Improving Diffusion-Based Image Synthesis with Context Prediction

Ling Yang, Jingwei Liu, Shenda Hong, Zhilong Zhang, Zhilin Huang, Zheming Cai, Wentao Zhang, Bin Cui

杨灵
北京 海淀

扫一扫上面的二维码图案，加我为朋友。

Contact: yangling0818@163.com

Peking University - China

Improving Diffusion-Based Image Synthesis with Context Prediction

Ling Yang, Jingwei Liu, Shenda Hong, Zhilong Zhang, Zhilin Huang, Zheming Cai, Wentao Zhang, Bin Cui

NEURAL INFORMATION PROCESSING SYSTEMS

Context information is critical for vision tasks

Representation Learning

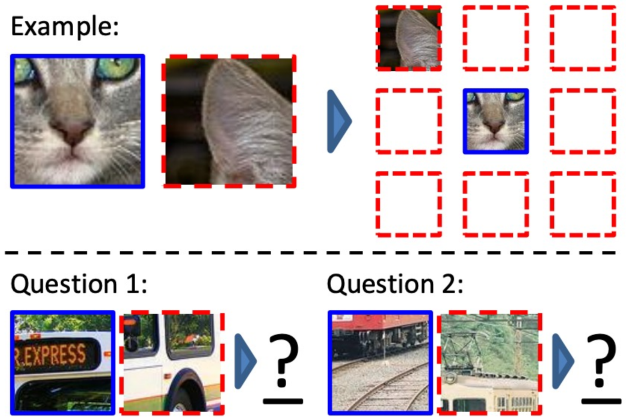


Figure 1. Our task for learning patch representations involves randomly sampling a patch (blue) and then one of eight possible neighbors (red). Can you guess the spatial configuration for the two pairs of patches? Note that the task is much easier once you have recognized the object!

Unsupervised Visual Representation Learning by Context Prediction

Image Inpainting

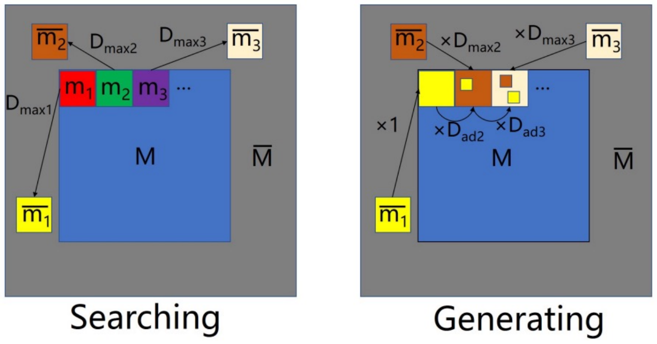


Figure 3. Illustration of the CSA layer. Firstly, we search the most similar contextual patch $\overline{m_i}$ of each generated patch m_i in the hole M, and initialize m_i with $\overline{m_i}$. Then, the previous generated patches and the most similar contextual patch are combined to generate the current one.

Coherent Semantic Attention for Image Inpainting

Peking University - China

Contact: yangling0818@163.com

Ling Yang, Jingwei Liu, Shenda Hong, Zhilong Zhang, Zhilin Huang, Zheming Cai, Wentao Zhang, Bin Cui

