FRÉCHET WAVELET DISTANCE: A DOMAIN-AGNOSTIC METRIC FOR IMAGE GENERATION

Anonymous authors
Paper under double-blind review

ABSTRACT

Modern metrics for generative learning like Fréchet Inception Distance (FID) and DINOv2-Fréchet Distance (FD-DINOv2) demonstrate impressive performance. However, they suffer from various shortcomings, like a bias towards specific generators and datasets. To address this problem, we propose the Fréchet Wavelet Distance (FWD) as a domain-agnostic metric based on the Wavelet Packet Transform ($\mathcal{W}_p$). FWD provides a sight across a broad spectrum of frequencies in images with a high resolution, preserving both spatial and textural aspects. Specifically, we use $\mathcal{W}_p$ to project generated and real images to the packet coefficient space. We then compute the Fréchet distance with the resultant coefficients to evaluate the quality of a generator. This metric is general-purpose and dataset-domain agnostic, as it does not rely on any pre-trained network while being more interpretable due to its ability to compute Fréchet distance per packet, enhancing transparency. We conclude with an extensive evaluation of a wide variety of generators across various datasets that the proposed FWD can generalize and improve robustness to domain shifts and various corruptions compared to other metrics.

1 INTRODUCTION

With the surge of generative neural networks, especially in the image domain, it becomes important to assess their performance in a robust and reliable way (Heusel et al., 2017a; Binkowski et al., 2018; Salimans et al., 2016; Kynkäänniemi et al., 2019; Stein et al., 2023). FID (Heusel et al., 2017a) has emerged as the de facto standard for comparing generative image synthesis approaches. However, it also shows various shortcomings, such as its reliance on a pre-trained classification backbone, i.e., InceptionV3 trained on ImageNet. This, by design, introduces a class dependency into FID leading to accidental distortions (Sauer et al., 2021). The FID scores actually improve if the evaluation set resembles ImageNet or if the use of ImageNet pretrained discriminator pushes the output distribution towards ImageNet, although the image quality remains the same in these cases (Kynkäänniemi et al., 2023).

To address the domain bias problem caused by the use of a pre-trained network, we propose an alternative metric based on the Wavelet Packet Transform ($\mathcal{W}_p$). In contrast to other pure frequency (Narwaria et al., 2012) or spatial (Wang et al., 2004; Horé & Ziou, 2010) metrics, wavelets have the advantage that they combine both frequency and spatial aspects in one metric. While frequency information is important (Durall et al., 2020; Dzanic et al., 2020; Rahaman et al., 2019; Schwarz et al., 2021; Wolter et al., 2022), it alone is insufficient to assess the quality of synthesized images without considering additional spatial information. Wavelets are thus an ideal representation for a metric comparing generative approaches for image synthesis. As FID, FWD utilizes the Fréchet distance of the real and generated set of images as a distance measure, but it is not computed based on InceptionV3 activation maps. Instead, it utilizes the wavelet-packet frequency band representations of $\mathcal{W}_p$ as illustrated in Figs. 1 and 3. To this end, we first use $\mathcal{W}_p$ to transform every image, where we use the wavelet transform at a fixed level. We then compute the Fréchet distance for each packet of the transform and average them over all packets. The proposed Fréchet Wavelet Distance (FWD) thus considers spatial information as well as all frequency bands.

To quantitatively assess those characteristics, we evaluate the proposed metric in terms of its domain bias and robustness. We further compare the proposed FWD to existing state-of-the-art metrics like FID, Kernel Inception Distance (KID) and DINOv2-Fréchet Distance (FD-DINOv2) on standard

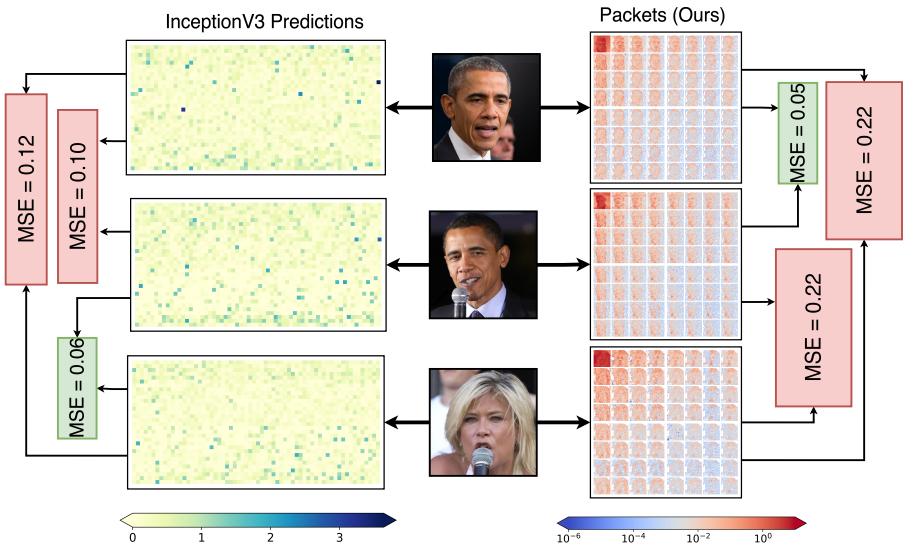


Figure 1: The first two images depict the same person, while the last image depicts a different person. Intuitively, the first two images are more similar than the other pairs of images. When computing the mean squared error between the images using the penultimate InceptionV3 activations or wavelet packets, we observe that the wavelet packets produce a low distance for the first two images, as expected. Surprisingly, according to InceptionV3, the last two images are similar since both images are classified as `'microphone'` whereas the first image as `'groom'`. Images from Fli (2024).

datasets. We show that FWD is a more robust metric that does not suffer from the domain bias and can thus be applied to any dataset. Kynkäänniemi et al. (2023) experimented with optimizing FID by selecting a subset of images from 250k generated images, where the subset's InceptionV3 activations are related to ImageNet classes. Building on this work, we observed a significant improvement in FID by $\approx 50\%$, when evaluated on this subset. FD-DINOv2 showed an unintended improvement of $\approx 2\%$ as well. This undesired improvement can likely be explained by the overlap between ImageNet and the DINOv2 training set. In contrast, FWD remains the same despite the manipulation. We also show that some unexpected FID results can be attributed to the dataset bias. Furthermore, FWD is significantly faster to compute. In an effort to produce reproducible work, we provide code for FWD as a part of the supplementary material.

In summary, this paper makes the following contributions:

1. We propose the Fréchet Wavelet Distance (FWD) as a dataset- and domain-agnostic metric for evaluation of generative approaches for image synthesis.
2. FWD is an interpretable metric, as the wavelet packet transform splits the frequency space into hierarchically organized, discrete subbands.
3. We show that the proposed method is computationally inexpensive and robust to corruption, perturbation, and distractors.
4. We show that FD-DINOv2 addresses the domain bias issue to some extent but at a very high computational cost. Furthermore, we provide evidence that it is still limited to the domains of the training data.

2 Related Work

2.1 Metrics for generative learning

A generative model should generate novel image samples that mirror the training set sample distribution, including data diversity. In a vision context, Salimans et al. (2016) proposed the Inception Score (IS) as a measure of image quality, independent of the target dataset statistics. The IS is

computed by measuring the entropy of the class probabilities of an InceptionV3. The score builds upon the assumption that a generative network that has converged to a meaningful solution will produce images that will allow InceptionV3 to make predictions with certainty. In other words, a certain InceptionV3 has a low prediction entropy. IS has been found to be sensitive to different ImageNet training runs (Barratt & Sharma, 2018). Furthermore, it does not use the statistics of the real data distribution a Generative Adversarial neural Network (GAN) is trained to model (Heusel et al., 2017a). In response Heusel et al. (2017a) proposes FID. Instead of measuring the entropy at the final layer FID is computed by evaluating the Fréchet distance (Dowson & Landau, 1982) the penultimate network activations computed on both the true and synthetic images. Today, comparing high-level InceptionV3 features using an FID-score (Heusel et al., 2017a) enjoys widespread adoption. Variants exist, Kernel Inception Distance (KID) (Binkowski et al., 2018), for example, relaxes the multi-variate Gaussian assumption of FID and measures the polynomial kernel distance between Inception features of generated and training dataset. Binkowski et al. (2018) kept the InceptionV3 backbone and replaced Fréchet distance with kernel distance. While FID captures general trends well, the literature also discusses its drawbacks. Kynkäänniemi et al. (2023) empirically studies the effect of ImageNet classes on FID for non-ImageNet datasets by using GradCAM. Furthermore, Kynkäänniemi et al. (2023) examines ImageNet bias using Projected Fast GAN (Proj. FastGAN) and StyleGAN2. Compared to StyleGAN2, Proj. FastGAN produces more accidental distortions like floating heads and artefacts Sauer et al. (2021). Surprisingly, Proj. FastGAN's FID is comparable to StyleGAN2's in their experiment. Chong & Forsyth (2020) found a generator-dependent architecture bias, which limits our ability to compare samples for smaller datasets with 50K or fewer images. Additionally, Parmar et al. (2022) found that both FID and KID are highly sensitive to resizing and compression. Barratt & Sharma (2018) reported FID sensitivity with respect to different InceptionV3 weights. While comparing Tensorflow and PyTorch implementations, Parmar et al. (2022) measured inconsistent scores due to differing resizing implementations. Finally, FID scores are hard to reproduce unless all details regarding its computation are carefully disclosed (Hug, 2024). Stein et al. (2023) proposed an alternative to over-reliance on InceptionV3, by replacing it with DINOv2-ViT-L/14 model. We observe this solves the domain bias somewhat at a significant computational cost. Unfortunately, DINOv2's training dataset is not publicly available. Consequently, gaps in the dataset remain hidden. This situation motivates the search for additional quality metrics.

