LDM-SDF: 3D Shape Synthesis using Latent Diffusion Model and Signed Distance Function Representation

Subhan Kamilov
Technical University of Munich
subhan.kamilov@tum.de

Abstract

Diffusion Models have recently demonstrated breakthrough results in 2D image generation. Building upon their success in 2D, we introduce LDM-SDF for 3D shape synthesis using a diffusion model trained on the vector of latent representations of 3D shapes. Our model utilizes the existing DeepSDF network for 3D shape reconstruction by mapping points in space to their signed distance value, thus allowing for high fidelity and expressiveness in generated shapes. We show how our method outperforms previous works while leaving space for improvements.

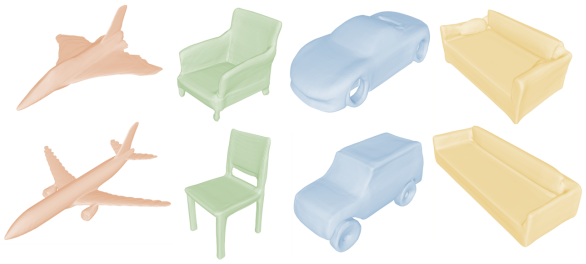

Figure 1. Novel 3D Shapes generated by LDM-SDF

1. Introduction

With the growing demand for 3D modeling across a wide range of industries, as it enables more immersive and interactive experiences, many deep learning methods have been proposed for the 3D shape reconstruction tasks [16, 18, 20]. Of particular interest, especially in rapidly developing industries like virtual and augmented reality, computer graphics, and animation, is the possibility of novel shapes generation from already existing ones. Thus, allowing to combine creativity with time efficiency in modeling the 3D world.

Diffusion Probabilistic Models have recently shown a state of the art results across many image-generation tasks [3, 7]. With the success of those diffusion models in the generation of high-quality and diverse 2D images, several works have been introduced attempting to utilize those models for the 3D shape generation [1, 4, 9].

While those works show promising results, most of them have efficiencies in the generation process and expressiveness of the created 3D shapes, which can be attributed to their use of point clouds for 3D shape representation. Other works [14, 22], on the other hand, are slow and memory consuming due to their high-dimensional feature representation of the models. Thus, restricting their usability in real-time applications.

In this work, we propose a new approach for the 3D shape synthesis based on the signed distance function representation of the 3D shapes and the latent diffusion model (LDM-SDF). Our method breaks down into a two-step process. First, we train an auto-decoder architecture in a supervised manner on the pairs composed of 3D point samples and their corresponding SDF values. In the second step, we apply a diffusion model on the latent shape encoding that we learned in the first step. This diffusion model can then be used at the inference time for the generation of novel and diverse 3D shapes.

Our contributions can be summarised as follows:

- We introduce a novel framework for the sythesis of 3D shapes based on the existing architecture for 3D reconstruction using implicit function representation, with a diffusion model trained on top of it for novel shape generation.
- We also demonstrate that our work can generate high-quality and diverse results and outperform current state-of-the-art results from other approaches without any hyperparameter tuning.
- Experiments show that our approach has minimum sampling and training time compared to existing works while also being less complex in the architecture.

2. Related Work

Signed Distance Function. An SDF is a continuous representation of the 3D shapes, which outputs a distance to the closest surface for a given point in the space. The sign of the signed distance function value defines whether the point is inside or outside of the surface. Point Cloud, on the other hand, is another representation for 3D shapes. Many works have been introduced on feature extraction from the point clouds and utilization of them for several downstream tasks such as segmentation, classification, and generation [12, 13]. However, they have limitations in describing continuous surfaces with complex topologies. Another representation of 3D scenes is an extension of 2D pixel representations to the voxels, which allows the extension of many well-known frameworks to 3D (e.g. convolutions). Nonetheless, compute and memory requirements that grow cubically limit them for high-resolution shape synthesis [19]. In contrast, SDF shows to be efficient and expressive continuous representations for 3D modeling [11].

