Diffusion Probabilistic Model Made Slim

Xingyi Yang[1] Daquan Zhou[2] Jiashi Feng[2] Xinchao Wang[1]*
National University of Singapore[1] ByteDance Inc.[2]
xyang@u.nus.edu, {daquanzhou, jshfeng}@bytedance.com, xinchao@nus.edu.sg

Abstract

Despite the recent visually-pleasing results achieved, the massive computational cost has been a long-standing flaw for diffusion probabilistic models (DPMs), which, in turn, greatly limits their applications on resource-limited platforms. Prior methods towards efficient DPM, however, have largely focused on accelerating the testing yet overlooked their huge complexity and sizes. In this paper, we make a dedicated attempt to lighten DPM while striving to preserve its favourable performance. We start by training a small-sized latent diffusion model (LDM) from scratch, but observe a significant fidelity drop in the synthetic images. Through a thorough assessment, we find that DPM is intrinsically biased against high-frequency generation, and learns to recover different frequency components at different time-steps. These properties make compact networks unable to represent frequency dynamics with accurate high-frequency estimation. Towards this end, we introduce a customized design for slim DPM, which we term as Spectral Diffusion (`SD`), for light-weight image synthesis. `SD` incorporates wavelet gating in its architecture to enable frequency dynamic feature extraction at every reverse step, and conducts spectrum-aware distillation to promote high-frequency recovery by inverse weighting the objective based on spectrum magnitude. Experimental results demonstrate that, `SD` achieves 8-18$\times$ computational complexity reduction as compared to the latent diffusion models on a series of conditional and unconditional image generation tasks while retaining competitive image fidelity.

1. Introduction

Diffusion Probabilistic Models (DPMs) [18, 57, 59] have recently emerged as a powerful tool for generative modeling, and have demonstrated impressive results in image synthesis [8, 45, 48], video generation [17, 20, 77] and 3D editing [43]. Nevertheless, the gratifying results come with a price: DPMs suffer from massive model sizes. In fact,

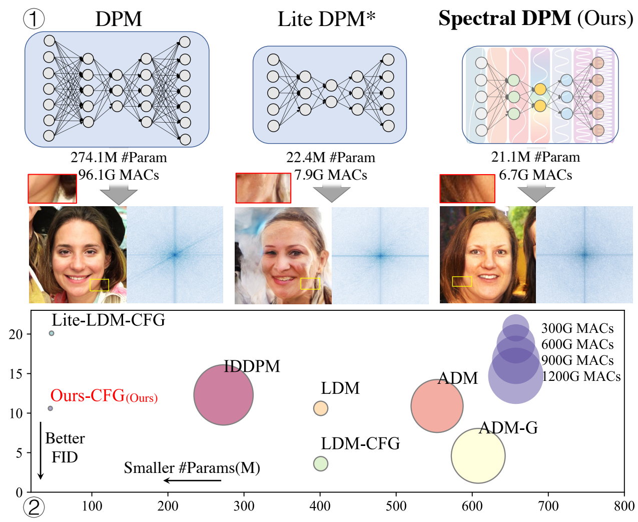

Figure 1. (1) Visualization of the frequency gap among generated images with the DPM [48], Lite DPM and our `SD` on FFHQ [27] dataset. Lite-DPM is unable to recover fine-grained textures, while `SD` can produce realistic patterns. (2) Model size, Multiply-Add cumulation (MACs) and FID score on ImageNet [7]. Our model achieves compelling visual quality with minimal computational cost. * indicates our re-implemented version.

state-of-the-art DPMs requires billions of parameters, with hundreds or even thousands of inference steps per image. For example, *DALL· E 2* [45], which is composed of 4 separate diffusion models, requires 5.5B parameters and 356 sampling steps in total. such an enormous model size, in turn, makes DPMs extremely cumbersome to be employed in resource-limited platforms.

However, existing efforts towards efficient DPMs have focused on model acceleration, but largely overlooked lightening of the model. For example, the approaches of [1, 32, 37, 38, 40, 52, 56] strive for faster sampling, while those of [13, 19, 48, 62] rely on reducing the input size. Admittedly, all of these methods give rise to shortened training or inference time, yet still, the large sizes prevent them from many real-world application scenarios.

In this paper, we make a dedicated efforts towards building compact DPMs. To start with, we train a lite version of the popular latent diffusion model (LDM) [48] by re-

*Corresponding author

ducing the channel size. We show the image generated by the original and and lite DPM in Figure 1. While the lite LDM sketches the overall structure of the faces, the high-frequency components, such as the skin and hair textures, are unfortunately poorly recovered. This phenomenon can be in fact revealed by the Discrete Fourier Transform (DFT) coefficient shown on the right column, indicating that the conventional design for DPMs leads to high-frequency deficiency when the model is made slim.

We then take an in-depth analysis on the DPMs through the lens of frequency, which results in two key observations. (1) *Frequency Evolution.* Under mild assumptions, we mathematically prove that DPMs learn different functionalities at different stages of the denoising process. Specifically, we show that the optimal denoiser in fact boils down to a cascade of wiener filters [66] with growing bandwidths. After recovering the low-frequency components, high-frequency features are added gradually in the later denoising stages. This evolution property, as a consequence, small DPMs fails to learn dynamic bandwidths with limited parameters. (2) *Frequency Bias.* DPM is biased towards dominant frequency components of the data distribution. It is most obvious when the noise amplitude is small, leading to inaccurate noise prediction at the end of the reverse process. As such, small DPMs struggle to recover the high-frequency band and image details.

Motivated by these observations, we propose a novel Spectral Diffusion (SD) model, tailored for light-weight image synthesis. Our core idea is to introduce the frequency dynamics and priors into the architecture design and training objective of the small DPM, so as to explicitly preserve the high-frequency details. The proposed solution consists of two parts, each accounting for one aforementioned observations. For the frequency evolution, we propose a wavelet gating operation, which enables the network to dynamically adapt to the spectrum response at different time-steps. In the upsample and downsample stage, the input feature is first decomposed through wavelet transforms and the coefficients are re-weighted through a learnable gating function. It significantly lowers the parameter requirements to represent the frequency evolution in the reverse process.

To compensate for the frequency bias for small DPMs, we distill high-frequency knowledge from a teacher DPM to a compact network. This is achieved by inversely weighting the distillation loss based on the magnitudes of the frequency spectrum. In particular, we give more weight to the frequency bands with small magnitudes, which strengthens the recovery of high-frequency details for the student model. By integrating both designs seamlessly, we build a slim latent diffusion model, called SD, which largely preserves the performance of LDM. Notably, SD inherits the advantages of DPMs, including superior sample diversity, training stability, and tractable parameterization. As shown in Figure 1, our model is $8 \sim 18\times$ times smaller and runs $2 \sim 5\times$ times faster than the original LDM, while achieving competitive image fidelity.

The contributions of this study are threefold:

1. This study investigates the task of diffusion model slimming, which remains largely unexplored before.
2. We identify that the key challenge lies in its unrealistic recovery for the high-frequency components. By probing DPMs from a frequency perspective, we show that there exists a spectrum evolution over different denoising steps, and the rare frequencies cannot be accurately estimated by small models.
3. We propose SD, a slim DPM that effectively restores imagery textures by enhancing high-frequency generation performance. SD achieves gratifying performance on image generation tasks at a low cost.

