Meeting Note

Author: Y.K. Lin
Date: 2024/09/16 – 2024/09/20

Weekly Summary

High-Resolution Image Synthesis with Latent Diffusion Models [1].

Plan for Next Week

An Introduction to Convolutional Neural Networks (arXiv: 1511.08458).

Details

Although the traditional diffusion model [2] shows impressive results on image generation, it typically generates the image on pixel level, which makes it hard to optimize and expensive to inference due to the sequential evaluation. Other methods such as generative adversarial networks (GANs) [3] encounter more problems like model collapse and difficulty of scaling up the complexity. Thus, Rombach et al. [1] propose the latent diffusion model (LDM) which reduces the required resources significantly for training and retains the quality and flexibility of diffusion model simultaneously. Furthermore, cross-attention layers are integrated into the model, providing the capability of generating images conditionally.

Figure 1 shows two main stages of LDM. The first is semantic compression, which focuses on reducing the semantically unimportant details, leading to more efficient compression without impacting the essential content of original data; the second is perceptual compression, which ensure the compressed image remains visible and perceptually important details using autoencoders and GANs [Re. 1].

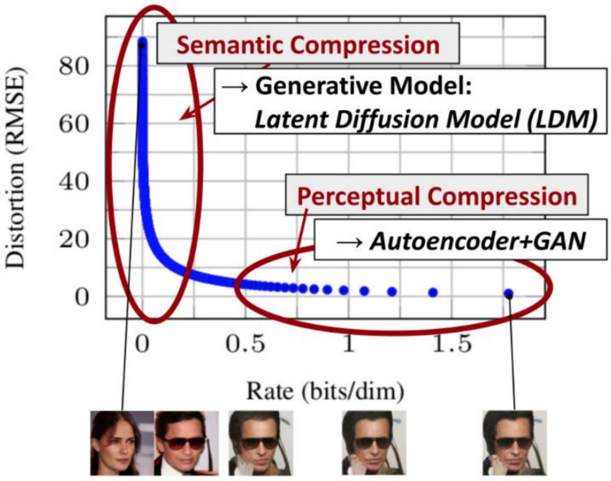


Figure 1. The illustration of semantic and perceptual compression.

Following the previous works, the authors separate the training process into two steps: the first is to train an autoencoder to obtain the low-dimensional representation (latent space) which is perceptually equal to the original data; and the second is to train a generative model (e.g. diffusion models) in the learned latent space to learn the sematic and conceptual information. This approach has some advantages: (1) the universal autoencoder only need to be trained once, and can be used to train different generative models, even to be utilized on other downstream tasks (e.g. CLIP-guided image generation); (2) the sampling of diffusion models is more computationally efficient because it is performed on a low-dimensional space.

As shown in Figure 2, the $\mathcal{E}$ denotes the encoder which encodes the image x into the latent representation $z = \mathcal{E}(x)$, and the decoder $\mathcal{D}$ reconstructs the image $\tilde{x} = \mathcal{D}(z) = \mathcal{D}(\mathcal{E}(x))$. This part, called perceptual compression, can be considered as a VQGAN [4] model, and is trained with perceptual loss [5] and batch-based [6] adversarial objective [4, 7, Re. 2], this ensures the reconstructions are confined to the image manifold by enforcing local realism and avoids blurriness introduced by relying solely on pixel-space losses such as L2 or L1 objectives [Re. 3]. As the result, we then can obtain a low-dimensional latent space with efficient information representation, where the high-frequency and imperceptible details are abstracted away.

