Image Generation at Different Detail Level: Scaling Skip Connections in ViT-based Diffusion Models

Chang Liu, **Karim Habashy**, **Peter Pan**, **Sirisha Rambhatla**, **Alexander Wong**
University of Waterloo
{chang.liu, karim.habashy, peter.pan, sirisha.rambhatla, alexander.wong}@uwaterloo.ca

Abstract

Denoising diffusion probabilistic models (DDPMs) excel in image generation, but users have limited control over the level of detail and semantic richness in generated images. Although prompt-based diffusion models can create more detailed images with descriptive prompts and utilize spatial masks to preserve unedited regions, diffusion models frequently overlook these constraints, leading to inconsistent image regions. Inspired by transformers, where each feature level encodes varying semantic information, we propose a feature scaling method at inference for a ViT-based diffusion model, U-ViT. Our experiments on CIFAR-10 indicate that this scaling approach effectively adjusts the level of detail in generated images.

1 Introduction

Denoising diffusion probabilistic models (DDPMs) have become the focal center in the research landscape due to their stability during training and superior image generation capabilities on image, 3D, video data, and beyond [1]. Compared to previous image generation frameworks such as variational autoencoders (VAE) [2] and generative adversarial networks (GAN) [3], diffusion models employ an image generation architecture involving a forward diffusion process and a reverse diffusion process. In the forward process, Gaussian noises are added to realistic sample images until the images become complete Gaussian noise. A neural network is trained in the reverse process to denoise at each step to map the Gaussian noise to the input sample.

However, controlling the level of detail in images generated by diffusion models is challenging. Some address this by re-generating images, but diffusion models are also notorious for incredibly slow image generation at inference due to the need to traverse the denoising reverse diffusion chain, which involves going through the same network hundreds or even thousands of times [4]. Conversely, prompt-based diffusion models can add detail using descriptive prompts and utilize spatial masks to preserve unedited areas. However, diffusion models frequently overlook these masked constraints, leading to inconsistent regions in generated images.

To enable control over fine-grained detail in generated images, we propose scaled skip connections for diffusion models at inference, a simple yet effective method that may shed light on how we can control level of details in image generation for diffusion models.This approach applies scale factors to the skip connections that introduce high-frequency information between shallow and deep feature layers, enabling us to adjust detail levels in the generated images. Applying this method to the CIFAR-100 dataset, we found that scaling skip connections can effectively modify the semantic content of generated images.

Our contributions are summarized below:

- To introduce high-frequency information in different feature levels in transformers-based architecture, we adapted scale factors from U-Net based DDPM [5] to U-ViT architecture [6] [1] of the same size.
- With no additional training, we can adjust the level of details in image generation at inference by weighing the feature maps in the denoising blocks and skip connections, as each component contributes to different levels of fine-grained detail during image generation.

2 Background

Denoising diffusion probabilistic models (DDPMs) are generative models typically used for image syn-

[1] As U-ViT outperforms the CNN based denoiser backbones while using less training data [6], we conducted our experiments using ViT based backbone

thesis. These models learn a conditional transition from pure Gaussian noise to examples in the image domain. They are a competitive method compared to Generative-Adversarial Networks (GANs) and Variational Autoencoders (VAEs), as DDPMs have been shown to generate higher quality images compared to VAEs, while not suffering the same instabilities encountered when training GANs [7].

To train DDPMs, a forward process is applied where noise is added iteratively to an input sample (usually an image) x_0 using a Markov Chain until it is no longer distinguishable from pure noise $x_T \sim \mathcal{N}(0, I)$. To recover the original image, a neural network is trained to sequentially predict the noise and remove it from the image using the same network. Effectively, this process parameterizes the reverse diffusion process by learning an adequate sequence of conditional distributions that lead to the distribution of the original data.

Typically, U-Net is the neural network architecture leveraged to predict and remove noise from images. U-Net is a convolutional network identified by its encoder-decoder architecture and its skip connections. Specifically, the encoder block downsamples the input image, effectively capturing its high-level semantics. The decoder is then tasked with upsampling the representation and returns the original dimensionality of the input. To assist with the recovery of fine-grained low-level details lost in the downsampling step, long skip connections from the encoder are concatenated with the denoising decoder features. This also stabilizes training by alleviating the vanishing gradient issue.

