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

Abstract

Image synthesis is one of the computer vision fields with the most spectacular recent development. Recently, diffusion models (DMs) and Latent Diffusion Probabilistic Models (LDM) achieve an impressive synthesis result. An interesting application built upon this structure is that the image synthesis can be conditioned on a 2D-layout given labels. However, based on our observations, a 2D-layout lacks information about the orientation and depth of the generated object images. To address this problem, we're going to propose a 3D-layout guided latent diffusion model making use of the 3D structures to synthesize 2D images and we demonstrate that this proposed method achieves excellent results on Objectron dataset. We believe that this model has the potential to be a valuable tool in a variety of applications, and we look forward to exploring its potential further in future work. Code: https://reurl.cc/VRba6N

1. Introduction

In recent years, researchers have shown an increased interest in image synthesis. 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.

Previous research, especially for the high-resolution synthesis of complex, natural scenes, likelihood-based methods [15, 16] are applied to model the training data distribution; GAN-based methods [2, 6, 10] implicitly learn the distribution via an extra discriminator. Among the approaches mentioned above, diffusion models (DMs) [19] 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, 8, 9, 20], super-resolution [18], colorization [20], or stroke-based synthesis [14] outperform other type s of generative models.

Besides, 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 [17] 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.

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

An interesting application built upon this structure is that the image synthesis can be conditioned on a 2D-layout given labels. However, 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.

To address this problem, we're going to propose a 3D-layout guided latent diffusion model making use of the 3D structures to synthesize 2D images and we demonstrate that this proposed method achieves excellent results on Objectron dataset [1]. We hope that our proposed 3D-layout guided latent diffusion model will demonstrate improved performance in image synthesis tasks. We believe that this model has the potential to be a valuable tool in a variety of applications, and we look forward to exploring its potential further in future work.

2. Related Work

2.1. Review of previous work

Image Synthesis is one of the computer vision fields with the most spectacular recent development. Previously, **Gen-**

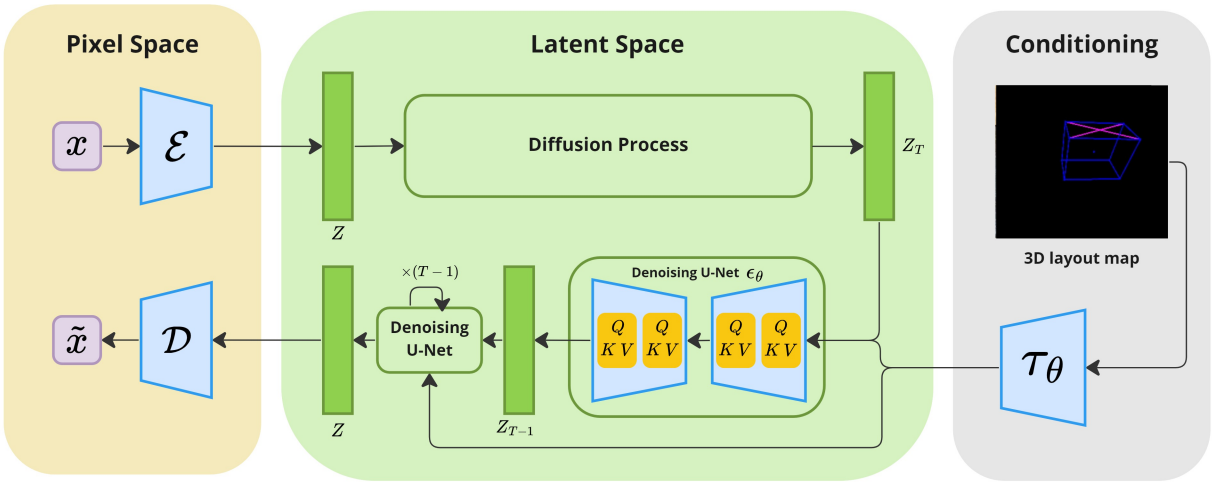


Figure 2. Overview of our framework.

erative Models are the most popular. There were two kinds of approaches: (1) Generative Adversarial Networks (GAN) based and (2) Likelihood-based. GAN [6] allows for efficient sampling of high-resolution images with good perceptual quality [2, 11], but are difficult to optimize [7] and struggle to capture the full data distribution. In contrast, likelihood-based methods emphasize good density estimation which renders optimization more well-behaved. Variational autoencoders(VAE) [13] and flow-based models [4,5] enable efficient synthesis of high-resolution images, but sample quality is not on par with GANs.

