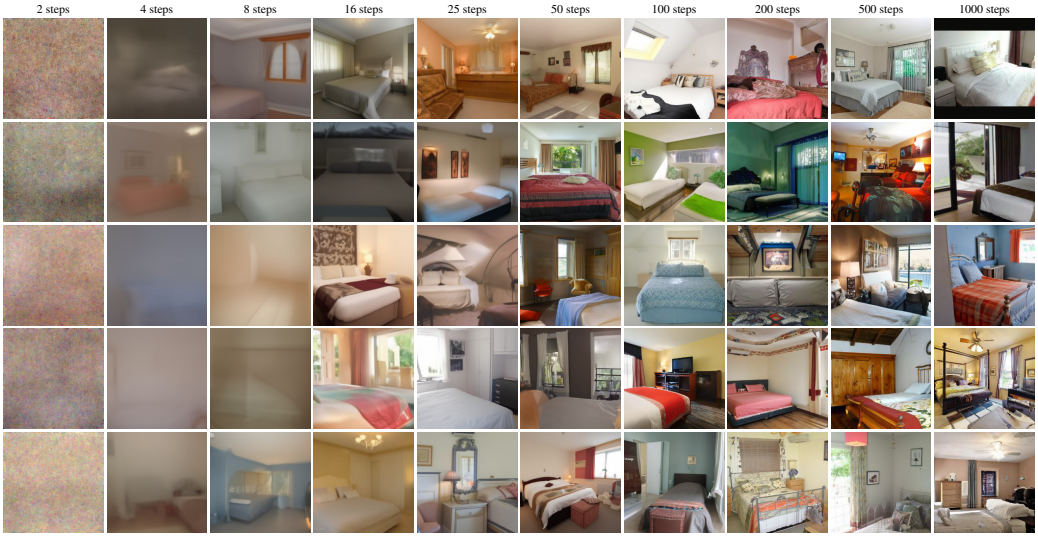

Figure 12: **Example images generated by ADM at different t.** When zoomed in, the high-frequency noise towards higher $t$ becomes apparent.



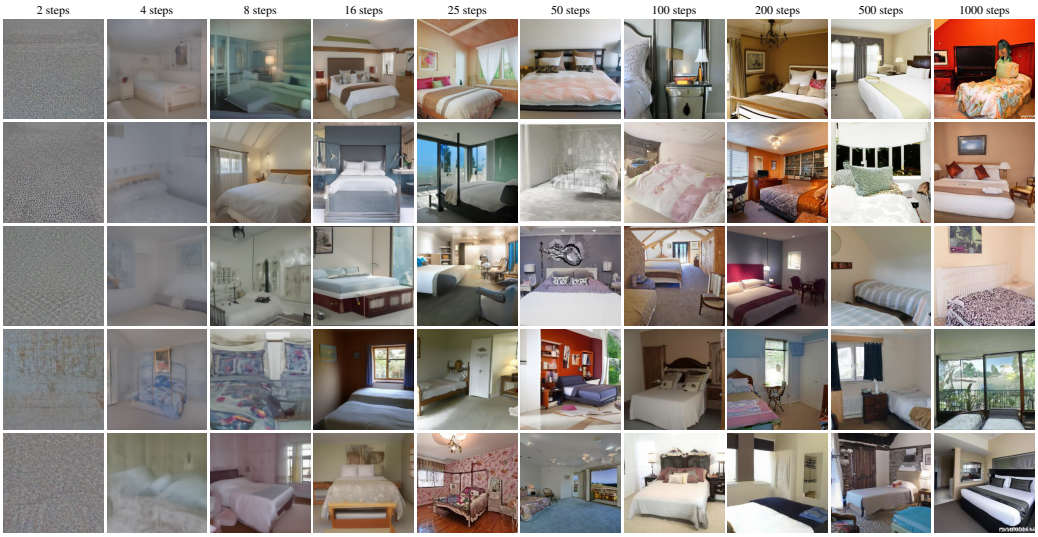

(a) Without DDIM

(b) With DDIM

Figure 13: **Accuracy of logistic regression for different numbers of sampling steps.** Note that the y-axis is not scaled linearly.