2. Related Work

Diffusion Probabilistic Models. DPMs [18, 55] are leading score-based generative models [58, 59, 65] with superior sample quality [8]. They use annealed noise scheduling [57] and are usually implemented as time-conditioned UNet [8, 50, 59] with attention mechanism [22, 48, 64]. Recent improvements in parameter moving average [42], objective [18], and scheduling [42] have greatly improved their visual quality. In this work, we focus on designing small-sized diffusion, which has rarely been studied before.

Efficient Diffusion. Efficient diffusion models for low-resource inferences is a trending topic. One approach is through reducing the sampling steps, which is either done by distilling multiple steps into a single step [38, 40, 52], or shortening the reverse steps while maintaining the image fidelity [1, 32, 37, 56]. Another possible solution is to diffuse in a lower dimensional space and then scale it up with a cascade structure [19] or in the latent space [48, 62]. In distinction from them, we build an efficient diffusion model using light-weight architecture and knowledge distillation.

Knowledge Transfer and Distillation. Knowledge Transfer (KT) [16, 71, 72, 74] refers to the process to transfer the knowledge from teacher models [25, 70, 73] to the student for model compression [14, 29, 60] and enhancing performance [9, 46, 47, 76]. Dataset distillation, on the other hand, focuses on learning compressed dataset [33, 34]. We make the first attempt to build slim DPM through distillation.

Frequency Analysis for Generative Model. In deep neural networks, the *frequency principle* is commonly observed, where low-frequency signals are fitted first before moving on to high-frequency components [2, 67, 68]. the frequency bias, is also evident in training deep generative models such as GANs [5, 11, 28, 54], where generators often struggle to produce natural high-frequency details.

In this paper, we examine the frequency behavior of DPMs. Taking advantage of its frequency properties, our

SD achieves realistic image generation at a low cost.

3. Background

3.1. Denoising Diffusion Probabilistic Models

Diffusion model reverses a progressive noise process based on latent variables. Given data $\mathbf{x}_0 \sim q(\mathbf{x}_0)$ sampled from the real distribution, we consider perturbing data with Gaussian noise of zero mean and β_t variance for T steps

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

where $t \in [1, T]$ and $0 < \beta_{1:T} < 1$ denote the noise scale scheduling. At the end of day, $\mathbf{x}_T \to \mathcal{N}(0, \mathbf{I})$ converge to isometric Gaussian noise. Although sampling from noise-perturbed distribution $q(\mathbf{x}_t) = \int q(\mathbf{x}_{1:t}|\mathbf{x}_0) d\mathbf{x}_{1:t-1}$ requires a tedious numerical integration over steps, the choice of Gaussian provides a close-form solution to generate arbitrary time-step $\mathbf{x}_t$ through

$$\mathbf{x}_t = \sqrt{\bar{\alpha}_t}\mathbf{x}_0 + \sqrt{1-\bar{\alpha}_t}\boldsymbol{\epsilon}, \quad \text{where} \quad \boldsymbol{\epsilon} \sim \mathcal{N}(0, \mathbf{I}) \quad (2)$$

where $\alpha_t = 1 - \beta_t$ and $\bar{\alpha}_t = \prod_{s=1}^{t} \alpha_s$. A variational Markov chain in the reverse process is parameterized as a time-conditioned denoising neural network $\mathbf{s}(\mathbf{x}, t; \boldsymbol{\theta})$ with $p_{\boldsymbol{\theta}}(\mathbf{x}_{t-1}|\mathbf{x}_t) = \mathcal{N}(\mathbf{x}_{t-1}; \frac{1}{\sqrt{1-\beta_t}}(\mathbf{x}_t + \beta_t \mathbf{s}(\mathbf{x}_t, t; \boldsymbol{\theta})), \beta_t \mathbf{I})$. The denoiser is trained to minimize a re-weighted evidence lower bound (ELBO) that fits the noise

$$\mathcal{L}_{\text{DDPM}} = \mathbb{E}_{t,\mathbf{x}_0,\boldsymbol{\epsilon}}\Big[||\boldsymbol{\epsilon} + \sqrt{1-\bar{\alpha}_t}\mathbf{s}(\mathbf{x}_t, t; \boldsymbol{\theta})||_2^2\Big] \quad (3)$$

$$\propto \mathbb{E}_{t,\mathbf{x}_0,\boldsymbol{\epsilon}}\Big[||\nabla_{\mathbf{x}_t} \log p(\mathbf{x}_t) - \mathbf{s}(\mathbf{x}_t, t; \boldsymbol{\theta})||_2^2\Big] \quad (4)$$

where the $\nabla_{\mathbf{x}_t} \log p(\mathbf{x}_t)$ are also called the score function [57]. Thus, the denoiser equivalently learns to recover the derivative that maximize the data log-likelihood [23,65]. With a trained $\mathbf{s}(\mathbf{x}, t; \boldsymbol{\theta}^*) \approx \nabla_{\mathbf{x}_t} \log p(\mathbf{x}_t)$, we generate the data by reversing the Markov chain

$$\mathbf{x}_{t-1} \leftarrow \frac{1}{\sqrt{1-\beta_t}}(\mathbf{x}_t + \beta_t \mathbf{s}(\mathbf{x}_t, t; \boldsymbol{\theta})) + \sqrt{\beta_t}\boldsymbol{\epsilon}_t \quad (5)$$

The reverse process could be understood as going along $\nabla_{\mathbf{x}_t} \log p(\mathbf{x}_t)$ from $\mathbf{x}_T$ to maximize the data likelihood.

3.2. Frequency Domain Representation of Images

Frequency domain analysis decomposes a image according to a set of basis functions. We focus on two discrete transformations: *Fourier* and *Wavelet* Transform.

Given a $H \times W$ input signal[1] $\mathbf{x} \in \mathbb{R}^{H \times W}$, Discrete Fourier Transform (DFT) $\mathcal{F}$ projects it onto a collection of sine and cosine waves of different frequencies and phases

$$\mathcal{X}(u, v) = \mathcal{F}[\mathbf{x}] = \sum_{x=1}^{H}\sum_{y=1}^{W} \mathbf{x}(x, y) e^{-j2\pi(\frac{u}{H}x + \frac{v}{W}y)}$$

[1]For simplicity, we only introduce the formulation for gray-image, while it is extendable to multi-channel inputs.

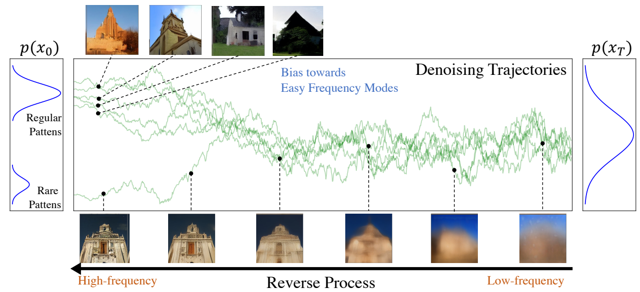

Figure 2. **Illustration of the Frequency Evolution and Bias for Diffusion Models.** In the reverse process, the optimal filters recover low-frequency components first and add on the details at the end. The predicted score functions may be incorrect for rare patterns, thus failing to recover complex and fine-grained textures.