2.2 Spectral Methods

Prior work found neural networks are spectrally biased (Rahaman et al., 2019). Many architectures favor low-frequency content (Durall et al., 2020; Gal et al., 2021; Wolter et al., 2022; Zhang et al., 2022). Related articles rely on the Fourier or Wavelet transform to understand frequency bias. Wavelet transforms as pioneered by Mallat (1989) and Daubechies (1992) have a solid track record in signal processing. The Fast Wavelet Transform (FWT) and the closely related Wavelet Packet Transform ($\mathcal{W}_p$), are starting to appear more frequently in the deep learning literature. Applications include Convolutional Neural Network (CNN) augmentation (Williams & Li, 2018), style transfer (Yoo et al., 2019), image denoising (Liu et al., 2020; Saragadam et al., 2023), image coloring (Li et al., 2022), face aging (Liu et al., 2019), video enhancement (Wang et al., 2020), face super-resolution (Huang et al., 2017), and generative machine learning (Gal et al., 2021; Guth et al., 2022; Zhang et al., 2022; Phung et al., 2023). Hernandez et al. (2019) uses the Fourier transform to measure the quality of human motion forecasting. Zhang et al. (2022) uses a FWT to remove artifacts from generated images. Phung et al. (2023) focuses on the FWT to increase the inference speed of diffusion models. This work proposes to use the Wavelet Packet Transform ($\mathcal{W}_p$) as an interpretable metric for generators.

3 Fréchet Wavelet Distance (FWD)

We want to tackle the problem of dataset-domain bias. To this end, we propose FWD, which in turn leverages the Wavelet Packet Transform ($\mathcal{W}_p$). We require two-dimensional filters for image processing where we use Haar wavelets. Consequently, we construct filter quadruples from the original single-dimensional filter pairs. The process uses outer products (Vyas et al., 2018):

$$\mathbf{h}_a = \mathbf{h}_\mathcal{L}\mathbf{h}_\mathcal{L}^T, \mathbf{h}_h = \mathbf{h}_\mathcal{L}\mathbf{h}_\mathcal{H}^T, \mathbf{h}_v = \mathbf{h}_\mathcal{H}\mathbf{h}_\mathcal{L}^T, \mathbf{h}_d = \mathbf{h}_\mathcal{H}\mathbf{h}_\mathcal{H}^T \tag{1}$$

With a for the approximation filter, h for the horizontal filter, v for the vertical filter, and d for the diagonal filter (Lee et al., 2019). We construct a $\mathcal{W}_p$-tree for images with these two-dimensional

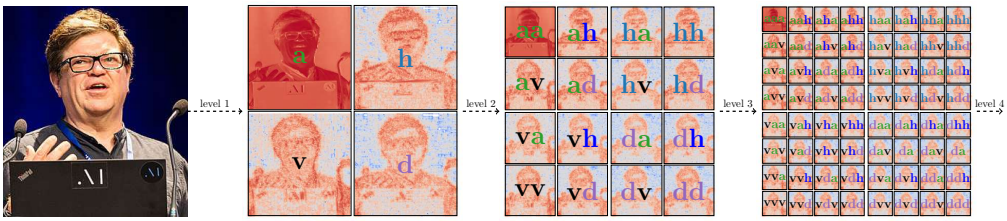


Figure 2: Illustration of the Wavelet Packet Transform ($\mathcal{W}_p$). For visualization purposes, we depict a level-3 transform. All later experiments use a level-4 transform. Image from Jérémy Barande (2024).

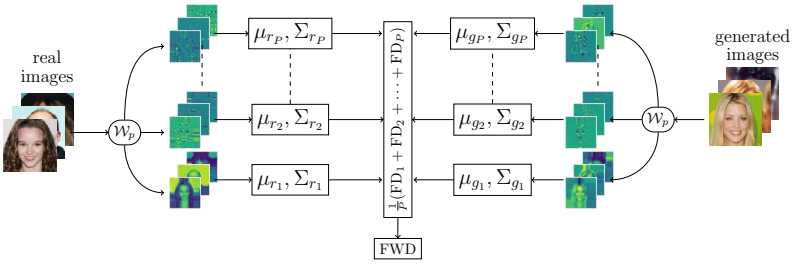


Figure 3: Fréchet Wavelet Distance (FWD) computation flow-chart. $\mathcal{W}_p$ denotes the wavelet-packet transform. Not all packet coefficients are shown, dashed lines indicate omissions. We compute individual Fréchet Distances for each packet coefficient and finally average across all the coefficients.

filters as illustrated in Fig. 2. Recursive convolution operations with the filter quadruples, i.e.,

$$\mathbf{C}_{F_l} * \mathbf{h}_j = \mathbf{C}_{\mathcal{F}_{l+1}} \tag{2}$$

at every recursion step where $*$ denotes a two-dimensional convolution with a stride of two. The filter codes $\mathcal{F}_{l+1}$ are constructed by applying all $j \in [a, h, v, d]$ filters to the previous filter codes $\mathcal{F}_l$. Initially, the set of inputs F_l will only contain the original image $\mathbf{C}_{F_0} = \{X\}$ as shown in Fig. 2. At level one, we obtain the result of all four convolutions with the input image and have $F_1 = [a, h, v, d]$. At level two, we repeat the process for all elements in F_1. F_2 now contains two-character keys $[aa, ah, av, ad, \ldots, dv, dd]$ as illustrated in Fig. 2. We typically continue this process until level 4 in this paper. We arrange the coefficients in $\mathbf{C}_{\mathcal{F}_l}$ as tensors $\mathbf{C}_l \in \mathbb{R}^{P,H_p,W_p}$ for the final layer. The total number of packages at every level is given by $P = 4^l$, and $H_p = \frac{H}{4^l}$ and $W_p = \frac{W}{4^l}$ where we denote the image height and width as H and W. We provide more details on $\mathcal{W}_p$ in the Supplementary.

Figure 3 illustrates how we compute the FWD. The process relies on the wavelet packet transform, as previously discussed. We process N images with C channels in parallel $\mathcal{W}_p : I_s \in \mathbb{R}^{N\times H\times W\times C} \rightarrow \mathbf{C} \in \mathbb{R}^{N\times P\times H_p\cdot W_p\cdot C}$. H, W denotes image height and width as before. To facilitate the ensuing metric evaluation, we flatten the last axes into $(H_p \cdot W_p \cdot C)$. Before computing the packets, all pixels are divided by 255 to re-scale all values to [0,1]. The metric is computed in three steps. First, we compute the individual packet mean via

$$\mu_p(I_N) = \frac{1}{N}\sum_{n=1}^{N} \mathcal{W}(I_n)_p, \tag{3}$$

where I_n is the nth image in the dataset and p represent the corresponding packet form P packets. Then we compute the covariance matrix as

$$\Sigma_p(I_N) = \frac{1}{N-1}\sum_{n=1}^{N}(\mathcal{W}(I_n)_p - \mu_p(I_N))(\mathcal{W}(I_N)_p - \mu_p(I_N))^T \tag{4}$$

Here $\mu \in \mathbb{R}^{P\times C\cdot H_p\cdot W_p}$ estimate the mean across the number of images, and $\Sigma \in \mathbb{R}^{P\times C\cdot H_p\cdot W_p\times C\cdot H_p\cdot W_p}$ represents the covariance among all the coefficients. Now we are ready to compute the distances given the packet mean and covariance values,

$$\mathrm{FD}_p(r, g) = d(\mathcal{N}(\mu_{r_p}, \Sigma_{r_p}), \mathcal{N}(\mu_{g_p}, \Sigma_{g_p}))^2 = ||\mu_{r_p} - \mu_{g_p}||_2^2 + \mathrm{tr}[\Sigma_{r_p} + \Sigma_{g_p} - 2\sqrt{\Sigma_{r_p}\Sigma_{g_p}}]. \tag{5}$$

(a) Proj. FastGAN on CelebA-HQ FID: 6.358 FWD: 1.388

(b) DDGAN on CelebA-HQ FID: 7.641 FWD: 0.408

Figure 4: Samples from (a) Proj. FastGAN, (b) DDGAN models on Large-scale Celeb Faces Attributes High Quality (CelebA-HQ) dataset. The FID prefers Proj. FastGAN irrespective of visual artefacts and floating heads, whereas our metric (FWD) ranks DDGAN higher than Proj. FastGAN.

With r and g denoting the real and generated images. tr denotes the trace operation. Utilising the above computed per-packet statistics for both real (μ_r, Σ_g) and generated samples (μ_r, Σ_g), we measure the mean of Fréchet Distance (Equation 5) across all packets

$$\text{FWD} = \frac{1}{P} \sum_{p=1}^{P} d(\mathcal{N}(\mu_{r_p}, \Sigma_{r_p}), \mathcal{N}(\mu_{g_p}, \Sigma_{g_p}))^2. \tag{6}$$

By averaging the distances of all frequency bands, the FWD captures frequency information across the spectrum.

