Towards Implicit Text-Guided 3D Shape Generation

Zhengzhe Liu\textsuperscript{1} Yi Wang\textsuperscript{2} Xiaojuan Qi\textsuperscript{3}\textsuperscript{*} Chi-Wing Fu\textsuperscript{1}\textsuperscript{*}
\textsuperscript{1}The Chinese University of Hong Kong \textsuperscript{2}Shanghai AI Laboratory \textsuperscript{3}The University of Hong Kong
{zzliu,cwfu}@cse.cuhk.edu.hk wangyi@pjlab.org.cn xjqi@eee.hku.hk

Abstract

In this work, we explore the challenging task of generating 3D shapes from text. Beyond the existing works, we propose a new approach for text-guided 3D shape generation, capable of producing high-fidelity shapes with colors that match the given text description. This work has several technical contributions. First, we decouple the shape and color predictions for learning features in both texts and shapes, and propose the word-level spatial transformer to correlate word features from text with spatial features from shape. Also, we design a cyclic loss to encourage consistency between text and shape, and introduce the shape IMLE to diversify the generated shapes. Further, we extend the framework to enable text-guided shape manipulation. Extensive experiments on the largest existing text-shape benchmark [10] manifest the superiority of this work. The code and the models are available at https://github.com/liuzhengzhe/Towards-Implicit-Text-Guided-Shape-Generation.

1. Introduction

3D shape creation has a wide range of applications, e.g., CAD, games, animations, computational design, augmented reality, etc. Significant progress has been made in recent years by exploiting neural networks and generative models to learn to produce 3D shapes. Yet, existing works [7,12,13,22,34,35,47,49,70,76,78] focus mostly on generating the overall shapes, whereas the more recent ones [11,14,23,44,59,79,80] attempt to generate shapes with more details.

In this work, we are interested in the challenging task of text-guided 3D shape generation—Given a sentence, e.g., “A comfortable red color chair with four legs,” we aim to develop a method to automatically generate a 3D shape that follows the text description; see Figure 1 (a) for our example results. This research direction has great potential for efficient 3D shape production, say by taking user speech/text input to guide or condition the process of generating 3D shapes. By this means, we can assist users to readily generate and edit 3D models for diverse applications.

*: Corresponding authors

While many methods [40,61–64,67,72,73,77,82,83] have been developed for generating 2D images from text, the task of generating 3D shapes from text is rather under-explored. Chen et al. [10] generate 3D shapes from natural language descriptions by learning joint text and shape embeddings, but the performance and visual quality are highly limited by the low-resolution 3D representations. Another very recent work [33] leverages semantic labels to guide the shape generation, but it requires predefined semantic labels and cannot directly deal with natural language inputs.

To enhance 3D shape generation from text, we propose a new solution by leveraging the implicit representation [13,48,54] to predict an occupancy field. Yet, several inherited challenges have not been addressed in the early works for properly adopting the implicit representation for the text-to-shape task. First, the above works generate shapes typically without colors, which are crucial in text-guided 3D shape generation, since text descriptions often contain colors; we empirically found that directly predicting shape and color with a single implicit decoder often lead to shape distortion and color blur. Second, text contains a large amount of spatial-relation information, e.g., “a wooden table on a metal base.” Still, spatial-relation local features are ignored in existing works, since the implicit decoder generally considers only the global feature from the auto encoder as input [13]. Third, the generated shapes are not all consistent with the
input texts, largely due to the semantic gap between text and 3D shape and also the lack of effective learning constraints. Last, text-to-shape generation is inherently one-to-many, i.e., diverse results may match the same input text. Yet, the existing regression-based approach outputs only a single shape.

This work presents a new approach for high-fidelity text-guided 3D shape generation. First, we decouple the shape and color predictions for feature learning in both texts and shapes to improve the generation fidelity; this strategy also aids the text-guided shape manipulation. Also, we introduce a word-level spatial transformer to learn to correlate the word features with the spatial domain in shapes. In addition, we design a cyclic loss to encourage the consistency between the generated 3D shape and the input text. Further, we propose a novel style-based latent shape-IMLE generator for producing diversified shapes from the same given text. Last, we extend the framework for text-guided 3D shape manipulation with a two-way cyclic loss. As shown in Figure 1 (b), we may modify the original text and our framework can produce new colored shapes according to the edited text, while keeping the other attributes unchanged.