Traditional Point-based Diffusion

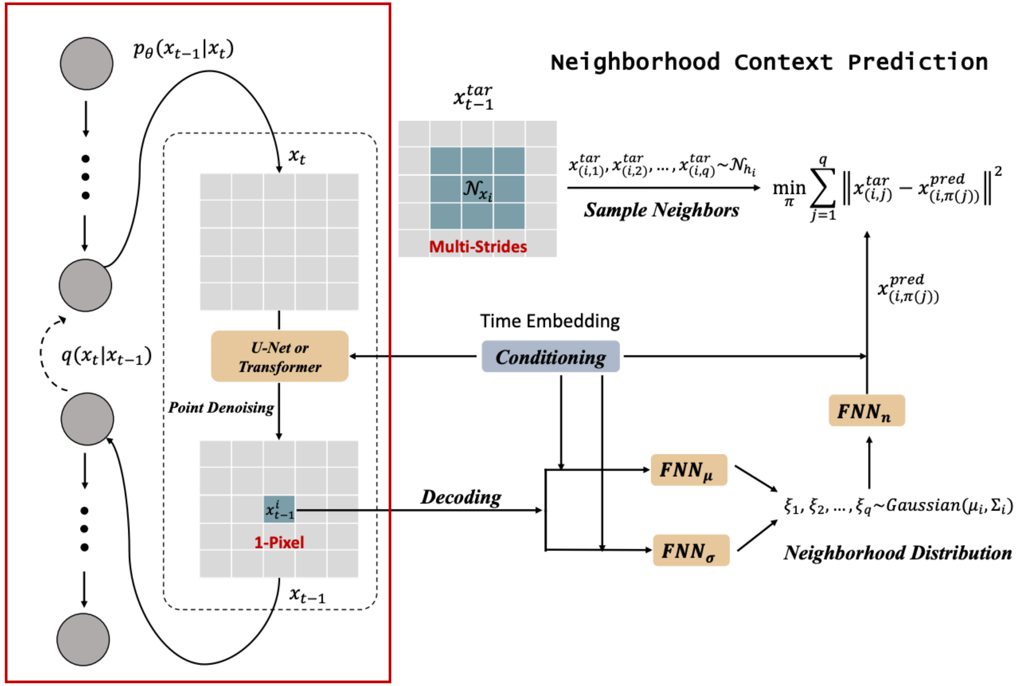


Figure 1: In training stage, ConPreDiff first performs self-denoising as standard diffusion models, then it conducts neighborhood context prediction based on denoised point x_{t-1}^i. In inference stage, ConPreDiff only uses its self-denoising network for sampling.

Peking University - China

Contact: yangling0818@163.com

Ling Yang, Jingwei Liu, Shenda Hong, Zhilong Zhang, Zhilin Huang, Zheming Cai, Wentao Zhang, Bin Cui

Training Objective

$$\mathcal{L}_{\textsc{ConPreDiff}} = \sum_{i=1}^{x \times y} \left[\underbrace{\mathcal{M}_p(\boldsymbol{x}_{t-1}^i, \hat{\boldsymbol{x}}^i)}_{point\ denoising} + \underbrace{\mathcal{M}_n(\boldsymbol{H}_{N_i^s}, \hat{\boldsymbol{H}}_{N_i^s})}_{context\ prediction} \right]$$

Mitigating Complexity Problem

$$\mathcal{L}_{\textsc{ConPreDiff}} = \sum_{i=1}^{x \times y} \left[\underbrace{\mathcal{M}_p(\boldsymbol{x}_{t-1}^i, \hat{\boldsymbol{x}}^i)}_{point\ denoising} + \underbrace{\mathcal{W}_2^2(\psi_n(\boldsymbol{x}_{t-1}^i, t), \mathcal{P}_{N_i^s})}_{neighborhood\ distribution\ prediction} \right]$$

Contact: yangling0818@163.com

Ling Yang, Jingwei Liu, Shenda Hong, Zhilong Zhang, Zhilin Huang, Zheming Cai, Wentao Zhang, Bin Cui

Efficient Large Context Decoding

$$\psi_n(\boldsymbol{x}_{t-1}^i, t) = \mathrm{FNN}_n(\xi),\ \xi \sim \mathcal{N}(\mu_i, \Sigma_i),$$
$$\mu_i = \mathrm{FNN}_\mu(\boldsymbol{x}_{t-1}^i), \Sigma_i = \mathrm{diag}(\exp(\mathrm{FNN}_\sigma(\boldsymbol{x}_{t-1}^i)))$$