4 Experiments

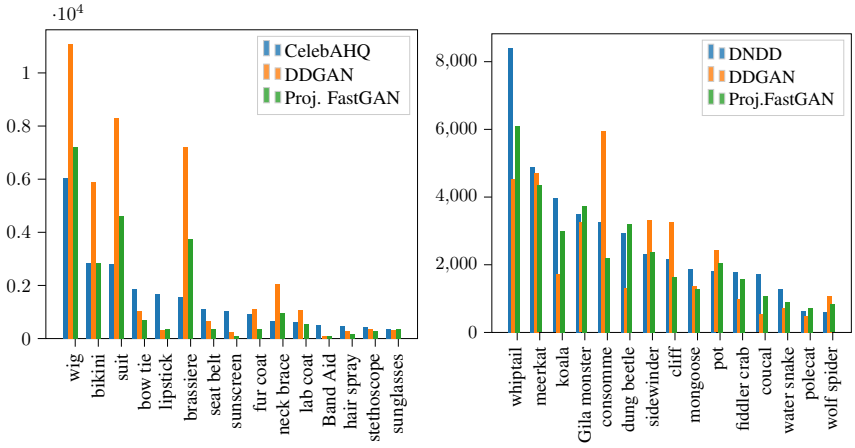
Our first series of experiments demonstrates the effect of domain bias on learned metrics, demonstrating the resilience of FWD to such bias. All experiments were implemented using the same code base.
Implementation We use PyTorch (Paszke et al., 2017) for neural network training and evaluation and compute FID using (Seitzer, 2020) as recommended by Heusel et al. (2017b). We work with the wavelet filter coefficients provided by PyWavelets (Lee et al., 2019). We chose the PyTorch-Wavelet-Toolbox (Wolter et al., 2024) software package for GPU support. FD-DINOv2 and KID are computed using the codebases from Stein et al. (2023) and Binkowski et al. (2018), respectively.

4.1 Effect of domain bias

Kynkäänniemi et al. (2023) observed that metrics based on ImageNet-trained network features emphasize ImageNet-related information. This behaviour is desired when we evaluate generators on ImageNet or similar datasets. When working with other datasets, this behaviour is misleading.
Datasets As datasets, we use Large-scale Celeb Faces Attributes High Quality (CelebA-HQ) (Karras et al., 2018), Flickr Faces High Quality (FFHQ), DNDD-Dataset (Yi et al., 2020), an agricultural dataset, and Sentinel (Schmitt et al., 2019), a remote sensing dataset. These datasets contain images that are very different from those in ImageNet. More information about the DNDD-Dataset and the Sentinel dataset can be found in the supplementary material.
Generators We study data-set domain bias effects using the Denoising Diffusion GAN (DDGAN), Proj. FastGAN and StyleGAN2 networks. Proj. FastGAN is particularly interesting. To improve training convergence, its discriminator relies on ImageNet weights (Sauer et al., 2021). Prior work found this architecture to improve FID on image datasets far from ImageNet, without substantially improving image quality (Kynkäänniemi et al., 2023).
Hyperparameters To examine the effect of data-set bias, we require generators, which are tuned to produce output that resembles our datasets' distribution. Specifically, we trained the Proj. FastGAN for 100 epochs on both CelebA-HQ dataset and DNDD-Dataset, respectively, using a learning rate of 1e-4 and batch-size of 64 with 8 A100 GPUs. For the Sentinel dataset, we trained Proj. FastGAN for 150 epochs, using the same hardware and hyperparameters. For FFHQ, pre-trained weights are available, as well as pre-trained weights for DDGAN on CelebA-HQ from Xiao et al. (2022). On DNDD-Dataset, we trained DDGAN for 150 epochs with a learning rate of 1e-4 and batch size of 8 on the same hardware. We also trained the

Table 1: Comparison of FID, FD-DINOv2 and FWD to depict domain bias. FID prefers Proj. FastGAN over DDGAN across all the datasets. Whereas FWD prefers DDGAN. We find that FD-DINOv2 agree with FWD across all datasets except DNDD-Dataset. This might be because agriculture data is not part of DINOv2's training set.

Dataset	Generator	FID ↓	FD-DINOv2↓	FWD(ours)↓
CelebA-HQ	Proj. FastGAN	**6.358**	685.889	1.388
	DDGAN	7.641	**199.761**	**0.408**
FFHQ	Proj. FastGAN	**4.106**	593.124	0.651
	StyleGAN2	4.282	**420.273**	**0.312**
DNDD-Dataset	Proj. FastGAN	**4.675**	**171.625**	1.442
	DDGAN	26.233	232.884	**1.357**
Sentinel	Proj. FastGAN	**8.96**	424.898	0.755
	DDGAN	23.615	**404.700**	**0.115**



(a) CelebA-HQ (b) DNDD-Dataset

Figure 5: Distribution of ImageNet Top-1 classes, predicted by InceptionV3 for real, DDGAN and Proj. FastGAN.(a) depict the distribution for CelebA-HQ dataset and (b) show distribution for DNDD-Dataset. Although irrelevant for FID computation, the Proj. FastGAN distribution aligns more closely with real distribution than DDGAN for both the datasets, contributing to lower FID.

DDGAN on the Sentinel dataset for 250 epochs, using a learning rate of 1e-4 and batch size of 4 on 4 A100 GPUs. For StyleGAN2, we use the pretrained weights with the code from Karras et al. (2020).

Table 1 presents the FID, KID, FD-DINOv2 and FWD values across all datasets for the aforementioned generators. Across all datasets, FID prefers Proj. FastGAN images.

Results Consider the CelebA-HQ-case in more detail. Figures 4a and 4b show images from the CelebA-HQ variants of Proj. FastGAN and DDGAN. Deformations are visible in the images from Proj. FastGAN on the left. Generally, we found more deformations in Proj. FastGAN images compared to DDGAN images. DDGAN, in other words produces more high-quality images. Supplementary Figures 12 and 13 illustrate this observation further. Consequently, it is surprising to see FID prefer Proj. FastGAN, as we would expect DDGAN to come out on top. We follow Kynkäänniemi et al. (2023), and investigate further. Figure 5a compares the InceptionV3 output label distribution of the original-CelebA-HQ images as well as their synthetic counterparts from DDGAN and Proj. FastGAN. We observe that InveptionV3 produces a label distribution for Proj. FastGAN, which resembles the distribution from InveptionV3 for the original CelebA-HQ images. The label distribution for images from DDGAN differs significantly. This discrepancy, also reported by Kynkäänniemi

Table 2: Comparison of computational efficiency between FID, FD-DINOv2 and FWD. FWD exhibit the lowest FLOPs and highest throughput. FD-DINOv2 has the highest FLOPs and lowest throughput because of deep network structure and FID in between. FLOPs are calculated over individual feature extractors on a single image and throughput is measured over 50k images.

Metric	GFLOPs↓	Throughput (imgs/sec) ↑
FID	1.114	526
FD-DINOv2	15.566	53
FWD	**0.006**	**1923**

Table 3: Evaluation of FID (ImageNet), FID (CelebA) and FWD on CelebA-HQ and FFHQ datasets. FID (ImageNet) prefers Proj. FastGAN in both datasets, whereas FID retrained on CelebA and FWD both prefer DDGAN in these datasets.

Dataset	Generator	FID (ImageNet) ↓	FID (CelebA) ↓	FWD(ours) ↓
CelebA-HQ	Proj. FastGAN	**6.358**	5.602	1.388
	DDGAN	7.641	**3.145**	**0.408**
FFHQ	Proj. FastGAN	**4.106**	2.204	0.651
	StyleGAN2	4.282	**0.897**	**0.312**

et al. (2023), explains why FID produces a misleading verdict. FWD, in contrast, prefers DDGAN, as we would expect.

The same pattern repeats in the results for our FFHQ-experiments, generally we see FID prefer Proj. FastGAN images, while FWD puts DDGAN on top. Our observations confirm the experiment in Kynkäänniemi et al. (2023). In a next step we study the effect of a larger network backbone for the neural Frećet distance computations. Stein et al. (2023) proposes to replace InveptionV3 with the much larger pretrained DINOv2 network. Table 1 lists the resulting distance metrics. For CelebA-HQ and FFHQ, FD-DINOv2 prefers DDGAN images. Here FD-DINOv2 and FWD agree.

To investigate further we consider the DNDD-Dataset of agricultural images (Yi et al., 2020) and the Sentinel (Schmitt et al., 2019) datasets. Samples from the Proj. FastGAN for DNDD-Dataset and Sentinel datasets are provided in Figures 15 and 17, respectively. Correspondingly, Figures 14 and 16 represent samples from DDGAN for DNDD-Dataset and Sentinel dataset, respectively. In both cases, FID consistently prefers Proj. FastGAN, which was also the case in all prior experiments. Histograms of the InceptionV3 label distribution are depicted in Figure 5b. The histograms indicate domain bias and resemble the observations reported above. On the DNDD-Dataset and Sentinel-datasets, the verdicts of FD-DINOv2 and FWD are particularly interesting. While both metrics correctly agree on the sentinel dataset, only FWD correctly prefers DDGAN on the agricultural images.