(a) Without DDIM




(b) With DDIM

Figure 14: **Example images generated by ADM with different numbers of sampling steps.**

### C.5 Potential Source of the Spectrum Discrepancies

Our analysis in the frequency domain (see Figures 3 and 18) suggests that current state-of-the-art DMs do not match the high-frequency content well. To further analyze these findings, we build on insights from the denoising autoencoder (DAE) literature. Roughly speaking, the task DAEs face is conceptually similar to the task of the noise predictor $\epsilon_\theta$ in DMs at a single time step: denoising a disturbed input at a fixed noise level. Note that while we believe that DMs and DAEs can be conceptually related, the concepts are distinct in several ways: DMs use parameter sharing to perform noise prediction at multiple noise levels to set up the generation as an iterative process. On the other hand, DAEs make use of a latent space to learn suitable representations for reconstruction, classically at a fixed noise level. Nevertheless, we believe it may be useful to take these insights into account.

A handy observation from DAEs relates the level of corruption to the learned feature scales: Denoising an image at small noise levels requires to accurately model fine granular details of the image, while coarse/large-scale features can still be recovered at high noise levels (see e.g., Vincent et al.

(2010); Geras & Sutton (2015; 2016)). Transferring these insight to DMs, we observe that the training objective guides the reconstruction performance across the different noise levels.

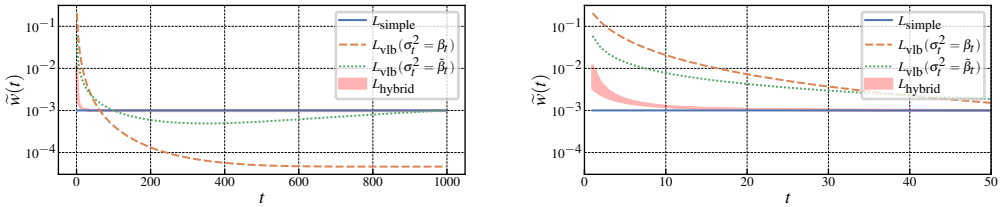
Recall that the objective of many prominent DMs can be stated as a weighted sum of mean squared error terms

$$L(\theta) = \sum_{t=0}^{T} w(t)\mathbb{E}_{t,x_0,\epsilon}[\|\epsilon - \epsilon_\theta(\mathbf{x}_t, t)\|^2] \tag{7}$$

usually with $T = 1000$. The theoretically derived variational lower bound $L_{\text{vlb}}$ (in the case of $\Sigma(t) = \sigma_t^2\mathbf{I}$) corresponds to the weighting scheme

$$w(t) = \frac{\beta_t^2}{2\sigma_t^2\alpha_t(1-\bar{\alpha}_t)} \tag{8}$$

with $\alpha_t$ and $\bar{\alpha}_t$ derived from the noise schedule $\beta_t$ (see Section 3). However, $L_{\text{vlb}}$ turns out to be extremely difficult to optimize in practice (see e.g., Dhariwal & Nichol (2021)), arguably due to large influence of the challenging denoising tasks near $t = 0$. To circumvent this issue, $L_{\text{simple}}$ was proposed by Ho et al. (2020) which corresponds to $w(t) = 1$, and turned out to be stable to optimize and already leads to remarkable perceptual quality. In order to achieve both, perceptual image quality and high likelihood values, Nichol & Dhariwal (2021) proposed $L_{\text{hybrid}} = L_{\text{simple}} + \lambda L_{\text{vlb}}$ with $\lambda = 0.001$ which increases the influence of low noise levels. The relative importance of reconstruction at different noise levels for the above discussed objectives is depicted in Figure 15.




(a) $0 \leq t \leq 1000$.

(b) $0 \leq t \leq 50$.