Final Loss for Both Discrete and Continuous Diffusion Backbones

$$\mathcal{L}_{\textsc{ConPreDiff}}^{dis} = \mathcal{L}_{t-1}^{dis} + \lambda_t \cdot \sum_{i=1}^{x \times y} \mathcal{W}_2^2(\psi_n(\boldsymbol{x}_{t-1}^i, t), \mathcal{P}_{N_i^s}),$$

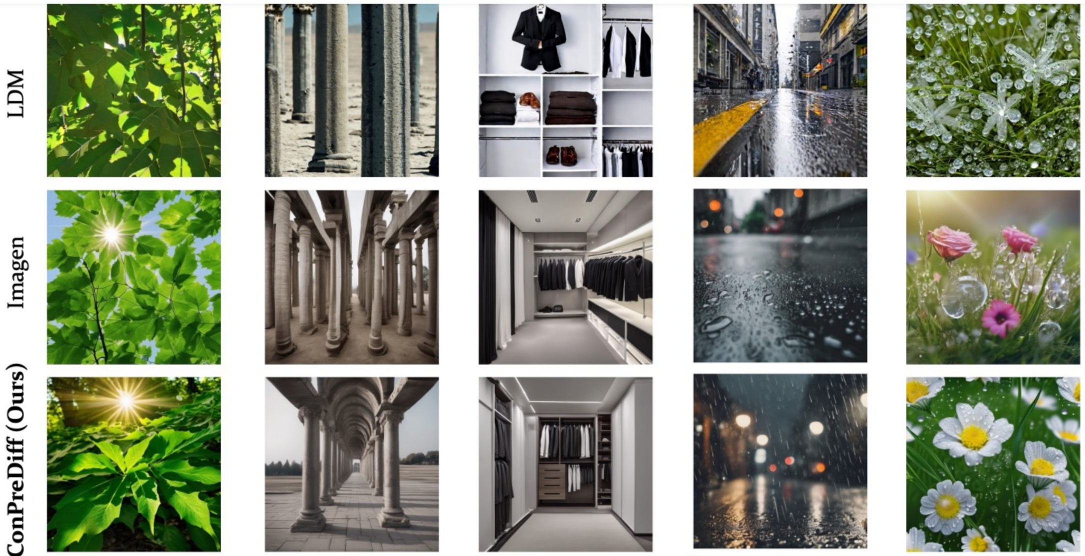
$$\mathcal{L}_{\textsc{ConPreDiff}}^{con} = \mathcal{L}_{t-1}^{con} + \lambda_t \cdot \sum_{i=1}^{x \times y} \mathcal{W}_2^2(\psi_n(\boldsymbol{x}_{t-1}^i, t), \mathcal{P}_{N_i^s}),$$

Improving Diffusion-Based Image Synthesis with Context Prediction

Ling Yang, Jingwei Liu, Shenda Hong, Zhilong Zhang, Zhilin Huang, Zheming Cai, Wentao Zhang, Bin Cui

