Diffusion-SDF: Text-to-Shape via Voxelized Diffusion

Muheng Li$^1$, Yueqi Duan$^1$$^2$, Jie Zhou$^1$, Jiwen Lu$^1$

$^1$Department of Automation, Tsinghua University

$^2$Department of Electronic Engineering, Tsinghua University

Abstract

With the rising industrial attention to 3D virtual modeling technology, generating novel 3D content based on specified conditions (e.g., text) has become a hot issue. In this paper, we propose a new generative 3D modeling framework called Diffusion-SDF for the challenging task of text-to-shape synthesis. Previous approaches lack flexibility in both 3D data representation and shape generation, thereby failing to generate highly diversified 3D shapes conforming to the given text descriptions. To address this, we propose a SDF autoencoder together with the Voxelized Diffusion model to learn and generate representations for voxelized signed distance fields (SDFs) of 3D shapes. Specifically, we design a novel UinU-Net architecture that implants a local-focused inner network inside the standard U-Net architecture, which enables better reconstruction of patch-independent SDF representations. We extend our approach to further text-to-shape tasks including text-conditioned shape completion and manipulation. Experimental results show that Diffusion-SDF generates both higher quality and more diversified 3D shapes that conform well to given text descriptions when compared to previous approaches. Code is available at: https://github.com/ttlmh/Diffusion-SDF.

1. Introduction

Exploring data representations for 3D shapes has been a fundamental and critical issue in 3D computer vision. Explicit 3D representations including point clouds [38, 39], polygon meshes [17, 24] and occupancy voxel grids [9, 53] have been widely applied in various 3D downstream applications [1, 35, 56]. While explicit 3D representations achieve encouraging performance, there are some primary limitations including not being suitable for generating watertight surfaces (e.g., point clouds), or being subject to topological constraints (e.g., meshes). On the other hand, implicit 3D representations have been widely studied more recently [3, 14, 37], with representative works including DeepSDF [37], Occupancy Network [32] and IM-Net [8]. In general, implicit functions encode the shapes by the isosurface of the function, which is a continuous field and can be evaluated at arbitrary resolution.

In recent years, numerous explorations have been conducted for implicit 3D generative models, which show promising performance on several downstream applications such as single/multi-view 3D reconstruction [23, 54] and shape completion [12, 34]. Besides, several studies have also explored the feasibility of directly generating novel 3D shapes based on implicit representations [15, 21]. However, these approaches are incapable of generating specified 3D shapes that match a given condition, e.g., a short text describing the shape characteristics as shown in Figure 1.
Strictly speaking, our approach employs a combined explicit-implicit representation in the form of voxelized signed distance fields.
3. Method

In this section, we introduce the methodology design for Diffusion-SDF. In general, we propose a two-stage pipeline as illustrated in Figure 2. Detailed information of each stage will be included in the following sections.

3.1. Autoencoding Signed Distance Fields

Signed distance fields (SDF) belong to a type of implicit 3D data representation which assigns the scalar signed distance value to the shape surface for each point \( p \) in the 3D space \( \mathbb{R}^3 \). Our objective is to generate the signed distance fields for the target 3D shapes that match the given text conditions. Specifically, the 3D data are represented by truncated signed distance functions (TSDF) in a regularly-spaced voxel grid as \([3, 22]\). Generating voxelized SDFs directly from denoising diffusion models is both cost-expensive and time-consuming. In addition, the local structural information of 3D shapes is not well-emphasized through direct voxel-based generation. To address this, we propose a patch-wise autoencoder to learn the latent representations for voxelized signed distance fields.