Generative Synthesis. Before diffusion models, autoregressive models [5], score-matching models [15], and GANs [6] were considered the most popular generative models, with the latter showing state-of-the-art results. However, with the recent breakthrough of diffusion models in beating GANs in 2D generative modeling [3], many methods have been proposed for their application in 3D. In this work, we show how they can be utilized in 3D generation by training them on the latent shape representations from the DeepSDF.

3. LDM-SDF Framework

Our approach can be described as a two-step process, in the first step we train a DeepSDF backbone to obtain latent shape representations of the 3D models, in the second step, we train our diffusion model to learn the density of obtained latent codes. Figure 1. Shows a high-level depiction of the two stages of the framework.

3.1. DeepSDF Backbone

We build our model upon the existing framework for 3D shape reconstruction via SDF representations [11]. The main purpose of this backbone is to simultaneously learn the decoder architecture and latent shape codes corresponding to each of the shapes in the training dataset. This can then be used to output an SDF value for a given 3D spatial point conditioned on the latent shape representation.

3.2. Latent Diffusion Model

We utilize the Diffusion model on the shape encoding learned during the first step of training as described in [7]. Diffusion Probabilistic Models, similar to other generative models like VAE, GANs try to convert noise from some simple distribution to a data sample. This is done by gradually denoising noisy samples until reaching a plausible generation similar to the actual examples used during training.

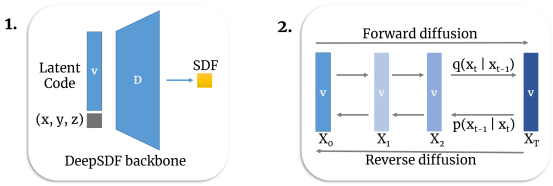

Figure 2. LDM-SDF Framework for 3D Shape Synthesis

In detail this process can be described in two stages:

1. A forward diffusion process q, which gradually adds gaussian noise to the original sample, until we end up with pure noise.

2. Learned reverse denoising diffusion process p_θ, where we train our neural network to gradually denoise the sample until we end up with the distribution representative of the original sample.

We train our model on the simplified version of the variational lower bound on negative log-likelihood as described from the original DDPM paper [7]:

$$L_{\text{simple}}(\theta) = \mathbb{E}_{t,\mathbf{x}_0,\boldsymbol{\epsilon}}\left[\left\|\boldsymbol{\epsilon} - \boldsymbol{\epsilon}_\theta(\sqrt{\bar{\alpha}_t}\mathbf{x}_0 + \sqrt{1-\bar{\alpha}_t}\boldsymbol{\epsilon}, t)\right\|^2\right] \quad (1)$$

Where we use L_1 loss function to compare the predicted noise $\boldsymbol{\epsilon}_\theta$ and actual applied noise $\boldsymbol{\epsilon}$ to the sample.

The architecture of the used denoising diffusion model for predicting the noise $\boldsymbol{\epsilon}_\theta$ is depicted in Fig. 3. It consists of 10 fully connected layers, which take 256-dimensional latent code as input, process it with a hidden size of 512, and output the added noise prediction.

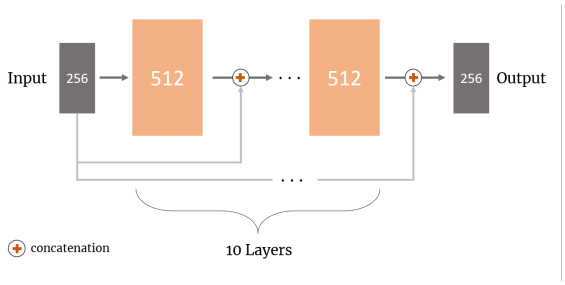

Figure 3. Used Denoising DIffusion Probabilistic Model

3.3. Sampling Novel 3D Shapes

At this stage, the unconditional generation of 3D shapes involves sampling a random Gaussian noise and gradually denoising it in T steps. Then, querying signed distance function values with our decoder network that we trained previously from the DeepSDF backbone. Following that, we can describe the shapes of 3D models as zero iso-surface decision boundaries learned by the MLP head. Finally, this implicit surface can be rendered through raycasting or rasterization of a mesh obtained with Marching Cubes.

