DiC: Rethinking Conv3x3 Designs in Diffusion Models

Supplementary Material

6. Additional Experiments

Further Details about Baselines. In Tab. 6, most baselines including PixArt-α, DiffiT [17], and DiT-LLaMA [5] are direct improvements over DiTs [30]; U-ViTs [1] are published earlier to DiTs, and we also find some coincidence between the hyperparameters for model setup. The major difference between the architecture of U-ViTs and DiTs is the use of skip connections. DiTs totally eliminate the skips, maintaining a clear isotropic architecture. The results of DiT-LLaMA is replicated by us because it fails to report an official result for 400K iterations under the DiT setting. This omission is strange to us because DiTs report the 400K results for smaller models when compared with DiT.

Credit. Baseline performance statistics in Tab. 6 are from [38], a work that measures the capability of Diffusion Transformers under the aligned standard setting of DiT.

DiT Combined with U-Net. We also conducted the experiment that combines DiT transformer block with the U-Net architecture, shown in Tab. 10. In contrast, U-Nets could bring more improvements to field-limited ConvNets.

ImageNet 256×256, 200K, cfg=1.5			
Model	G FLOPs	FID↓	IS↑
DiT-XL/2	118.6	12.96	94.26
DiT+U-Net	117.5	**11.03**	**104.92**

Table 10. **Improvements of U-Net on DiT.** The improvements of U-Net on transformers are not as large as on ConvNets.

More Traing Iterations on ImageNet 512x512. We have extended the training iterations for some limited number of iterations, shown in Tab. 11. Both DiC-XL and DiC-H could outperform DiT-XL/2 at much fewer training iterations while maintaining a speed advantage.

ImageNet 512×512, Scale Up			
Model	Training Steps	FID↓	IS↑
DiT-XL/2	1.3M	13.78	-
DiC-XL	600K	13.64	**102.63**
DiC-H	400K	**12.89**	101.78

Table 11. **Fast Convergence of DiC.** DiC models could achieve better performance at much fewer training iterations on ImageNet 512x512.

Visual Results from ImageNet 512x512. In Fig. 6, we also present the samples generated by DiC-XL, trained for 1M iterations. The samples are generated with the setting of DiT.