$\mathbf{x}(x, y)$ is the pixel value at (x, y); $\mathcal{X}(u, v)$ represents complex value at frequency (u, v); e and j are Euler's number and the imaginary unit.

On the other hand, Discrete Wavelet Transform (DWT) projects it onto multi-resolution wavelets functions. In a single-scale case, $\mathbf{x}$ is decomposed into 4 wavelet coefficients $\mathbf{x}_{\text{LL}}, \mathbf{x}_{\text{LH}}, \mathbf{x}_{\text{HL}}, \mathbf{x}_{\text{HH}} = \text{DWT}(\mathbf{X})$ by halving the scale, where $\mathbf{x}_{\{\text{LL,LH,HL,HH}\}} \in \mathbb{R}^{\frac{H}{2} \times \frac{W}{2}}$. $\mathbf{x}_{\text{LL}}$ represents low-frequency component and $\mathbf{x}_{\{\text{LH,HL,HH}\}}$ are high-frequency components that contains the textural details. The coefficients could then be inverted and up-sampled to the original input $\mathbf{x} = \text{IDWT}(\mathbf{x}_{\text{LL}}, \mathbf{x}_{\text{LH}}, \mathbf{x}_{\text{HL}}, \mathbf{x}_{\text{HH}})$.

4. Frequency Perspective for Diffusion

In general signal processing, denoising is often performed in frequency space. Similar to Figure 1, Table 1 compares Low-freq and High-freq error[2] for different DPMs on FFHQ dataset. Lite-LDM performs poorly due to its lack of high-frequency generation.

Method	#Param	FID↓	Low-freq Error↓	High-freq Error↓
LDM	274.1M	5.0	0.11	0.75
Lite-LDM	22.4M	17.3	0.28(+0.17)	3.35(+2.17)

Table 1. Low-freq and High-freq error for different model size.

Thus, we examine DPM's behavior in the frequency domain. As illustrated in Figure 2, we make two findings: (1) *Frequency Evolution.* Diffusion model learns to recover the low-frequency components at first, and gradually adds in photo-realistic and high-frequency details. (2) *Frequency Bias.* Diffusion model makes biased recovery for the minority frequency band.

4.1. Spectrum Evolution over Time

DPM optimizes a time-conditioned network to fit the noise at multiple scales, which gives rise to a denoising trajectory over time-steps. We examine this trajectory closely

[2]The error computed as the $\mathbb{E}_f[\mathbb{E}[|\mathcal{F}_{real}|] - \mathbb{E}[|\mathcal{F}_{gen}|]]$ over 300 real and generated samples, with the low-high cut-off frequency of 28Hz.

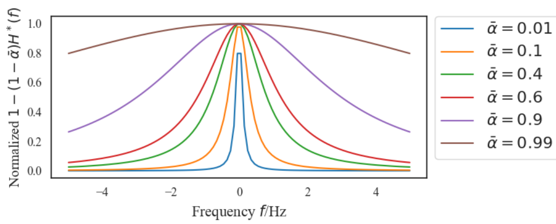

Figure 3. $1-(1-\bar{\alpha})|H^*(f)|^2$ of the optimal linear denoising filter with different $\bar{\alpha}$.

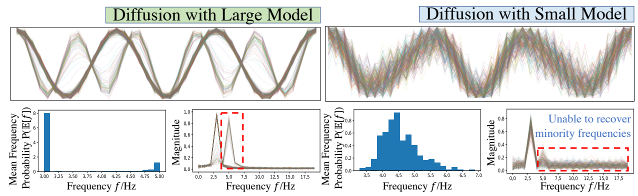

Figure 4. Toy example for 1D signal fitting. Small DPM is unable to recover minority frequency components.

from a frequency perspective. When assuming the network is a linear filter, we give the optimal filter in terms of its spectrum response at every timestep. This filter is commonly known as **Wiener filter** [66].

Proposition 1. *Assume $\mathbf{x}_0$ is a wide-sense stationary signal and $\boldsymbol{\epsilon}$ is white noise of variance $\sigma^2=1$. For $\mathbf{x}_t=\sqrt{\bar{\alpha}}\mathbf{x}_0+\sqrt{1-\bar{\alpha}}\boldsymbol{\epsilon}$, the optimal linear denoising filter h_t at time t that minimize $J_t=\|h_t*\mathbf{x}_t-\boldsymbol{\epsilon}\|^2$ has a closed-form solution*

$$\mathcal{H}_t^*(f)=\frac{1}{\bar{\alpha}|\mathcal{X}_0(f)|^2+1-\bar{\alpha}} \tag{6}$$

where $|\mathcal{X}_0(f)|^2$ is the power spectrum of $\mathbf{x}_0$ and $\mathcal{H}_t^*(f)$ is the frequency response of h_t^*.

Although the linear assumption poses a strong restriction on the model architecture, we believe it provides valuable insights into how the reverse process has been performed.

DPM goes from structure to details. In this study, we make a widely accepted assumption about the power spectra of natural images follows a power law [3, 10, 61, 63], $\mathbb{E}[|X_0(f)|^2]=A_s(\theta)/f^{\alpha_S(\theta)}$. $A_s(\theta)$ is called an amplitude scaling factor and $\alpha_S(\theta)$ is the frequency exponent. If we set $A_s(\theta)=1$ and $\alpha_S(\theta)=2$, the frequency response of the signal reconstruction filter $1-\sqrt{1-\bar{\alpha}}h$ is in Figure 3.

In the reverse process, t goes from $T\to 0$, and $\bar{\alpha}$ increases from $0\to 1$. Therefore, DPM displays a spectrum-varying behavior over time. In the beginning, we have a narrow-banded filter ($\bar{\alpha}=0.1$ and $\bar{\alpha}=0.01$) that only restores the low-frequency components that control the rough structures. t goes down and $\bar{\alpha}$ gradually increases, with more details and high-frequency components restored in the images, like the human hairs, wrinkles, and pores.

We plot the denoised predictions $\hat{\mathbf{x}}_0$ at different steps using pre-trained LDM [48] in Figure 2, which shows that DPM generates low-frequency first and transits into high-frequency. The same empirical observation that DPM goes from rough to details has been shown in [6,18,39,48], while we are the first to give its numerical solutions.

4.2. Frequency Bias in Diffusion Model

Another challenge in diffusion-based model is the inaccurate noise estimation in low-density regions [57]. It results from the expectation over $p(\mathbf{x}_0)$ in the loss function

$$\mathcal{L}_{\mathrm{DDPM}}\propto\int p(\mathbf{x}_0)\mathbb{E}_{t,\boldsymbol{\epsilon}}\Big[||\nabla_{\mathbf{x}_t}p(\mathbf{x}_t)-\mathbf{s}(\mathbf{x}_t,t;\boldsymbol{\theta})||_2^2\Big]\mathrm{d}\mathbf{x}_0 \tag{7}$$

Weighting the denoising objective by $p(\mathbf{x}_0)$ can introduce bias in the trained diffusion model, causing it to prioritize high-density regions while ignoring rare patterns.

One example of a long-tail pattern in image generation tasks is the frequency bias, where high-frequency components are rare. Consequently, training small diffusion-based models on the biased distribution can make it challenging to generate such high-frequency patterns, as the model tends to overemphasize low-frequency images. This issue can significantly impact the quality of generated images.