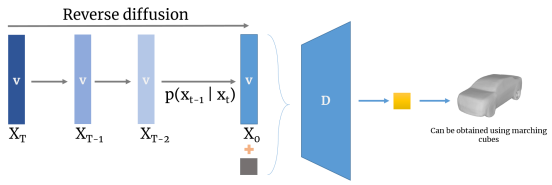

Figure 4. Process of unconditional sampling of the novel 3D shapes

4. Experiments

We trained our model on 4 object classes from the ShapeNet dataset [2] individually. The categories we trained on are Planes, Chairs, Sofas, and Cars. To compare our method to the previous works on 3D shape synthesis we refer to the latest obtained state-of-the-art results in [14] work and evaluate the generative capacity of our model on 3 object classes of cars, planes, and chairs respectively.

To do this, we first randomly sampled 500 objects from Chairs, 400 from Cars, and 400 from Planes object categories and trained them for 600 epochs each in DeepSDF backbone using default hyperparameters provided in the DeepSDF paper, except for decreasing the batch size from 32 to 16 in order to fit in our K80 GPU provided in Google Colab. Following this, we trained our diffusion model on learned latent shape encodings of each object class. We utilized the Adam optimization algorithm in the training, with a batch size of 32 and a learning rate of 10^{-4}. Denoising diffusion model was then trained for 1000 epochs for each of the shapes respectively.

4.1. Evaluation metrics

We use modified Fréchet Inception Distance (FID) as suggested in [21] to evaluate the quality of the generated samples. In the described method [21], shading images of each shape are rendered from 20 distinct views, which are then used for calculation of FID across each view and averaging it to obtain a final results:

$$\text{FID} = \frac{1}{20}\left[\sum_{i=1}^{20} \|\mu_g^i - \mu_r^i\|^2 + \text{Tr}\left(\Sigma_g^i + \Sigma_r^i - 2(\Sigma_r^i\Sigma_g^i)^{\frac{1}{2}}\right)\right], \tag{2}$$

where r and g stand for the real and generated examples, while μ^i , Σ^i define the mean and covariance matrices for the shaded images rendered from the i^{th} view.

Furthermore, metrics such as Precision and Recall were also evaluated as described in [8] for measuring the fidelity and diversity of the generated shapes.

4.2. Quantitative Results

We evaluate our model against current state-of-the-art methods for 3D shapes generation (see Table 1.). It can be seen that our method outperforms previous methods in all metrics on Cars class. Additionally, we achieve best results in Precision for the Planes category and relatively close results to the NFD [14] in FID scores for Planes and Chairs.

Data	Method	FID ↓	Precision ↑	Recall ↑
Cars	PVD*	335.8	0.1	0.2
	SDF-StyleGAN	98.0	35.9	36.2
	NFD	83.6	49.5	50.5
	Ours	**38.78**	**91.0**	**53.5**
Chairs	PVD*	305.8	0.2	1.7
	SDF-StyleGAN	36.5	90.9	87.4
	NFD	**26.4**	**92.4**	**94.8**
	Ours	43.65	87.6	52.1
Planes	PVD*	244.4	2.7	3.8
	SDF-StyleGAN	65.8	64.5	72.8
	NFD	**32.4**	70.5	**81.1**
	Ours	42.13	**78.2**	57.9

Table 1. Quality metrics on ShapeNet datasets across the baselines. We outperform previous methods in all evaluation metrics for the Cars category and achieve best precision in the Planes category. Metrics are calculated on shaded rendering images of shapes.