NEURAL INFORMATION PROCESSING SYSTEMS

Text-to-Image Synthesis



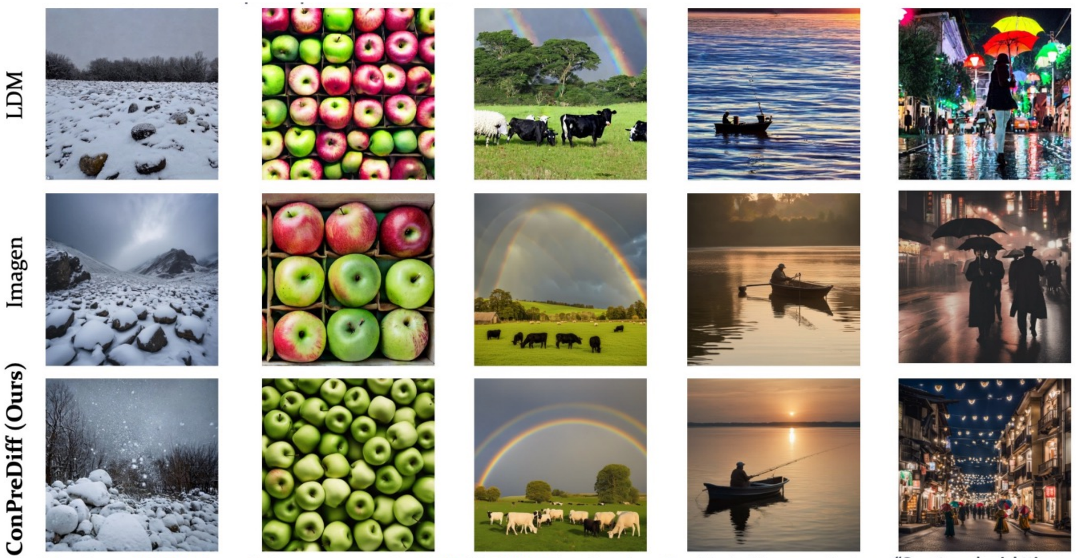
"The sunlight shines on the leaves, and every green leaf reflects light."

"Some ancient stone pillars stand upon the earth, arranged in an orderly manner and spaced apart from each other."

"There are black suits and white pants hanging in the wardrobe."

"On a rainy day street, the air near the ground is filled with raindrops."

"A bunch of flowers bloomed in the grass, with dewdrops on each petal."



"The sky was blanketed with thick snow, while the ground lay adorned with stones."

"A box contains apples, each displaying a touch of green hue."

"A rainbow rise after the rain, a group of cows and sheep graze in the fields. Cows are black and sheep are white."

"An elderly fisherman sits in a boat, his fishing rod set aside, gazing towards the distant horizon. The last rays of the sun reflect off the ripples on the water's surface."

"On a town's nighttime streets, light strips were pulled up on the busy streets. People dressed in various attire and holding umbrellas strolled along the streets."

Peking University - China

Contact: yangling0818@163.com

Improving Diffusion-Based Image Synthesis with Context Prediction

Ling Yang, Jingwei Liu, Shenda Hong, Zhilong Zhang, Zhilin Huang, Zheming Cai, Wentao Zhang, Bin Cui

NEURAL INFORMATION PROCESSING SYSTEMS

Table 1: Quantitative evaluation of FID on MS-COCO for 256 × 256 image resolution.

Approach	Model Type	FID-30K	Zero-shot FID-30K
AttnGAN [96]	GAN	35.49	-
DM-GAN [113]	GAN	32.64	-
DF-GAN [86]	GAN	21.42	-
DM-GAN + CL [100]	GAN	20.79	-
XMC-GAN [107]	GAN	9.33	-
LAFITE [112]	GAN	8.12	-
Make-A-Scene [22]	Autoregressive	7.55	-
DALL-E [61]	Autoregressive	-	17.89
LAFITE [112]	GAN	-	26.94
LDM [65]	Continuous Diffusion	-	12.63
GLIDE [54]	Continuous Diffusion	-	12.24
DALL-E 2 [62]	Continuous Diffusion	-	10.39
Improved VQ-Diffusion [85]	Discrete Diffusion	-	8.44
Simple Diffusion [31]	Continuous Diffusion	-	8.32
Imagen [69]	Continuous Diffusion	-	7.27
Parti [104]	Autoregressive	-	7.23
Muse [7]	Non-Autoregressive	-	7.88
eDiff-I [3]	Continuous Diffusion	-	6.95
ConPreDiff$_{dis}$	Discrete Diffusion	-	**6.67**
ConPreDiff$_{con}$	Continuous Diffusion	-	**6.21**

Peking University - China

Contact: yangling0818@163.com

Improving Diffusion-Based Image Synthesis with Context Prediction

Ling Yang, Jingwei Liu, Shenda Hong, Zhilong Zhang, Zhilin Huang, Zheming Cai, Wentao Zhang, Bin Cui

Image Inpainting

Figure 3: Inpainting examples generated by our CONPREDIFF.