Example 1. We fit a toy diffusion model to 1D functions $f(x)=cos(\alpha 2\pi x)$, where $P(\alpha=3)=0.8$ and $P(\alpha=5)=0.2$. We adopt a two-layer feed-forward neural network, with 1000 denoising steps and hidden units $M=\{64,1024\}$. More details is in Supplementary.

We plot the 300 generated signals in Figure 4 (Top), their DFT magnitudes in (Button Right), and the mean frequency histogram in (Button Left). Small model ($M=64$) faces difficulty recovering the minority frequencies other than $\alpha=3$, while large model ($M=1024$) achieves smooth denoised results over all freq bands, especially when $\alpha=5$.

It provides concrete evidence that small DPMs have intrinsic defects in recovering the high frequencies.

5. Spectral Diffusion Model

As explained above, our goal is to reduce the size of the DPMs by incorporating frequency dynamics and priors into the architecture design and training objectives. We start with the LDM [48] as our baseline and then design a wavelet-gating module that enables time-dynamic inference for the light-weight model. A spectrum-aware distillation is applied to enhance the high-frequency generation performance. Both modifications allow us to achieve photo-realistic image generation with minimal model size and computational effort.

5.1. Dynamic Wavelet Gating

As depicted in Section 4.1, the reverse process requires a cascade of filters with a dynamic frequency response. While Vanilla UNet [50] is effective in reconstructing image details, it is incapable of incorporating dynamic spectrum into a single set of parameters. As a result, the small-size DPM cannot compensate for the changing bandwidth.

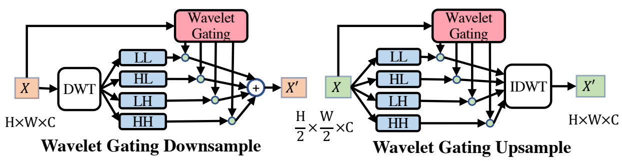

Figure 5. `WG-Down` and `WG-Up` with wavelet gating.

To address this problem, we propose inserting the **Wavelet Gating** (WG) module into the network to adapt it to varying frequency response automatically. WG decomposes the feature map into wavelet bands and selectively attends to the proper frequency at different reverse steps, which is uniquely tailored for the diffusion model.

Gating over the Wavelet Coefficients. We replace all down-sample and up-sample in UNet with DWT and IDWT [12, 69], and pose a soft gating operation on wavelet coefficients to facilitate step-adaptive image denoising. We call them `WG-Down` and `WG-Up`, as shown in Figure 5.

Similar to channel attention [21, 44, 78], information from input feature $\mathbf{X}$ is aggregated to produce a soft gating

$$g_{\{\mathsf{LL},\mathsf{LH},\mathsf{HL},\mathsf{HH}\}} = \text{Sigmoid}(\text{FFN}(\text{Avgpool}(\mathbf{X}))) \quad (8)$$

where g_i is the gating score of each wavelet band; FFN is a 2 layer feed-forward network and Avgpool stands for the average pooling. The coefficients are then gated with g_i to produce the output $\mathbf{X}'$.

In the `WG-Down`, we apply WG after the DWT operation to fuse the sub-band coefficients with weighted summation $\mathbf{X}' = \sum_{i \in \{\mathsf{LL},\mathsf{LH},\mathsf{HL},\mathsf{HH}\}} g_i \odot \mathbf{X}_i$, where $\odot$ is the element-wise multiplication. In the `WG-Up`, the input feature is splitted into 4 chunks as the wavelet coefficients. Then, WG is carried out to re-weight each sub-band before $\mathbf{X}' = \mathsf{IDWT}(g_{\mathsf{LL}} \odot \mathbf{X}_{\mathsf{LL}}, g_{\mathsf{LH}} \odot \mathbf{X}_{\mathsf{LH}}, g_{\mathsf{HL}} \odot \mathbf{X}_{\mathsf{HL}}, g_{\mathsf{HH}} \odot \mathbf{X}_{\mathsf{HH}})$. In this paper, we apply Haar wavelet by default.

5.2. Spectrum-Aware Knowledge Distillation

The diffusion model encounters challenges in modeling the high-frequency components (in Section 4.2), especially for efficient requirements. In combat with spectrum deficiency in image generation, we distill the prediction of a large pre-trained teacher model to a compact WG-Unet student. Beyond spatial output matching, we apply **Spectrum-Aware Distillation** to guide the student model in synthesizing naturalistic image details.Our approach involves re-weighting the distillation loss based on the spectrum magnitude. We increase the error penalty for components with low magnitudes, such as high-frequency bands, while reducing the weight for low-frequency elements.

Given a teacher diffusion model $\boldsymbol{s}_T(\cdot;\boldsymbol{\theta}_T)$, we would like to distill a student $\boldsymbol{s}_T(\cdot;\boldsymbol{\theta}_T)$ by mimicking the outputs and features. At time-step t, the perturbed image $\mathbf{x}_t$ is fed into both networks to produce the outputs and features. A L2 loss [35, 49] is use to quantify their spatial distance

$$\mathcal{L}_{\text{spatial}} = \sum_i \|\mathbf{X}_T^{(i)} - \mathbf{X}_S^{(i)}\|_2^2 \quad (9)$$

where $\mathbf{X}_T^{(i)}$ and $\mathbf{X}_S^{(i)}$ stand for a pair of teacher/student's output features or outputs of the same scale. A single 1×1 Conv layer is used to align the dimensions.

In addition to spatial distillation, we draw inspiration from imbalanced learning [4, 24, 30] to design a distillation loss that promotes the recovery of minority frequencies. The proposed method involves taking a pair of model predictions and a clean image $\mathbf{x}_0$, interpolating $\mathbf{x}_0$ to the same size as the feature map, and then computing their 2D DFT

$$\mathcal{X}_T^{(i)} = \mathcal{F}[\mathbf{X}_T^{(i)}], \mathcal{X}_S^{(i)} = \mathcal{F}[\mathbf{X}_S^{(i)}], \mathcal{X}^{(i)} = \mathcal{F}[\text{Resize}(\mathbf{x}_0)] \quad (10)$$

The $\mathcal{X}_0$ is then applied to modulate the difference between $\mathcal{X}_T^{(i)}$ and $\mathcal{X}_S^{(j)}$

$$\mathcal{L}_{\text{freq}} = \sum_i \omega_i \|\mathcal{X}_T^{(i)} - \mathcal{X}_S^{(j)}\|_2^2, \text{where } \omega = |\mathcal{X}^{(i)}|^{\alpha} \quad (11)$$

with a scaling factor $\alpha < 0$ ($\alpha = -1$ in our experiment), $\mathcal{L}_{\text{freq}}$ pushes the student towards learning the minority frequencies yet down-weights the majority components. Together with the DDPM objective in Eq. 3, our training objective becomes $\mathcal{L} = \mathcal{L}_{\text{DDPM}} + \lambda_s \mathcal{L}_{\text{spatial}} + \lambda_f \mathcal{L}_{\text{freq}}$ with weighting factors $\lambda_s = 0.1$ and $\lambda_f = 0.1$.

Note that our method aims to learn accurate score prediction at each denoising step, which is orthogonal to existing distillation on sampling step reduction [40, 52].