4.3. Qualitative Results

We also compare 3D shape generations from our model against baseline methods [14, 21, 22] in Figure 5. Overall, shapes generated by our model show higher fidelity compared to the other methods, which is also reflected in the precision metric. We can also note high quality of the generated shapes from our model. It can be seen that our model fails to represent the discs of the cars. This can be attributed to the lack of full training of DeepSDF backbone and complete optimization of the learned latent codes due to our time and compute limitations in this work.

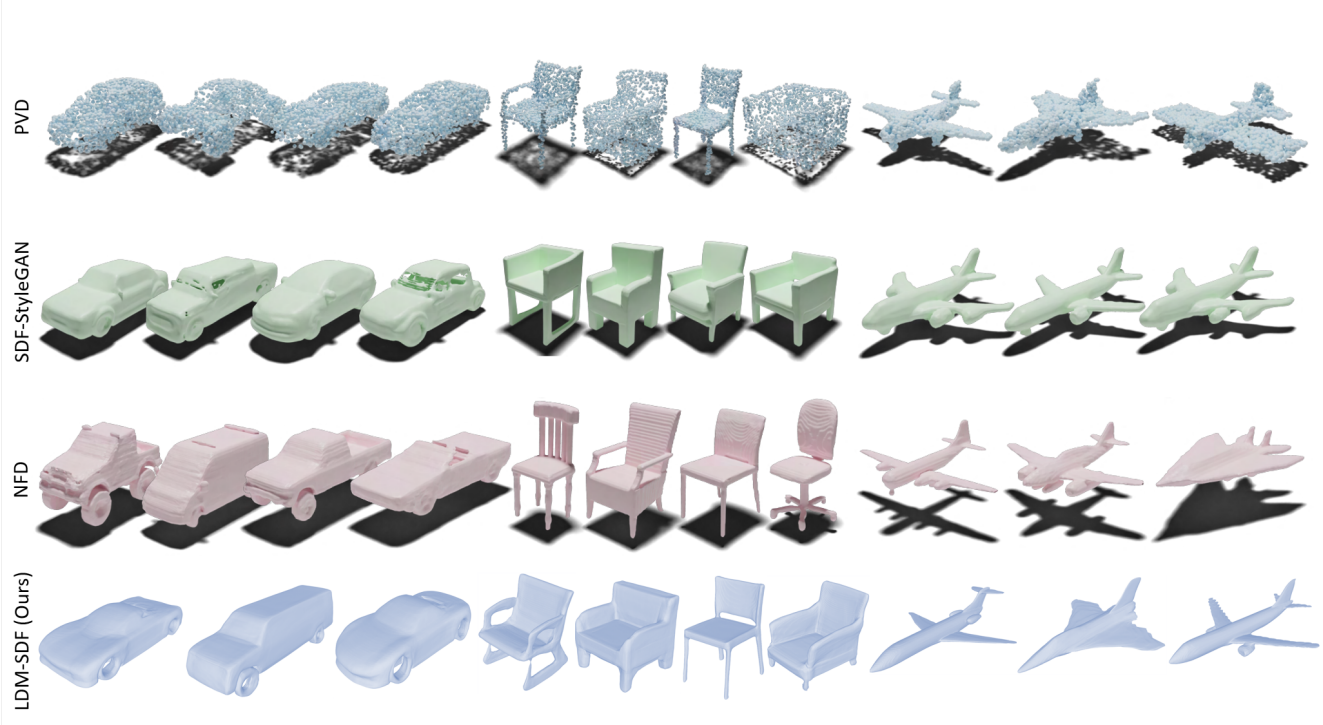

Figure 5. Comparison of the generated 3D shapes of our model and the state-of-the-art baselines on Cars, Chairs, and Planes ShapeNet object classes

5. Conclusion & Future Work

In this work, we have presented a novel framework for the synthesis of 3D shapes utilizing diffisuion model on the latent shape encodings obtained from the DeepSDF backbone. Our method outperformed previous state-of-the-art results in all metrics in Cars category and showed the highest precision in Planes category, while demonstrating close to the best results in FID and Precision metrics for other classes. Low dimension of the used shape encoding (256 in our case) in DeepSDF architecture also results in faster training and sampling time of the diffusion model, whereas previous state-of-the-art method [14] that models 3D shapes via triplane features requires many times higher memory and compute resources to train diffusion model on its $128 \times 128 \times 32 \times 3$ dimensional triplane representations.