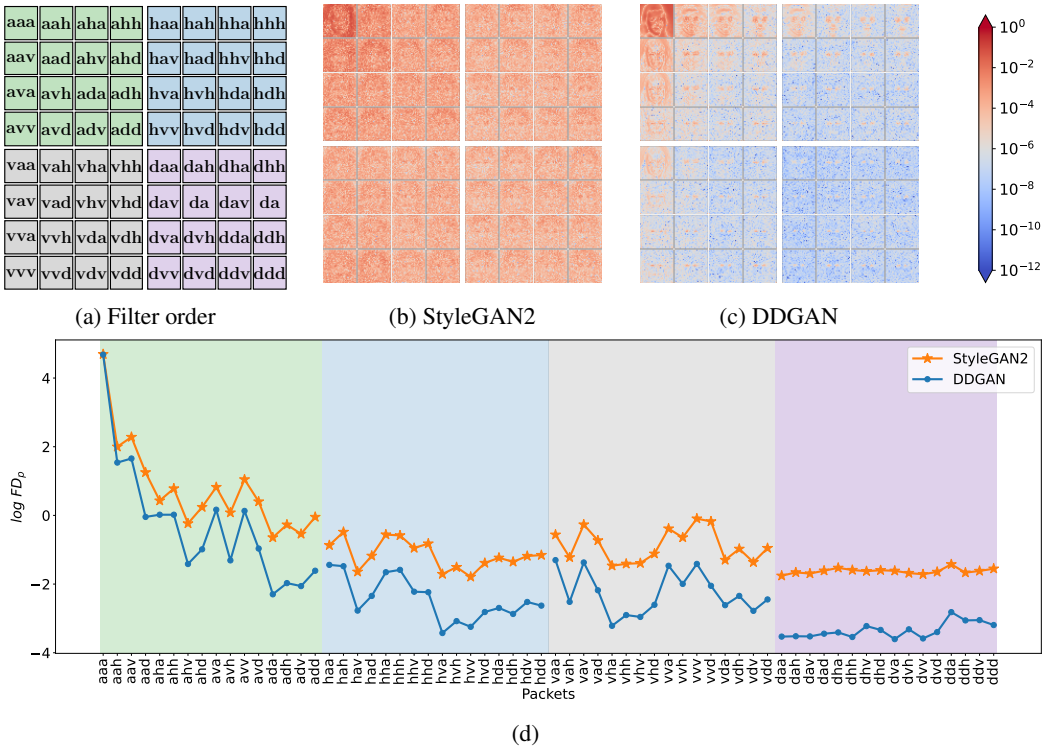
We carefully chose the DNDD-Dataset dataset, as agriculture images are not commonly used and the dataset does not resemble ImageNet. We speculate that the LVD-142M dataset may include satellite imagery, contributing to a consistent ranking. Unfortunately, the closed source of the LVD-142M dataset used for training DINOv2 (Oquab et al., 2023) complicates a deeper investigation into this domain bias. In this first set of experiments, we observed that while FD-DINOv2 provides a partial remedy to the domain bias problems it still produced an inconsistent ordering for the DNDD-Dataset-images. Furthermore, this partial remedy comes at a great computational cost. Table 2 shows that FWD is over 36 times faster to compute than FD-DINOv2.

In a second series of experiments, we investigate the effect of retraining which is another expensive solution to the domain bias problem. To this end, we train InceptionV3 on Large-scale Celeb Faces Attributes (CelebA). CelebA comes with 40 facial attributes, which we use to train a classifier. After convergence, we see an exact match ratio of 90% and recalculate FID using this new backbone. The FID (CelebA) column of Table 3 lists the corresponding scores, and FID (CelebA) and FWD provide the same order.

However, in the case of the agricultural data set, the retrained FID (DNDD) in Table 4 remains biased, while FWD produces meaningful domain agnostic results. DNDD-Dataset contains 3600 images

Table 4: Evaluation of FID (ImageNet) and FID (DNDD) and FWD on theDNDD-Dataset. FID trained (DNDD) still prefer Proj. FastGAN in this case, because of limited data availability to extract meaningful representations from the InceptionV3 network. Whereas FWD rank DDGAN images better.

Dataset	Generator	FID (ImageNet) ↓	FID (DNDD) ↓	FWD(ours) ↓
DNDD-Dataset	Proj. FastGAN	**4.675**	**20.937**	1.442
	DDGAN	26.233	52.521	**1.357**



(a) Filter order (b) StyleGAN2 (c) DDGAN

(d)

Figure 6: Interpretation of FWD. (a) represents the blueprint for level-3 $\mathcal{W}_p$ transformation. (b) and (c) depict the mean absolute packet difference between CelebA-HQ dataset and StyleGAN2, and DDGAN respectively. (d) shows the per-packet Fréchet distances for StyleGAN2 in orange and DDGAN in blue.

with 7 classes. Networks are tasked with detecting nutrient deficiency in the soil, such as Sodium, Calcium, unfertilized, and 4 others. Once more, we use a re-trained InceptionV3 backbone for the FID computation. In comparison to CelebA or ImageNet, this is a small data set and the re-trained network does not provide meaningful features. We believe this is a very interesting use case since it illustrates that the FWD is not just free from data bias, it also provides meaningful feedback for low resource tasks, where retraining InceptionV3 is not feasible.

In conclusion, experiments in this section indicate that metrics like FID, and FD-DINOv2, while useful, are prone to domain bias when applied to datasets beyond ImageNet. Contrary, FWD offers a more consistent, domain-agnostic evaluation along with computation efficiency.

4.2 FWD INTERPRETABILITY

A generative metric is interpretable if and only if we can understand the underlying mechanics that produce the ranking. This section explains the decisions made by FWD in one specific case.

Dataset and Generators We use CelebA-HQ and we focus on samples from DDGAN and StyleGAN2.

Table 5: Matching fringe features for 250k images of FFHQ dataset generated using StyleGAN2. By optimizing FID, FD-DINOv2 gets optimized indirectly, because ImageNet is part of the DINOv2 training set. Whereas FWD stays resilient to FID optimization.

Metric	Random Images	FID-Optimized Images	Change
FID	4.278 ± 0.019	$\mathbf{2.031 \pm 0.005}$	-52.53%
FD-DINOv2	420.223 ± 0.563	$\mathbf{414.048 \pm 0.905}$	-1.47%
FWD	$\mathbf{0.338 \pm 0.017}$	0.398 ± 0.009	+17.75%

Table 6: Comparing various generative models using Fréchet Wavelet Distance (FWD),Fréchet Inception Distance (FID), DINOv2-Fréchet Distance (FD-DINOv2), and Kernel Inception Distance (KID) on CelebA-HQ, LSUN-Churches, LSUN-Bedrooms and ImageNet datasets.

Dataset	Image Size	Method	FID↓	KID↓	FD-DINOv2↓	FWD↓ (ours)
CelebAHQ	256	DDIM Song et al. (2021)	32.333	0.0313	654.482	12.317
		DDPM (Ho et al., 2020)	19.101	0.0152	341.838	4.697
		StyleSwin (Zhang et al., 2022)	23.257	0.0264	255.404	1.528
		StyleGAN2 (Karras et al., 2020)	15.439	0.0155	593.344	0.476
		DDGAN (Xiao et al., 2022)	**7.203**	**0.0034**	**199.761**	**0.408**
Churches	256	DDIM (Song et al., 2021)	11.775	0.0043	538.400	4.919
		DDPM (Ho et al., 2020)	9.484	0.0036	454.402	3.546
		StyleSwin (Zhang et al., 2022)	**3.187**	**0.0005**	**435.967**	2.835
		StyleGAN2 (Karras et al., 2020)	4.309	0.0007	444.044	**0.753**
Bedrooms	256	DDIM (Song et al., 2021)	25.857	0.0094	452.419	9.521
		DDPM (Ho et al., 2020)	**16.251**	**0.0058**	**392.481**	**5.187**
ImageNet	64	Imp. Diff. (VLB) (Nichol & Dhariwal, 2021)	33.522	0.0264	670.952	2.182
		EDM (Karras et al., 2024)	12.295	0.0108	113.704	1.160
		BigGAN (Brock et al., 2019)	5.128	0.0024	170.601	0.441
		Imp. Diff. (Hybrid) (Nichol & Dhariwal, 2021)	**3.091**	**0.0006**	**96.208**	**0.392**

Section 3 formulates FWD as an average of per packet FWD scores. This design choice allows us to understand the overall FWD-score in terms of the individual packet coefficients for each frequency band. Figures 6b and 6c depict the mean absolute difference per packet between the original CelebA-HQ and generated samples from StyleGAN2 and DDGAN, respectively. Figure 6d presents both generators' per-packet FWD. Figure 6(d) shows that DDGAN has a lower Fréchet distance. Overall, we observe that the mean absolute packets translate into per packet Fréchet distances, which validates the FWD overall.

4.3 Evaluation of Robustness

The section follows up on prior work by Kynkäänniemi et al. (2023). The authors generate a large set of samples and find a specific combination of images with an optimal FID. Weighted sampling produces possible combinations. The process chooses images according to a corresponding weight. The weights or drawing probabilities are optimized with FID as the objective function. We follow this process and sample 50k images from a large set with optimized weights as probabilities. We employ generated images from StyleGAN2 and real-world images from the FFHQ dataset. Table 5 lists the resulting FID, FD-DINOv2 and FWD values. We observe that FWD is robust to FID optimization, whereas FD-DINOv2 showed a little reduction by optimizing FID.

In addition to FID optimization, we study the impact of image perturbation in supplementary Figure 7. We find that FWD and FD-DINOv2 is closer to a bijective mapping in the presence of perturbation than FID. This behaviour is desirable since we would always expect a larger distance if for example more noise is added. This is not always the case for FID. Consider for example the last quarter of the uniform noise intensity in (b), where FID falls even though more noise is added.

4.4 Comparison to State-of-the-art

To understand the spectral qualities of existing generative methods for image synthesis, we evaluated various Diffusion and GAN models across a wide range of benchmark datasets.
Datasets We compare common metrics and our FWD on CelebA-HQ (Karras et al., 2018), the Church and Bedroom subsets of the Large-scale Scene UNderstanding (LSUN) dataset (Yu et al., 2015), and finally ImageNet (Russakovsky et al., 2015). In order to retain consistent spatial and frequency characteristics across various image sizes, we use level 4 packet transform for 256x256 images. Furthermore, we reduce the transformation level by a factor of 1 for reduction in image size by half.
Generators For the evaluation, we use the diffusion approaches Denoising Diffusion Probabilistic Models (DDPM) (Ho et al., 2020), Denoising Diffusion Implicit Models (DDIM) (Song et al., 2021), Improved Diffusion (Nichol & Dhariwal, 2021), DDGAN (Xiao et al., 2022), EDM (Karras et al., 2024), as well as the GAN approaches like StyleGAN2 (Karras et al., 2019), StyleSwin (Zhang et al., 2022) and BIGGAN (Brock et al., 2019).
Hyperparameters All generators are evaluated with pretrained weights as provided by the respective paper codebases.
Metrics We compute FID, KID, FD-DINOv2 and finally our own FWD. Table 6 lists all numbers. FID-scores are from the standard implementation by Seitzer (2020). For CIFAR10, we use 50k images to evaluate all metrics. The ImageNet numbers are computed with 50k images from the validation set. For CelebAHQ and LSUN we work with 30k images.