Figure 15: **Relative importance of the reconstruction tasks for prominent DMs.** We show the relative influence $\widetilde{w}(t) = w(t)/\sum_{t=0}^{T} w(t)$ for (a) the whole denoising-diffusion process and (b) a close-up on the lowest noise levels. For $L_{\text{vlb}}$ we depict both, lower and upper bound for the denoising variance $\sigma_t^2$ (given as $\sigma_t^2 = \beta_t$ and $\sigma_t^2 = \tilde{\beta}_t$ as defined in Ho et al. (2020)). In the case of $L_{\text{hybrid}}$, used by IDDPM and ADM, we use $\sigma_t^2\mathbf{I}$ for visualization purposes instead of a diagonal covariance matrix and plot $\widetilde{w}(t)$ for the range of admissible variances $\sigma_t^2$. All plots assume the linear noise schedule $\beta_t$ by Ho et al. (2020).

To put the above discussion in a nutshell, we hypothesize that the ability of a DM to match the real frequency spectrum is governed by the reconstruction performance at the corresponding noise levels. Importantly, successful error prediction at low noise scales, i.e., near $t = 0$, requires to capture the high-frequency content of an image. We deduce that the relative down-weighting of the influence of low noise levels when using $L_{\text{simple}}$ or $L_{\text{hybrid}}$ (compared to the theoretically derived $L_{\text{vlb}}$ which would account for high likelihoods) results in the observed mismatch of high frequencies.

The mean reduced spectra of various DMs (Figure 3) support this hypothesis: Both Nichol & Dhariwal (2021) and Dhariwal & Nichol (2021) train their model with $L_{\text{hybrid}}$ (which incorporates a relatively higher weight on low noise levels), and are able to reduce the gap to the real spectrum when compared to the baseline trained with $L_{\text{simple}}$ (Ho et al., 2020). Clearly, we believe that not only the weighting scheme $w(t)$ accounts for the resulting spectral properties, but more importantly the model's capabilities to successfully predict the low level noise in the first place. Nevertheless, the weighting scheme acts as a proxy that encourages the model to focus on specific noise levels.

We conclude that the objectives of DMs are well designed to guide the model to high perceptual quality (or benchmark metrics such as FID), while falling short on providing sufficient information to accurately model the high frequency content of the target images, which would be better captured by a likelihood-based objective like $L_{\text{vlb}}$.

# D RESULTS ON ADDITIONAL DATASETS

To demonstrate the generalization of our findings, we perform classification and frequency analysis on additional data from ADM, PNDM, and LDM. Note that ADM-G-U refers to the two-stage up-sampling stack in which images are generated at a resolution of 64x64 and subsequently up-sampled to 256×256 pixels using a second model (Dhariwal & Nichol, 2021). The generated images are obtained according to the instructions given in Section A. 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 results on other GANs, we refer to the original publications of the detectors (Wang et al., 2020; Gragnaniello et al., 2021; Mandelli et al., 2022a).

## D.1 PERFORMANCE OF UNIVERSAL DETECTORS

Similar to Table 2, we report AUROC, Pd@5%, and Pd@1% for the additional datasets in Table 4. Overall, the results do not invalidate the claim that universal GAN-detectors are incapable of effectively detecting images generated by DMs. Both PNDM and LDM trained on LSUN Church, however, are detected significantly better by all classifiers except Mandelli et al. (2022a). This might be caused by the relatively high FID (compared to the state of the art) both models achieve on this dataset (8.69 for PNDM, 4.02 for LDM).

