Pretraining is All You Need for Image-to-Image Translation

Tengfei Wang et al., HKUST, Microsoft Research Asia

Preprint (NeurIPS 2022 under review)

https://arxiv.org/abs/2205.12952

Presented by Minho Park

Contributions

1. Adapt pre-trained generative model (i.e., GLIDE[1]) to accommodate various kind of image-to-image translation and achieve state-of-the-art results.
2. Propose adversarial and perceptual loss for fine-tuning diffusion upsampler.
3. Propose normalized guidance sampling (similar to dynamic thresholding in Imagen[2])

1) Nichol, Alex, et al. "Glide: Towards photorealistic image generation and editing with text-guided diffusion models." *arXiv preprint arXiv:2112.10741* (2021) (ICML 2022).
2) Saharia, Chitwan, et al. "Photorealistic Text-to-Image Diffusion Models with Deep Language Understanding." *arXiv preprint arXiv:2205.11487* (2022).

Diffusion Model (Method)

- While GANs mainly work for specific domains (e.g., faces), diffusion models emerge to show impressive expressivity of synthesizing a wide variety of images.

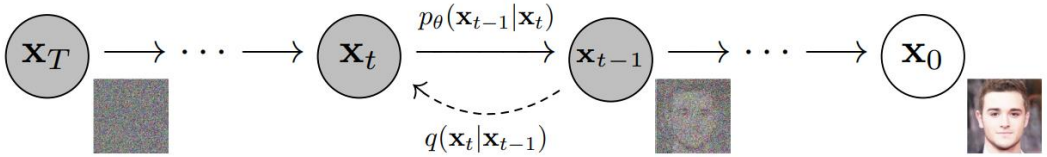
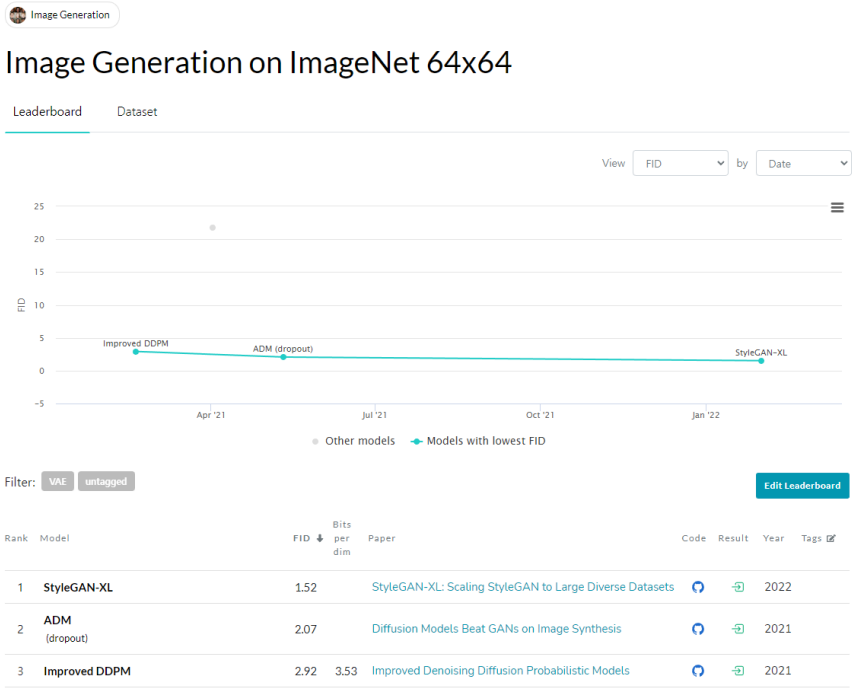


Figure 2: The directed graphical model considered in this work.

Denoising Diffusion Probabilistic Models: Method

Figure 3: Samples from an unconditional diffusion model with classifier guidance to condition on the class "Pembroke Welsh corgi". Using classifier scale 1.0 (left; FID: 33.0) does not produce convincing samples in this class, whereas classifier scale 10.0 (right; FID: 12.0) produces much more class-consistent images.