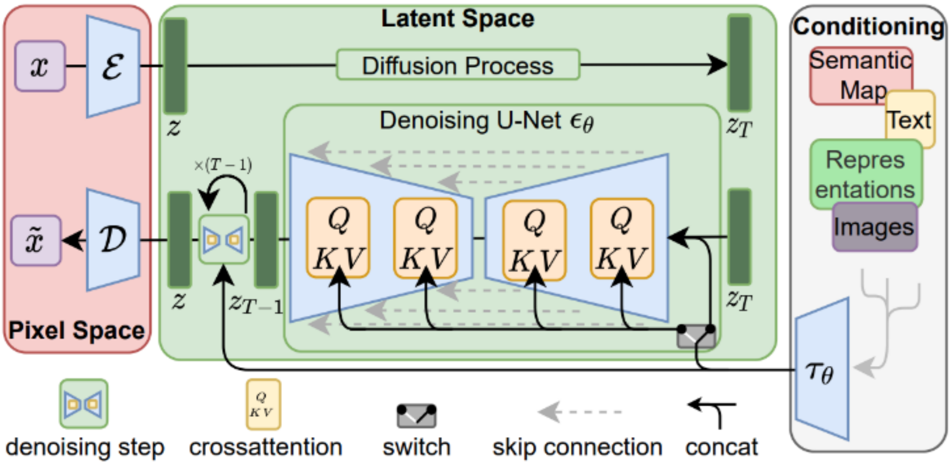


Figure 2. Overall architecture of LDM.

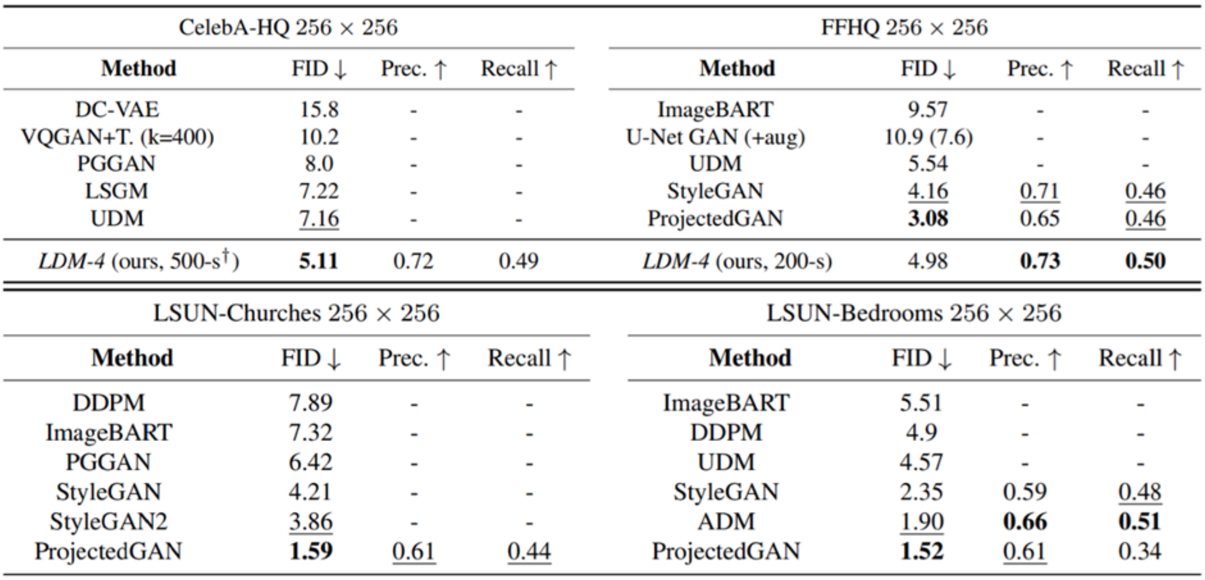
After we obtain the latent representation z from encoder $\mathcal{E}$, it then be fed into a fixed forward process [Re. 4] and outputted as z_T. Like DDPM [8], LDM also uses U-Net [9] as the backbone of denoising model. To give the capability of conditional generation to LDM, the authors use cross-attention [10] to integrate the conditional input $\tau_\theta(y)$, where y is the conditional data (e.g. texts, sematic maps, etc.) and τ_θ is a domain specific encoder that project y to another vector space. The attention can be formed as follows:

$$Attention(Q,K,V) = softmax\left(\frac{QK^T}{\sqrt{d}}\right) \cdot V$$

where $Q = W_Q^{(i)} \cdot \varphi_i(z_t)$, $K = W_K^{(i)} \cdot \tau_\theta(y)$, $V = W_V^{(i)} \cdot \tau_\theta(y)$, and $\varphi_i(z_t)$ denotes the flattened intermediate representation output from U-Net.