Building on this work, Latent Diffusion Models (LDMs) [8], embed the U-Net into the latent space of a pre-trained AutoEncoder (AE). This shift to a lower-dimensional latent space means that latent diffusion models (LDMs) need significantly less computation and time to generate images. The AE allows for modelling more complex statistics of the data, further improving image quality generation. Beyond this, the latent space enables cross-modality encoding, allowing for class and text conditioning of the outputs. This is is shown in Figure 1.

Drawing inspiration from the original U-Net, Bao et al. pushes further in the direction of less reliance on the U-Net by proposing U-ViT, a transformer based denoising backbone with long skip connections between the shallow and deep layers [6]. By employing these long skip connections, low level feature information is able to propagate through the transformer layers of the denoising U-ViT, easing the pixel-level prediction objective in diffusion models.

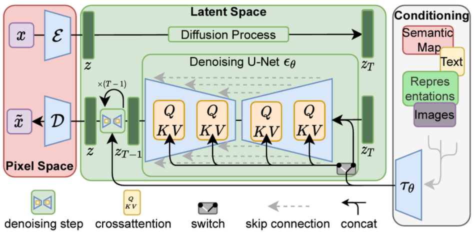
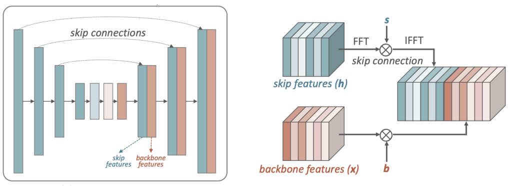


Figure 1: **Training mechanism of a Latent Diffusion Model.** We can see that the input and output use an AutoEncoder to go to and from a latent space, where the diffusion process is applied. Optionally, the reverse process can leverage multi-modal conditioning in this joint latent space using pre-trained frozen encoders



(a) UNet Architecture

(b) FreeU Operations

Figure 2: Modulating factors proposed in FreeU [9]

In FreeU [9], the authors employ a study of the U-Net architecture and point out the significance of the information propagated through the denoising blocks and the long skip connections [9]. They note that the denoising block contributes to the generation of high-level (low frequency) components of the generated samples, where these generated features embody the global/smooth characteristics of an image. Conversely, the skip connections carry over the low-level (high-frequency) information to later layers for denoising - once the global features of the image have already been resolved. Equipped with this knowledge, the authors propose a method that, when applied during inference, can lead to improved image generation quality with no addition of *any* trainable parameters. They introduce two modulating factors for the skip connections and denoising blocks, depicted in Figure 2. The first is used to downscale the low frequency information present in skip connections, as the authors argue that low frequency present in the skip connection features may be attenuating the efficacy of the denoising blocks. Due to this removal of low frequency information from the reverse diffusion process, the second factor is employed to upscale the denoising decoder blocks.

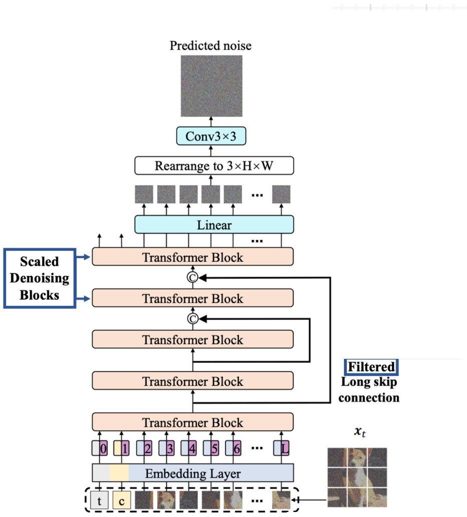


Figure 3: **Our method.** As shown in the dark blue rectangles, we scaled the skip connections and backbone features at inference for the U-ViT architecture