ADM: Conditional sampling for DDIM

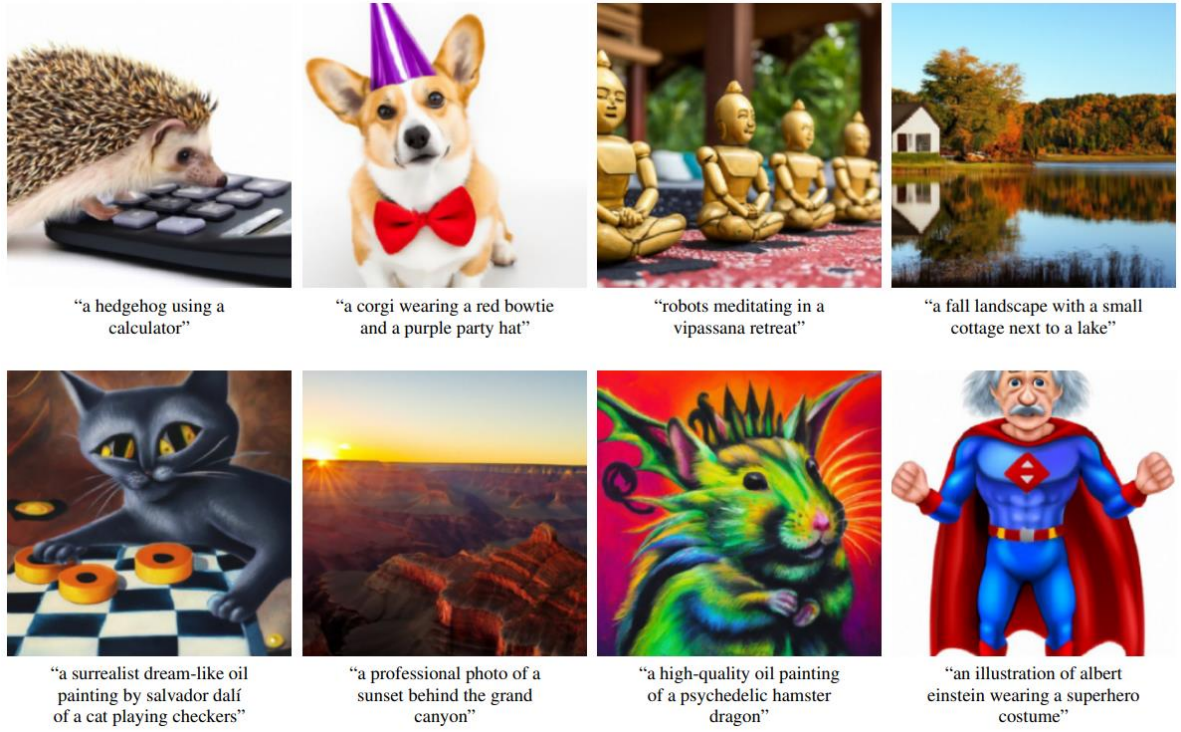


Unconditional benchmark (ImageNet 64X64)

Ho, Jonathan, Ajay Jain, and Pieter Abbeel. "Denoising diffusion probabilistic models." *Advances in Neural Information Processing Systems* 33 (2020): 6840-6851.
Dhariwal, Prafulla, and Alexander Nichol. "Diffusion models beat gans on image synthesis." *Advances in Neural Information Processing Systems* 34 (2021).

Diffusion Model (GLIDE)

- Classifier-free guidance for CLIP.
  - It makes us do not have to train CLIP with noised dataset.



Selected samples from GLIDE using classifier-free guidance.

Iteratively creating a complex scene using GLIDE. The green region is erased, and the model fills it in conditioned on the given prompt.

Examples of text-conditional SDEdit with GLIDE.

Nichol, Alex, et al. "Glide: Towards photorealistic image generation and editing with text-guided diffusion models." *arXiv preprint arXiv:2112.10741* (2021) (ICML 2022).

PITI

- PITI provides reasonable qualitative results on complex datasets leveraging GLIDE.
  - Public GLIDE is trained on approximately 67M text-image pairs with people and violent objects removed.