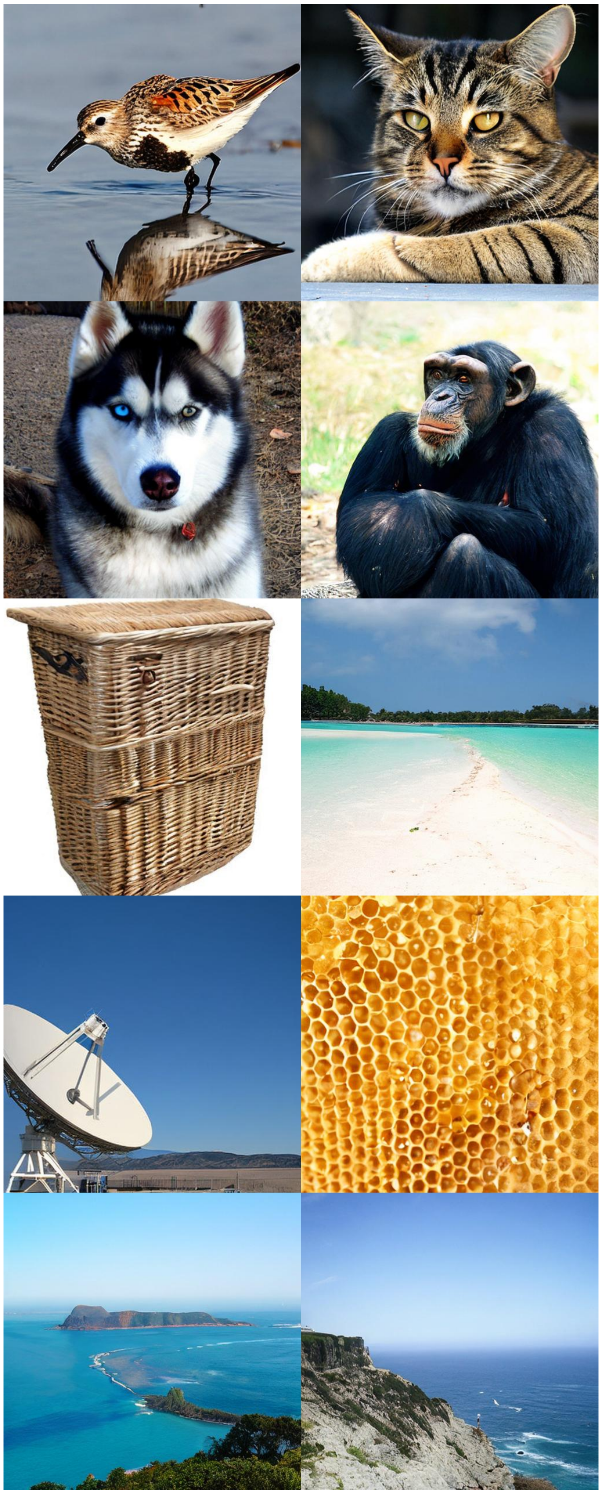

Figure 6. **512x512 Samples generated by DiC at 1M iterations.** The samples are generated following the setting of DiT, at $cfg = 4$. Best viewed on screen.

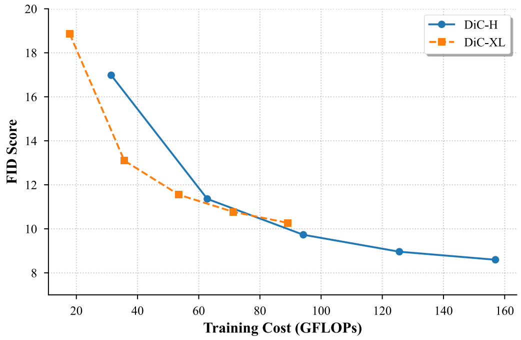

Figure 7. **The scaling curve of DiC training.** The FID scores are recorded once every 200K iterations in the first 1M training iterations. The scaling effect of DiC as model gets larger is obvious from the plot.

Scaling plots. We visualize the scaling curve of DiC-XL and DiC-H as shown in Fig. 7.

Comparing with more architectures. Apart from the architectures mentioned in Tab. 4, we are aware of some other competitive architectures including Simple Diffusion [20], RIN [21], EDM2 [22], and HDiT [6]. However, we find difficulty in comparing these methods with DiC: DiC is mainly focused on 256-sized latent diffusion, following DiT [30] and SiT [29]; these work, on the other hand, focuses on large (mostly pixel-space) diffusion; and they require large training costs to reach SOTA FIDs (e.g. EDM2 requires the training cost of 939.5-2147.5M img, which is around 4M to 8M iterations in our setting). We try to align HDiT, RIN, and EDM2 to our setting (under the training framework of DiT; in order to keep FLOPs aligned with DiC for fair comparison, we increase the depths and widths of these models). Results turns out that these methods either converges slowly (RIN, EDM2) or completely fails (HDiT).

Details regarding Representation Alignment on DiC. We consider applying Representation Alignment (REPA) [46] (using its variant U-REPA [39] tailored for U-Net) to achieve faster convergence. We use the standard training hyperparameter for REPA. For sampling, we use $cfg = 1.8$ for SDE, and $cfg = 2$ for ODE, both equipped with guidance interval (following the default setting of the official codebase). Amazingly, DiC-XL could reach an FID of 1.74 after 1M training steps with the help of REPA, as shown in Tab. 12.

ImageNet 256×256, REPA			
Model	Training iter	Sampling	FID↓
DiC-XL+U-REPA	1M	ODE	1.74
DiC-XL+U-REPA	1M	SDE	1.75

Table 12. **Fast Convergence of DiC with the help of REPA.** DiC-XL could achieve 1.75 FID after 1M training iterations.