Table 4: **Detection performance of pre-trained universal detectors on additional datasets.** For Wang et al. (2020) and Gragnaniello et al. (2021) we consider two different variants, respectively. The best score (determined by the highest Pd@1%) for each generator is highlighted in **bold**.

| AUROC / Pd@5% / Pd@1% | Wang et al. (2020) | | Gragnaniello et al. (2021) | | Mandelli et al. (2022a) |
|---|---|---|---|---|---|
| | Blur+JPEG (0.5) | Blur+JPEG (0.1) | ProGAN | StyleGAN2 | |
| ADM (LSUN Cat) | 58.4 / 8.4 / 2.5 | 58.1 / 8.5 / 3.3 | **60.2 / 9.3 / 4.2** | 51.7 / 5.5 / 1.8 | 55.6 / 6.0 / 1.3 |
| ADM (LSUN Horse) | 55.5 / 6.7 / 1.5 | 53.4 / 6.0 / 2.2 | **56.1 / 7.5 / 2.7** | 50.2 / 4.8 / 1.4 | 44.2 / 2.4 / 0.5 |
| ADM (ImageNet) | 69.1 / 13.9 / 4.1 | 71.7 / 15.5 / 4.5 | 72.1 / 13.2 / 3.5 | **83.9 / 38.5 / 16.6** | 60.1 / 6.8 / 0.9 |
| ADM-G-U (ImageNet) | 67.2 / 11.7 / 3.7 | 62.3 / 6.2 / 1.2 | 66.8 / 6.0 / 1.6 | **78.9 / 26.7 / 10.2** | 60.0 / 7.6 / 1.0 |
| PNDM (LSUN Church) | 76.9 / 25.7 / 10.2 | 77.6 / 28.4 / 12.0 | 90.9 / 54.1 / 24.5 | **99.3 / 96.5 / 85.8** | 56.4 / 6.9 / 1.9 |
| LDM (LSUN Church) | 86.3 / 42.1 / 19.8 | 82.2 / 33.9 / 14.2 | 98.8 / 93.7 / 75.5 | **99.5 / 98.2 / 90.2** | 58.9 / 5.8 / 1.3 |
| LDM (FFHQ) | 69.4 / 14.2 / 3.6 | 71.0 / 14.8 / 3.6 | **91.1 / 54.2 / 25.4** | 67.2 / 10.2 / 2.1 | 63.0 / 5.6 / 0.6 |

## D.2 FREQUENCY ANALYSIS

We analyze the DFT (Figure 16), DCT (Figure 17), and reduced spectra (Figure 18) using a similar process as in Section 6.1. Regarding frequency artifacts, the results are consistent with that from LSUN Bedroom, LDM exhibits grid-like artifacts while ADM and PNDM do not. The spectra of ADM on LSUN Cat and LSUN Horse do contain irregular, vertical structures, which we did not observe for any other model and dataset. However, these are substantially different and not as pronounced as GAN artifacts.

The reduced spectra have largely the same characteristics as for LSUN Bedroom, with the exception of ImageNet. Here we observe an overestimation towards higher frequencies, which is the opposite of what we see for ADM on other datasets. A possible explanation could be that the authors sampled images from LSUN using 1000 and from ImageNet using only 250 steps. We suppose that the number of spectral discrepancies is highly training dependent.