Given a 3D shape, we first sample its truncated signed distance field \( x \) as a voxel grid of size \( D^3 \). Before the shape \( x \) is encoded, it is first split into \( N \) local patches of size \( P^3 \), producing a sequence of shape patches \( X_p = [x_{p1}, \ldots, x_{pN}] \), where \( N = (D/P)^3 \) is the resulting number of patches. Then, the local shape encoder \( \mathcal{E}_{loc} \) encodes each shape patches into latent representations \( Z_p = [z_{p1}, \ldots, z_{pN}] \), where \( z_{pn} = \mathcal{E}_{loc}(x_{pn}) \in \mathbb{R}^c \), and \( c \) is the number of latent channels. Here, since each patch is encoded independently, the local structural information can be explored well through the local shape encoder. Meanwhile, the input data scale can also be downsampled by the factor \( P \). However, when reconstructing the shape from local patches, it is also necessary to consider the spatial location of each patch in the global shape, and the interrelationship between adjacent patches. To preserve both patch-to-global and patch-to-patch information while decoding the latent embeddings, the patch embeddings are first rearranged into a voxel grid embedding \( z \), and then sent into the patch-joint decoder \( D \) that reconstructs the SDF field from the latent patches, giving \( \tilde{x} = D(z) \).

In detail, we adopt a VAE \([26, 44]\)-like autoencoder, which encodes each shape patch into a normal distribution. We utilize a combination of both \( L_1 \) reconstruction loss together with the KL-regularization loss at the training stage for the SDF autoencoder. The latter one forces a mild KL-penalty towards a standard normal distribution on each learned patch latent. The final SDF representations will be in the form of patch-independent Gaussian distributions.

3.2. Voxelized Diffusion Models (VDMs)

Diffusion models (DMs) \([19, 49]\) are a class of probabilistic generative models that learn to fit a certain distribution by gradually denoising a Gaussian variable through a fixed Markov Chain of length \( T \). Given a data sample \( x_0 \sim q(x_0) \), DMs describe two different processes in the opposite direction: a forward process \( q(x_{0:T}) \) that gradually transform a data sample into pure Gaussian noise, and a reverse process \( p_\theta (x_{0:T}) \) that gradually denoise a pure
Gaussian sample into real data,

\[
q(x_0:T) = q(x_0) \prod_{t=1}^{T} q(x_t | x_{t-1}), \\
p_\theta(x_0:T) = p(x_T) \prod_{t=1}^{T} p_\theta(x_{t-1} | x_t),
\]

where \(q(x_t | x_{t-1})\) and \(p_\theta(x_{t-1} | x_t)\) are both Gaussian transition probabilities in the forms of

\[
q(x_t | x_{t-1}) = \mathcal{N}(x_t; \sqrt{1 - \beta_t} x_{t-1}, \beta_t I), \\
p_\theta(x_{t-1} | x_t) = \mathcal{N}(x_{t-1}; \mu_\theta(x_t, t), \beta_I).
\]

Based on [19], the mean variable \(\mu_\theta(x_t, t)\) for the reverse transition \(p_\theta(x_{t-1} | x_t)\) can be expressed in the form of

\[
\mu_\theta(x_t, t) = \frac{1}{\sqrt{\alpha_t}} \left( x_t - \frac{\beta_t}{\sqrt{1 - \alpha_t}} \epsilon(x_t, t) \right),
\]

where \(\alpha_t = 1 - \beta_t\), \(\alpha_t = \Pi_{i=1}^{t} \alpha_i\), and \(\beta_t\) decreases to 0 gradually as \(t\) approaches 0. The evidence lower bound (ELBO) is maximized at the training stage of DMs, eventually leading to the loss function as the form of

\[
\mathcal{L}_{DM} = \mathbb{E}_{x,t,\epsilon \sim \mathcal{N}(0,1)} \left[ \| \epsilon - \epsilon_\theta(x_t, t) \|^2 \right],
\]

where \(x_t = \sqrt{\alpha_t} x_0 + \sqrt{1 - \alpha_t} \epsilon\), \(\epsilon\) is a noise variable, and \(t\) is uniformly sampled from \(\{1, \ldots, T\}\). The neural network-based score estimator \(\epsilon_\theta(x_t, t)\) is the core function of denoising diffusion models, which is implemented as a timestep-conditioned denoising autoencoder in [19].