Considering CelebA-HQ, FID, KID, FD-DINOv2 and FWD agree most of the time. Considering FID and FWD, only DDPM and StyleSwin are swapped. FWD therefore delivers a comparable quality metric in this case. The ordering remains largely unchanged for LSUN churches. In terms of FWD, we observe a stable ranking across the two datasets. Except DDPM and StyleSwin, which are swapped. We observe larger changes in magnitude for FID-scores when making the switch from CelebA-HQ to LSUN churches. The switch moves us towards ImageNet, since "church" is an ImageNet class, but "face" is not. Supplementary Figure 11b depicts the histograms of top-1 classes classified by InceptionV3 on LSUN churches for DDPM and StyleSwin. We observe that StyleSwin matches the activation histograms of LSUN churches more accurately than the histograms of DDPM. This observation is a manifestation of domain bias and explains the FID inconsistency for both generators.

We also consider the LSUN-Bedrooms and ImageNet 64 datasets, where FID and FWD agree. We expect pristine performance for FID on ImageNet since this setting is perfectly in its data-domain. Yet, FD-DINOv2 places EDM (Karras et al., 2024) ahead of BigGAN, which is surprising since this does not match with the ranking from FID. FID and FWD agree and arrive at the same ranking.

5 Conclusion

Modern generative models exhibit frequency biases (Durall et al., 2020), while commonly used metrics such as FID, KID and FD-DINOv2 are affected by domain bias (Kynkäänniemi et al., 2023). To address these limitations, FWD accounts for frequency information without introducing a domain-specific bias. Even though FD-DINOv2 offers a partial solution to this issue, it comes at a very high computational cost and has thus a negative environmental impact. In response, this paper introduced FWD a novel metric based on the wavelet packet transform. Our metric allows consistent, domain-agnostic evaluation. At the same time, its formulation is computationally efficient. Our findings show that FWD is robust to input perturbations and interpretable through the analysis of individual frequency bands. Optimizing FID or FD-DINOv2 metrics can negatively impact reproducibility, if optimized samples are not provided. In such cases, the use of FWD in conjunction with traditional metrics ensures a comprehensive and accurate evaluation of generative models while also helping to detect and mitigate domain bias.

References

Flickr image data base. `https://www.flickr.com`, 2024. Accessed: 2024-09-30, Images are distributed using various Creative Commons Licenses.

Evaluating diffusion models. https://huggingface.co/docs/diffusers/conceptual/evaluation, 2024. Accessed: 2024-04-11.

Shane T. Barratt and Rishi Sharma. A note on the inception score. *ICML 2018 workshop on Theoretical Foundations and Applications of Deep Generative Models.*, 2018. URL http://arxiv.org/abs/1801.01973.

Mikolaj Binkowski, Danica J. Sutherland, Michael Arbel, and Arthur Gretton. Demystifying MMD gans. In *6th International Conference on Learning Representations, ICLR 2018, Vancouver, BC, Canada, April 30 - May 3, 2018, Conference Track Proceedings*, 2018.

Andrew Brock, Jeff Donahue, and Karen Simonyan. Large scale GAN training for high fidelity natural image synthesis. In *International Conference on Learning Representations*, 2019. URL https://openreview.net/forum?id=B1xsqj09Fm.

Min Jin Chong and David A. Forsyth. Effectively unbiased FID and inception score and where to find them. In *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2020, Seattle, WA, USA, June 13-19, 2020*, pp. 6069–6078. Computer Vision Foundation / IEEE, 2020.

Ingrid Daubechies. *Ten Lectures on Wavelets*. Society for Industrial and Applied Mathematics, 1992.

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

DC Dowson and BV666017 Landau. The fréchet distance between multivariate normal distributions. *Journal of multivariate analysis*, 12(3):450–455, 1982.

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*, pp. 7890–7899, 2020.

Tarik Dzanic, Karan Shah, and Freddie Witherden. Fourier spectrum discrepancies in deep network generated images. *Advances in neural information processing systems*, 33:3022–3032, 2020.

Rinon Gal, Dana Cohen Hochberg, Amit Bermano, and Daniel Cohen-Or. Swagan: A style-based wavelet-driven generative model. *ACM Trans. Graph.*, 40(4), July 2021.

Florentin Guth, Simon Coste, Valentin De Bortoli, and Stephane Mallat. Wavelet score-based generative modeling. *Advances in Neural Information Processing Systems*, 35:478–491, 2022.

Alejandro Hernandez, Jurgen Gall, and Francesc Moreno-Noguer. Human motion prediction via spatio-temporal inpainting. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 7134–7143, 2019.

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. *Advances in neural information processing systems*, 30, 2017a.

Martin Heusel, Thomas Unterthiener, Wendy Kan, Mark Hamilton, Zejian Li, Marc Uecker, Wang Penghui, and Partik Joshi. Two time-scale update rule for training gans. https://github.com/bioinf-jku/TTUR, 2017b.

Jonathan Ho, Ajay Jain, and Pieter Abbeel. Denoising diffusion probabilistic models. In Hugo Larochelle, Marc'Aurelio Ranzato, Raia Hadsell, Maria-Florina Balcan, and Hsuan-Tien Lin (eds.), *Advances in Neural Information Processing Systems 33: Annual Conference on Neural Information Processing Systems 2020, NeurIPS 2020, December 6-12, 2020, virtual*, 2020.

Alain Horé and Djemel Ziou. Image quality metrics: PSNR vs. SSIM. In *20th International Conference on Pattern Recognition, ICPR 2010, Istanbul, Turkey, 23-26 August 2010*, pp. 2366–2369. IEEE Computer Society, 2010. doi: 10.1109/ICPR.2010.579. URL https://doi.org/10.1109/ICPR.2010.579.

Huaibo Huang, Ran He, Zhenan Sun, and Tieniu Tan. Wavelet-srnet: A wavelet-based cnn for multi-scale face super resolution. In *Proceedings of the IEEE international conference on computer vision*, pp. 1689–1697, 2017.

Arne Jensen and Anders la Cour-Harbo. *Ripples in mathematics: the discrete wavelet transform*. Springer Science & Business Media, 2001.

via Wikimedia Commons Jérémy Barande. Wikimedia commons: Yann lecun - 2018 (cropped). https://en.wikipedia.org/wiki/Yann_LeCun#/media/File:Yann_LeCun_-_2018_(cropped).jpg, 2024. Accessed: 2024-05-22.

Tero Karras, Timo Aila, Samuli Laine, and Jaakko Lehtinen. Progressive growing of gans for improved quality, stability, and variation. In *6th International Conference on Learning Representations, ICLR 2018, Vancouver, BC, Canada, April 30 - May 3, 2018, Conference Track Proceedings*, 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*, pp. 4401–4410, 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*, pp. 8110–8119, 2020.

Tero Karras, Miika Aittala, Samuli Laine, Erik Härkönen, Janne Hellsten, Jaakko Lehtinen, and Timo Aila. Alias-free generative adversarial networks. *Advances in Neural Information Processing Systems*, 34:852–863, 2021.

Tero Karras, Miika Aittala, Jaakko Lehtinen, Janne Hellsten, Timo Aila, and Samuli Laine. Analyzing and improving the training dynamics of diffusion models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 24174–24184, 2024.

Diederik P. Kingma and Max Welling. Auto-encoding variational bayes. In Yoshua Bengio and Yann LeCun (eds.), *2nd International Conference on Learning Representations, ICLR 2014, Banff, AB, Canada, April 14-16, 2014, Conference Track Proceedings*, 2014.

Tuomas Kynkäänniemi, Tero Karras, Samuli Laine, Jaakko Lehtinen, and Timo Aila. Improved precision and recall metric for assessing generative models. In Hanna M. Wallach, Hugo Larochelle, Alina Beygelzimer, Florence d'Alché-Buc, Emily B. Fox, and Roman Garnett (eds.), *Advances in Neural Information Processing Systems 32: Annual Conference on Neural Information Processing Systems 2019, NeurIPS 2019, December 8-14, 2019, Vancouver, BC, Canada*, pp. 3929–3938, 2019.

Tuomas Kynkäänniemi, Tero Karras, Miika Aittala, Timo Aila, and Jaakko Lehtinen. The role of imagenet classes in fréchet inception distance. In *The Eleventh International Conference on Learning Representations, ICLR 2023, Kigali, Rwanda, May 1-5, 2023*. OpenReview.net, 2023. URL https://openreview.net/pdf?id=4oXTQ6m_ws8.

Gregory Lee, Ralf Gommers, Filip Waselewski, Kai Wohlfahrt, and Aaron O'Leary. Pywavelets: A python package for wavelet analysis. *Journal of Open Source Software*, 4(36):1237, 2019.

Jin Li, Wanyun Li, Zichen Xu, Yuhao Wang, and Qiegen Liu. Wavelet transform-assisted adaptive generative modeling for colorization. *IEEE Transactions on Multimedia*, 2022.

Lin Liu, Jianzhuang Liu, Shanxin Yuan, Gregory Slabaugh, Aleš Leonardis, Wengang Zhou, and Qi Tian. Wavelet-based dual-branch network for image demoiréing. In *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part XIII 16*, pp. 86–102. Springer, 2020.

Yunfan Liu, Qi Li, and Zhenan Sun. Attribute-aware face aging with wavelet-based generative adversarial networks. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 11877–11886, 2019.

Stéphane Mallat. A theory for multiresolution signal decomposition: The wavelet representation. *IEEE Trans. Pattern Anal. Mach. Intell.*, 11(7):674–693, 1989.

Stéphane Mallat. *A wavelet tour of signal processing*. Elsevier, 1999.

Manish Narwaria, Weisi Lin, Ian Vince McLoughlin, Sabu Emmanuel, and Liang-Tien Chia. Fourier transform-based scalable image quality measure. *IEEE Trans. Image Process.*, 21(8):3364–3377, 2012.