3 Methodology

Our method adapts modulating factors for skip connections and denoising blocks from UNet-based Diffusion models to ViT-based diffusion models. By using a ViT-based LDM, we can effectively adjust the high-frequency and low-frequency information in shallow and deep layers of ViT during the diffusion process to edit the semantic richness of generated images. First, we apply a high pass filter to the skip connection features. To do this, we compute the Fourier Transformer of the content of the skip connection h_l to obtain the frequency information, where l is outlines the layer in question. Because the rationale for using skip connections at inference time is to supply the later layers with high-frequency information, we downscale all features below some threshold value r_{thresh} by a factor s_l.

$$h'_l = \text{IFFT}(\text{FFT}(h_l) \odot \beta_l) \tag{1}$$

$$\beta_l(r) = \begin{cases} s_l, & \text{if } r < r_{\text{thresh}} \\ 1, & \text{otherwise} \end{cases} \tag{2}$$

To make up for lost information in the skip connection filtering, we amplify the scaling of the denoiser transformer blocks concatenated with the skip connections. Because we're working with a vision transformer, the features propagated through the network are not output maps from convolutional kernels, but rather fixed-sized patches. Furthermore, the U-ViT model appends time and class conditioning tokens to the network as patches for simplicity. To deal with this, we omit these first 2 tokens from the scaling operation. We then determine the scaling factor α_l using a normalized average of the features of the transformer block and β_l.

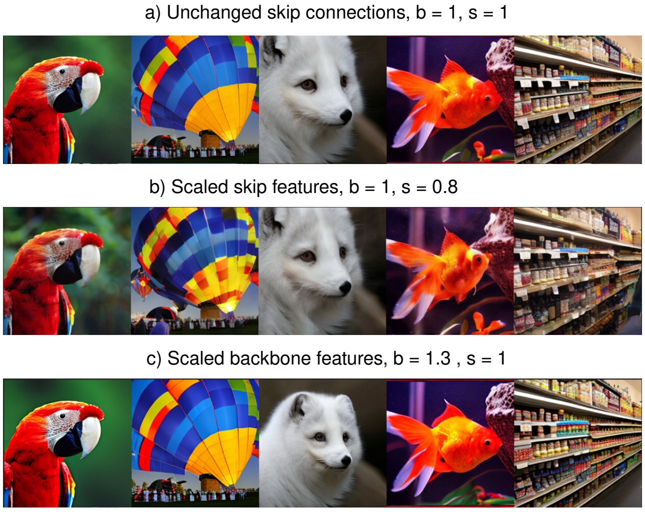
$$\bar{x}_l = \frac{1}{N}\sum_{i=1}^{N} x_{l,i} \tag{3}$$

$$\alpha_l = (b_l - 1) \cdot \frac{\bar{x}_l - \min(\bar{x}_l)}{\max(\bar{x}_l) - \min(\bar{x}_l)} + 1 \tag{4}$$

$$x'_{l,i} = x_{l,i} \odot \alpha_l \tag{5}$$

4 Result

The experimental results reveal an interplay between two key parameters, s and b, within the model. Notably, when s is systematically decreased while maintaining b at a constant level, there is an increase in semantic information of the generated image. As seen in Figure 4b), there are more details on the face of the arctic fox as well as the background of the parrot. This trend suggests that reducing the scaling factor s independently (i.e. significant filtering out of lower frequency information in the skip connections) accentuates the high-level features and introduces more details into the generated images.


a) Unchanged skip connections, b = 1, s = 1

b) Scaled skip features, b = 1, s = 0.8

c) Scaled backbone features, b = 1.3 , s = 1

Figure 4: **Qualitative Result: Influence of parameters b and s on image synthesis**. Scaling down the skip feature provided richer semantic information, as seen in the detailed trees behind the parrot and the details on the hot air balloons. Similarly, scaling up the backbone feature provided more vibrant colors of the parrot and more details on the face of the arctic fox.

Conversely, when b is increased (more denoising per step) while s is held constant, an interesting pattern emerges. Initially, the results show an improvement in image quality, suggesting that higher values of b contribute to generating sharper and more defined images. However, beyond a certain threshold, the images become excessively sharp, potentially at the cost of losing essential details. This observation highlights the balance required when tuning the b parameter, as high values compromise the quality of the generated images.