**UnU-Net Architecture.** In our scenario, we are designing a novel diffusion-based architecture to generate latent SDF representations with regard to the target text conditions. The corresponding TSDF fields can be then reconstructed through the patch-joint decoder which is pre-trained in the first stage. As we mentioned in the above section, the key part of generative diffusion models is to estimate the function approximator \(\epsilon_\theta\) from the data distribution of the training set. Since our model is designed to generate latent samples, the score estimator can be expressed as \(\epsilon_\theta(z_t, t)\), as the loss function becomes

\[
\mathcal{L}_{Diffusion-SDF} = \mathbb{E}_{z,t,\epsilon \sim \mathcal{N}(0,1)} \left[ \| \epsilon - \epsilon_\theta(z_t, t) \|^2 \right].
\]

To incorporate the 3D positional prior of different latent shape patches, we adopt a 3D U-Net [10]-based autoencoder as the neural function approximator \(\epsilon_\theta\) that directly learns to denoise the voxel grid embedding \(z_t\) that is obtained from the pre-trained local shape encoder \(\mathcal{E}_{loc}\). Besides, another crucial prior for our design is the patch-based encoding process, through which all the shape patches are encoded into independent Gaussian distributions. Thus, when generating the local shape embeddings, an important point is to recover the independent distribution for each shape patch as well. To address this, we propose a novel UnU-Net architecture for the implementation of the autoencoder-based neural score estimator as shown in Figure 3. The standard 3D U-net [10] adopts a series of \(3 \times 3 \times 3\) convolutional layers to construct a downsampling- Upsampling network architecture with the information shared through skip connections. This design takes into account the patch-to-patch and patch-to-global features via hierarchical receptive fields. Given that our input data are compressed 3D latent representations, and each patch is encoded independently, we propose to implant another inner network inside the outer U-Net architecture where the embedded resolution is equal to the original latent resolution \(D/P\). Specifically, we adopt a \(1 \times 1 \times 1\) convolution-based ResNet [18] structure to learn independent patch-focused information. We also introduce the spatial Transformer [51] network that accepts positional embedded patch representations as input tokens, following self-attention layers to capture the relational information between independent local patch embeddings. The inner synthesis paths are skip-connected to the outer ones, ensuring information transmission from inner to outer networks. This architecture is designed to capture the intra-patch information along with both patch-to-patch and patch-to-global relations.

**Text-Guided Shape Generation** So far, we have discussed the generative process without text conditions. To synthe-
size latent SDF representations based on given conditions, we introduce a conditioning mechanism based on classifier-free guidance diffusion [20]. We adopt a text-conditioned score estimator. In detail, to fit the target conditions into the 3D denoising autoencoder, we utilize the cross-attention mechanism as proposed in [45]. The input caption is first encoded as text embeddings through a pre-trained text encoder \( \tau \). Thus, the conditioned score can be formed as \( \epsilon_\theta (z_t|\tau(c)) \), where \( c \) is the input caption. Meanwhile, we also obtain an unconditioned score \( \epsilon_\theta (z_t|\emptyset) \) with an empty sequence as input caption. Within the guided diffusion sampling process, the output of the model is extrapolated further in the direction of \( \epsilon_\theta (z_t|\tau(c)) \) and away from \( \epsilon_\theta (z_t|\emptyset) \) as:

\[
\hat{\epsilon}_\theta (z_t|\tau(c)) = \epsilon_\theta (z_t|\emptyset) + s \cdot (\epsilon_\theta (z_t|\tau(c)) - \epsilon_\theta (z_t|\emptyset)),
\]

where \( s \geq 1 \) refers to the guidance scale.