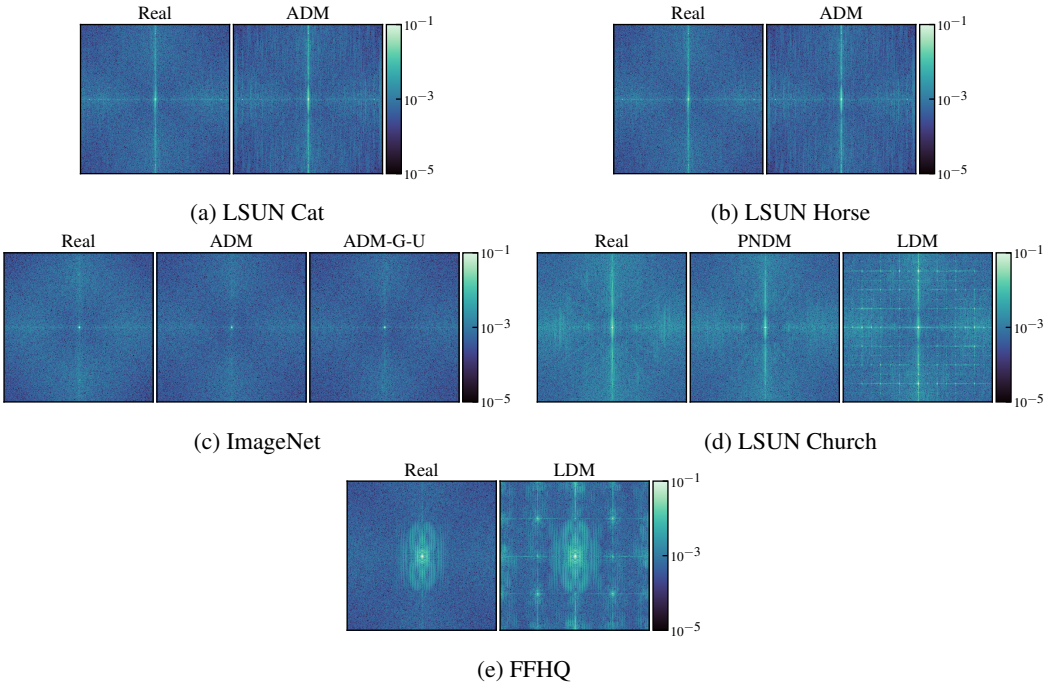



Figure 16: **Mean of DFT spectrum from real and generated images from additional datasets.** To increase visibility, the color bar is limited to $[10^{-5}, 10^{-1}]$, with values lying outside this interval being clipped.