In addition, understanding the distinct roles played by skip connections versus denoising blocks here seems to be important. Skip connections and denoising blocks constitute integral components influencing the model's ability to capture both high and low-frequency information during the denoising process. While skip connections contribute to the propagation of low-level details and facilitate the recovery of fine-grained features, denoising blocks play a crucial role in synthesizing high-level, global characteristics of the generated images. The delicate interplay between these components is essential for achieving a balance between sharpness, quality, and the preservation of details. Further exploration into the interactions and individual contributions of skip connections and denoising blocks may contain the potential to fine-tune their functionalities and enhance the generated images.

5 Discussions

Our preliminary results suggest that scaling feature connections holds promise for controlling detail levels in image generation. Further work is needed to confirm our observation. For example, the next step includes calculating the Fréchet Inception Distance (FiD) as a quantitative metric for evaluation.

Future direction involves extending the applicability of these scaling factors to the training phase. Specifically, during training, the skip connections currently utilize a downscaling operation on frequencies below a predefined threshold. However, we can replace this fixed downscaling with a dynamically learned low-pass filter, potentially leveraging mathematical models such as Tschebyscheff [10] or Butterworth [11]. This adaptive approach could optimize the model's ability to capture relevant frequency information, offering a trainable refinement to the denoising process.

References

[1] L. Yang, Z. Zhang, Y. Song, S. Hong, R. Xu, Y. Zhao, W. Zhang, B. Cui, and M.-H. Yang, "Diffusion models: A comprehensive survey of methods and applications," *ACM Computing Surveys*, vol. 56, no. 4, pp. 1–39, 2023.

[2] D. P. Kingma and M. Welling, "Auto-encoding variational bayes," *arXiv preprint arXiv:1312.6114*, 2013.

[3] I. Goodfellow, J. Pouget-Abadie, M. Mirza, B. Xu, D. Warde-Farley, S. Ozair, A. Courville, and Y. Bengio, "Generative adversarial networks," *Communications of the ACM*, vol. 63, no. 11, pp. 139–144, 2020.

[4] Z. Wang, Y. Jiang, H. Zheng, P. Wang, P. He, Z. Wang, W. Chen, and M. Zhou, "Patch diffusion: Faster and more data-efficient training of diffusion models," *arXiv preprint arXiv:2304.12526*, 2023.

[5] R. Rombach, A. Blattmann, D. Lorenz, P. Esser, and B. Ommer, "High-resolution image synthesis with latent diffusion models," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2022, pp. 10 684–10 695.

[6] F. Bao, S. Nie, K. Xue, Y. Cao, C. Li, H. Su, and J. Zhu, "All are worth words: A vit backbone for diffusion models," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2023, pp. 22 669–22 679.

[7] J. Ho, A. Jain, and P. Abbeel, "Denoising diffusion probabilistic models," *ArXiv*, vol. abs/2006.11239, 2020. [Online]. Available: https://api.semanticscholar.org/CorpusID:219955663

[8] R. Rombach, A. Blattmann, D. Lorenz, P. Esser, and B. Ommer, "High-resolution image synthesis with latent diffusion models," *2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 10 674–10 685, 2021. [Online]. Available: https://api.semanticscholar.org/CorpusID:245335280

[9] C. Si, Z. Huang, Y. Jiang, and Z. Liu, "Freeu: Free lunch in diffusion u-net," *ArXiv*, vol. abs/2309.11497, 2023. [Online]. Available: https://api.semanticscholar.org/CorpusID:262064720

[10] R. J. Stegen, "Excitation coefficients and beamwidths of tschebyscheff arrays," *Proceedings of the IRE*, vol. 41, no. 11, pp. 1671–1674, 1953.

[11] I. W. Selesnick and C. S. Burrus, “Generalized digital butterworth filter design,” *IEEE Transactions on signal processing*, vol. 46, no. 6, pp. 1688–1694, 1998.