**Text-Guided Shape Completion** Current shape completion approaches mainly focused on recovering the full shapes from partial input shapes or single view images [1, 57, 59]. Besides, we have explored the feasibility of text-conditioned shape completion. According to the patch-based design of our SDF encoder, information for the given shape patches can be encoded even if some of the patches are missing. Motivated by the success of diffusion model-based design of our SDF encoder, information into each denoising step of the reverse process. For a full shape representation, we transform it into a partial shape representation given the mask \( m_z \). For each estimation step \( z_{t-1} \), we combine the estimated result through the reverse process \( \hat{z}_{t-1} \) with the forward sampling result of the unmasked patches \( \hat{z}_{t-1} \),

\[
z_{t-1} = (1 - m) \odot \hat{z}_{t-1} + m \odot \hat{z}_{t-1},
\]

where \( \odot \) refers to element-wise product, and

\[
\hat{z}_{t-1} \sim \mathcal{N} (\mu_\theta (z_t, t, c), \beta_t I)
\]

\[
\hat{z}_{t-1} = \sqrt{\alpha_t} z_0 + \sqrt{1 - \alpha_t} \epsilon.
\]

By such a strategy, the unknown region can be recovered based on both the given shape and text conditions.

**Text-Guided Shape Manipulation** Visual content editing is one of the most valuable applications for vision-language generation. We propose a diffusion-based text-guided shape manipulation approach. For a given shape representation \( z_{init} \), our goal is to transform it into another shape representation \( z_{goal} \) based on text \( c \). Inspired by image manipulation techniques [4, 25], we use a cycle-sampling strategy based on a pre-trained VDM. Firstly, we forward-sample the given shape \( z_{init} \) for \( t_{mid} \) steps, where \( 0 < t_{mid} < T \). This operation will lead to an intermediate output \( z_{mid} \)

\[
z_{mid} = \sqrt{\alpha_{t_{mid}}} z_{init} + \sqrt{1 - \alpha_{t_{mid}}} \epsilon.
\]

Then, the reverse process will start from \( z_{mid} \), and another \( t_{mid} \) denoising steps will be conducted conditioned on \( c \). Such a design will make the generated shapes correspond to the target text descriptions while maintaining the original shape characteristics.

### 4. Experiments

We extensively experimented with our proposed **Diffusion-SDF** on various text-conditioned 3D shape synthesis tasks, such as text-to-shape generation, text-conditioned shape completion, and text-guided shape manipulation. The following subsections describe details of experimental settings, results, and analyses.

**Dataset.** We mainly trained and evaluated our approach on the current largest text-shape dataset Text2Shape (T2S) [6]. T2S gathered data in the form of shape-text pairs based on two object classes (chairs, tables) in ShapeNet [5]. T2S contains about 75K shape-text pairs in total (~30K for chairs, ~40K for tables), with an average of ~16 words per description. T2S was originally designed to contain both color and shape information within the text descriptions, while we mainly focus on text-shape correspondence.

**Implementation details.** For the training stage of the SDF autoencoder, to improve the model’s generalization ability, we trained the autoencoder across the 13 categories of ShapeNet [5] dataset. For the input data, we sampled the voxelized SDF from original shapes with \( 64 \times 64 \times 64 \) grid points, which is also the data size adopted for follow-up operations. For the diffusion stage, to facilitate the generative sampling speed, we adopt the DDIM [50] sampler to reduce the original DDPM sampling steps from \( T = 1000 \) to 50. For the conditioning mechanism, we adopt a freeze CLIP [40] text embedder as text encoder \( \tau \).

#### 4.1. Text-Conditioned Shape Generation

The most immediate application of our approach involves generating novel shapes that are conditioned on textual descriptions. To showcase the efficacy of our approach, we conduct a comparative evaluation of our generation results against two state-of-the-art supervised text-to-shape synthesis methods that are based on implicit 3D representations [29, 34]. Besides, there are some other works regarding text-to-shape generation, which have different settings compared to our approach. For instance, [48] proposed a zero-shot shape generation approach conditioned on categorical texts, and [15, 33] proposed to generate textured shapes based on provided/pre-generated meshes. As a result, these prior works are not directly compared against our proposed approach.

