3LG-LDM: A 3D-layout guided Latent Diffusion Model

Pu Ching 110062577 | Hsiao-Wei Chen 110062573 | Yen-Pin Cheng 110065521 | Jie-Ying Li 110065502 | Pin-Xuan Liu 110062534

1. Introduction

Image synthesis is one of the computer vision fields with the most spectacular recent development. A learned latent code or sampled noise (e.g., Gaussian) can be utilized to generate a novel image by learning the data distribution. To address the problem, especially for the high-resolution synthesis of complex, natural scenes, likelihood-based methods [11, 12] are applied to model the training data distribution; GAN-based methods [2, 4, 8] implicitly learn the distribution via an extra discriminator. Among the approaches mentioned above, diffusion models (DMs) [15] achieve an impressive synthesis result. Through the denoising auto-encoder, a pure noise map can be gradually transform ed into a realistic image. Recently, incredible applications built upon DMs like image synthesis [3, 5, 6, 16], super-resolution [14], colorization [16], or stroke-based synthesis [10] outperform other type s of generative models. However, the diffusion process upon pixel space is computationally demanding and takes hundreds of GPU days. Aiming to reduce the calculation complexity while maintaining the high sampling quality, latent diffusion [13] is proposed to operate the diffusion process on the pretrained latent space. Except for the dimension reduction on the denoising auto-encoder, proposed cross-attention layers provide a flexible solution to fuse multi-modality conditions. An interesting application built upon this structure is that the image synthesis can be conditioned on a 2D-layout given labels. Nevertheless, based on our observations, a 2D-layout lacks information about the orientation and depth of the generated object images, e.g. the horse faces in different directions in Fig 1. In this work , we're going to propose a 3D-layout guided latent diffusion model making use of the 3D structures to synthesize 2D images. We believe such improvement may provide an interface for applications like a virtual tour or synthetic data generation.

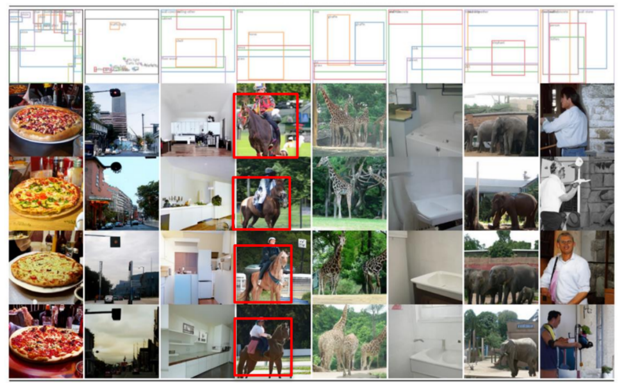


Figure 1. Layout-to-image synthesis results from LDM

2. Technical part

In practical terms, we'll separate the whole framework into 2 stages: 1) 3D object detection and 2) Latent diffusion.

For the initial experiments on 3D object detection, we're going to apply the ground truth 3D bounding box annotated from the Objectron dataset [1]. To explore the effects of noisy bounding boxes, we'll also apply the on-the-shelf detector [7] to estimate the scales, rotation, and translation of detected objects on the COCO dataset [9]. Given the 3D bounding box, we can obtain a projected 2D layout map (See Fig 2) that implies the 3D geometry of each object. Following the pre processing mentioned above, we can automatically prepare the paired data between the ground truth images and the layout maps. Note that we discovered that providing the whole 3D representation (e.g., voxel) may be costly on memory space, so we assume that using the projected map can provide sufficient information but is computationally efficient.

