All are Worth Words: a ViT Backbone for Score-based Diffusion Models

Anonymous Author(s)
Affiliation
Address
`email`

Abstract

Vision transformers (ViT) have shown promise in various vision tasks including low-level ones while the U-Net remains dominant in score-based diffusion models. In this paper, we perform a systematical empirical study on the ViT-based architectures in diffusion models. Our results suggest that adding extra long skip connections (like the U-Net) to ViT is crucial to diffusion models. The new ViT architecture, together with other improvements, is referred to as U-ViT. On several popular visual datasets, U-ViT achieves competitive generation results to SOTA U-Net while requiring comparable amount of parameters and computation if not less.

1 Introduction

Along with the development of algorithms, the revolution of backbones plays a central role in the success of (score-based) diffusion models. A representative example is the U-Net architecture employed in prior work [15, 5], which remains dominant in diffusion models for image generation tasks. A very natural question is whether the reliance of the U-Net is necessary in such models.

On the other hand, vision transformers (ViT) [3] have shown promise in various vision tasks [1, 4] including low-level ones [17, 19]. Compared to CNN, ViT is preferable at a large scale because of its scalability and efficiency [3]. Although the score-based diffusion models have been scaled up dramatically [12], it is still not clear whether ViT is suitable for score modeling or not.

In this paper, we perform a systematical empirical study on the ViT-based architectures in diffusion models. We modify the standard ViT as follows:

1. adding extra long skip connections (like the U-Net),
2. adding an extra 3x3 convolutional block before output, and
3. treating everything including the time embedding, label embedding and patches of the noisy image as tokens.

The resulting architecture is referred to as *U-ViT*.

On several popular visual datasets, U-ViT achieves competitive generation results to SOTA U-Net architectures while requires comparable amount of parameters and computation if not less. Our results suggest that

1. ViT is promising for score-based diffusion models;
2. the long skip connections play a central role in the success of diffusion models; and
3. the down-sampling and up-sampling operators are not necessary for diffusion models.

We believe that future diffusion models on large scale or cross-modality datasets potentially benefit from U-ViT.

Under review at the NeurIPS 2022 Workshop on Score-Based Methods. Do not distribute.

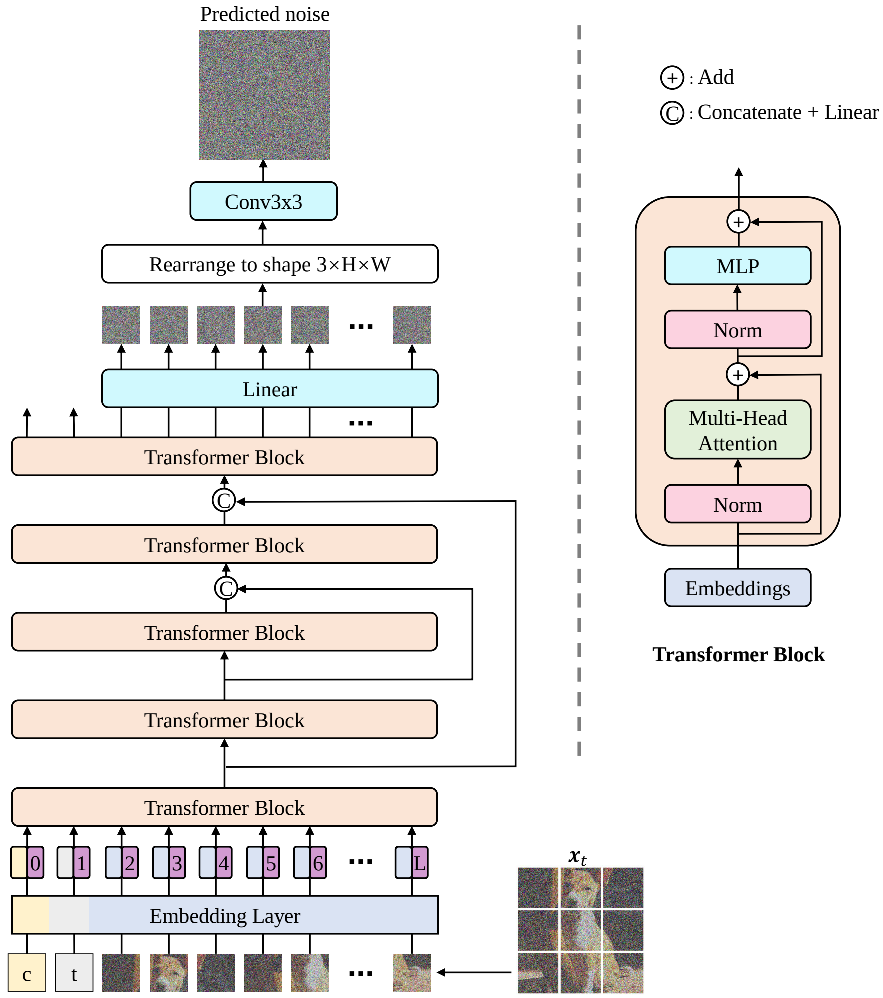


Figure 1: The U-ViT architecture.