Extensive experiments on the largest existing text-shape dataset [10] demonstrate the superiority of our approach over the existing works, both qualitatively and quantitatively.

2. Related Work

Text-to-image generation. Remarkable progress has been made for generating images from text [40, 41, 61–63, 72, 77, 82, 83]. Recently, approaches [56, 64, 66, 67, 73, 75, 81] based on the unconditional GAN [6, 36, 37] were also proposed.

Compared with text-to-image, it is more challenging to generate 3D shapes from texts. First, unlike 2D images, 3D shapes are unstructured and irregular without well-defined grid structures. Also, the text-to-shape task requires a comprehensive prediction of the whole 3D shape, while the text-to-image task addresses image generation, which is a projection of the 3D shape. Further, there are plenty of large-scale image datasets [46, 51, 71] to support text-to-image. Yet, as far as we know, the largest dataset for text-to-shape was proposed in [10], which has 75k texts and 15k shapes of 128^3 resolution. The lack of large-scale and high-quality training data makes the text-to-shape task even harder.

3D shape representations, generation, and manipulation. Unlike images, 3D shapes can be represented as, e.g., voxel grids [18, 24], point clouds [2, 60], and meshes [21]. Also, various methods [30, 38, 39, 45, 68] have been proposed for generating and manipulating shapes for different 3D representations. Yet, the generated shapes are limited by the resolution and quality of the training set. To generate shapes of arbitrary resolution, recent works [13, 16, 47, 48, 54] start to explore implicit functions, which in fact have been used in many tasks, e.g., single-view reconstruction [44, 50, 76], 3D scene reconstruction [29, 35, 58], and 3D texture generation [17, 52, 53]. In existing works, a typical approach is to leverage an auto-encoder (AE) to adopt to multiple 3D generation tasks and map the input modalities into the AE’s learned feature space, e.g., single-image 3D reconstruction [13, 76], point-cloud-based shape generation [7, 15], and 3D completion [74].

Following the above works, a straightforward approach for text-guided 3D shape generation is to map the text feature into the AE’s feature space then adopt an implicit decoder to generate the 3D shape. This simple approach, however, has several drawbacks, as discussed in Section 1.

Recently, several works make it possible to manipulate implicit 3D shapes [19, 27, 32, 84] using a reference box or reference points as guidance. Yet, none of them enables 3D shape manipulation with natural language descriptions.

3D shape generation from text. A series of works are proposed to address the tasks on texts and 3D shapes, including learning the text-shape correspondence [3], cross-modal retrieval [26, 69], shape-to-text generation [25], text-guided shape composition [31], and 3D object localization [9].

As far as we are aware of, there are only few works [10, 32] that address the challenging text-to-shape task. Chen et al. [10] propose to directly predict colored voxels with adversarial learning on top of a jointly-learned text-shape embedding. Though plausible shapes can be produced, the shape resolution and texture quality are still far from being satisfactory. Also, the generated shapes may not be consistent with the input texts due to the large semantic gap between text and shape. Jahan et al. [33] propose a semantic-label guided shape generation approach; however, it can only take one-hot semantic keywords as input and the generated shapes are also unsatisfying in quality, without color and texture.

This work presents a new framework, capable of generating high-fidelity 3D shapes with good semantic correspondence between the text and shape. Also, our framework enables text-guided 3D shape manipulation for both shape and color, outperforming the existing works by a large margin, as demonstrated in the experiments.

Diversified generation. Besides GANs, IMLE (Implicit Maximum Likelihood Estimation) is another approach to aid multi-modal generation, e.g., super-resolution [42], semantic-layout-guided image synthesis [43], image de-compression [57], and shape completion [4]. Compared with GANs, IMLE mitigates the mode collapse of GANs and boosts the result diversity. In this work, we leverage IMLE for generating multiple shapes from the same text input.

3. Methodology

3.1. Overview

Given text T, we aim to generate high-quality 3D shape S with colors, following the description of T. To generate high-quality results, we exploit the implicit occupancy
representation other than the explicit voxel/point/mesh representations to characterize shapes with color. Specifically, the predicted shape with color is denoted as $S \in \mathbb{R}^{N \times (1+3)}$, including the shape $\in \mathbb{R}^{N \times 1}$ (a set of occupancy values in the voxels) and the color $\in \mathbb{R}^{3N}$ (the associated set of RGB values), respectively, where $N$ is the number of sample points, concerning the generation quality.

