Project Proposal - Text-guided Multi-instance Shape Synthesis

Mustafa Sercan Amaçr
sercan.amac@tum.de

Youssef Youssef
youssef.youssef@tum.de

Abstract

*3D shape synthesis from text is mostly done in a single pass without any modifications. The recent **ShapeCrafter** model architecture introduced recursive queries for shape generation. We aim to implement this architecture and train it as a baseline and then extend it by improving visual quality results. We also aim to reverse the process by introducing text shape descriptions from given 3D shapes.*

1. Introduction

2D image synthesis from input text has been an ongoing research topic for the past years. While there have been many notable breakthroughs such as diffusion models for the 2D case, 3D shape synthesis from text received much less attention. Existing models for the 3D case generate the shape from text in a single step. This might be easier to handle and obtain results. However, it is not how humans normally think and generate queries. Generally, humans explain their thoughts in a **recursive** and **iterative** manner. A person might not know exactly what is required or needed from the beginning, but will be able to identify what is **not** required or what is missing in each time step and accordingly modify or add to the query to get closer to the desired goal iteratively.

2. Related Work

The ideas explained before form the main basis behind [3]. They present a new novel model, **ShapeCrafter**, that can iteratively generate and modify 3D shapes according to the given text at every time step. They begin by creating a new data set **Text2Shape++** based on Text2Shape [2] to be able to train the models on recursive input queries. **ShapeCrafter** consists of 3 models (trained separately) to handle the text-guided 3D shape synthesis problem: first the text Feature Extraction Model, which extracts the text feature and projects it to voxel grids of probability distribution. Secondly, the text and shape feature concatenation model, which concatenates extracted text features and shape features. Finally, the shape feature refinement model, which generates the shape from the concatenated features.

3. Goal

We first begin by obtaining the **Text2Shape++** Dataset. The dataset is not publicly available. However, there is a publicly available script that transforms the **Text2Shape** dataset to **Text2Shape++**. There is neither code nor pre-trained models publicly available, so we aim to implement, train and improve the architecture mentioned in [3] using the obtained dataset as a baseline model.

4. Stretch Goals

We have several ideas and novel applications to tackle depending on the timing constraints and main goal progress. First, we aim to generate text description from 3D shapes, where the process is reversed, we input 3D shapes and get text descriptions. We also aim to improve the visual quality of the generated images, since many images presented in [3] have deformations and holes, namely Fig.1, Fig.3 and Fig.4. Finally, we can look into the generality of the current architecture. The current implementation only handles 2 object categories (limited to **Text2Shape** dataset). So we can try to introduce 2 more 3D shapes, for example sofas and beds. Depends on if we can generate text descriptions from **ShapeNet** [1].

5. Timeline

Period/Milestones	Tasks
23-4 to 3-5	Proposal, Background reading, Dataset, Sandbox
8-5 to 28-5	Proposal Feedback, Baseline implementation, have good initial results
28-5 to 31-5	Presentation 1
1-6 to 25-6	Work on improvements, Text Description generation model and generalization
25-6 to 28-6	Presentation 2
28-6 to 12-7	Final changes and improvements, Poster presentation and model deployment if there is enough time

References

[1] Angel X Chang, Thomas Funkhouser, Leonidas Guibas, Pat Hanrahan, Qixing Huang, Zimo Li, Silvio Savarese, Manolis Savva, Shuran Song, Hao Su, et al. Shapenet: An information-rich 3d model repository. *arXiv preprint arXiv:1512.03012*, 2015. 1

[2] Kevin Chen, Christopher B Choy, Manolis Savva, Angel X Chang, Thomas Funkhouser, and Silvio Savarese. Text2shape: Generating shapes from natural language by learning joint embeddings. *arXiv preprint arXiv:1803.08495*, 2018. 1

[3] Rao Fu, Xiao Zhan, YIWEN CHEN, Daniel Ritchie, and Srinath Sridhar. Shapecrafter: A recursive text-conditioned 3d shape generation model. In S. Koyejo, S. Mohamed, A. Agarwal, D. Belgrave, K. Cho, and A. Oh, editors, *Advances in Neural Information Processing Systems*, volume 35, pages 8882–8895. Curran Associates, Inc., 2022. 1