Recently, **Diffusion Probabilistic Models (DM)** [19] have achieved advanced results in density estimation [12] as well as in sample quality [3]. However, DM has the downside of low inference speed and very high training costs.

Fortunately, the above two drawbacks are addressed with **Latent Diffusion Probabilistic Models (LDM)** [17], which work on a compressed latent space of lower dimensionality. This renders training computationally cheaper and speeds up inference with almost no reduction in synthesis quality.

2.2. Contributions

As shown in Figure 1, current image synthesis task can generate images within the constraints of a given 2D layout. However, such input information can only ensure that the generated image can be in the "corresponding position" to generate the desired object, but there is no way to generate a specific orientation. To solve this limitation that model can generate a more controlled image. We propose using the 3D layout guided to generate images. In summary, our contributions are as follows:

- We are the first to use the 3D layout as a guide to generate images, which allows the generated images not only to be generated in the correct position but also to control the orientation of the generated objects.
- In experiments, we demonstrate that using the proposed 3D layout guide method performs better than the 2D layout guide on Objectron dataset [1].

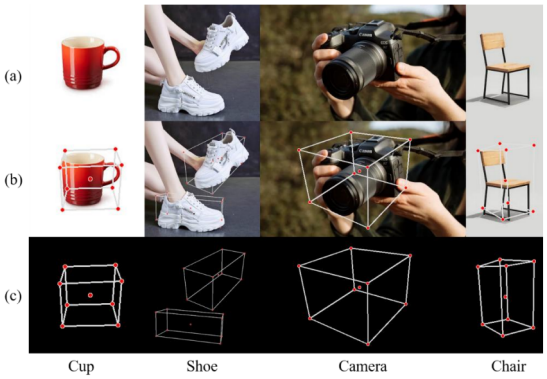


Figure 3. projected 3D layout map, (a) original image. (b) annotated image. (c) 3D layout map

3. Technical part

3.1. Summary of the technical solution

The framework of 3LG-LDM is shown as Fig 2. This model is based on previous work [17] and simply divide into three spaces: pixel space, latent space, and conditioning. First, the encoder compresses input into a low-dimensional feature vector. Then the latent diffusion model in latent space apply conditional denoising. By decoding the output feature vector, we can get the generated image.

Considering most of works use the 2D-layout guided method to condition the image synthesis, and this way will lose some information about the object. In order to avoid this problem, we modify the conditioning input from the 2D-layout map to the 3D-layout map to keep more information about object orientation in generated images.

3.2. Details of the technical solution

We apply the VQ-VAE model as backbone. Given the noise map x as input, the encoder $\mathcal{E}$ encodes x into a latent representation $Z = \mathcal{E}(x)$. To control the synthesis process through 3D layout map, we pass the latent representation Z into the conditional latent diffusion process. In this process, the encoder τ_θ compress the 3D layout map into an embedding vector to add condition on latent diffusion process. After passing through several denoising process, we can produce an output representation. Then it can be decoded to image $\tilde{x}$ through the decoder $\mathcal{D}$. In the following section, we introduce two parts:3D Layout Map Generation and Conditional Latent Diffusion Model, in detail.

3D Layout Map Generation. In practice we use the ground-truth 3D bounding box annotated from the Objectron dataset [1]. Given the 3D bounding box, we can obtain a projected 2D layout map (See Fig 3) 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.

Conditional Latent Diffusion Model. 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.