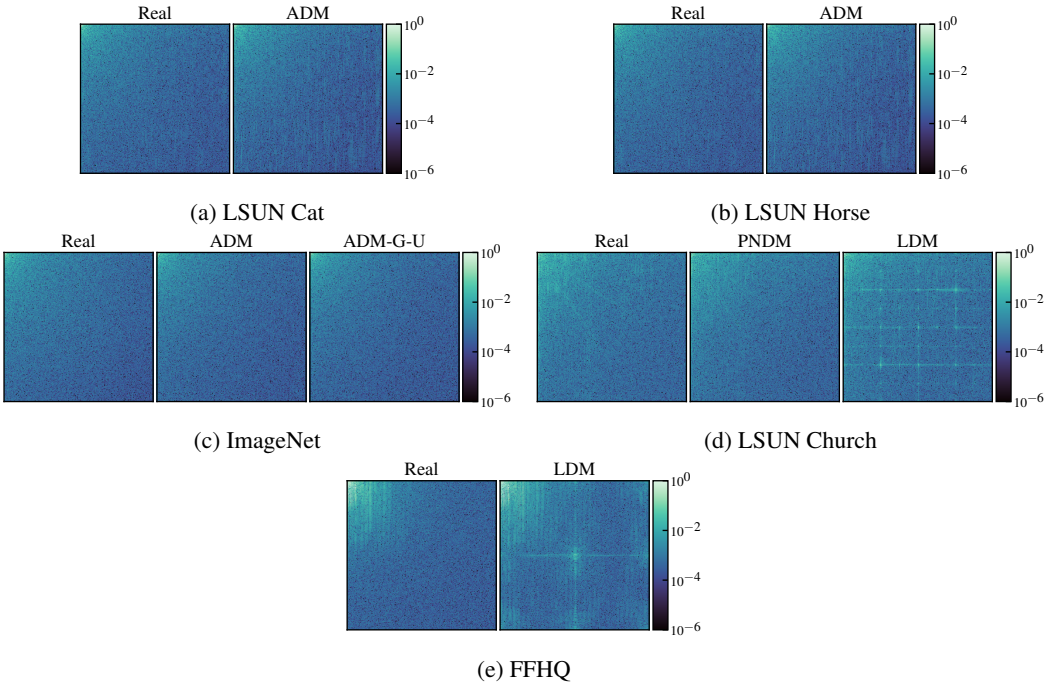
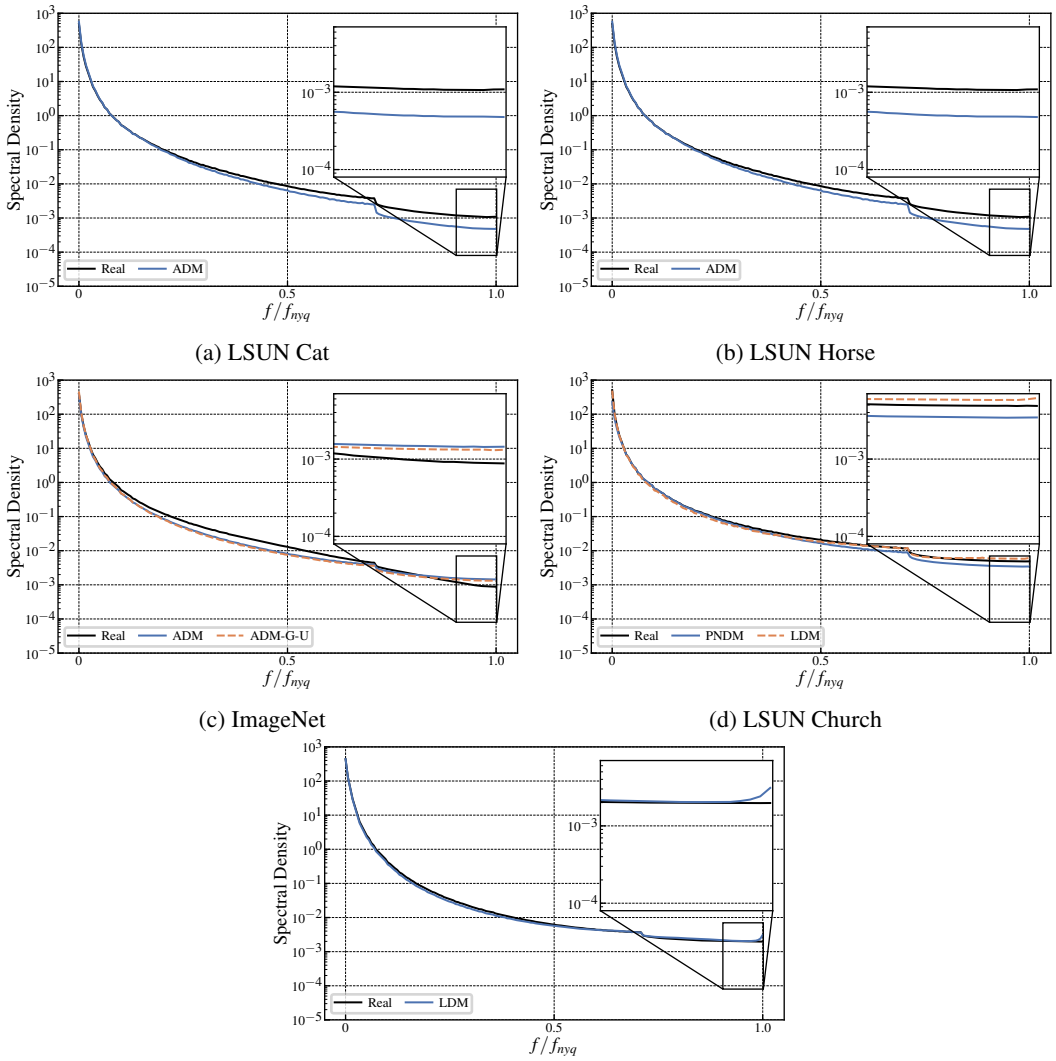



Figure 17: **Mean of DCT spectrum from real and generated images from additional datasets.** To increase visibility, the color bar is limited to $[10^{-6}, 10^{0}]$, with values lying outside this interval being clipped.




(a) LSUN Cat

(b) LSUN Horse

(c) ImageNet

(d) LSUN Church

(e) FFHQ

Figure 18: **Mean reduced spectrum from real and generated images from additional datasets.** The higher end of the spectrum where GAN-characteristic discrepancies occur is magnified.