Contact: yangling0818@163.com

Ling Yang, Jingwei Liu, Shenda Hong, Zhilong Zhang, Zhilin Huang, Zheming Cai, Wentao Zhang, Bin Cui

Table 2: Quantitative evaluation of image inpainting on CelebA-HQ and ImageNet.

CelebA-HQ Method	Wide LPIPS ↓	Narrow LPIPS ↓	Super-Resolve 2× LPIPS ↓	Altern. Lines LPIPS ↓	Half LPIPS ↓	Expand LPIPS ↓
AOT [105]	0.104	0.047	0.714	0.667	0.287	0.604
DSI [56]	0.067	0.038	0.128	0.049	0.211	0.487
ICT [91]	0.063	0.036	0.483	0.353	0.166	0.432
DeepFillv2 [103]	0.066	0.049	0.119	0.049	0.209	0.467
LaMa [84]	0.045	0.028	0.177	0.083	**0.138**	0.342
RePaint [49]	0.059	0.028	0.029	**0.009**	0.165	0.435
ConPreDiff	**0.042**	**0.022**	**0.023**	0.022	**0.139**	**0.297**

ImageNet Method	Wide LPIPS ↓	Narrow LPIPS ↓	Super-Resolve 2× LPIPS ↓	Altern. Lines LPIPS ↓	Half LPIPS ↓	Expand LPIPS ↓
DSI [56]	0.117	0.072	0.153	0.069	0.283	0.583
ICT [91]	0.107	0.073	0.708	0.620	0.255	0.544
LaMa [84]	0.105	0.061	0.272	0.121	**0.254**	0.534
RePaint [49]	0.134	0.064	0.183	**0.089**	0.304	0.629
ConPreDiff	**0.098**	**0.057**	**0.129**	0.107	0.285	**0.506**

Contact: yangling0818@163.com

Improving Diffusion-Based Image Synthesis with Context Prediction

Ling Yang, Jingwei Liu, Shenda Hong, Zhilong Zhang, Zhilin Huang, Zheming Cai, Wentao Zhang, Bin Cui

NEURAL INFORMATION PROCESSING SYSTEMS

FFHQ 256 × 256			
Method	FID ↓	Prec. ↑	Recall ↑
ImageBART[16]	9.57	-	-
U-Net GAN (+aug) [72]	7.6	-	-
UDM [39]	5.54	-	-
StyleGAN [36]	4.16	0.71	0.46
ProjectedGAN [71]	3.08	0.65	0.46
LDM [65]	4.98	0.73	0.50
ConPreDiff	**2.24**	**0.81**	**0.61**

LSUN-Bedrooms 256 × 256			
Method	FID ↓	Prec. ↑	Recall ↑
ImageBART [16]	5.51	-	-
DDPM [28]	4.9	-	-
UDM [39]	4.57	-	-
StyleGAN [36]	2.35	0.59	0.48
ADM [14]	1.90	0.66	0.51
ProjectedGAN [71]	1.52	0.61	0.34
LDM-4 [65]	2.95	0.66	0.48
ConPreDiff	**1.12**	**0.73**	**0.59**

CelebA-HQ 256 × 256			
Method	FID ↓	Prec. ↑	Recall ↑
DC-VAE [55]	15.8	-	-
VQGAN+T. [17] (k=400)	10.2	-	-
PGGAN [43]	8.0	-	-
LSGM [87]	7.22	-	-
UDM [39]	7.16	-	-
LDM [65]	5.11	0.72	0.49
ConPreDiff	**3.22**	**0.83**	**0.57**