6. Experiments

This section verifies the efficacy of the `SD` on high-resolution image synthesis in Section 6.1, and validates the individual contributions of each module via ablation study in Section 6.2.

Datasets and Evaluation. We evaluate our model on 4 unconditional and 2 conditional benckmarks. Specially, we train unconditional `SD` models on LSUN-Churches/Bedrooms [75], FFHQ [27], and CelebA-HQ [26]. Furthermore, the model's performance is also assessed in the context of class-conditioned ImageNet [7] and MS-COCO [31] text-to-image generation. For the text-to-image task, we first train on LAION-400M [53] and test on MS-COCO directly.

Training and Evaluation Details. We build our model on the LDM [48] frameworks[3]. For fair comparison[4], we implement a lite-version of LDM, with a channel dimension of 64 as our baseline model. We call it Lite-LDM.

[3] `https://github.com/CompVis/latent-diffusion`

[4] Generative models from other families (e.g. GAN, VAE, and Flow) are excluded intentionally for fair computation comparison.

FFHQ 256 × 256			
Model	#Param	MACs	FID↓
DDPM [18]	113.7M	248.7G	8.4
P2 [6]	113.7M	248.7G	7.0
LDM [48]	274.1M	96.1G	5.0
Lite-LDM	22.4M(12.2×)	7.9G(12.2×)	17.3(−12.3)
Ours	21.1M(13.0×)	6.7G(14.3×)	10.5(−5.5)

CelebA-HQ 256 × 256			
Model	#Param	MACs	FID↓
Score SDE [59]	65.57M	266.4G	7.2
DDGAN [62]	39.73M	69.9G	7.6
LDM [48]	274.1M	96.1G	5.1
Lite-LDM	22.4M(12.2×)	7.9G(12.2×)	14.3(−9.2)
Ours	21.1M(13.0×)	6.7G(14.3×)	9.3(−4.2)

LSUN-Bedroom 256 × 256			
Model	#Param	MACs	FID↓
DDPM [18]	113.7M	248.7G	4.9
IDDPM [42]	113.7M	248.6G	4.2
ADM [8]	552.8M	1114.2G	1.9
LDM [48]	274.1M	96.1G	3.0
Lite-LDM	22.4M(12.2×)	7.9G(12.2×)	10.9(−7.9)
Ours	21.1M(13.0×)	6.7G(14.3×)	5.2(−2.2)

LSUN-Church 256 × 256			
Model	#Param	MACs	FID↓
DDPM [18]	113.7M	248.7G	4.9
IDDPM [42]	113.7M	248.6G	4.3
ADM [8]	552.8M	1114.2G	1.9
LDM [48]	295.0M	18.7G	4.0
Lite-LDM	32.8M(9.0×)	2.1G(8.9×)	13.6(−9.6)
Ours	33.8M(8.7×)	2.1G(8.9×)	8.4(−4.4)

Table 2. Unconditional generation results comparison to prior DPMs. The results are taken from the original paper, except that DDPM is take from the [6].

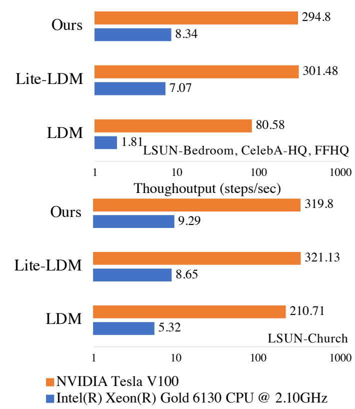

Figure 6. Throughput for unconditional image generation.

Our proposed SD is trained on 4 unconditional benchmarks for a duration of 150k iterations, using a mini-batch size of either 512 or 256. We employ the AdamW [36] optimizer with an initial learning rate of 1.024×10^{-3} and linear learning rate decay. For class- and text-conditioned generation, we set the initial learning rate to 5.12×10^{-4} while keeping other parameters constant. The synthesized image quality is evaluated based on the FID score [15] using 50k generated samples at a resolution of 256. We utilize a 200-step DDIM [56] sampling by default. We compare the model size and computational cost in terms of parameter count and Multiply-Add cumulation (MACs), and report the throughput for the speed comparision. All experiments are conducted on 8x NVIDIA Tesla V100 GPUs. Additional details can be found in the Supplementary Material.

6.1. Image Generation Results

Unconditional Image Generation. We evaluate the sample quality on LSUN-Churches/Bedrooms [75] FFHQ [27], and CelebA-HQ [26]. The results, as presented in Table 2, indicate that directly training small-sized diffusion models results in significant performance deterioration, with Lite-LDM showing an FID drop of 12.3 on FFHQ and 13.2 on CelebA-HQ. In contrast, our SD achieves a $8 \sim 14$ times computation reduction compared to the official LDM while maintaining comparable image fidelity. For instance, with a 21.1M Unet model and 6.7G MACs, our approach achieves an FID score of 5.2, which is very close to the 4.9 FID in DDPM, but with only $\frac{1}{37}$ of its computation cost.

Figure 6 displays the throughput, which indicates the number of time steps executed by the model per second. It is measured by averaging over 30 runs with a batch-size of 64. We see that, Lite-LDM, while being fast, has inferior visual quality. In comparison, our SD is 4.6× faster on CPU and 3.6× on GPU compared to LDM on 3 of the 4 datasets.

In Figure 7, rows 1-4, we evaluate the visual quality of the synthesized samples. Despite having fewer parameters and less computation, SD model generates realistic samples with high-frequency details and decent sample diversity.

Method	#Param	MACs	FID↓
IDDPM [42]	273.1M	1416.3G	12.3
ADM [8]	553.8M	1114.2G	10.9
LDM [48]	400.9M	99.8G	10.6
ADM-G [8]	553.8+54.1M	1114.2+72.2G	4.6
LDM-CFG [48]	400.9M	99.8G	3.6
Lite-LDM-CFG	47.0M(8.5×)	11.1G (9.0×)	20.1(−16.5)
Ours-CFG	45.4M(8.8×)	9.9G (10.1×)	10.6(−7.0)

Table 3. Comparison of class-conditional image generation methods on ImageNet [7] with recent state-of-the-art methods. "G" stands for the classifier guidance and "CFG" refers to the classifer-free guidance for conditional image generation.

Class-conditional Image Generation. Our class-conditioned image generation performance on ImageNet is validated and presented in Table 3. With super-mini architecture and classifier-free guidance of $w = 3.0$, our SD achieves an FID score of 10.6. As the comparison, the ADM [8] only gets FID=10.9, but with 553.8M parameters and 1114.2 MACs. Lite-LDM, though being comparably fast, suffers from its inability for high-frequency generation, gets a high FID score of 20.1.

Generated results are visualized in Figure 7 row 5-10. Our SD is able to produce diverse images of different categories, particularly good at animal generation like `corgi` and `bear`. Nonetheless, we observe some failure cases where faces and shapes are distorted. Additionally, our model struggles in generating crowded instances, as exemplified in the `banana` category.