**Quantitative comparison.** To compare the generation performance of our approach to the previous methods quantitatively, we propose to use several metrics to evaluate the generated results:
The chair is retro and round with little to no arm rest. A sofa with thick legs. A solid locking chair made of grey tiles. Fabric chair with single-piece gray metal legs. An old fashioned rocking chair with very high back.


Figure 4. Qualitative results of text-to-shape generation. Our approach can generate shapes that conform to diverse text descriptions.

- **IoU** (Intersection over Union) measures the occupancy similarity between the generated shape to the ground truth. To compute the IoU score, we downsample all the generated SDFs to occupancy voxel grids of size $32^3$. IoU metric is used to measure the conformity of the generated samples with the ground truth shapes.

- **Acc** (Classification Accuracy) introduces a voxel-based classifier, which is pre-trained to classify the 13 categories in ShapeNet [5] according to [48]. Acc metric is used to measure the semantic authenticity of the generated samples.

- **CLIP-S** (CLIP Similarity Score) introduces the pre-trained vision-language model CLIP [40] for further evaluation. The CLIP can be used to measure the correspondence between given images and texts. We render 5 view images for each generated shape, and compute the cosine similarity score between rendered images and the given text. The highest score is reported for each text query. CLIP-S metric is used to measure the semantic conformance between the generated samples and input texts.

- **TMD** (Total Mutual Difference). For each given text description, we generate $k = 10$ different samples. Then, we compute the average IoU score for each generated shape to other $k - 1$ shapes. The average IoU score for all text queries is reported. TMD assesses the generation diversity for each given text query.

To keep consistency with previous approaches, we limit the comparisons to the chair category in Text2Shape [6] dataset. Since the officially released AutoSDF [34] model was trained on another dataset [2], we re-trained the model following the original settings on the training set of Text2Shape for the fair comparison. The results are shown in Table 1. From the results, it can be seen that our approach is able to achieve relatively high generation quality with high IoU and Acc scores, and robust text-shape conformance with good CLIP-S performance, while achieving much better generation diversity with a much lower TMD performance.

**Qualitative results.** The quantitative experiment shows the generation performance of our approach on the limited given subset from Text2Shape dataset. Due to the generalization of natural language, our approach is capable of generating different forms of shapes based on various text descriptions. Figure 4 shows the generated results of Diffusion-SDF from different text inputs. The results show that our method can generate eligible results from distinguished conditions. Besides, our method is capable of synthesizing various shape outputs from the same input text description. Figure 5 compares our approach with [29, 34] based on same text queries. The results have shown that our method has a great advantage in the diversity of generated samples, while the previous approaches are unable to generate highly-diversified shapes.

### 4.2. Ablation Studies

We have conducted extensive ablation studies to demonstrate the effectiveness of our special architecture design for Diffusion-SDF.

<table>
<thead>
<tr>
<th>Methods</th>
<th>IoU↑</th>
<th>Acc↑</th>
<th>CLIP-S↑</th>
<th>TMD↓</th>
</tr>
</thead>
<tbody>
<tr>
<td>Liu et al. [29]</td>
<td>0.160</td>
<td>34.79</td>
<td>29.94</td>
<td>0.891</td>
</tr>
<tr>
<td>AutoSDF [34]</td>
<td>0.187</td>
<td>83.88</td>
<td>29.10</td>
<td>0.581</td>
</tr>
<tr>
<td>Diffusion-SDF (Ours)</td>
<td><strong>0.194</strong></td>
<td><strong>88.56</strong></td>
<td><strong>30.88</strong></td>
<td><strong>0.169</strong></td>
</tr>
</tbody>
</table>
Effectiveness of the \textit{UinU-Net} architecture. In order to validate the benefit of our proposed modifications to the conventional U-Net architecture, we additionally compare our proposed \textit{UinU-Net} architecture to the simpler U-Net architecture used in [45]. The quantitative results for text-to-shape generation are shown in Table 2. Qualitative comparisons are shown in Figure 6. From the results, it can be found that our proposed \textit{UinU-Net} architecture generally improves the quality of generated shapes. Since the shape representations are repositioned based on separate patch embeddings extracted by the SDF encoders, the key to recovering the authentic shape SDFs from the joint decoder is to ensure that the generated SDF representations are also distributed independently in patches. From the qualitative results, it can be inferred that the inner network is conducive to recovering the independently distributed patch representations.