Alexander Quinn Nichol and Prafulla Dhariwal. Improved denoising diffusion probabilistic models. In *Proceedings of the 38th International Conference on Machine Learning*, volume 139 of *Proceedings of Machine Learning Research*, pp. 8162–8171. PMLR, 18–24 Jul 2021.

Maxime Oquab, Timothée Darcet, Théo Moutakanni, Huy Vo, Marc Szafraniec, Vasil Khalidov, Pierre Fernandez, Daniel Haziza, Francisco Massa, Alaaeldin El-Nouby, et al. Dinov2: Learning robust visual features without supervision. *arXiv preprint arXiv:2304.07193*, 2023.

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.

Adam Paszke, Sam Gross, Soumith Chintala, Gregory Chanan, Edward Yang, Zachary DeVito, Zeming Lin, Alban Desmaison, Luca Antiga, and Adam Lerer. Automatic differentiation in pytorch. In *31st Conference on Neural Information Processing Systems (NIPS 2017)*, 2017.

William Peebles and Saining Xie. Scalable diffusion models with transformers. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 4195–4205, 2023.

Hao Phung, Quan Dao, and Anh Tran. Wavelet diffusion models are fast and scalable image generators. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2023, Vancouver, BC, Canada, June 17-24, 2023*, pp. 10199–10208. IEEE, 2023. doi: 10.1109/CVPR52729.2023.00983. URL https://doi.org/10.1109/CVPR52729.2023.00983.

Nasim Rahaman, Aristide Baratin, Devansh Arpit, Felix Draxler, Min Lin, Fred Hamprecht, Yoshua Bengio, and Aaron Courville. On the spectral bias of neural networks. In *International Conference on Machine Learning*, pp. 5301–5310. PMLR, 2019.

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 (IJCV)*, 115 (3):211–252, 2015. doi: 10.1007/s11263-015-0816-y.

Tim Salimans, Ian Goodfellow, Wojciech Zaremba, Vicki Cheung, Alec Radford, and Xi Chen. Improved techniques for training gans. *Advances in neural information processing systems*, 29, 2016.

Vishwanath Saragadam, Daniel LeJeune, Jasper Tan, Guha Balakrishnan, Ashok Veeraraghavan, and Richard G Baraniuk. Wire: Wavelet implicit neural representations. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 18507–18516, 2023.

Axel Sauer, Kashyap Chitta, Jens Müller, and Andreas Geiger. Projected gans converge faster. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2021.

Michael Schmitt, Lloyd Haydn Hughes, Chunping Qiu, and Xiao Xiang Zhu. Sen12ms–a curated dataset of georeferenced multi-spectral sentinel-1/2 imagery for deep learning and data fusion. *arXiv preprint arXiv:1906.07789*, 2019.

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

Maximilian Seitzer. pytorch-fid: FID Score for PyTorch. https://github.com/mseitzer/pytorch-fid, August 2020. Version 0.3.0.

Jascha Sohl-Dickstein, Eric Weiss, Niru Maheswaranathan, and Surya Ganguli. Deep unsupervised learning using nonequilibrium thermodynamics. In *International conference on machine learning*, pp. 2256–2265. PMLR, 2015.

Jiaming Song, Chenlin Meng, and Stefano Ermon. Denoising diffusion implicit models. In *9th International Conference on Learning Representations, ICLR 2021, Virtual Event, Austria, May 3-7, 2021*, 2021.

George Stein, Jesse Cresswell, Rasa Hosseinzadeh, Yi Sui, Brendan Ross, Valentin Villecroze, Zhaoyan Liu, Anthony L Caterini, Eric Taylor, and Gabriel Loaiza-Ganem. Exposing flaws of generative model evaluation metrics and their unfair treatment of diffusion models. In *Advances in Neural Information Processing Systems*, volume 36, 2023.

Gilbert Strang and Truong Nguyen. *Wavelets and filter banks*. SIAM, 1996.

Aaron Van Den Oord, Oriol Vinyals, et al. Neural discrete representation learning. *Advances in neural information processing systems*, 30, 2017.

Aparna Vyas, Soohwan Yu, and Joonki Paik. *Multiscale transforms with application to image processing*. Springer, 2018.

Jianyi Wang, Xin Deng, Mai Xu, Congyong Chen, and Yuhang Song. Multi-level wavelet-based generative adversarial network for perceptual quality enhancement of compressed video. In *European Conference on Computer Vision*, pp. 405–421. Springer, 2020.

Zhou Wang, Alan C. Bovik, Hamid R. Sheikh, and Eero P. Simoncelli. Image quality assessment: from error visibility to structural similarity. *IEEE Trans. Image Process.*, 13(4):600–612, 2004.

Travis Williams and Robert Li. Wavelet pooling for convolutional neural networks. In *International conference on learning representations*, 2018.

Moritz Wolter, Felix Blanke, Raoul Heese, and Jochen Garcke. Wavelet-packets for deepfake image analysis and detection. *Machine Learning*, 111(11):4295–4327, 2022.

Moritz Wolter, Felix Blanke, Jochen Garcke, and Charles Tapley Hoyt. ptwt - the pytorch wavelet toolbox. *Journal of Machine Learning Research*, 25(80):1–7, 2024. URL `http://jmlr.org/papers/v25/23-0636.html`.

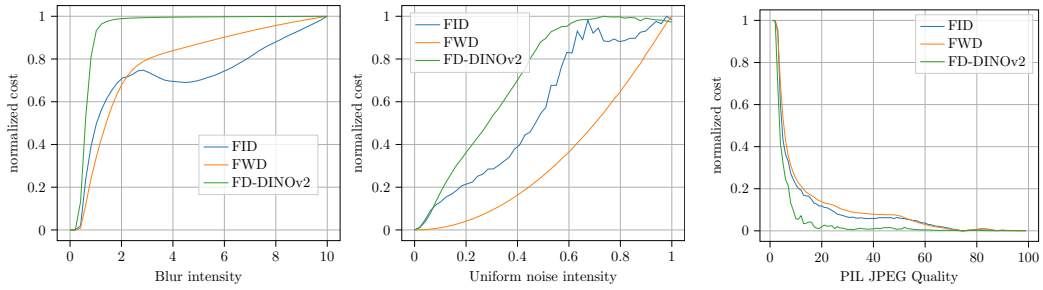
Zhisheng Xiao, Karsten Kreis, and Arash Vahdat. Tackling the generative learning trilemma with denoising diffusion gans. In *International Conference on Learning Representations*, 2022.

Jinhui Yi, Lukas Krusenbaum, Paula Unger, Hubert Hüging, Sabine J Seidel, Gabriel Schaaf, and Juergen Gall. Deep learning for non-invasive diagnosis of nutrient deficiencies in sugar beet using rgb images. *Sensors*, 20:5893, 2020.

Jaejun Yoo, Youngjung Uh, Sanghyuk Chun, Byeongkyu Kang, and Jung-Woo Ha. Photorealistic style transfer via wavelet transforms. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 9036–9045, 2019.

Fisher Yu, Yinda Zhang, Shuran Song, Ari Seff, and Jianxiong Xiao. Lsun: Construction of a large-scale image dataset using deep learning with humans in the loop. *ArXiv*, abs/1506.03365, 2015. URL `https://api.semanticscholar.org/CorpusID:8317437`.

Bowen Zhang, Shuyang Gu, Bo Zhang, Jianmin Bao, Dong Chen, Fang Wen, Yong Wang, and Baining Guo. Styleswin: Transformer-based gan for high-resolution image generation. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 11304–11314, 2022.



(a) (b) (c)

Figure 7: Figures depicting the effect of perturbations like (a) Gaussian blurs and (b) uniform noise corruption as well as (c) jpeg-compression on FID, FWD and FD-DINOv2.

A Supplementary

A.1 Acronyms

$\mathcal{W}_p$ Wavelet Packet Transform

CelebA Large-scale Celeb Faces Attributes

CelebA-HQ Large-scale Celeb Faces Attributes High Quality

CNN Convolutional Neural Network

DDGAN Denoising Diffusion GAN

DDIM Denoising Diffusion Implicit Models

DDPM Denoising Diffusion Probabilistic Models

DNDD-Dataset Deep Nutrient Deficiency Dikopshof Dataset

FD-DINOv2 DINOv2-Fréchet Distance

FFHQ Flickr Faces High Quality

FID Fréchet Inception Distance

FWD Fréchet Wavelet Distance

FWT Fast Wavelet Transform

GAN Generative Adversarial neural Network

IS Inception Score

KID Kernel Inception Distance

LSUN Large-scale Scene UNderstanding

MSE Mean Squared Error

Proj. FastGAN Projected Fast GAN

VAE Variational AutoEncoder

A.2 The fast wavelet and wavelet packet transforms

This supplementary section summarizes key wavelet facts as a convenience for the reader. See, for example, Strang & Nguyen (1996); Mallat (1999) or Jensen & la Cour-Harbo (2001) for excellent detailed introductions to the topic.