2 Development of the U-ViT Architecture

We first attempt to train a diffusion model using a vanilla ViT [3] on CIFAR10. For simplicity, we treat everything including the time embedding, label embedding and patches of the noisy image as tokens. With carefully tuned hyperparameters, a 13-layer ViT of size 41M achieves a FID 5.97, which is significantly better than 20.20 of the prior ViT-based diffusion models [18]. We conjecture that this is mainly because our model is larger. However, this is clearly worse than 3.17 of the U-Net [5] of a similar size.

The importance of the skip connections in U-Net has been realized for a long time in low-level vision tasks [13]. Since all local information are also crucial in score modeling (or noise prediction), we hypothesize that the skip connections play a central role in such tasks as well. Therefore, we add extra skip connections to ViT and obtain a FID of 4.24.

Finally, we add a 3x3 convolutional block before the output to avoid potential artifacts between patches and obtain a FID of 3.11, which is competitive to the results of DDPM [5]. The overall architecture is illustrated in Fig. 1 and the ablation results are summarized in Table 1 for clarity.

Table 1: Ablation study on the architecture design on CIFAR10.

Skip connection	Conv3x3	FID
✓	✓	**3.11**
✓	×	4.24
×	✓	7.37
×	×	5.97

3 Experiments

We evaluate U-ViT on CIFAR10 [7], CelebA 64x64 [8], and ImageNet 64x64 [2]. We provide detailed experimental settings in Table 2.

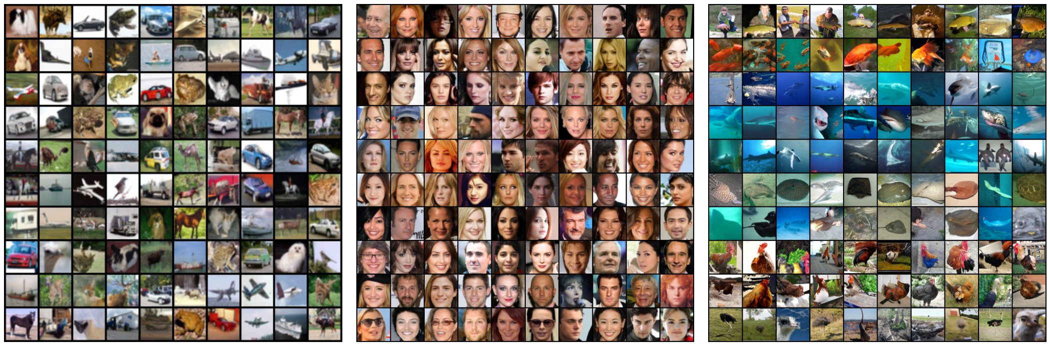
Dataset	CIFAR10	CelebA 64x64	ImageNet 64x64
Patch size	2	4	4
Layers	13	13	17
Hidden size	512	512	768
MLP size	2048	2048	3072
Heads	8	8	12
Params	44M	44M	131M
Noise schedule	VP SDE [16]	VP SDE	VP SDE
Batch size	128	128	1024
Training steps	500K	500K	300K
Warm-up steps	5K	5K	5K
Optimizer	AdamW [9]	AdamW	AdamW
Learning rate	2e-4	2e-4	3e-4
Weight decay	0.03	0.03	0.03
Betas	(0.99, 0.999)	(0.99, 0.99)	(0.99, 0.99)
Sampler	EM	EM	DPM-Solver [10]
Sampling steps	1K	1K	50

Table 2: The experimental settings. EM represents the Euler-Maruyama sampler.

We compare U-ViT with commonly used U-Net in diffusion models [5, 11, 16]. We also compare with GenViT [18], a smaller ViT which does not employ long skip connections and the 3x3 convolutional block, and incorporates time before normalization layers. As shown in Table 3, the FID results on CIFAR10 and CelebA 64x64 are comparable to U-Net. As shown in Table 4, on ImageNet 64x64, U-ViT is comparable to IDDPM U-Net (small), which has a comparable number of parameters. Note that there is still a gap between U-ViT and IDDPM U-Net (large), which could potentially be narrowed by further increasing the U-ViT size or increasing the batch size and training steps. We provide generated samples of U-ViT in Figure 2, which have good quality and clear semantics.

