Supplementary Materials: Relational Diffusion Distillation For Efficient Image Generation

Anonymous Authors

1 DERIVATING DENOISED IMAGE FOR DISTILLATION

DDIM [1] proposes an implicit sampling to speed up the inference process by the following equation.

$$\mathbf{z}_s = \alpha_s \theta(\mathbf{z}_t, t) + \sigma_s \frac{\mathbf{z}_t - \alpha_t \theta(\mathbf{z}_t, t)}{\sigma_t} \tag{1}$$

Assuming that we have an N-step teacher, and the current sampling time is t, then we can get $t' = t - 1/N$ and $t'' = t - 2/N$. Based on Eq. 1, $\mathbf{z}_{t'}$ and $\mathbf{z}_{t''}$ are calculated as

$$\mathbf{z}_{t'} = \alpha_{t'} \eta(\mathbf{z}_t) + \sigma_{t'} \frac{\mathbf{z}_t - \alpha_t \eta(\mathbf{z}_t)}{\sigma_t} \tag{2}$$

$$\mathbf{z}_{t''} = \alpha_{t''} \eta(\mathbf{z}_{t'}) + \sigma_{t''} \frac{\mathbf{z}_{t'} - \alpha_{t'} \eta(\mathbf{z}_{t'})}{\sigma_{t'}} \tag{3}$$

where η is the teacher model.

Assuming the student has the denoised image $\mathbf{x}^S$ and gets noisy image $\widetilde{\mathbf{z}}_{t''}$ in one step. We want to align $\mathbf{z}_{t''}$ and $\widetilde{\mathbf{z}}_{t''}$, then we have

$$\mathbf{z}_{t''} = \widetilde{\mathbf{z}}_{t''} = \alpha_{t''} \mathbf{x}^S + \sigma_{t''} \frac{\mathbf{z}_t - \alpha_t \mathbf{x}^S}{\sigma_t} \tag{4}$$

Based on Eq. 4, we can actually align $\mathbf{x}^T$ and $\mathbf{x}^S$, thus we can get the distillation target $\mathbf{x}^T$ by

$$\mathbf{x}^T = \mathbf{x}^S = \frac{\mathbf{z}_{t''} - (\sigma_{t''}/\sigma_t)\mathbf{z}_t}{\alpha_{t''} - (\sigma_{t''}/\sigma_t)\alpha_t} \tag{5}$$

2 EXPERIMENT DETAILS

2.1 Model Architecture

For the CIFAR-10 dataset, we use the same architecture as described in RCFD [2]. The U-Net includes four feature map resolutions (32 × 32 to 4 × 4), and it has two convolutional residual blocks per resolution level and self-attention blocks at 8 × 8 resolution. Diffusion time t is embedded into each residual block. The initial channel number is 128 and is multiplied by 2 at the last three resolutions.

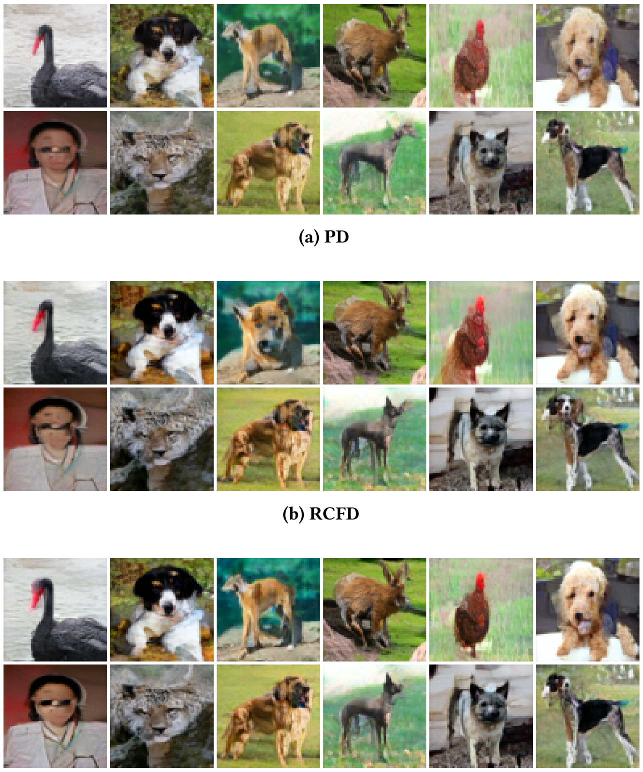
For the ImageNet 64×64 dataset, we slightly modify the architecture to fit the resolution. The U-Net includes four feature map resolutions (64 × 64 to 8 × 8), and it has three convolutional residual blocks per resolution level and self-attention blocks at 16 × 16 and 8 × 8 resolution. Diffusion time t and class label y are embedded into each residual block. The initial channel number is 128 and is multiplied by 2, 3, and 4 at the corresponding last three resolutions.

2.2 Training details

In training basic model process using PD. For CIFAR-10, we set the learning rate to 0.0002 (with a warmup period of 5000 iterations), dropout to 0.1, batch size to 128, exponential moving average (EMA) decay to 0.9999, gradient clipping to 1, and the total number of iterations to 800,000. For ImageNet, we set the learning rate to 0.0001 (with a warmup period of 5000 iterations), dropout to 0, batch size to 512, EMA decay to 0.9999, gradient clipping to 1, and the total number of iterations to 1,000,000.

During the distillation process using different methods, for CIFAR-10, we set the learning rate (using cosine annealing) to 5e-5, batch size to 128, gradient clipping to 1, and the total number of iterations to 20,000 for 8 to 2-step distillation and 40,000 for 2 to 1-step distillation. For ImageNet, we set the learning rate (using cosine annealing) to 0.0001, batch size to 256, gradient clipping to 1, and the total number of iterations to 50,000 for 8 to 2-step distillation and 100,000 for 2 to 1-step distillation.

3 MORE VISUALIZATION RESULTS



(a) PD

(b) RCFD

(c) RDD (Ours)

Figure 1: Samples generated in one step by (top) PD, (middle) RCFD, and (bottom) our proposed RDD on ImageNet 64×64. All corresponding images are generated from the same initial noise.

In Fig. 1 we visualize some of the generated results on ImageNet 64×64. Among them, our method RDD is superior to PD and RCFD in generating details and color. This shows that our method can further improve the effect of distillation and improve the quality of image generation.

REFERENCES

[1] Jiaming Song, Chenlin Meng, and Stefano Ermon. 2020. Denoising Diffusion Implicit Models. In *International Conference on Learning Representations*.

[2] Wujie Sun, Defang Chen, Can Wang, Deshi Ye, Yan Feng, and Chun Chen. 2023. Accelerating diffusion sampling with classifier-based feature distillation. In *2023 IEEE International Conference on Multimedia and Expo (ICME)*. IEEE, 810–815.