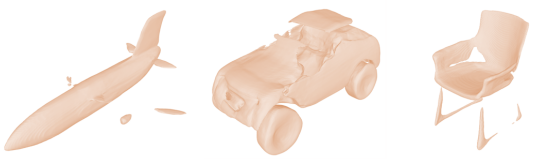

Figure 6. Failure cases

Despite the demonstrated high-quality results, our model also had failure cases in the generation task (see Fig. 6). We think that those failed cases can be attributed for partial training of the DeepSDF backbone due to time and compute constraints. Which results in limitations in the capacity of decoder architecture as well as learned latent shape representations.

Since we have not fully utilized the hyperparameter space of our model and present it with a base setting, we believe that our method can be further improved with hyperparameter tuning. Other than that, some novel approaches can be used to improve our DeepSDF backbone, for example, Fourier feature mapping [17] can be added for better reconstruction of high-frequency details in 3D models. As a result, this will also improve the quality of the generated shapes from the diffusion model and make them close to the real ones. Finally, recent improvements on the diffusion models [10] can be employed for faster sampling and higher quality generations.

6. Acknowledgements

We would like to thank Dr. Yinyu Nie for supervising us in this project and for fruitful discussions and insights.

References

[1] Miguel Angel Bautista, Pengsheng Guo, Samira Abnar, Walter Talbott, Alexander Toshev, Zhuoyuan Chen, Laurent Dinh, Shuangfei Zhai, Hanlin Goh, Daniel Ulbricht, Afshin Dehghan, and Josh Susskind. Gaudi: A neural architect for immersive 3d scene generation, 2022. 1

[2] Angel X. Chang, Thomas Funkhouser, Leonidas Guibas, Pat Hanrahan, Qixing Huang, Zimo Li, Silvio Savarese, Manolis Savva, Shuran Song, Hao Su, Jianxiong Xiao, Li Yi, and Fisher Yu. Shapenet: An information-rich 3d model repository, 2015. cite arxiv:1512.03012. 3

[3] Prafulla Dhariwal and Alex Nichol. Diffusion models beat gans on image synthesis. *CoRR*, abs/2105.05233, 2021. 1, 2

[4] Emilien Dupont, Hyunjik Kim, S. M. Ali Eslami, Danilo J. Rezende, and Dan Rosenbaum. From data to functa: Your data point is a function and you should treat it like one. *CoRR*, abs/2201.12204, 2022. 1

[5] Patrick Esser, Robin Rombach, and Björn Ommer. Taming transformers for high-resolution image synthesis. *CoRR*, abs/2012.09841, 2020. 2

[6] Ian J. Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. Generative Adversarial Networks. June 2014. 2

[7] Jonathan Ho, Ajay Jain, and Pieter Abbeel. Denoising diffusion probabilistic models. *CoRR*, abs/2006.11239, 2020. 1, 2

[8] Tuomas Kynkäänniemi, Tero Karras, Samuli Laine, Jaakko Lehtinen, and Timo Aila. Improved precision and recall metric for assessing generative models, 2019. 3

[9] Shitong Luo and Wei Hu. Diffusion probabilistic models for 3d point cloud generation. *CoRR*, abs/2103.01458, 2021. 1

[10] Alex Nichol and Prafulla Dhariwal. Improved denoising diffusion probabilistic models. *CoRR*, abs/2102.09672, 2021. 4

[11] Jeong Joon Park, Peter Florence, Julian Straub, Richard A. Newcombe, and Steven Lovegrove. Deepsdf: Learning continuous signed distance functions for shape representation. *CoRR*, abs/1901.05103, 2019. 2