Table 3: FID $\downarrow$ results on unconditional datasets.

Architecture	CIFAR10	CelebA 64x64
DDPM U-Net [5]	3.17	3.26 [14]
IDDPM U-Net [11]	2.90	-
DDPM++ U-Net [16]	2.55	1.90 [6]
GenViT [18]	20.20	-
U-ViT (ours)	3.11	3.13



(a) CIFAR10 (b) CelebA 64x64 (c) ImageNet 64x64

Figure 2: Generated samples of U-ViT.

Table 4: FID ↓ results on class-conditional ImageNet 64x64 and comparison of experimental setting.

Architecture	FID ↓	Params	Batch size	Training steps
IDDPM U-Net (small) [11]	6.92	100M	2048	1700K
IDDPM U-Net (large) [11]	2.92	270M	2048	250K
U-ViT (ours)	6.75	131M	1024	300K

3.1 Efficiency Comparison

We compare efficiency of U-Net and U-ViT on CIFAR10 in Table 5. U-ViT has fewer parameters. When the computation resource is unsaturated, e.g., using a batch size of 1, U-ViT has a much higher throughput than U-Net. When the computation resource is saturated, e.g., using a large batch size of 500, U-ViT has a slightly lower throughput than U-Net. This means that U-ViT has a slightly larger computational cost, but meanwhile enjoys a better parallelism than U-Net.

Table 5: Efficiency comparison on CIFAR10 in one A40 GPU. Throughput is measured by the number of processed inputs in a second.

Method	FID ↓	Params	Throughput (batch size=1)	Throughput (batch size=500)
IDDPM U-Net [11]	2.90	53M	22/s	1297/s
U-ViT (ours)	3.11	44M	55/s	1125/s

References

[1] Xinlei Chen, Saining Xie, and Kaiming He. An empirical study of training self-supervised vision transformers. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 9640–9649, 2021.

[2] 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.

[3] Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, et al. An image is worth 16x16 words: Transformers for image recognition at scale. *arXiv preprint arXiv:2010.11929*, 2020.

[4] Kaiming He, Xinlei Chen, Saining Xie, Yanghao Li, Piotr Dollár, and Ross Girshick. Masked autoencoders are scalable vision learners. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 16000–16009, 2022.

[5] Jonathan Ho, Ajay Jain, and Pieter Abbeel. Denoising diffusion probabilistic models. *arXiv preprint arXiv:2006.11239*, 2020.

[6] Dongjun Kim, Seungjae Shin, Kyungwoo Song, Wanmo Kang, and Il-Chul Moon. Soft truncation: A universal training technique of score-based diffusion model for high precision score estimation. In *International Conference on Machine Learning*, pages 11201–11228. PMLR, 2022.

[7] Alex Krizhevsky, Geoffrey Hinton, et al. Learning multiple layers of features from tiny images. 2009.

[8] Ziwei Liu, Ping Luo, Xiaogang Wang, and Xiaoou Tang. Deep learning face attributes in the wild. In *2015 IEEE International Conference on Computer Vision, ICCV 2015, Santiago, Chile, December 7-13, 2015*, pages 3730–3738. IEEE Computer Society, 2015.

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

[10] 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.

[11] Alex Nichol and Prafulla Dhariwal. Improved denoising diffusion probabilistic models. *arXiv preprint arXiv:2102.09672*, 2021.

[12] 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.

[13] 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.

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

[15] Yang Song and Stefano Ermon. Generative modeling by estimating gradients of the data distribution. *arXiv preprint arXiv:1907.05600*, 2019.

[16] Yang Song, Jascha Sohl-Dickstein, Diederik P Kingma, Abhishek Kumar, Stefano Ermon, and Ben Poole. Score-based generative modeling through stochastic differential equations. *arXiv preprint arXiv:2011.13456*, 2020.

[17] Robin Strudel, Ricardo Garcia, Ivan Laptev, and Cordelia Schmid. Segmenter: Transformer for semantic segmentation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 7262–7272, 2021.

[18] Xiulong Yang, Sheng-Min Shih, Yinlin Fu, Xiaoting Zhao, and Shihao Ji. Your vit is secretly a hybrid discriminative-generative diffusion model. *arXiv preprint arXiv:2208.07791*, 2022.

[19] Sixiao Zheng, Jiachen Lu, Hengshuang Zhao, Xiatian Zhu, Zekun Luo, Yabiao Wang, Yanwei Fu, Jianfeng Feng, Tao Xiang, Philip HS Torr, et al. Rethinking semantic segmentation from a sequence-to-sequence perspective with transformers. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 6881–6890, 2021.