Scaled Skip Connections for Semantic Detail Control in ViT-Based Diffusion Models

Chang Liu, Karim Habashy, Peter Yuchen Pan, Sirisha Rambhatla, Alexander Wong

University of Waterloo

{chang.liu, karim.habashy, peter.pan, sirisha.rambhatla, alexander.wong}@uwaterloo.ca

Motivation

- 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.
- 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 preliminary experiments on CIFAR-10 indicate that this scaling approach effectively adjusts the level of detail in generated images.

Proposed Methods

To edit the semantic richness of generated images, we adjust the high-frequency and low-frequency information in shallow and deep layers of ViT during the diffusion process.

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.
- 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 by a factor s

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

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

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.

We determine the scaling factor α_l using a normalized average of the features of the transformer block and β_l

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

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

$$x'_{l,i} = x_{l,i} \odot \alpha_l \quad (5)$$

Model Architecture

Figure 1: Scaling Skip Connections for Diffusion Models [1, Fig. 4]

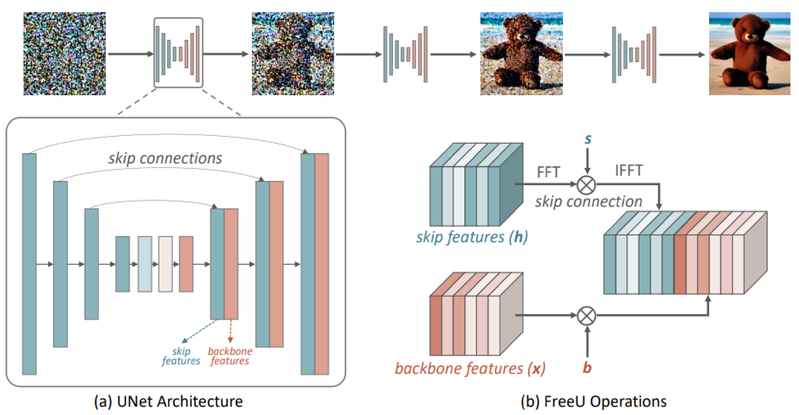
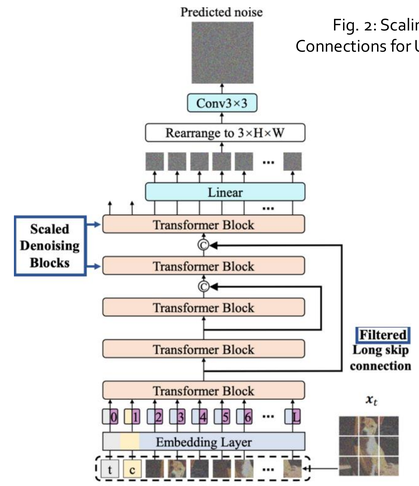


Fig. 2: Scaling Skip Connections for U-ViT (Ours)



Experiments

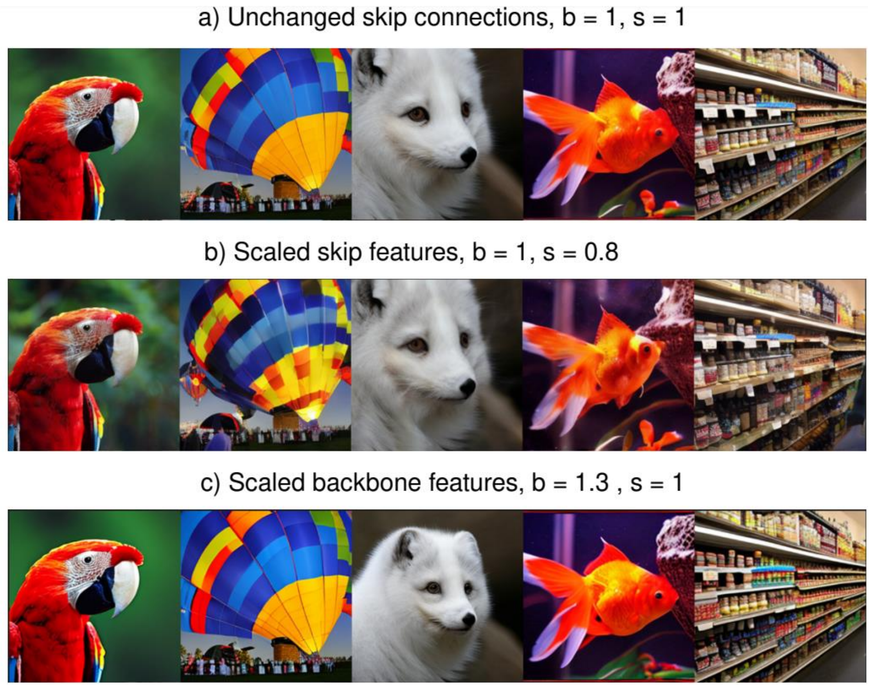


Figure 3: Qualitative Result: Influence of parameters *b* and *s* on image synthesis

- 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 3b), 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.
- 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 Frechet Inception Distance (FiD) as a quantitative metric for evaluation.

References

[1] C. Si, Z. Huang, Y. Jiang, and Z. Liu, "FreeU: Free Lunch in Diffusion U-Net," *https://arxiv.org/pdf/2309.11497.pdf*, Sep. 2023.

[2] F. Bao *et al.*, "All are worth words: A VIT backbone for diffusion models," *2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2023. doi:10.1109/cvpr52729.2023.02171