Our framework consists of a text encoder $B$, feature generator $G$, spatial aware decoder $D'$, and shape encoder $E$. Its overall architecture is given in Figure 2. In inference, $B$ extracts text feature $\bar{f}$ from text $T$ (where $\bar{f}$ and $\bar{f}_c$ are the shape and color portions of $\bar{f}$, respectively), and $G$ produces multiple instances of such feature $\{\bar{f}_i\}_{i=1}^{m}$ based on $\bar{f}$ conditioned on various random vectors $\{z_i\}$. Then, $D'$ generates diverse shapes $\{S_i \in \mathbb{R}^{N \times (1+3)}\}_{i=1}^{m}$ with color.

The model training of our method is non-trivial. We train the overall framework in three stages (see again Figure 2): (a) shape auto-encoder, (b) text-guided shape generation, and (c) diversified shape generation with IMLE. Specifically,

- First, as shown in Figure 2(a), we train shape encoder $E$ and implicit decoder $D$. As shown in the top middle, unlike existing works [13, 33] that ignore colors in the shape generation, $D$ composes of $D_s$ and $D_c$ that account for the decoding of shape and color, respectively, when $D$ predicts the output shape.
- Then, we adopt BERT-based text encoder $B$ [20] to help extract text feature $\bar{f} = \{\bar{f}_s, \bar{f}_c\}$ and word-level feature $\bar{f}_w = \{\bar{f}_{s,w}, \bar{f}_{c,w}\}$ from input text $T$ (see Figure 2(b)), and map $\bar{f}$ into the joint text-shape feature space to reduce the domain gap between the text and the shape. Further, we propose the spatial-aware decoder $D'$ to leverage local feature $\hat{f}_l$ extracted by the word-level spatial transformer (WLST), which explicitly correlates the spatial and word features to improve the fidelity of $S$. Also, we formulate cyclic loss $L_{cyk}$ to encourage the consistency between shape $S$ and text $T$.

- Lastly, we propose to adopt style-based shape generator $G$ that conditions on a set of random noise vectors $\{z_i\}_{i=1}^{m}$ to enable diversified 3D shape generation with feature $\bar{f}_i$, as shown in Figure 2(c).

In the following, we will detail each component of the framework and the associated losses.

### 3.2. Shape Auto-Encoder

We extend the auto-encoder in [13] to jointly reconstruct the shape and color. As shown in Figure 2(a), our shape auto-encoder aims to map the input voxel-based shape $I \in \mathbb{R}^{64 \times 64 \times 64}$ into a compact feature space. Specifically, encoder $E$ [13] extracts the shape and color features $\tilde{f} = \{f_s, f_c\}$ from $I$, whereas decoder $D$ reconstructs the shape and color through $D_s$ and $D_c$, respectively. Inside
D, we concatenate a sample (or query) point coordinate \( p = (x, y, z) \) with each feature vector \( \{f_w, f_s\} \) as input to \( D_s \) or \( D_s' \). \( D_s \) and \( D_s' \) have the same architecture with seven fully-connected and leaky-ReLU layers, except in the last layer, \( D_s \) outputs a single occupancy value and \( D_s \) outputs three values for RGB color, both at the sample point \( p \).

The shape auto-encoder is trained to reconstruct the shape and color of the input shape \( \mathbb{R}^N \) with an \( L_2 \) regression:

\[
L_{ae} = \sum_{p} \left\| D_s(f_s + p) - I(p) \right\|_2^2 + \lambda_s \sum_{k \in \{R, G, B\}} \sum_{p} \left\| D_v(f_c + p)[k] - I(p)[k] \right\|_2^2 + \lambda_c \sum_{p} \left\| \right\|_2^2 + \lambda_c \sum_{p} \left\| \right\|_2^2,
\]

where \( I(p) \) and \( I(k, p) \) denote the ground-truth occupancy and color values, respectively, at point \( p \); \( \oplus \) denotes concatenation; \( I \) is an indicator function of value 1 if \( p \) is inside the input shape, and 0, otherwise; and \( \lambda_s \) and \( \lambda_c \) are weights for the shape and color reconstructions, respectively.

3.3. Text-Guided Shape Generation

As shown in Figure 2 (b), the text-guided shape generation network consists of three modules: shape encoder \( E \), BERT-based text encoder \( B \), and spatial-aware decoder \( D' \). With \( E \) and \( D' \) \((D'_s \) and \( D'_s') \) initialized by the corresponding components in the shape auto-encoder, our goal here is to train the whole network end-to-end to obtain \( B \) and \( D' \).