Text-to-Image Generation. We trained our text-conditioned SD using a fixed CLIP encoder on LAION-400M, as done in prior work [48]. Then, we performed zero-shot inference on MS-COCO using $w = 2.0$. Our evaluation metric is the *zero-shot FID-30K* score from GLIDE [41]. This score measures the similarity of 30K randomly selected prompts from the validation set to the entire

Figure 7. Randomly sampled 256 × 256 images generated by our models trained on CelebA-HQ [26], FFHQ [27], LSUN-Bedroom and LSUN-Church [75], ImageNet [7]. All images are sampled with 200 DDIM steps.

Method	#Param	FID↓
GLIDE [41]	5.0B	12.24
DALLE2 [45]	5.5B	10.39
Imagen [51]	3.0B	7.27
LDM [48]	1.45B	12.63
Ours	77.6M(18.7×)	18.87

Table 4. Zero-Shot FID on MS-COCO text-to-image generation.

MS-COCO validation set using generated images.

Table 4 presents the evaluation results. Our 77.6M model achieves a FID score of 18.87, which is 18.7× smaller than LDM. We also provide qualitative analysis for text-to-image generation with new prompts, in Figure 8. Although the image quality is inferior to those large-sized diffusion models, our model is capable of producing vivid drawings based on descriptions, with minimal computational cost and portable model size. Our SD is good at abstract or carton style paintings. However, it is still challenging to generate human body and faces, as in the "basketball player" example.

6.2. Ablation Study and Analysis

In this section, we validate the effectiveness of wavelet gating and spectrum-aware distillation, on whether and how they help to improve the image fidelity.

Wavelet Gating. We validate the effectiveness of the Wavelet Gating by replacing our WG operation with the nearest neighbor resizer in LDM [48] and train on the FFHQ dataset. The results, presented in Table 5, demonstrate that removing WG significantly increases the FID from 10.5 to 12.4. Furthermore, using WG alone improves Lite-LDM's FID by 2.6. Both results indicate that WG promote the sam-

Figure 8. Selected samples from Spectral Diffusion using classifier-free guidance $w = 5.0$ for text-to-image generation.

Method	**FFHQ** 256×256							
+ Wavelet Gating		✓			✓		✓	✓
+ Spatial Distill			✓		✓	✓		✓
+ Freq Distill				✓		✓	✓	✓
FID↓	17.3	14.7	16.6	15.3	12.3	12.4	11.4	10.5

Table 5. Ablation study on FFHQ dataset.

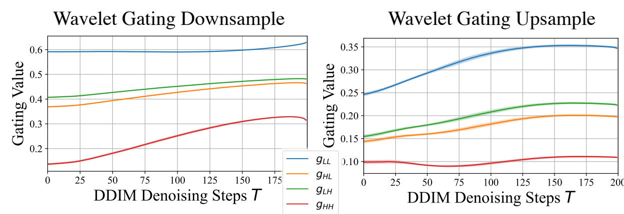

Figure 9. Wavelet gating function values at different t. We plot the mean$\pm$std for 100 generated images.

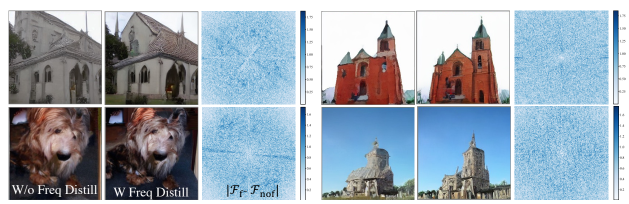

Figure 10. Generated images W or W/O the freq term, as well as their DFT difference $|\mathcal{F}_{\mathrm{f}} - \mathcal{F}_{\mathrm{nof}}|$. Zoom in for better view.

ple quality of the small DPMs.

Furthermore, we analyzed the gating functions at different denoising steps for a pre-trained text-to-image SD model, as shown in Figure 9. Each curve represents the average gating coefficient for 100 generated images. The trends of the downsample and upsample operations diverge, with high-frequency details emerging in $\hat{\mathbf{x}}t$ towards the end of denoising (large t). The WG-Down thus enhances the high-frequency signals with increased $g_{\{\mathsf{HL},\mathsf{LH},\mathsf{HH}\}}$ while keeping the low-frequency part constant. On the other hand, the WG-Up promotes g_{LL} in the late stage of denoising. Predicted noises boost its low-frequency components, resulting in high-frequency recovery in the $\hat{\mathbf{x}}_0 = \frac{\mathbf{x}_t - \sqrt{1-\bar{\alpha}}\boldsymbol{\epsilon}}{\sqrt{\bar{\alpha}}}$.

Spectrum-Aware Distillation. To evaluate the effectiveness of SA-Distillation, we conducted an ablation study by sequentially removing each loss term. Our findings, presented in Table 5, show that the spatial term contributes only 0.9 FID improvement, while the frequency term accounts for 1.8 FID. It highlights the importance of the frequency term in achieving high-quality image generation.

We also visualize the images generated by trained models with (W) or without (W/O) the frequency term in Figure 10, with their DFT difference. The model without $\mathcal{L}_{freq}$ makes smoother predictions, while our method recovers the details like hair or architectural textures. Our method prioritizes high-frequency distillation, resulting in improvements in high-frequency components in $|\mathcal{F}f - \mathcal{F}nof|$.

7. Conclusion

In the study, we focus on reducing the computation cost for diffusion models. The primary obstacle to training small DPMs is their inability to provide high-frequency realistically, which results from the frequency evolution and bias of diffusion process. In order to resolve these problems, we propose Spectral Diffusion (SD) for efficient image generation. It performs spectrum dynamic denoising by using a wavelet gating operation, which automatically enhances different frequency bands at different reverse steps. A large pre-trained network helps to improve the performance of high-frequency generation by knowledge distillation. By seamlessly integrating both modifications, our model is 8-18 $\times$ slimmer and runs 2-5$\times$ faster than the latent diffusion model, with negligible performance drop.

Acknowledgment

This research is supported by the Ministry of Education, Singapore, under its Academic Research Fund Tier 2 (Award Number: MOE-T2EP20122-0006).

References

[1] Fan Bao, Chongxuan Li, Jun Zhu, and Bo Zhang. Analytic-dpm: an analytic estimate of the optimal reverse variance in diffusion probabilistic models. *arXiv preprint arXiv:2201.06503*, 2022. 1, 2

[2] Ronen Basri, Meirav Galun, Amnon Geifman, David Jacobs, Yoni Kasten, and Shira Kritchman. Frequency bias in neural networks for input of non-uniform density. In *International Conference on Machine Learning*, pages 685–694. PMLR, 2020. 2

[3] Geoffrey J Burton and Ian R Moorhead. Color and spatial structure in natural scenes. *Applied optics*, 26(1):157–170, 1987. 4

[4] Kaidi Cao, Colin Wei, Adrien Gaidon, Nikos Arechiga, and Tengyu Ma. Learning imbalanced datasets with label-distribution-aware margin loss. *Advances in neural information processing systems*, 32, 2019. 5

[5] Yuanqi Chen, Ge Li, Cece Jin, Shan Liu, and Thomas Li. Ssd-gan: Measuring the realness in the spatial and spectral domains. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 35, pages 1105–1112, 2021. 2

[6] 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*, pages 11472–11481, 2022. 4, 6

[7] Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. Imagenet: A large-scale hierarchical image database. In *2009 IEEE conference on computer vision and pattern recognition*, pages 248–255. Ieee, 2009. 1, 5, 6, 7