[12] Charles Ruizhongtai Qi, Hao Su, Kaichun Mo, and Leonidas J. Guibas. Pointnet: Deep learning on point sets for 3d classification and segmentation. *CoRR*, abs/1612.00593, 2016. 2

[13] Charles Ruizhongtai Qi, Li Yi, Hao Su, and Leonidas J. Guibas. Pointnet++: Deep hierarchical feature learning on point sets in a metric space. *CoRR*, abs/1706.02413, 2017. 2

[14] J. Ryan Shue, Eric Ryan Chan, Ryan Po, Zachary Ankner, Jiajun Wu, and Gordon Wetzstein. 3d neural field generation using triplane diffusion, 2022. 1, 3, 4

[15] Jascha Sohl-Dickstein, Eric A. Weiss, Niru Maheswaranathan, and Surya Ganguli. Deep unsupervised learning using nonequilibrium thermodynamics. *CoRR*, abs/1503.03585, 2015. 2

[16] Xingyuan Sun, Jiajun Wu, Xiuming Zhang, Zhoutong Zhang, Chengkai Zhang, Tianfan Xue, Joshua B. Tenenbaum, and William T. Freeman. Pix3d: Dataset and methods for single-image 3d shape modeling. *CoRR*, abs/1804.04610, 2018. 1

[17] Matthew Tancik, Pratul P. Srinivasan, Ben Mildenhall, Sara Fridovich-Keil, Nithin Raghavan, Utkarsh Singhal, Ravi Ramamoorthi, Jonathan T. Barron, and Ren Ng. Fourier features let networks learn high frequency functions in low dimensional domains. *CoRR*, abs/2006.10739, 2020. 4

[18] Yue Wang, Yongbin Sun, Ziwei Liu, Sanjay E. Sarma, Michael M. Bronstein, and Justin M. Solomon. Dynamic graph CNN for learning on point clouds. *CoRR*, abs/1801.07829, 2018. 1

[19] Zhirong Wu, Shuran Song, Aditya Khosla, Xiaoou Tang, and Jianxiong Xiao. 3d shapenets for 2.5d object recognition and next-best-view prediction. *CoRR*, abs/1406.5670, 2014. 2

[20] Li Yi, Lin Shao, Manolis Savva, Haibin Huang, Yang Zhou, Qirui Wang, Benjamin Graham, Martin Engelcke, Roman Klokov, Victor S. Lempitsky, Yuan Gan, Pengyu Wang, Kun Liu, Fenggen Yu, Panpan Shui, Bingyang Hu, Yan Zhang, Yangyan Li, Rui Bu, Mingchao Sun, Wei Wu, Minki Jeong, Jaehoon Choi, Changick Kim, Angom Geetchandra, Narasimha Murthy, Bhargava Ramu, Bharadwaj Manda, M. Ramanathan, Gautam Kumar, P. Preetham, Siddharth Srivastava, Swati Bhugra, Brejesh Lall, Christian Häne, Shubham Tulsiani, Jitendra Malik, Jared Lafer, Ramsey Jones, Siyuan Li, Jie Lu, Shi Jin, Jingyi Yu, Qixing Huang, Evangelos Kalogerakis, Silvio Savarese, Pat Hanrahan, Thomas A. Funkhouser, Hao Su, and Leonidas J. Guibas. Large-scale 3d shape reconstruction and segmentation from shapenet core55. *CoRR*, abs/1710.06104, 2017. 1

[21] Xin-Yang Zheng, Yang Liu, Peng-Shuai Wang, and Xin Tong. Sdf-stylegan: Implicit sdf-based stylegan for 3d shape generation, 2022. 3

[22] Linqi Zhou, Yilun Du, and Jiajun Wu. 3d shape generation and completion through point-voxel diffusion. *CoRR*, abs/2104.03670, 2021. 1, 3