Conditioned on the processed data, we can project the layout map into an embedding vector using τ_θ and then pass it to the cross-attention layers implementing $Attention(Q, K, V) = softmax((\frac{QK^T}{\sqrt{d}}) \cdot V)$, with $Q = W_Q^{(i)} \cdot \rho_i(z_t)$, $K = W_K^{(i)} \cdot \tau_\theta(y)$, $V = W_V^{(i)} \cdot \tau_\theta(y)$. The detailed loss function is listed in equation 1, where ϵ is the sampled noise, ϵ_θ is our noise estimator, t is the timestep, z_t is the latent map at timestep t, and y is our layout map. To implement the complete LDM structure, here we choose LDM-4 as our pretrained image auto-encoder.

$$L_{LDM} := \mathbb{E}_{\varepsilon(x),y,\epsilon\sim N(0,1),t}[\|\epsilon - \epsilon_\theta(Z_t, t, \tau_\theta(y))\|_2^2] \quad (1)$$

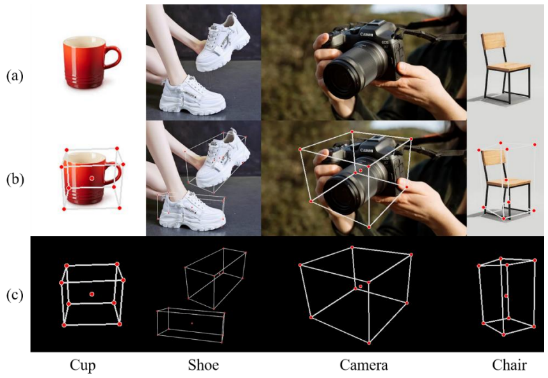


Figure 2. projected 3D layout map, (a) original image. (b) annotated image. (c) 3D layout map

3. Milestones

We already finished 2D and 3D layout data generation by mediapipe. The 3D layout map is shown in the image Fig 2 (c); the 2D layout map is generated with the top-left and the bottom right coordinates.

Furthermore, we also trained the diffusion model on Objectron dataset with 3D and 2D layout as input. In addition, we conducted the analysis of qualitative and quantitative results.

As for qualitative result, it shows that the image generated with 3D layout map can align the orientation with ground truth image (See Column (b) in Fig 4), and the image generated with 2D layout map is quite misaligned with the ground truth image.

As for quantitative result, we compared the 3D bounding box IOU between 3D layout guided image and 2D layout guided image (See Fig 5), and we expect that the IOU from 3D layout guided image is better than the other one. The results in Table 1 demonstrate that the IOU from 3D layout guided method is 8% better than the 2D layout guided method. In addition, we compare the successful object detection rate and FID of the image generated with 2D layout guided and 3D layout guided because we want to prove that the quality of 3D layout guided image is better than the ones from 2D. The results in Table 2 demonstrate that the images generated by the 3D guided layout can be more successfully detected than the images generated by the 2D guided layout. Moreover, the FID of the 3D guided layout method is lower than the 2D guided layout method. Therfore, we can claim that the image generated by 3D guided layout is more qualitative than the image generated by 2D guided layout.

4. Remaining milestones

Next, we will complete the last part, training a latent diffusion model using 2D and 3D layout data generated from the COCO dataset. Then we will evaluate the performance of both models using successful detection rate and FID metrics and compare the two dataset results. Furthermore, to test the generality of our model, we plan to randomly generate some 2D and 3D layouts as input and analyze the quantitative results. Through these experiments, it can be confirmed whether 3D-layout guided images can better rep-

Guided methods	Mean IOU
2D guided layout	28.54%
3D guided layout	36.53%

Table 1. Mean bounding box IOU of each guided method

Guided methods	Successful Detection Rate	FID ↓
2D guided layout	26.1%	36.03
3D guided layout	36.4%	22.95

Table 2. Successful detection rate and FID

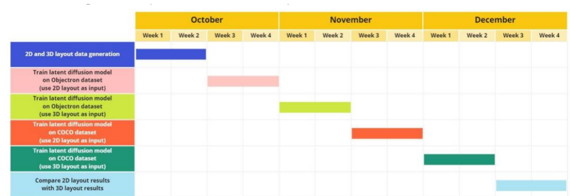
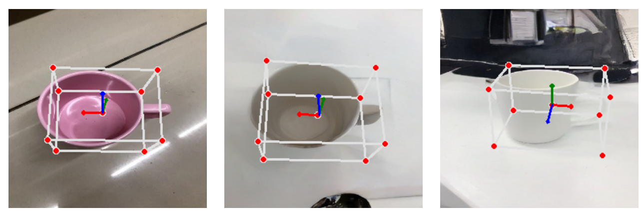
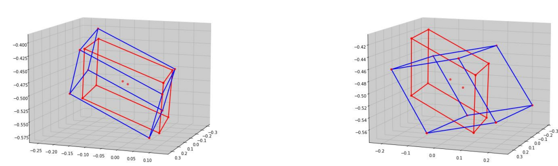