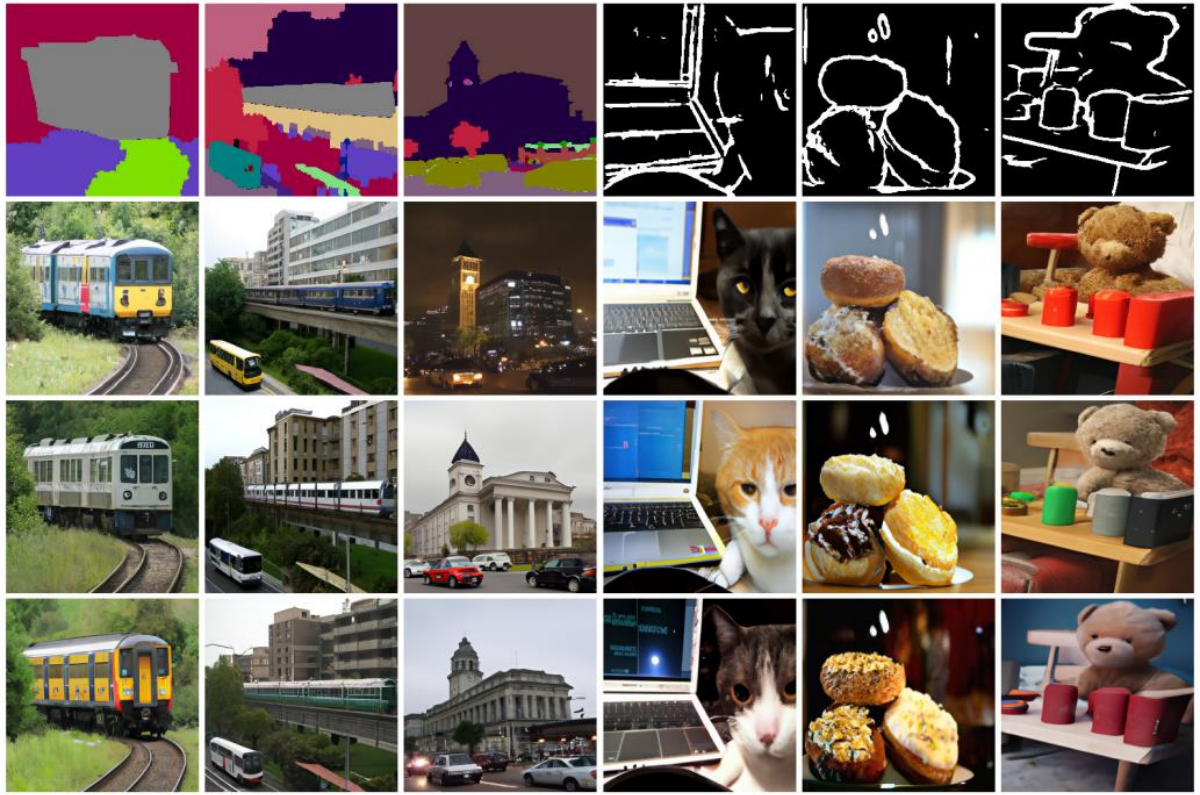


Figure 1: Diverse images sampled by our method given semantic layouts or sketches.

Figure 3: Visual comparisons on COCO and ADE20K. More results are shown in **Appendix**.

Public GLIDE: https://github.com/openai/glide-text2im

Overall Framework

- As opposed to taking images from the same domain as for discriminate tasks, the pretrained model for generative tasks consumes vastly different kinds of images in distinct downstream tasks.
- They expect the diffusion model to generate images from a latent space that is later shared to use for all the downstream tasks.