\textbf{Inner network architecture.} As for the \textit{UinU-Net} architecture, we have conducted ablation experiments on several variants for the inner network to validate the effectiveness of the final design. The results have shown in Table 3. According to the results, it can be inferred that the inner outer concatenation mechanism and spatial attention layer can generally improve the generation quality. Without the inner-outer concatenation mechanism, the accuracy score and CLIP-S have significant decreases, indicating that the information transmission mechanism has affected the semantic authenticity and conformance of the generated shapes. Accuracy score has dropped without the spatial attention module, indicating that the introduction of patch-to-patch information helps preserve the full shape semantics while recovering the patch-wise representations.

### 4.3. Text-Guided Shape Completion

In this section, we validate the effectiveness of our approach on the text-guided shape completion task. Specifically, given a partial input shape, our goal is to generate the missing part conditioned on the input text descriptions. For
instance, we can generate the body of a chair based on the given chair legs or generate the missing chair legs based on the chair body. The qualitative results of our approach are displayed in Figure 7. The results show that our approach is capable of generating the missing part of a shape while being well-blended with the given shape, as evident in the generated examples such as revolving throne, revolving toilet, and revolving sofa. Moreover, our completion performance is not limited by the shape geometry in the training set. The core limitation of our approach is the cut region of patch grids, which can be addressed by increasing resolution and decreasing patch size.

5. Limitations and Conclusion

Limitations. This paper demonstrates the superiority of our approach on the Text2Shape [6] dataset, which is limited to only two categories from ShapeNet, restricting our approach’s generalization to other categories. Moreover, the lack of current shape-text datasets prevents us from validating our approach on additional benchmarks. To address this, we propose introducing more datasets and exploring zero-shot text-to-shape generation by leveraging knowledge from 2D vision-language models. While some studies have attempted to tackle this challenge [48], achieving flexible text-to-shape generation remains a significant obstacle. Our future work will seek a possible solution to this problem.

Conclusion. Our paper presents a framework for text-to-shape synthesis called Diffusion-SDF, which utilizes a diffusion model to generate voxelized SDFs conditioned on texts. The approach comprises a two-stage pipeline: a patch-wise autoencoder for generating Gaussian SDF representations, followed by a Voxelized Diffusion model with UinU-Net denoisers for generating patch-independent SDF representations. We evaluate our approach on various text-to-shape synthesis tasks, and the results demonstrate that it can generate highly diverse and high-quality 3D shapes.

Acknowledgments This work was supported in part by the National Key Research and Development Program of China under Grant 2017YFA0700802, in part by the National Natural Science Foundation of China under Grant 62206147 and Grant 62125603, and in part by a grant from the Beijing Academy of Artificial Intelligence (BAAI).
References


[27] Zhifeng Kong, Wei Ping, Jiaji Huang, Kexin Zhao, and Bryan Catanzaro. Diffwave: A versatile diffusion model for audio synthesis. In ICLR, 2021. 2


[34] Paritosh Mittal, Yen-Chi Cheng, Maneesh Singh, and Shubham Tulsiiani. Autosdf: Shape priors for 3D completion, reconstruction and generation. In CVPR, pages 306–315, 2022. 1, 2, 5, 6


[41] Aditya Ramesh, Prafulla Dhariwal, Alex Nichol, Casey Chu, and Mark Chen. Hierarchical text-conditional image generation with CLIP latents. 2022. 2


[54] Guandao Yang, Xun Huang, Zekun Hao, Ming-Yu Liu, Serge J. Belongie, and Bharath Hariharan. Pointflow: 3D point cloud generation with continuous normalizing flows. In ICCV, pages 4540–4549, 2019. 1, 2