Figure 3. Milestone



(a) (b) (c)

Figure 4. Visualization of image generated with 3D/2D layout map, (a)ground truth image, (b)image generated with 3D layout map, (c)image generated with 2D layout map



(a) 3D bounding box interaction (b) 2D bounding box interaction

Figure 5. Visualization of bounding interaction

resent the orientation and depth of objects than 2D-layout guided images.

References

[1] Adel Ahmadyan, Liangkai Zhang, Artsiom Ablavatski, Jianing Wei, and Matthias Grundmann. Objectron: A large scale dataset of object-centric videos in the wild with pose annotations. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 7822–7831, 2021. 1

[2] Andrew Brock, Jeff Donahue, and Karen Simonyan. Large scale gan training for high fidelity natural image synthesis. *arXiv preprint arXiv:1809.11096*, 2018. 1

[3] Prafulla Dhariwal and Alexander Nichol. Diffusion models beat gans on image synthesis. *Advances in Neural Information Processing Systems*, 34:8780–8794, 2021. 1

[4] Ian Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. Generative adversarial networks. *Communications of the ACM*, 63(11):139–144, 2020. 1

[5] Jonathan Ho, Ajay Jain, and Pieter Abbeel. Denoising diffusion probabilistic models. *Advances in Neural Information Processing Systems*, 33:6840–6851, 2020. 1

[6] Jonathan Ho, Chitwan Saharia, William Chan, David J Fleet, Mohammad Norouzi, and Tim Salimans. Cascaded diffusion models for high fidelity image generation. *J. Mach. Learn. Res.*, 23:47–1, 2022. 1

[7] Tingbo Hou, Adel Ahmadyan, Liangkai Zhang, Jianing Wei, and Matthias Grundmann. Mobilepose: Real-time pose estimation for unseen objects with weak shape supervision. *arXiv preprint arXiv:2003.03522*, 2020. 1

[8] Tero Karras, Samuli Laine, and Timo Aila. A style-based generator architecture for generative adversarial networks. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 4401–4410, 2019. 1

[9] Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C Lawrence Zitnick. Microsoft coco: Common objects in context. In *European conference on computer vision*, pages 740–755. Springer, 2014. 1

[10] Chenlin Meng, Yang Song, Jiaming Song, Jiajun Wu, Jun-Yan Zhu, and Stefano Ermon. Sdedit: Image synthesis and editing with stochastic differential equations. *arXiv preprint arXiv:2108.01073*, 2021. 1

[11] Aditya Ramesh, Mikhail Pavlov, Gabriel Goh, Scott Gray, Chelsea Voss, Alec Radford, Mark Chen, and Ilya Sutskever. Zero-shot text-to-image generation. In *International Conference on Machine Learning*, pages 8821–8831. PMLR, 2021. 1

[12] Ali Razavi, Aaron Van den Oord, and Oriol Vinyals. Generating diverse high-fidelity images with vq-vae-2. *Advances in neural information processing systems*, 32, 2019. 1

[13] Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. High-resolution image synthesis with latent diffusion models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 10684–10695, 2022. 1

[14] Chitwan Saharia, Jonathan Ho, William Chan, Tim Salimans, David J Fleet, and Mohammad Norouzi. Image super-resolution via iterative refinement. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2022. 1

[15] Jascha Sohl-Dickstein, Eric Weiss, Niru Maheswaranathan, and Surya Ganguli. Deep unsupervised learning using nonequilibrium thermodynamics. In *International Conference on Machine Learning*, pages 2256–2265. PMLR, 2015. 1

[16] Yang Song, Jascha Sohl-Dickstein, Diederik P Kingma, Abhishek Kumar, Stefano Ermon, and Ben Poole. Score-based generative modeling through stochastic differential equations. *arXiv preprint arXiv:2011.13456*, 2020. 1