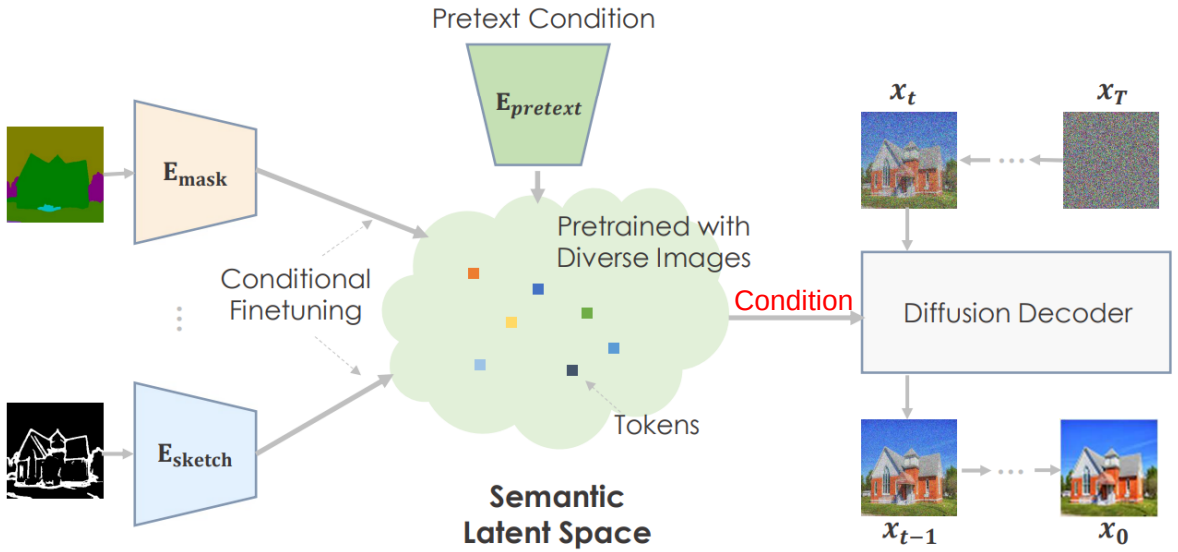


Figure 2: The overall framework. We can perform pretraining on huge data via different pretext tasks and learn a highly semantic latent space that models general and high-quality image statistics. For downstream tasks, we perform conditional finetuning to map the task-specific conditions to this pretrained semantic space. By leveraging the pretrained knowledge, our model renders plausible images based on different conditions.

Downstream Adaptation (Base Model)

- Stage-wise training is **helpful to cultivate the pretrained knowledge** as much as possible and is proven crucial for much improved quality.

1. Train the task-specific encoder and leave the pretrained decoder intact.
   - The outputs will roughly match the semantics of the input, but without accurate spatial alignment.
2. Finetune both the encoder and decoder altogether.
   - After this, we obtain much improved spatial semantic alignment.

Figure 5: Fixing the decoder generates high-quality images but fails to align with the condition.

Table 3: Ablation study of the proposed PITI on ADE20K dataset.
(a) Finetune strategy.

Finetune strategy	FID
Fixed decoder	12.6
One-stage finetune	13.3
Two-stage finetune	8.9

Downstream Adaptation (Upsampler)

1. Finetune the diffusion upsampler for high-resolution generation.
   - Apply random degradation (the real-world BSR degradation).
   - They also introduce L_0 filter[1] to mimic the oversmoothed effect.
   - However, they still observe oversmoothed results.
2. Besides computing a standard mean square error loss for noise prediction, we propose to impose perceptual loss and adversarial loss to improve the perceptual realism of local image structures.

$$\begin{aligned}\mathcal{L}_{\text{perc}} &= \mathbb{E}_{t,\boldsymbol{x}_0,\boldsymbol{\epsilon}}\|\boldsymbol{\psi}_m(\hat{\boldsymbol{x}}_0^t) - \boldsymbol{\psi}_m(\boldsymbol{x}_0)\|, \\ \mathcal{L}_{\text{adv}} &= \mathbb{E}_{t,\boldsymbol{x}_0,\boldsymbol{\epsilon}}\left[\log D_\theta(\hat{\boldsymbol{x}}_0^t)\right] + \mathbb{E}_{\boldsymbol{x}_0}\left[\log(1 - D_\theta(\boldsymbol{x}_0))\right],\end{aligned}$$

Table 3: Ablation study of the proposed PITI on ADE20K dataset.
(b) Upsampling strategy.

Degradation	$\mathcal{L}_{\text{perceptual}}$	$\mathcal{L}_{\text{adversarial}}$	FID
			14.5
✓			12.1
✓	✓		9.8
✓	✓	✓	8.9