[8] Prafulla Dhariwal and Alexander Nichol. Diffusion models beat gans on image synthesis. *Advances in Neural Information Processing Systems*, 34:8780–8794, 2021. 1, 2, 6

[9] Gongfan Fang, Xinyin Ma, Mingli Song, Michael Bi Mi, and Xinchao Wang. Depgraph: Towards any structural pruning. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2023. 2

[10] David J Field. Relations between the statistics of natural images and the response properties of cortical cells. *Josa a*, 4(12):2379–2394, 1987. 4

[11] Joel Frank, Thorsten Eisenhofer, Lea Schönherr, Asja Fischer, Dorothea Kolossa, and Thorsten Holz. Leveraging frequency analysis for deep fake image recognition. In *International conference on machine learning*, pages 3247–3258. PMLR, 2020. 2

[12] Minghan Fu, Huan Liu, Yankun Yu, Jun Chen, and Keyan Wang. Dw-gan: A discrete wavelet transform gan for nonhomogeneous dehazing. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 203–212, 2021. 5

[13] Shuyang Gu, Dong Chen, Jianmin Bao, Fang Wen, Bo Zhang, Dongdong Chen, Lu Yuan, and Baining Guo. Vector quantized diffusion model for text-to-image synthesis. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 10696–10706, 2022. 1

[14] Song Han, Huizi Mao, and William J Dally. Deep compression: Compressing deep neural networks with pruning, trained quantization and huffman coding. *International Conference on Learning Representations (ICLR)*, 2016. 2

[15] 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, 2017. 6

[16] Geoffrey Hinton, Oriol Vinyals, Jeff Dean, et al. Distilling the knowledge in a neural network. *arXiv preprint arXiv:1503.02531*, 2(7), 2015. 2

[17] Jonathan Ho, William Chan, Chitwan Saharia, Jay Whang, Ruiqi Gao, Alexey Gritsenko, Diederik P Kingma, Ben Poole, Mohammad Norouzi, David J Fleet, et al. Imagen video: High definition video generation with diffusion models. *arXiv preprint arXiv:2210.02303*, 2022. 1

[18] Jonathan Ho, Ajay Jain, and Pieter Abbeel. Denoising diffusion probabilistic models. *Advances in Neural Information Processing Systems*, 33:6840–6851, 2020. 1, 2, 4, 6

[19] Jonathan Ho, Chitwan Saharia, William Chan, David J Fleet, Mohammad Norouzi, and Tim Salimans. Cascaded diffusion models for high fidelity image generation. *J. Mach. Learn. Res.*, 23:47–1, 2022. 1, 2

[20] Jonathan Ho, Tim Salimans, Alexey Gritsenko, William Chan, Mohammad Norouzi, and David J Fleet. Video diffusion models. *arXiv:2204.03458*, 2022. 1

[21] Jie Hu, Li Shen, and Gang Sun. Squeeze-and-excitation networks. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 7132–7141, 2018. 5

[22] Zilong Huang, Xinggang Wang, Lichao Huang, Chang Huang, Yunchao Wei, and Wenyu Liu. Ccnet: Criss-cross attention for semantic segmentation. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 603–612, 2019. 2

[23] Aapo Hyvärinen and Peter Dayan. Estimation of non-normalized statistical models by score matching. *Journal of Machine Learning Research*, 6(4), 2005. 3

[24] Liming Jiang, Bo Dai, Wayne Wu, and Chen Change Loy. Focal frequency loss for image reconstruction and synthesis. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 13919–13929, 2021. 5

[25] Yongcheng Jing, Yiding Yang, Xinchao Wang, Mingli Song, and Dacheng Tao. Amalgamating knowledge from heterogeneous graph neural networks. In *CVPR*, 2021. 2

[26] Tero Karras, Timo Aila, Samuli Laine, and Jaakko Lehtinen. Progressive growing of gans for improved quality, stability, and variation. *arXiv preprint arXiv:1710.10196*, 2017. 5, 6, 7

[27] 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*, pages 4401–4410, 2019. 1, 5, 6, 7

[28] Mahyar Khayatkhoei and Ahmed Elgammal. Spatial frequency bias in convolutional generative adversarial networks. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 36, pages 7152–7159, 2022. 2

[29] Yanxi Li, Minjing Dong, Yunhe Wang, and Chang Xu. Neural architecture search via proxy validation. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2022. 2

[30] Tsung-Yi Lin, Priya Goyal, Ross Girshick, Kaiming He, and Piotr Dollár. Focal loss for dense object detection. In *Proceedings of the IEEE international conference on computer vision*, pages 2980–2988, 2017. 5

[31] Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C Lawrence Zitnick. Microsoft coco: Common objects in context. In *European conference on computer vision*, pages 740–755. Springer, 2014. 5

[32] Luping Liu, Yi Ren, Zhijie Lin, and Zhou Zhao. Pseudo numerical methods for diffusion models on manifolds. In *International Conference on Learning Representations*, 2022. 1, 2

[33] Songhua Liu, Kai Wang, Xingyi Yang, Jingwen Ye, and Xinchao Wang. Dataset distillation via factorization. In *Conference on Neural Information Processing Systems*, 2022. 2

[34] Songhua Liu, Jingwen Ye, Runpeng Yu, and Xinchao Wang. Slimmable dataset condensation. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2023. 2

[35] Yifan Liu, Ke Chen, Chris Liu, Zengchang Qin, Zhenbo Luo, and Jingdong Wang. Structured knowledge distillation for semantic segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 2604–2613, 2019. 5

[36] Ilya Loshchilov and Frank Hutter. Decoupled weight decay regularization. *arXiv preprint arXiv:1711.05101*, 2017. 6

[37] Cheng Lu, Yuhao Zhou, Fan Bao, Jianfei Chen, Chongxuan Li, and Jun Zhu. Dpm-solver: A fast ode solver for diffusion probabilistic model sampling in around 10 steps. *arXiv preprint arXiv:2206.00927*, 2022. 1, 2

[38] Eric Luhman and Troy Luhman. Knowledge distillation in iterative generative models for improved sampling speed. *arXiv preprint arXiv:2101.02388*, 2021. 1, 2

[39] Hengyuan Ma, Li Zhang, Xiatian Zhu, and Jianfeng Feng. Accelerating score-based generative models with preconditioned diffusion sampling. In *European Conference on Computer Vision*, 2022. 4

[40] Chenlin Meng, Ruiqi Gao, Diederik P Kingma, Stefano Ermon, Jonathan Ho, and Tim Salimans. On distillation of guided diffusion models. *arXiv preprint arXiv:2210.03142*, 2022. 1, 2, 5

[41] Alex Nichol, Prafulla Dhariwal, Aditya Ramesh, Pranav Shyam, Pamela Mishkin, Bob McGrew, Ilya Sutskever, and Mark Chen. Glide: Towards photorealistic image generation and editing with text-guided diffusion models. *arXiv preprint arXiv:2112.10741*, 2021. 6, 7

[42] Alexander Quinn Nichol and Prafulla Dhariwal. Improved denoising diffusion probabilistic models. In *International Conference on Machine Learning*, pages 8162–8171. PMLR, 2021. 2, 6