Several experiments are done for LDM, such as unconditional image generation, text-to-image generation, layout-to-image generation, super-resolution, LDM achieves the state-of-art perfo-

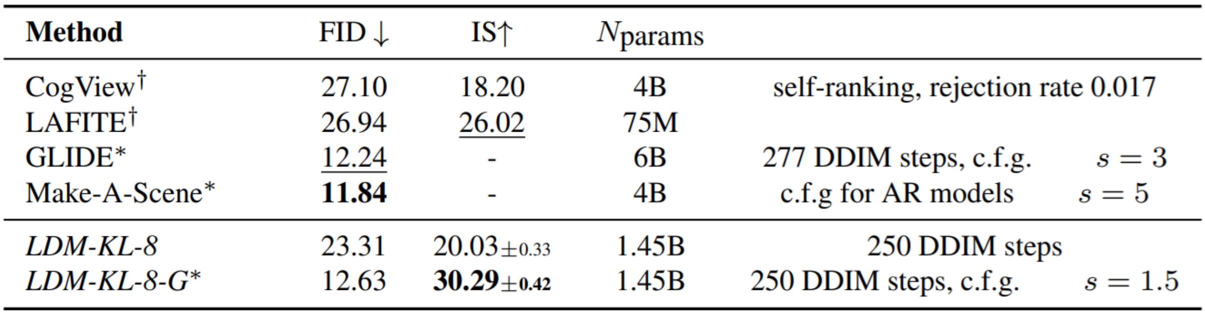
Table 1. The experimental results for unconditional image generation.



CelebA-HQ 256 × 256				FFHQ 256 × 256			
Method	FID ↓	Prec. ↑	Recall ↑	**Method**	FID ↓	Prec. ↑	Recall ↑
DC-VAE	15.8	-	-	ImageBART	9.57	-	-
VQGAN+T. (k=400)	10.2	-	-	U-Net GAN (+aug)	10.9 (7.6)	-	-
PGGAN	8.0	-	-	UDM	5.54	-	-
LSGM	7.22	-	-	StyleGAN	4.16	0.71	0.46
UDM	7.16	-	-	ProjectedGAN	**3.08**	0.65	0.46
LDM-4 (ours, 500-s†)	**5.11**	0.72	0.49	*LDM-4* (ours, 200-s)	4.98	**0.73**	**0.50**

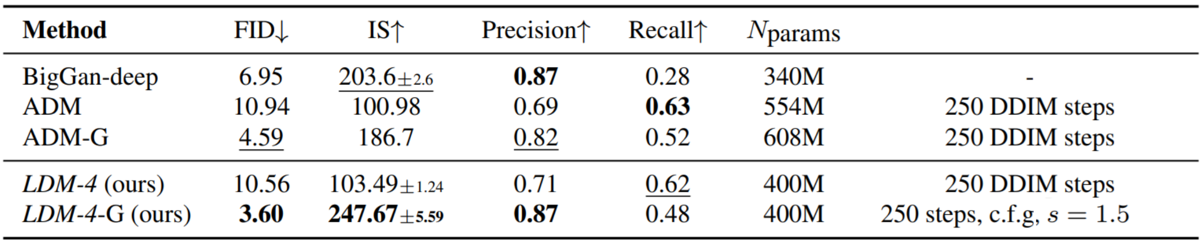
LSUN-Churches 256 × 256				LSUN-Bedrooms 256 × 256			
Method	FID ↓	Prec. ↑	Recall ↑	**Method**	FID ↓	Prec. ↑	Recall ↑
DDPM	7.89	-	-	ImageBART	5.51	-	-
ImageBART	7.32	-	-	DDPM	4.9	-	-
PGGAN	6.42	-	-	UDM	4.57	-	-
StyleGAN	4.21	-	-	StyleGAN	2.35	0.59	0.48
StyleGAN2	3.86	-	-	ADM	1.90	**0.66**	**0.51**
ProjectedGAN	**1.59**	0.61	0.44	ProjectedGAN	**1.52**	0.61	0.34

Table 2. The experimental results for text-conditional image generation.



Method	FID ↓	IS↑	N_{params}		
CogView†	27.10	18.20	4B	self-ranking, rejection rate 0.017	
LAFITE†	26.94	26.02	75M		
GLIDE*	12.24	-	6B	277 DDIM steps, c.f.g.	$s = 3$
Make-A-Scene*	**11.84**	-	4B	c.f.g for AR models	$s = 5$
LDM-KL-8	23.31	20.03±0.33	1.45B	250 DDIM steps	
*LDM-KL-8-G**	12.63	**30.29**±**0.42**	1.45B	250 DDIM steps, c.f.g.	$s = 1.5$

Table 3. The experimental results for class-conditional image generation.