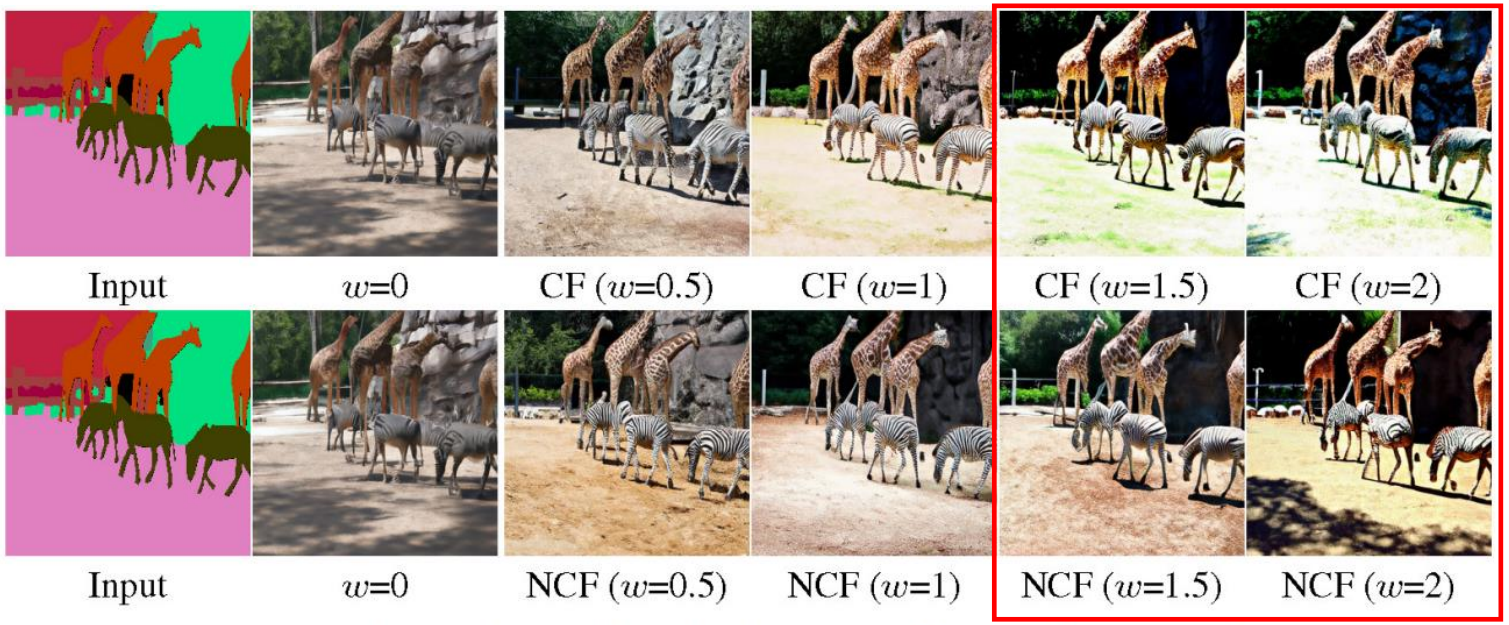
1) Xu, Li, et al. "Image smoothing via L 0 gradient minimization." *Proceedings of the 2011 SIGGRAPH Asia conference*. 2011.

Normalized Classifier-free Guidance

- Classifier-free guidance (GLIDE): $\hat{\epsilon}_\theta\left(\boldsymbol{x}_t|\boldsymbol{y}\right) = \boldsymbol{\epsilon}_\theta\left(\boldsymbol{x}_t|\boldsymbol{y}\right) + w \cdot \left(\boldsymbol{\epsilon}_\theta\left(\boldsymbol{x}_t|\boldsymbol{y}\right) - \boldsymbol{\epsilon}_\theta\left(\boldsymbol{x}_t|\boldsymbol{\emptyset}\right)\right),$
- Not only μ has been shifted, but also the variance of noise sample σ has been increased that leading to over-saturated images with over-smoothed textures.

Similar approach has been proposed in Imagen

- Then they propose normalized classifier-free guidance (NCF): $\tilde{\boldsymbol{\epsilon}}_\theta(\boldsymbol{x}_t|\boldsymbol{y}) = \frac{\sigma}{\hat{\sigma}}(\hat{\boldsymbol{\epsilon}}_\theta(\boldsymbol{x}_t|\boldsymbol{y}) - \hat{\mu}) + \mu.$



Is NCF better than CF?

Figure 6: Effect of normalized classifier-free (NCF) guidance sampling.

Saharia, Chitwan, et al. "Photorealistic Text-to-Image Diffusion Models with Deep Language Understanding." *arXiv preprint arXiv:2205.11487* (2022).

Quantitative Results

- FID, mIoU (missing), and user study.

FID (CLIP)

Table 1: Comparison of FID on various image translation tasks with the best score highlighted.

Method	ADE20K	COCO (Mask)	Flickr (Mask)	COCO (Sketch)	Flickr (Sketch)	DIODE
Pix2PixHD [50]	35.3	37.5	26.1	27.1	16.8	18.2
SPADE [36]	18.9	15.0	17.4	48.9	29.5	17.0
OASIS [44]	14.8	8.8	10.5	-	-	-
Ours (from scratch)	16.3	13.0	10.6	13.0	9.4	13.9
Ours	**8.9**	**5.2**	**6.1**	**8.8**	**6.0**	**11.5**

Table 2: User study on COCO. We report the preference rate of our approach over baselines.