**Text encoder \( B \).** We employ the BERT structure [20] to build text encoder \( B \) for extracting text feature \( f \) from input text \( T \) and mapping \( f \) to the joint text-shape feature space.

**Spatial-aware decoder \( D' \).** \( D' \) aims to transform text feature \( f \) to the predicted shape \( S \) with color. Instead of simply using the trained implicit decoder \( D \), we construct the spatial-aware decoder \( D' \) with the word-level spatial transformer (WLST). In short, we take the local features from WLST to improve the spatial correlation implied from \( T \).

The right side of Figure 2(b) shows the architecture of the spatial-aware decoder \( D' \). First, we concatenate \( f_s \) and \( p \) and transform the result \( \{f_s + p\} \in \mathbb{R}^{N \times (d + 3)} \) using a fully-connected layer, where \( N \) is the number of sample points for shape reconstruction and \( d \) is the channel dimension of \( f_s \). Then, we transform the word-level BERT features \( \{f_w\} \in \mathbb{R}^{K \times d_W} \) (where \( K \) is the number of words in input text and \( d_W \) is the channel dimension of each word feature \( f_w \)) from \( B \) using a fully-connected layer. The transformed spatial and word features are denoted as \( R \in \mathbb{R}^{N \times d_R} \) and \( W \in \mathbb{R}^{K \times d_W} \), respectively, where \( R_i \in \mathbb{R}^{d_R} \) is the \( i \)-th row in \( R \) that corresponds to the \( i \)-th sample point and \( W_j \in \mathbb{R}^{d_W} \) is the \( j \)-th row in \( W \) that corresponds to the \( j \)-th word in input text. Importantly, we formulate the WLST to learn the correlation between \( \{R_i\} \) and \( \{W_j\} \); see the next paragraph for details. After that, \( D'_s \) takes the global feature \( f_s \), sample point coordinate \( p_i \), and local feature \( f_{s,i} \) from WLST as inputs to predict the occupancy value at \( p_i \) for shape reconstruction.

Figure 3 shows the architecture of the WLST. With the spatial features \( R \) and word features \( W \), we first establish an attention map \( A \) to explicitly correlate each word feature \( W_j \) with each sample point \( p_i \); see Figure 4 for example visualizations of \( A \), revealing how it captures the spatial regions in a shape for different words in the input text. Next, we use the softmax function to process \( A \) to generate the normalized attention matrix \( a \). The output local shape feature \( f_{s,i} \) of point \( p_i \) is the weighted aggregation of the word-level features \( W_j \) across the whole input text. Hence, our WLST can be formulated as

\[
f_{s,i} = \sum_{j} \text{softmax}(\frac{f_Q(W_j)F_K(R_i)}{\sqrt{d}})F_V(R_i),
\]

where \( F_Q, F_K, \) and \( F_V \) are fully-connected layers; see Figure 3 for the architecture of the WLST. Similarly, \( D'_s \) also leverages a WLST for extracting local color feature \( f_{c,i} \).

With the WLST, we can extend the implicit decoder \( D \) to take into account the extra local feature \( \tilde{f}_i = \{\tilde{f}_{s,i}, \tilde{f}_{c,i}\} \) (see Figure 2), which is produced by explicitly learning the correlation between the word-level spatial descriptions and the 3D shape. Hence, we can make every single word in the input text accessible to the shape decoder and enhance the fidelity (or local details) of the generated shape.

**Cyclic consistency loss.** To reduce the semantic gap between the text and shape, we propose a cyclic consistency loss to encourage the consistency between input text \( T \) and output shape \( S \) from \( D' \). To form a cycle, we first grid-sample \( 64 \times 64 \times 64 \) points to use \( D' \) to generate \( S \), and utilize encoder \( E \) from the trained shape auto-encoder to extract features \( f_{cyc} \) from \( S \); see Figure 2(b). Then, we define the cyclic consistency loss to operate on the semantic meaningful feature space instead of the low-level occupancy or color values, such that it can regularize the shape generation.
in a closed loop by encouraging the high-level features $f_{\text{cyc}}$ to be similar to $f = \{f_s, f_c\}$ from the shape encoder.