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

4. Experiments

4.1. Dataset

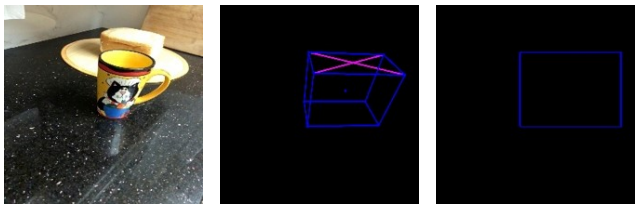
We adopt Objectron dataset [1] as our training dataset. The Objectron dataset consists of 4 categories: shoes, chairs, cameras, and cups. Each image in the dataset includes the 3D bounding box information. Thus, we can generate 3D and 2D layout map from the Objectron dataset like Fig 4. In this work, We choose cups to test our ideas. There are total 55676 images of the cup, and we sample 80% images for training, then the rest 20% images are for testing. In training and testing stage, we resize our input images to size 256×256.

4.2. Implementation Details

In our experiments, we adopt VQ-VAE pretrained on LAION-400M with scale factor $f = 4$ as backbone of the Latent Diffusion Model. To obtain a comparable test-field, we fix the computational resources to a single NVIDIA 3090 for all experiments in this section and train all models for the same number of steps and with the same number of parameters.

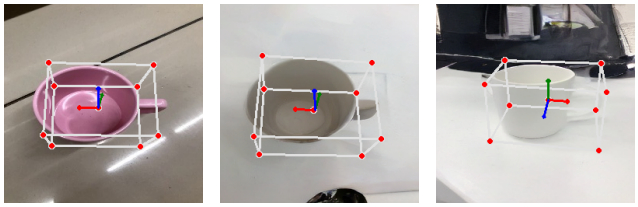
4.3. Qualitative Analysis

The purpose of this section is to show that the 3D layout guided LDM can generate more accurate orientation than the 2D layout guided LDM. Fig 5 presents that the image generated with 3D layout map (column b) is more aligned than image generated with 2D layout map (column c) with ground truth image orientation. Furthermore, in Fig 6, we can also discover that the bounding box interaction of 3D layout guided LDM is more overlapped than 2D layout guided LDM, which can prove that the 3D layout guided LDM can generate more accurate orientation than the 2D one.



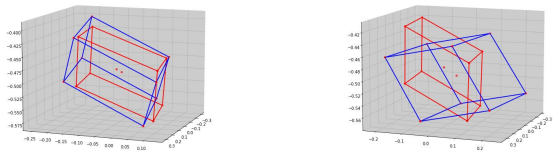
(a) Raw image (b) 3d layout map (c) 2D layout map

Figure 4. Objectron Dataset



(a) (b) (c)

Figure 5. 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 layout guided LDM (b) 2D layout guided LDM

Figure 6. Visualization of bounding interaction

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

Table 1. Mean bounding box IOU of each guided method

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

Table 2. Successful detection rate and FID

4.4. Quantitative Analysis

In this section, we calculate the IOU to measure the bounding box interaction. We expect that the 3D layout guided LDM can achieve better IOU than the 2D one because we think that the 3D layout map can guide the orientation of the generated images. The results in Table.1 demonstrate that the IOU from 3D layout guided method is 8% better than the 2D layout guided method, which is correspond to our thought. Moreover, we measure the successful object detection rate and FID of the image generated with 2D layout guided and 3D layout guided for evaluating the generated image quality. The results in Table.2 demonstrate that the images generated by the 3D layout guided can be more successfully detected than the images generated by the 2D layout guided. Moreover, the FID of the 3D layout guided method is lower than the 2D layout guided method. Therefore, we can claim that the image generated by 3D layout guided is more qualitative than the image generated by 2D layout guided.

4.5. Discussion

To test the limitation of the 3D layout guided latent diffusion model, we are curious about that what if we synthesize images with randomly generated 3D layout map? The experimental results in Fig. 7 reveal that the generated images do not follow the guided condition because the scale of boxes and the orientation is limited in the real-world data. In addition, the cups in our training data are almost in the center region of the pictures. Therefore, the 3D layout region around the edge is different with the training 3D layout map. We think that the augmentation may resolve the second problem.

5. Conclusions

In this work, we propose 3D layout map guided latent diffusion model to address the orientation problem of the images generated by 2D layout map guided latent diffusion model. The experimental results demonstrate that the 3D layout map guided latent diffusion model can generate images with more accurate direction. In addition, the experimental results also shows that the generated images guided from 3D layout map can be detected more successfully and have lower FID than the other one. In conclusion, we are the first one to use the 3D layout as a guide to generate images, and the results demonstrate that the the 3d layout map can assist the orientation and the quality of the generated images. However, there is still a limitation, e.g. if we random generate the 3D layout map, the generated images are not in corresponding area. We think that we can do some augmentation to fix this problem in future work.