[43] Ben Poole, Ajay Jain, Jonathan T. Barron, and Ben Mildenhall. Dreamfusion: Text-to-3d using 2d diffusion. *arXiv*, 2022. 1

[44] Zequn Qin, Pengyi Zhang, Fei Wu, and Xi Li. Fcanet: Frequency channel attention networks. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 783–792, 2021. 5

[45] Aditya Ramesh, Prafulla Dhariwal, Alex Nichol, Casey Chu, and Mark Chen. Hierarchical text-conditional image generation with clip latents. *arXiv preprint arXiv:2204.06125*, 2022. 1, 7

[46] Sucheng Ren, Zhengqi Gao, Tianyu Hua, Zihui Xue, Yonglong Tian, Shengfeng He, and Hang Zhao. Co-advise: Cross inductive bias distillation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 16773–16782, June 2022. 2

[47] Sucheng Ren, Fangyun Wei, Zheng Zhang, and Han Hu. Tinymim: An empirical study of distilling mim pre-trained models. June 2023. 2

[48] 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*, pages 10684–10695, 2022. 1, 2, 4, 5, 6, 7

[49] Adriana Romero, Nicolas Ballas, Samira Ebrahimi Kahou, Antoine Chassang, Carlo Gatta, and Yoshua Bengio. Fitnets: Hints for thin deep nets. *arXiv preprint arXiv:1412.6550*, 2014. 5

[50] Olaf Ronneberger, Philipp Fischer, and Thomas Brox. U-net: Convolutional networks for biomedical image segmentation. In *International Conference on Medical image computing and computer-assisted intervention*, pages 234–241. Springer, 2015. 2, 4

[51] 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, et al. Photorealistic text-to-image diffusion models with deep language understanding. *arXiv preprint arXiv:2205.11487*, 2022. 7

[52] Tim Salimans and Jonathan Ho. Progressive distillation for fast sampling of diffusion models. In *International Conference on Learning Representations*, 2022. 1, 2, 5

[53] Christoph Schuhmann, Richard Vencu, Romain Beaumont, Robert Kaczmarczyk, Clayton Mullis, Aarush Katta, Theo Coombes, Jenia Jitsev, and Aran Komatsuzaki. LAION-400M: open dataset of clip-filtered 400 million image-text pairs. *CoRR*, abs/2111.02114, 2021. 5

[54] Katja Schwarz, Yiyi Liao, and Andreas Geiger. On the frequency bias of generative models. *Advances in Neural Information Processing Systems*, 34:18126–18136, 2021. 2

[55] Jascha Sohl-Dickstein, Eric Weiss, Niru Maheswaranathan, and Surya Ganguli. Deep unsupervised learning using nonequilibrium thermodynamics. In *International Conference on Machine Learning*, pages 2256–2265. PMLR, 2015. 2

[56] Jiaming Song, Chenlin Meng, and Stefano Ermon. Denoising diffusion implicit models. *arXiv preprint arXiv:2010.02502*, 2020. 1, 2, 6

[57] Yang Song and Stefano Ermon. Generative modeling by estimating gradients of the data distribution. *Advances in Neural Information Processing Systems*, 32, 2019. 1, 2, 3, 4

[58] Yang Song, Sahaj Garg, Jiaxin Shi, and Stefano Ermon. Sliced score matching: A scalable approach to density and score estimation. In *Uncertainty in Artificial Intelligence*, pages 574–584. PMLR, 2020. 2

[59] 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*, 2021. 1, 2, 6

[60] Xiu Su, Shan You, Jiyang Xie, Fei Wang, Chen Qian, Changshui Zhang, and Chang Xu. Searching for network width with bilaterally coupled network. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2022. 2

[61] David J Tolhurst, Yoav Tadmor, and Tang Chao. Amplitude spectra of natural images. *Ophthalmic and Physiological Optics*, 12(2):229–232, 1992. 4

[62] Arash Vahdat, Karsten Kreis, and Jan Kautz. Score-based generative modeling in latent space. *Advances in Neural Information Processing Systems*, 34:11287–11302, 2021. 1, 2, 6

[63] van A Van der Schaaf and JH van van Hateren. Modelling the power spectra of natural images: statistics and information. *Vision research*, 36(17):2759–2770, 1996. 4

[64] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. *Advances in neural information processing systems*, 30, 2017. 2

[65] Pascal Vincent. A connection between score matching and denoising autoencoders. *Neural computation*, 23(7):1661–1674, 2011. 2, 3

[66] Norbert Wiener, Norbert Wiener, Cyberneticist Mathematician, Norbert Wiener, Norbert Wiener, and Cybernéticien Mathématicien. *Extrapolation, interpolation, and smoothing of stationary time series: with engineering applications*, volume 113. MIT press Cambridge, MA, 1949. 2, 4

[67] Zhi-Qin John Xu, Yaoyu Zhang, Tao Luo, Yanyang Xiao, and Zheng Ma. Frequency principle: Fourier analysis sheds light on deep neural networks. *arXiv preprint arXiv:1901.06523*, 2019. 2

[68] Zhi-Qin John Xu, Yaoyu Zhang, and Yanyang Xiao. Training behavior of deep neural network in frequency domain. In *International Conference on Neural Information Processing*, pages 264–274. Springer, 2019. 2

[69] Mengping Yang, Zhe Wang, Ziqiu Chi, and Wenyi Feng. Wavegan: Frequency-aware gan for high-fidelity few-shot image generation. *arXiv preprint arXiv:2207.07288*, 2022. 5

[70] Xingyi Yang, Zhou Daquan, Songhua Liu, Jingwen Ye, and Xinchao Wang. Deep model reassembly. In *Conference on Neural Information Processing Systems*, 2022. 2

[71] Xingyi Yang, Jingwen Ye, and Xinchao Wang. Factorizing knowledge in neural networks. In *European Conference on Computer Vision*, pages 73–91. Springer, 2022. 2

[72] Yiding Yang, Jiayan Qiu, Mingli Song, Dacheng Tao, and Xinchao Wang. Distilling knowledge from graph convolutional networks. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2020. 2

[73] Jingwen Ye, Yixin Ji, Xinchao Wang, Kairi Ou, Dapeng Tao, and Mingli Song. Student becoming the master: Knowledge amalgamation for joint scene parsing, depth estimation, and more. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2019. 2

[74] Jingwen Ye, Songhua Liu, and Xinchao Wang. Partial network cloning. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2023. 2

[75] 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. *arXiv preprint arXiv:1506.03365*, 2015. 5, 6, 7

[76] Daquan Zhou, Xiaojie Jin, Xiaochen Lian, Linjie Yang, Yujing Xue, Qibin Hou, and Jiashi Feng. Autospace: Neural architecture search with less human interference. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 337–346, 2021. 2

[77] Daquan Zhou, Weimin Wang, Hanshu Yan, Weiwei Lv, Yizhe Zhu, and Jiashi Feng. Magicvideo: Efficient video generation with latent diffusion models. *arXiv preprint arXiv:2211.11018*, 2022. 1

[78] Daquan Zhou, Zhiding Yu, Enze Xie, Chaowei Xiao, Animashree Anandkumar, Jiashi Feng, and Jose M Alvarez. Understanding the robustness in vision transformers. In *International Conference on Machine Learning*, pages 27378–27394. PMLR, 2022. 5