To reduce the memory consumption and training time, we firstly grid-sample $16 \times 16 \times 16$ points to form a low-resolution voxelized shape $S_i$, then tri-linearly upsample $S_i$ to $S$ of the same resolution as $I (64 \times 64 \times 64)$.

**Network training.** Initially with the shape auto-encoder, we train the text-guided shape generation network end-to-end with the shape auto-encoder loss $L_{ae}$ on $D'$,

$$
L_{ae} = \lambda_s \Sigma_p \|D'_s(f_s, p, \hat{f}_c, R_s) - I(p)\|_2^2 \\
+ \lambda_s \Sigma_{k \in \{R, G, B\}} \|D'_s(f_s, p, \hat{f}_c)[k] - I(p)[k]\|_2^2 (I(p)),
$$

(3)

$$
L_{\text{reg}} = \lambda_r \|\hat{f} - f\|_2^2,
$$

(4)

$$
L_{\text{cyc}} = \lambda_{\text{cyc}} \|f_{\text{cyc}} - f\|_2^2,
$$

(5)

where $\lambda_s$, $\lambda_c$, $\lambda_r$, and $\lambda_{\text{cyc}}$ are weights.

### 3.4. Diversified 3D Shape Generation

To enable diversified 3D shape generation for the same input text, we propose a style-based latent shape-IMLE generator $G$, namely shape IMLE, which operates in the latent space; see Figure 2(c). Taking text feature $\hat{f} = \hat{f}_s \oplus \hat{f}_c$ from $\mathcal{B}$ as input, $G$ generates $\{f_s = \hat{f}_s \oplus \hat{f}_c, f_c\}_{i=1}^{m}$ conditioned on a set of random vectors $Z = \{z_i\}_{i=1}^{m}$. Different from GANs, which encourage the generated samples to be similar to the target data, IMLE inversely encourages each target data to have a similar generated sample to avoid mode collapse [43]. Figure 5 shows the architecture of the shape IMLE $G$.

For the training of $G$, it is optimized as follows:

$$
\min_{\theta} \mathbb{E}_{Z} \left[ \min_{k \in \{1, \ldots, m\}} d(G_{\theta}(\hat{f}; z_k), f) \right]
$$

(6)

where $\theta$ denotes the weights of generator $G$; $d(\cdot, \cdot)$ is a distance metric; and $z_k \sim N(0, 1)$.

With each input $f$, we randomly sample $m$ random noise vectors $\{z_i\}$ to generate $m$ different outputs $\{\hat{f}_i\}$. Among them, the one that is most similar to the ground truth $f$, say $\hat{f}_k$, is trained to be closer to $f$ with an $L_2$ regression. So, we can encourage every ground truth $f$ to have a similar generated sample to avoid the mode-collapse issue in GANs, while promoting diversified shape generation [42, 43]. We train the shape IMLE $G$ with all the other modules $E, B, D'$ frozen (see Figure 2) using an $L_2$ loss on $\hat{f}_k = G(\hat{f}, z_k)$:

$$
L_G = \min_{k \in \{1, \ldots, m\}} \|G(\hat{f}, z_k), f\|_2^2.
$$

(7)

During the inference, we feed every feature of $\{\hat{f}_1, \ldots, \hat{f}_m\}$ into $D$ for generating diversified shapes, without using the ground truth $f$ to select the nearest $\hat{f}_k$.

### 3.5. Text-Guided Shape Manipulation

Next, we extend our framework for text-guided shape manipulation, i.e., to generate shape $\hat{S}$ that matches text $T_2$ that is slightly modified from original text $T_1$ by replacing/inserting/removing one or a few words, and without other attributes unchanged for the same random noise $z$.

Taking shape manipulation (with color unchanged) as an example, we may directly feed feature $\hat{f}_s = \{\hat{f}_{s, a}, \hat{f}_{s, e}\}$ from the edited text to $D'$ to generate the new edited shape. Yet, it could cause drastic changes in the unedited region and colors (Figure 7(b)). Considering the decoupled shape and color features, we may mix $\hat{f}_s$ from edited text and $\hat{f}_c$ from original text as input to $D'$. This simple approach ensures the consistency of the unedited attributes but shape and color may not well align with the edited shape (Figure 7(c)), since $\hat{f}_s$ and $\hat{f}_c$ actually come from different texts.