The Fast Wavelet Transform (FWT) relies on convolution operations with filter pairs. Figure 8 illustrates the process. The forward or analysis transform works with a low-pass $\mathbf{h}_{\mathcal{L}}$ and a high-pass filter $\mathbf{h}_{\mathcal{H}}$. The analysis transform repeatedly convolves with both filters,

$$\mathbf{x}_s *_1 \mathbf{h}_k = \mathbf{c}_{k,s+1} \tag{7}$$

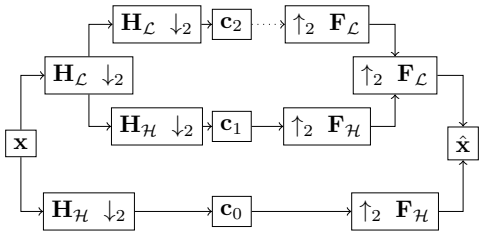


Figure 8: Overview of the Fast Wavelet Transform (FWT) computation. $\mathbf{h}_\mathcal{L}$ denotes the analysis low-pass filter and $\mathbf{h}_\mathcal{H}$ the analysis high pass filter. $\mathbf{f}_\mathcal{L}$ and $\mathbf{f}_\mathcal{H}$ the synthesis filter pair. $\downarrow_2$ denotes downsampling with a factor of two, $\uparrow_2$ means upsampling. The analysis transform relies on stride two convolutions. The synthesis or inverse transform on the right works with stride two transposed convolutions. $\mathbf{H}_k$ and $\mathbf{F}_k$ with $k \in [\mathcal{L}, \mathcal{H}]$ denote the corresponding convolution operators.

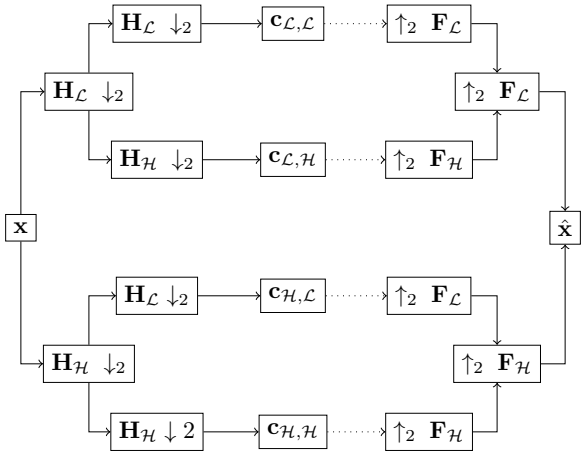


Figure 9: Scematic drawing of the full Wavelet Packet Transform ($\mathcal{W}_p$) in a single dimension. Compared to Figure 8, the high-pass filtered side of the tree is expanded, too.

with $k \in [\mathcal{L}, \mathcal{H}]$ and $s \in \mathbb{N}_0$ the set of natural numbers, where $\mathbf{x}_0$ is equal to the original input signal $\mathbf{x}$. At higher scales, the FWT uses the low-pass filtered result as input, $\mathbf{x}_s = \mathbf{c}_{\mathcal{L},s}$ if $s > 0$. And $*_1$ as the 1d-convolution operation. The dashed arrow in Figure 8 indicates that we could continue to expand the FWT tree here.

The Wavelet Packet Transform ($\mathcal{W}_p$) additionally expands the high-frequency part of the tree. A comparison of Figures 8 and 9 illustrates this difference. Whole expansion is not the only possible way to construct a wavelet packet tree. See Jensen & la Cour-Harbo (2001) for a discussion of other options. In both figures, capital letters denote convolution operators. These may be expressed as Toeplitz matrices Strang & Nguyen (1996). The matrix nature of these operators explains the capital boldface notation. Coefficient subscripts record the path that leads to a particular coefficient.

We construct filter quadruples from the original filter pairs to process two-dimensional inputs. The process uses outer products Vyas et al. (2018):

$$\mathbf{h}_a = \mathbf{h}_\mathcal{L}\mathbf{h}_\mathcal{L}^T, \mathbf{h}_h = \mathbf{h}_\mathcal{L}\mathbf{h}_\mathcal{H}^T, \mathbf{h}_v = \mathbf{h}_\mathcal{H}\mathbf{h}_\mathcal{L}^T, \mathbf{h}_d = \mathbf{h}_\mathcal{H}\mathbf{h}_\mathcal{H}^T \tag{8}$$

With a for approximation, h for horizontal, v for vertical, and d for diagonal Lee et al. (2019). We can construct a $\mathcal{W}_p$-tree for images with these two-dimensional filters. Figure 10 illustrates the computation of a full two-dimensional wavelet packet tree. More formally, the process initially evaluates

$$\mathbf{x}_0 * \mathbf{h}_j = \mathbf{c}_{j,1} \tag{9}$$

with $\mathbf{x}_0$ equal to an input image $\mathbf{X}$, $j \in [a, h, v, d]$, and $*$ for two-dimensional convolution. At higher scales, all resulting coefficients from previous scales serve as inputs. The four filters are repeatedly convolved with all outputs to build the full tree. The inverse transforms work analogously. We refer

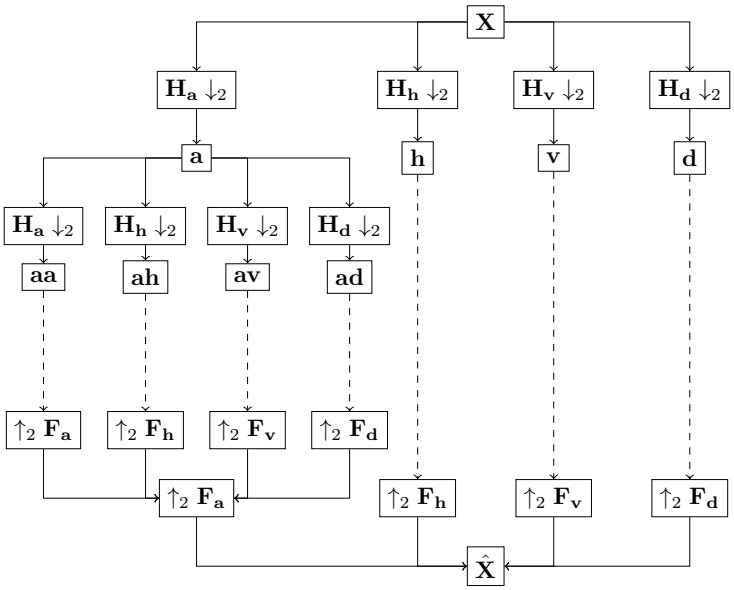


Figure 10: Two dimensional Wavelet Packet Transform ($\mathcal{W}_p$) computation overview. $\mathbf{X}$ and $\hat{\mathbf{X}}$ denote input image and reconstruction respectively. We compute the Fréchet Wavelet Distance (FWD) using the wavelet packet coefficients $\mathbf{p}$. The transform is invertible, the distance computation is therefore based on a lossless representation.

to the standard literature Jensen & la Cour-Harbo (2001); Strang & Nguyen (1996) for an extended discussion.

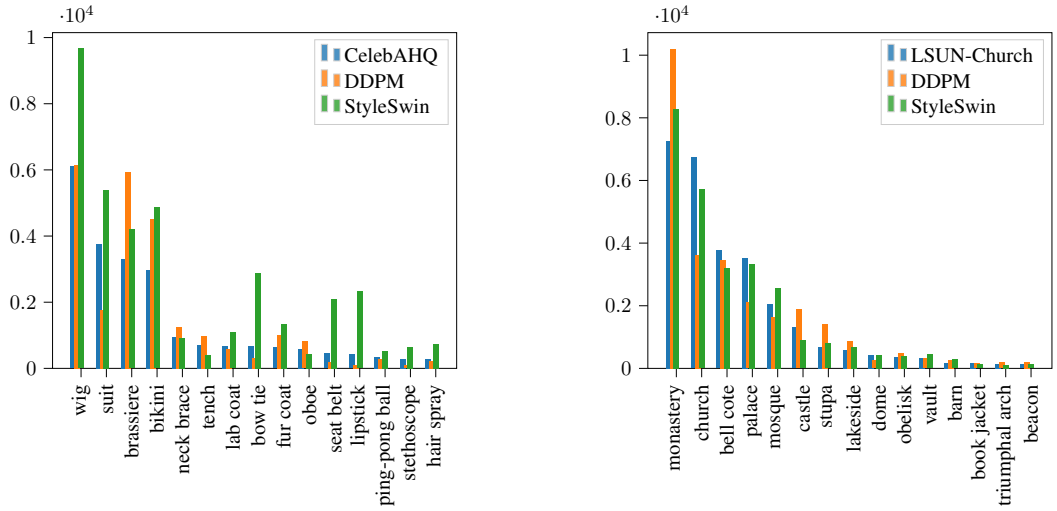
Compared to the FWT, the high-frequency half of the tree is subdivided into more bins, yielding a fine-grained view of the entire spectrum. We always show analysis and synthesis transforms to stress that all wavelet transforms are lossless. Synthesis transforms reconstruct the original input based on the results from the analysis transform.

A.3 Histogram matching - InceptionV3

A.4 Generative Architectures

Prior work mainly falls into the three GAN, Diffusion, and Variational AutoEncoder (VAE) architecture groups. The StyleGAN architecture family Karras et al. (2019; 2020; 2021) is among the pioneering architectures in generative vision. GAN's allow rapid generation of high-quality images but suffer from training instability and poor mode coverage Salimans et al. (2016). Sauer et al. (2021) proposed the Projected Fast GAN (Proj. FastGAN)-architecture, which stabilizes and improves training convergence by introducing ImageNet pre-trained weights into the discriminator. The upgraded discriminator pushes the output distribution towards ImageNet. VAE models, on the other hand, Kingma & Welling (2014); Van Den Oord et al. (2017) enable the generation of diverse image sets, but are unable to produce high-quality images.