Method	FID↓	IS↑	Precision↑	Recall↑	N_{params}	
BigGan-deep	6.95	203.6±2.6	**0.87**	0.28	340M	-
ADM	10.94	100.98	0.69	**0.63**	554M	250 DDIM steps
ADM-G	4.59	186.7	0.82	0.52	608M	250 DDIM steps
LDM-4 (ours)	10.56	103.49±1.24	0.71	0.62	400M	250 DDIM steps
LDM-4-G (ours)	**3.60**	**247.67**±**5.59**	**0.87**	0.48	400M	250 steps, c.f.g, $s = 1.5$

rmance on each of the tasks, as Table 1, 2 and 3 show.

Remarks

[Re. 1]

Basically, we can say that the LDM is the combination of VQGAN and DDPM.

[Re. 2]

The adversarial objective is used to enhance the quality of generated images.

[Re. 3]

Which means that by ensuring reconstructions remain on the “image manifold,” the model produces more realistic images. The term “local realism” refers to preserving image details that make them appear natural. Traditional pixel-based losses like L2 or L1 (which minimize pixel differences) can introduce blurriness because they don't always account for fine image structures. This approach avoids that by using a more sophisticated method to keep the generated images sharp and realistic.

[Re. 4]

Please refer to Meeting Nore 2024-08-30, Re. 2.

References

[1] Rombach, R., Blattmann, A., Lorenz, D., Esser, P., & Ommer, B. (2022). High-resolution image synthesis with latent diffusion models. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition* (pp. 10684-10695).

[2] Sohl-Dickstein, J., Weiss, E., Maheswaranathan, N., & Ganguli, S. (2015, June). Deep unsupervised learning using nonequilibrium thermodynamics. In *International conference on machine learning* (pp. 2256-2265). PMLR.

[3] Goodfellow, I., Pouget-Abadie, J., Mirza, M., Xu, B., Warde-Farley, D., Ozair, S., ... & Bengio, Y. (2014). Generative adversarial nets. *Advances in neural information processing systems*, *27*.

[4] Esser, P., Rombach, R., & Ommer, B. (2021). Taming transformers for high-resolution image synthesis. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition* (pp. 12873-12883).

[5] Zhang, R., Isola, P., Efros, A. A., Shechtman, E., & Wang, O. (2018). The unreasonable effectiveness of deep features as a perceptual metric. In *Proceedings of the IEEE conference on computer vision and pattern recognition* (pp. 586-595).

[6] Isola, P., Zhu, J. Y., Zhou, T., & Efros, A. A. (2017). Image-to-image translation with conditional adversarial networks. In Proceedings of the IEEE conference on computer vision

and pattern recognition (pp. 1125-1134).

[7] Esser, P., Rombach, R., Blattmann, A., & Ommer, B. (2021). Imagebart: Bidirectional context with multinomial diffusion for autoregressive image synthesis. *Advances in neural information processing systems*, *34*, 3518-3532.

[8] Ho, J., Jain, A., & Abbeel, P. (2020). Denoising diffusion probabilistic models. *Advances in neural information processing systems*, *33*, 6840-6851.

[9] Ronneberger, O., Fischer, P., & Brox, T. (2015). U-net: Convolutional networks for biomedical image segmentation. In *Medical image computing and computer-assisted intervention–MICCAI 2015: 18th international conference, Munich, Germany, October 5-9, 2015, proceedings, part III 18* (pp. 234-241). Springer International Publishing.

[10] Vaswani, A. (2017). Attention is all you need. *Advances in Neural Information Processing Systems*.

Comments

This week, we improved our meeting note by:

(1) using the font with 12-point size, page margin with 25.4 mm on all sides (top, bottom, left, and right).

(2) adding page numbers on the bottom center of each page.

(3) using [Re. 1] instead of [Re 1] to refer the remarks.

(4) setting the maximum width of images to 127.36 mm.

(5) allowing using dashes (-) to break words.