	Ours > SPADE	Ours > OASIS	Ours > Scratch
Preference Rate	93.6%	84.1%	87.4 %

- Smaller training dataset.

Table 4: Comparison on FID with different training image sizes.

Training Size	Pix2PixHD	SPADE	OASIS	Ours (From Scratch)	Ours
25%	44.1	27.1	26.5	24.0	**16.2**
50%	38.4	24.6	20.4	18.8	**12.7**
100%	35.3	18.9	14.8	16.3	**8.9**

Qualitative Results

Figure 4: Visual comparisons on other datasets. More results are shown in **Appendix**.

Limitation

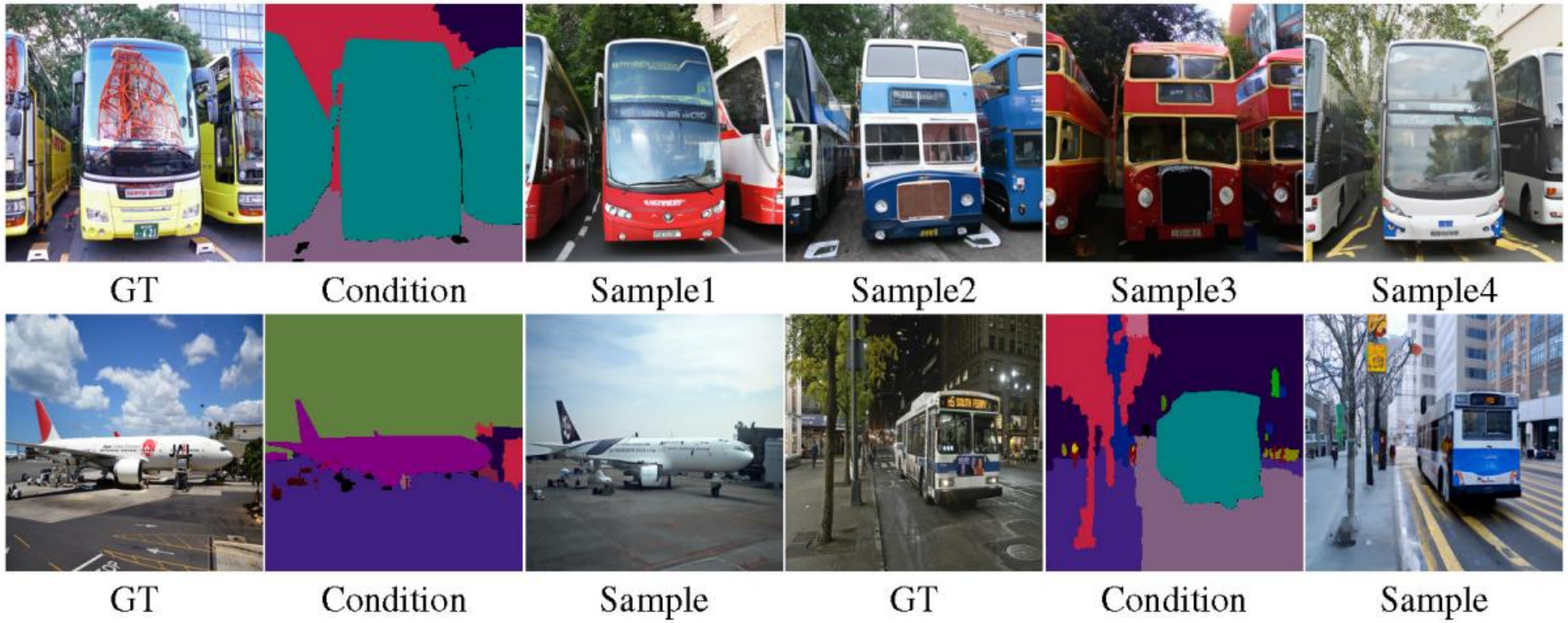


Figure 21: Limitation. In the first row, though our model can produce diverse buses in different samples, we found buses within a single sample tend to exhibit similar style and color. Another limitation of the proposed method is that the samples do not always perfectly align with the input conditions, with some small objects missed. In the left example of the second row, trees in front of the airplane are not synthesized properly. Similarly, the lighting in the right example is missed.