To encourage shape-color alignment, we propose to feed shape feature $\hat{f}_{s, a}$ (extracted from text $T_2$) and color feature $\hat{f}_{c, a}$ (extracted from text $T_1$) to $G_3$ to predict the manipulated feature $\hat{f}_{s, a, e}$, then, we can feed $\hat{f}_{s, a, e}, \hat{f}_{c, a}$ to $D'$ to produce the edited shape $\hat{S}$. Yet, this approach could still lead to certain changes in the unedited attributes (Figure 7(d)). Figure 6 shows our full framework further with the two-way cyclic loss, i.e., $L_{\text{cyc}, a}$ and $L_{\text{cyc}, e}$. Here, we use shape encoder $E$ to extract manipulated feature $\hat{f} = \{\hat{f}_{s, e}, \hat{f}_{c, e}\}$ from $\hat{S}$ and formulate $L_{\text{cyc}, e}$ for shape consistency $(\hat{f}_{s, e}$ and $\hat{f}_{s, a, e})$ and $L_{\text{cyc}, e}$ for color consistency $(\hat{f}_{c, e}$ and $\hat{f}_{c, a}$). Then, we can formulate the overall loss:

$$
L_{\text{mani}} = (\|\hat{f}_s - \hat{f}_{s, e}\|_2^2 + \|\hat{f}_c - \hat{f}_{c, e}\|_2^2) I(\text{IoU}(I_1, I_2) > t) \\
+ L_{G_1} + L_{G_2},
$$

(8)

where the first term is the two-way cyclic consistency loss, which takes effect only when the Intersection over Union (IoU) between the associated ground-truth shapes $I_1$ and $I_2$ is larger than threshold $t$. The last two terms fine-tune the shape IMLE for a diversified generation (see Eq. (7)).

To train the framework, we initialize its weights from shape IMLE then finetune $G$ using $L_{\text{mani}}$ with all other modules $E, B, D'$ frozen. Also, we randomly sample two unpaired texts $T_1, T_2$ to simulate the original and edited texts. With the two-way cyclic loss, the shape IMLE can learn to generate edited shapes with other attributes unchanged,
Figure 6. Overview of our text-guided shape manipulation framework (with color unchanged). Given two pieces of text $T_1$, $T_2$, shape IMLE $G_1$ and $G_2$ use the same random noise $z_i$ to generate shapes. $G_3$ takes $\{f_{2s}, f_{1c}\}$ and $z_i$ as input to generate shape $S$ with feature $\{f_s, f_c\}$ (encoded by $E$), such that $f_s$ and $f_c$ should be similar to $f_{2s}$ and $f_{1c}$, respectively. Hence, we propose a two-way cyclic loss ($L_{\text{cyc}_s}$ and $L_{\text{cyc}_c}$) to encourage shape consistency between $S$ and $T_2$, and color consistency between $S$ and $T_1$. $G_1$, $G_2$, $G_3$ share the same weights.

Figure 7. (a) The original shape from the unedited text. (b) The shape from the edited text. It shows that even editing just a color-unrelated word may influence the generated color. (c) Replacing $f_{2c}$ with the original color feature $f_{1c}$, can cause misalignment between generated shape and color. (d) Our approach without the two-way cyclic loss, the unedited attributes may still change. (e) Our full approach with the two-way cyclic loss produces an edited shape that better preserves the unedited attributes.

Figure 8. Our text-guided shape and color manipulation results.

while better aligning the shape and color. Please see the supplementary material for the details on the color manipulation framework. Besides Figures 1(b,c) and 7(a,e), Figure 8 shows two more text-guided manipulation results.

4. Experiments

4.1. Dataset and Implementation Details

Our approach is evaluated on the largest text-shape dataset ShapeNet 3D models with natural language descriptions [10]. The dataset contains 15,038 shapes from the table and chair classes of ShapeNet [8]; 75,344 natural language descriptions, 16.3 words per description on average, and 8,147 unique words in the whole dataset [10].

We implement our framework in PyTorch [55]. To train the shape auto-encoder, we sample 4,096 points with the strategy in [13] and train the network for 500 epochs in $16^3$ resolution, then continue the training for another 500 epochs in $32^3$ resolution with learning rate $1e^{-4}$. For text-guided shape generation, we train the network end-to-end for 200 epochs, then fine-tune it end-to-end in $64^3$ resolution for another 200 epochs. For diversified shape generation, we train the shape IMLE for 100 epochs with learning rate $1e^{-3}$ and the other network modules frozen. Lastly, we fine-tune the shape IMLE for another 100 epochs with the two-way cyclic consistency loss to enable manipulation. We set hyper-parameters $d, d_t, \lambda_s, \lambda_c, \lambda_{\text{reg}}, \lambda_{\text{cyc}},$ and $t$ as 256, 32, 2, 1, 1, 0.005, and 0.01, respectively, using a small validation set.