Diffusion models Sohl-Dickstein et al. (2015); Ho et al. (2020); Peebles & Xie (2023) have emerged as a very promising alternative and produce high-quality images Ho et al. (2020); Dhariwal & Nichol (2021) in an autoregressive style. DDPMs, for example, are Markovian processes that learn to gradually separate added noise from data during training. During infernce images are generated from Gaussian noise via a reverse process that requires iterating through all steps to generate an image. Song et al. (2021) reduced the number of sampling steps by introducing DDIM, which rely on a deteministic non-Markovian sampling process. Furthermore, Nichol & Dhariwal (2021) proposed the use of strided sampling, to reduce the sampling timesteps and also provide a performance improvement by using cosine- instead of linear sampling. Moreover, Nichol & Dhariwal (2021) adopt a weighted variational lower bound to supplement the Mean Squared Error (MSE) loss. In an attempt to solve the generative learning trilemma (image quality, diversity and fast sampling), Xiao



(a) CelebAHQ

(b) LSUN-Church

Figure 11: Histogram matching using top-1 classes from InceptionV3 network to explain the change in FID ranking changes between StyleSwin and DDPM on both (a) CelebAHQ and (b) LSUN church datasets.

et al. (2022) proposed Denoising Diffusion GAN (DDGAN). The paper parameterizes a conditional GAN for the reverse diffusion process and demonstrates faster generation speed.

A.5 Compute details

In principle, our experiments run on single GPUs. For some experiments, we used up to 16 nodes with 4 Nvidia A40 GPUs each at a large scientific computing centre.

A.6 DNDD-Dataset

DNDD-Dataset contains 3600 images with 7 classes. Networks are tasked with detecting nutrient deficiency in the soil, such as Sodium, Calcium, unfertilized, and 4 others. The images of sugar beet are captured over the 2019 growth period. Capturing occurred at the long-term fertilizer experiment (LTFE) Dikopshof near Bonn. The images were annotated with seven types of fertilizer treatments. The dataset is used for image classification and domain adaptation.

We train the Proj. FastGAN and DDGAN models on this dataset. We further preprocessed the dataset by splitting the 1000x1000 resolution image to 256x256 resolutions. This resulted in 57600 images overall. The training details are further provided in the main paper.

A.7 Sentinel Dataset

The Sentinel dataset consists of 180,662 triplets of Synthetic Aperture Radar (SAR) image patches collected from Sentinel-1 and Sentinel-2 missions. From these, we only use ROIs_2017_Winter subset images, which contain 31,825 images. We train the Proj. FastGAN and DDGAN on this subset. The original images are in "tif" format and conversion to "jpg" is made using the official codebase provided in Schmitt et al. (2019).

A.8 Additional samples

Figure 12: Samples from CelebA-HQ generated from DDGAN.

Figure 13: Samples from CelebA-HQ generated from Proj. FastGAN.

Figure 14: Samples from DNDD-Dataset generated from DDGAN.

Figure 15: Samples from DNDD-Dataset generated from Proj. FastGAN.

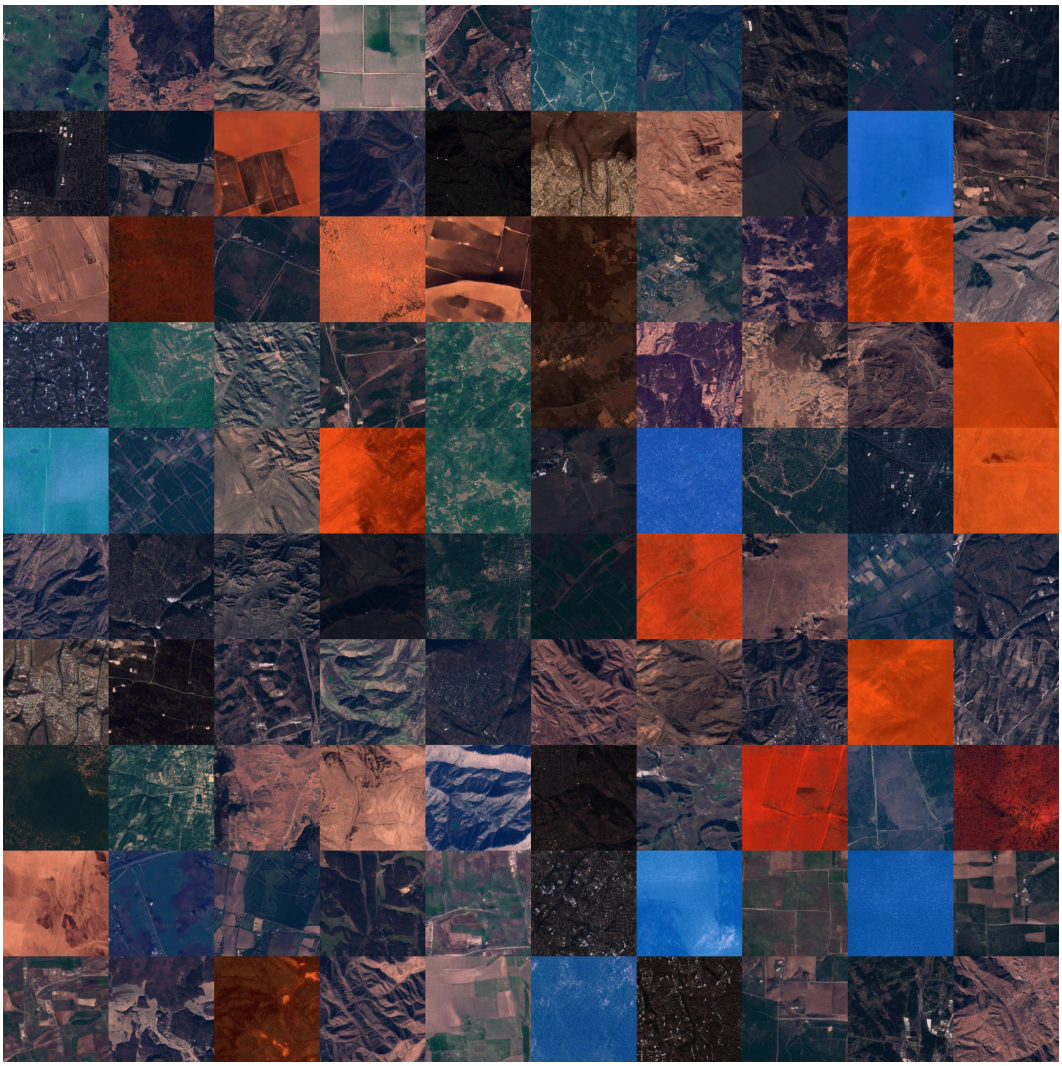


Figure 16: Samples from Sentinel dataset generated from DDGAN.

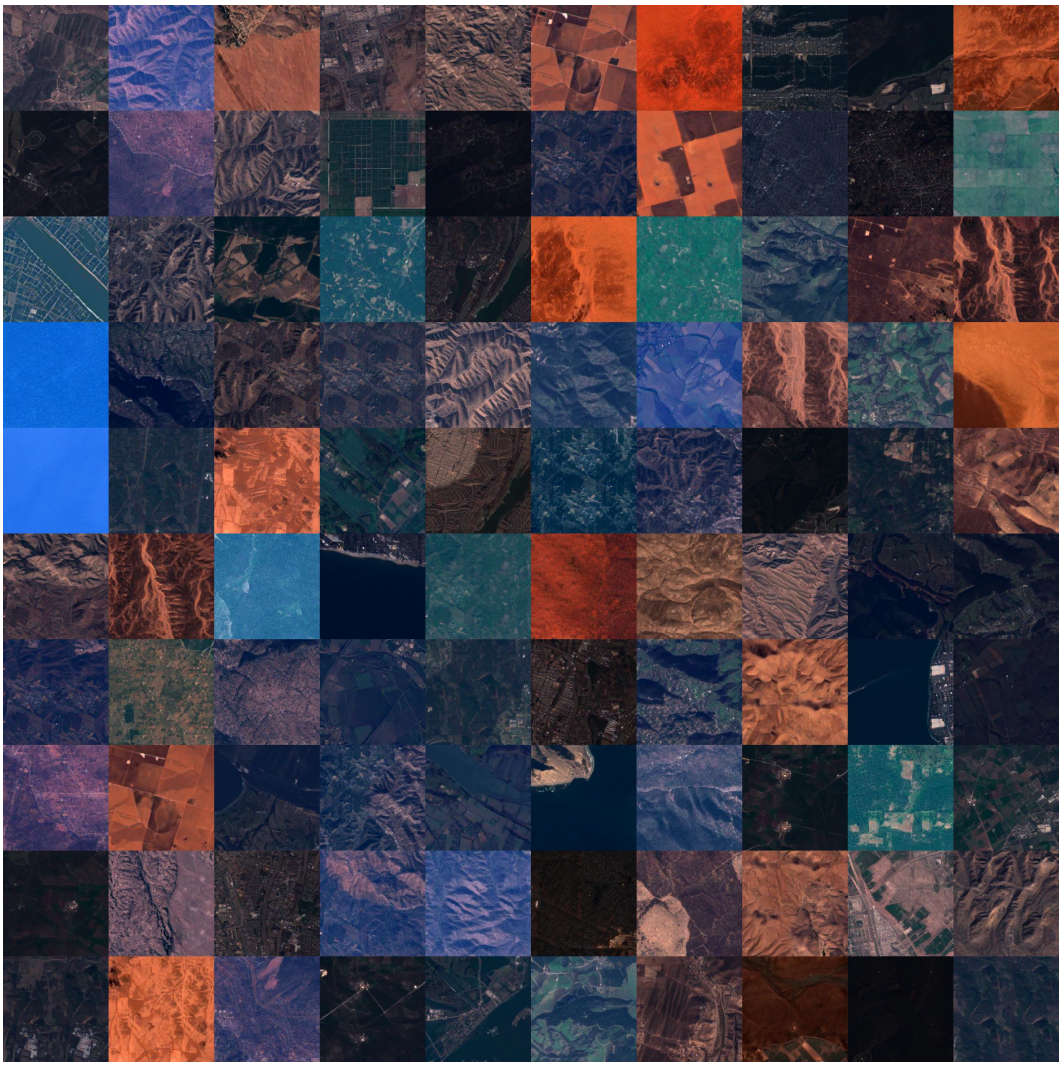


Figure 17: Samples from Sentinel dataset generated from Proj. FastGAN.