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, 3

[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, 2

[3] Prafulla Dhariwal and Alexander Nichol. Diffusion models beat gans on image synthesis. *Advances in Neural Information Processing Systems*, 34:8780–8794, 2021. 1, 2

[4] Laurent Dinh, David Krueger, and Yoshua Bengio. Nice: Non-linear independent components estimation. *arXiv preprint arXiv:1410.8516*, 2014. 2

[5] Laurent Dinh, Jascha Sohl-Dickstein, and Samy Bengio. Density estimation using real nvp. *arXiv preprint arXiv:1605.08803*, 2016. 2

[6] 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, 2

[7] Ishaan Gulrajani, Faruk Ahmed, Martin Arjovsky, Vincent Dumoulin, and Aaron C Courville. Improved training of wasserstein gans. *Advances in neural information processing systems*, 30, 2017. 2

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

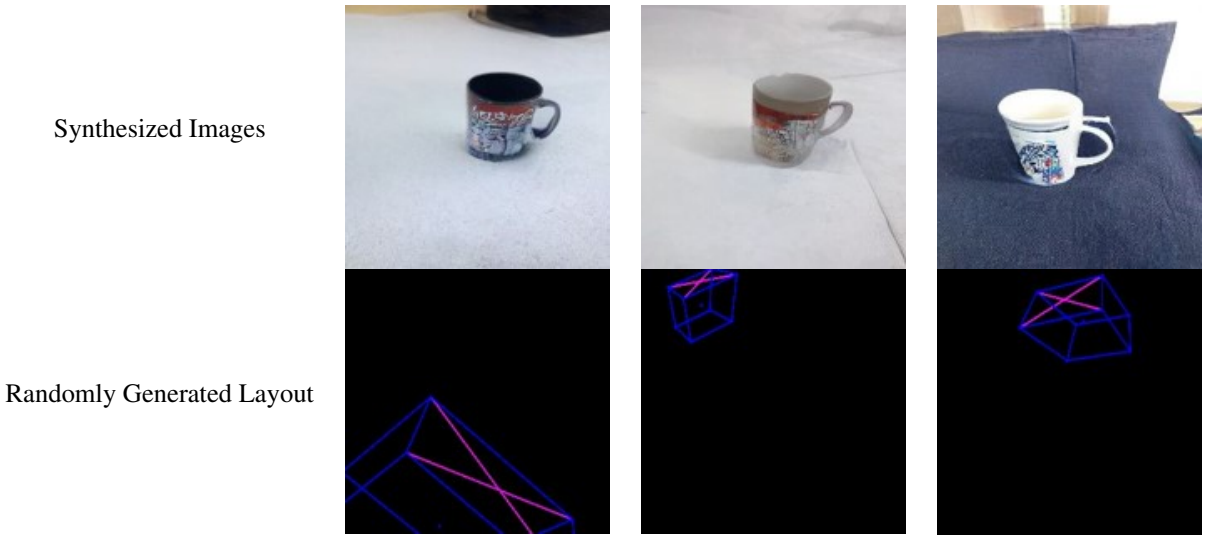


Figure 7. Visualization of synthesized images generated with random 3D layout

[9] 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

[10] 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

[11] Tero Karras, Samuli Laine, Miika Aittala, Janne Hellsten, Jaakko Lehtinen, and Timo Aila. Analyzing and improving the image quality of stylegan. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 8110–8119, 2020. 2

[12] Diederik Kingma, Tim Salimans, Ben Poole, and Jonathan Ho. Variational diffusion models. *Advances in neural information processing systems*, 34:21696–21707, 2021. 2

[13] Diederik P Kingma and Max Welling. Auto-encoding variational bayes. *arXiv preprint arXiv:1312.6114*, 2013. 2

[14] 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

[15] 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

[16] 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

[17] 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, 2

[18] 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

[19] 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, 2

[20] 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