4.2. Comparison with the Existing Works

We compare our method with two existing works [10,33] (see also Section 2) on text-guided shape generation.

For a fair comparison with [10], we transform our generated results into voxels in the same resolution as [10], i.e., $32^3$. Also, we follow its train/val/test (80%/10%/10%) split and its evaluation metrics, i.e., IoU, EMD, IS, and Acc (Err=1-Acc), and directly compare our results with the numbers in [10]. Table 1 reports the results, showing that our method outperforms [10] for all evaluation metrics, manifesting its effectiveness. Note that “IS” ranges [0, 2], as it is built upon a two-category classification model, so both methods (1.96 vs. 1.97) already achieve satisfying performance in this respective. The qualitative comparisons in Figure 9 also demonstrate the superiority of our approach, which is able to generate much better shapes and colors (see Figure 9 (b, c)) in comparison with [10] (see Figure 9 (a)).

The other work [33] focuses on generating shapes from phrase descriptions; see the left side of Figure 10 (a). Since its setting is very different from ours, we only compare with it qualitatively. To do so, we first prepare sentence
4.3. Ablation Studies

We conduct extensive ablation studies to validate the effectiveness of the key components in “text-guided shape generation” and “diversified generation.” To measure the diversity and quality of the generated shapes, we formulate two new metrics, PS and FPD, based on Inception Score (IS) [65] and Fréchet Inception Distance (FID) [28]; please see the supplementary material for the details. To evaluate the text-shape consistency, we adopt R-Precision [77]. To reduce the training time, we train all models in 32³ resolution.

**Text-guided shape generation.** We evaluate the effectiveness of the following major components in this module (Figure 2 (a,b)): joint training with a pre-trained auto-encoder (AE), decoupled shape-color decoder (DSCD), WLST module (WLST), and cyclic loss (CL). Please refer to the supplementary material for the details of each setup.

Quantitative and qualitative results are shown in Table 2 and Figure 11, respectively. Note that all models in this setting achieve satisfying R-Precision (> 98% except “Without AE”), so we report R-Precision only in the next “Diversified generation” setting. First, auto-encoder joint training (AE) is crucial for model convergence. Without AE, the baseline approach fails to converge, leading to unreasonable results (see Figure 11 (a)) of very low quality. Second, decoupling shape and color in the decoder structure (DSCD) improves both PS and FPD by a large margin, manifesting its effectiveness in promoting high-fidelity and diversified synthesis. This is also verified in the qualitative comparison shown in Figure 11 (b). Third, empowered by the word-level correlation, we can enrich the local details; see “red back and seat cushion” in Figure 11 (c). Lastly, cyclic loss (CL) improves the consistency between the generated shape and input text; see the visual comparison in Figure 11 (d). Note that both WLST and CL benefit the model more in the “Diversified generation” component to be detailed below.

**Diversified generation.** Next, we evaluate the major modules for diversified generation (Figure 2 (c)). First, we re-

Table 1. Quantitative comparisons with the existing work [10].

<table>
<thead>
<tr>
<th>Method</th>
<th>IoU (↑)</th>
<th>IS (↑)</th>
<th>EMD (↓)</th>
<th>Err (↓)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Text2Shape [10]</td>
<td>12.21</td>
<td>1.97</td>
<td>0.2071</td>
<td>2.52</td>
</tr>
<tr>
<td>Ours</td>
<td>9.64</td>
<td>1.96</td>
<td>0.4443</td>
<td>2.63</td>
</tr>
</tbody>
</table>

Table 2. Ablation studies on text-guided shape generation.