LSUN-Churches 256 × 256			
Method	FID ↓	Prec. ↑	Recall ↑
DDPM [28]	7.89	-	-
ImageBART [16]	7.32	-	-
PGGAN [43]	6.42	-	-
StyleGAN [36]	4.21	-	-
StyleGAN2 [37]	3.86	-	-
ProjectedGAN [71]	**1.59**	0.61	0.44
LDM [65]	4.02	0.64	0.52
ConPreDiff	1.78	**0.74**	**0.61**

Ling Yang, Jingwei Liu, Shenda Hong, Zhilong Zhang, Zhilin Huang, Zheming Cai, Wentao Zhang, Bin Cui

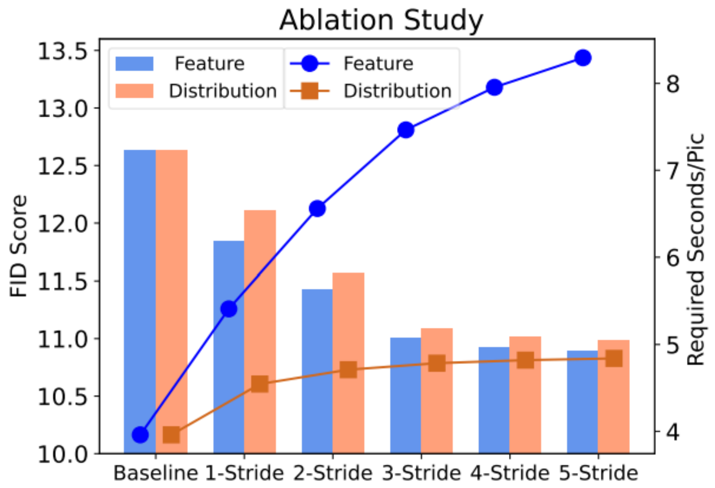


Figure 4: Bar denotes FID and line denotes time cost.

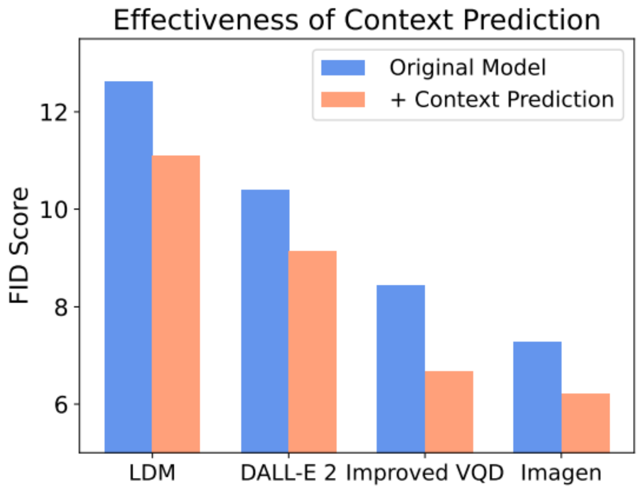


Figure 5: Equip diffusion models with our context prediction.

Contact: yangling0818@163.com

PEKING UNIVERSITY 1898

Improving Diffusion-Based Image Synthesis with Context Prediction

Ling Yang, Jingwei Liu, Shenda Hong, Zhilong Zhang, Zhilin Huang, Zheming Cai, Wentao Zhang, Bin Cui

NEURAL INFORMATION PROCESSING SYSTEMS

"A photo of a dark Goth house"

"A teddy bear sitting on a chair. "

"A person holding a bunch of bananas on a table."

"A group of elephants walking in muddy water. "

"Green frog on green grass"

"The plane wing above the clouds. "

" Trees on African grassland "

" Cat fell asleep on the owner's bed "

"A red hydrant sitting in the snow."

" Pancakes with ketchup "

"A photo of an adult lion."

"A photo of an white garlic ice cream"

Peking University - China

Contact: yangling0818@163.com

Ling Yang, Jingwei Liu, Shenda Hong, Zhilong Zhang, Zhilin Huang, Zheming Cai, Wentao Zhang, Bin Cui

Thanks for Listening !

Contact: yangling0818@163.com