<table>
<thead>
<tr>
<th>Method</th>
<th>IoU (↑)</th>
<th>PS (↑)</th>
<th>FPD (↓)</th>
<th>R-Precision (↑)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Without AE</td>
<td>0.03</td>
<td>1.01±0.00</td>
<td>67.37</td>
<td></td>
</tr>
<tr>
<td>+AE</td>
<td>12.04</td>
<td>2.95±0.03</td>
<td>35.05</td>
<td></td>
</tr>
<tr>
<td>further +DSCD</td>
<td>12.00</td>
<td>3.16±0.04</td>
<td>31.09</td>
<td></td>
</tr>
<tr>
<td>further +WLST</td>
<td>12.24</td>
<td>3.21±0.05</td>
<td><strong>30.34</strong></td>
<td></td>
</tr>
<tr>
<td>further +CL (full)</td>
<td><strong>12.33</strong></td>
<td><strong>3.26±0.06</strong></td>
<td>30.80</td>
<td></td>
</tr>
</tbody>
</table>

Table 3. Ablation studies on diversified shape generation.

<table>
<thead>
<tr>
<th>Method</th>
<th>PS (↑)</th>
<th>FPD (↓)</th>
<th>R-Precision (↑)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Latent GAN</td>
<td>3.31±0.02</td>
<td>30.70</td>
<td>21.20±0.11</td>
</tr>
<tr>
<td>FC IMLE</td>
<td>2.93±0.02</td>
<td>29.53</td>
<td>25.97±0.09</td>
</tr>
<tr>
<td>Shape IMLE</td>
<td>3.39±0.02</td>
<td>29.65</td>
<td>27.60±0.39</td>
</tr>
<tr>
<td>further +WLST</td>
<td>3.39±0.03</td>
<td>28.41</td>
<td>34.37±0.09</td>
</tr>
<tr>
<td>+WLST+CL (full)</td>
<td><strong>3.45±0.02</strong></td>
<td><strong>27.26</strong></td>
<td><strong>40.71±0.10</strong></td>
</tr>
</tbody>
</table>

Figure 9. Results by Text2shape [10] (a) vs. ours (b,c) vs. GT (d).

Figure 10. Results generated by [33] (a) vs. ours (b).

Figure 11. Qualitative ablation studies on shape generation.
place the style-based shape IMLE with two different components for shape generation: a Latent GAN and a fully-connected IMLE (FC IMLE). Besides, we explore models without the proposed WLST module and cyclic loss (CL), which benefit diversified shape generation. Please refer to the supplementary material for the details of each setup.

Quantitative and qualitative results are shown in Table 3 and Figure 12, respectively. Note that this setting focuses on shape diversity and quality, so we do not adopt “IoU,” which measures the shape similarity to the ground truths. In comparison with “Latent GAN,” the IMLE model can synthesize diversified colors and avoid generating collapsed invalid shapes (see Figure 12 (a) vs. (b)), while attaining better quantitative results. Further, the proposed style-based generator (“shape IMLE”) consistently improves on all metrics and yields higher quality shapes with better completeness in comparison with “FC IMLE,” as shown in Figure 12 (c). Lastly, the WLST module and cyclic loss further help improves the generation fidelity and text-shape consistency by a large margin as shown in the last two rows of Table 3, manifesting their effectiveness (see Figure 12 (d,e)).

4.4. Text-Guided Shape and Color Manipulation

More text-guided manipulation results are shown in Figures 13 and 14, in addition to Figures 1(b,c), 7(a,e), and 8. Thanks to our two-way cyclic loss, our model enables text-guided modification of colors and shapes in the generated results, while trying to keep the other attributes intact. For instance, we are able to modify a “square” table to become “circular,” while keeping the other irrelevant regions unchanged, e.g., the legs of the chair; see Figure 13 (a). If we change the word “pink” to “blue,” only the associated parts in the shape are changed accordingly; see Figure 14 (a). More comparisons with the existing work [10] and further ablation study on manipulation can be found in the supplementary material.

5. Conclusion

We have presented a novel framework capable of generating diversified 3D shapes with colors from text descriptions, while allowing flexible text-guided manipulations. Besides the framework, we propose to decouple the shape and color predictions for learning both shape and color features from texts and design the word-level spatial transformer to explicitly correlate words with spatial locations to enhance the local details. Also, we develop the cyclic consistency loss to enhance the text-shape consistency and introduce the style-based shape-IMLE generator for diversifying the generated shapes. Further, we extend the framework for text-guided shape manipulation with the novel two-way cyclic loss. Extensive experimental studies manifest the effectiveness of our framework. Limitation analysis and future works are elaborated in the supplementary material.

Acknowledgement. This work is supported by the Research Grants Council of the Hong Kong Special Administrative Region (Project No. 14201921 and 27209621).
References


