Sculpting with Words

Victor Perez
Morphogen Research
`vipermu97@gmail.com`

Joel Simon
Morphogen Research
`joel@morphogen.io`

Tal Shiri
Morphogen Research
`toasty.tal@gmail.com`

Abstract

In this work we present an interactive framework for text-guided 3D mesh generation. Our tool integrates the power of large multimodal pre-trained models (i.e. CLIP) and differentiable rendering into a traditional 3D sculpting environment. Our approach for sculpting consists of a user-guided optimization of the values of a 3D Signed Distance Field using gradient descent and a CLIP-based loss.

1   Introduction

The design of interfaces for the collaboration between humans and machine learning models has the potential of augmenting our creative power. With this work, we tackle this challenge in the context of the generation of 3D forms, a task that generally requires substantial technical skill and where the interaction between the user and the artwork is essential. Our framework proposes a flexible interface that enables a real-time collaboration with a novel approach for text-guided generation of 3D meshes, allowing us to expand our capabilities in this task.

Our approach leverages the power of CLIP [8], in particular its visual and textual encoders to guide the generation of images with text. This approach has proved to be effective when used to optimize pre-trained generative models such as GANs [6] or VAEs [1] as well as in settings where other kinds of parameters were used to generate images like in [4] or [2]. As far as we know, the extension of the use of CLIP to optimize 3D shapes has only been used to optimize pre-trained 3D generative models [3] of human bodies [7].

A current limitation of the aforementioned text-guided generative methods is the lack of control of its outputs, something important in a 2D context but essential when dealing with 3D shape generations. The primary creative act is the clever selection of text to feed as input which then runs autonomously and is deterministic to the prompt and the random seed. Accordingly, the text input is arguably the primary object of authorship. Converting black box optimizes into interactive processes offers users far more accordance and ultimately sense of authorship.

2   Interactive Sculpting Process

Our interface consists of a 3D interactive environment similar to most common mesh editing software. The user begins by loading an arbitrary starting mesh and optimization prompt. The prompt can be selectively applied to any region of the mesh by positioning a sphere of influence (Green sphere in figure1 A, B). The sphere can be applied locally or globally (figure 1C). Once the optimization has begun the user may dynamically control various parameters such as learning rate, sculpting resolution, and the relative strengths of the loss functions. All parameters including the prompt can be modified at any time, enabling multiple overlapping optimization results. Crucially, the camera of the user determines the perspective that it is optimized by (with some degree of random augmentation which is also controllable). The user can directly undo any optimization process or directly manipulate the mesh to encourage the optimization to go in a new direction.
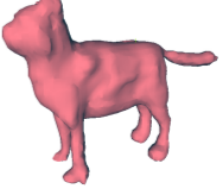
Prompt	A: "A cat with ribs"	B: "Big ears"	C: "A muscular animal"
Base mesh with mask			
Result			

Figure 1: Examples of single prompt edits

3 Mesh Optimization using CLIP

Our goal is to maximize the resemblance between the shape of a 3D mesh and the content described in a text prompt. Extending the optimization process proposed in SDFDiff [5], we exploit the use of 3D Signed Distance Fields (SDFs) to represent 3D meshes. These representations consist of a 3D grid where the absolute value of each voxel represents whether distance to the mesh and its sign represents if it is contained inside or outside the limits of the mesh.

Our optimization process starts with a pre-defined 3D SDF. We use a differentiable renderer to compute an image that represent a certain view of the encoded mesh. Then, using CLIP, we encode this image along with a text prompt obtaining a visual and textual representations respectively that share a common latent space. Finally, we compute their negative cosine similarity and we use gradient descent to optimize the SDF values, hence obtaining an updated mesh. In practice, we mask the previous gradients in order to only update a certain region of the mesh. For each optimization step, we also compute the loss for multiple images rendered from random augmentations around a viewpoint to add robustness to our optimizations.
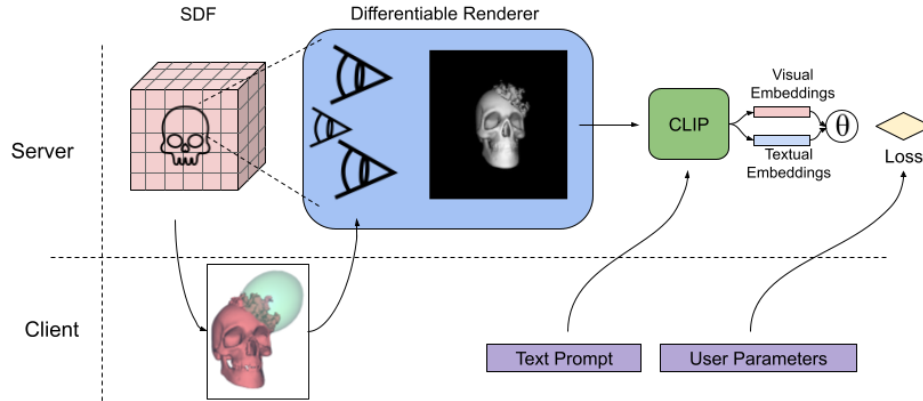

Figure 2: Overview of the mesh editing process

2

References

[1] Clip guided diffusion hq 256x256.ipynb - colaboratory. `https://colab.research.google.com/drive/12a_Wrfi2_gwwAuN3VvMTwVMz9TfqctNj`. (Accessed on 09/24/2021).

[2] eps696/aphantasia: Clip + fft/dwt = text-to-image. `https://github.com/eps696/aphantasia`. (Accessed on 09/24/2021).

[3] Nikolay jetchev on twitter: ""orangutan burnt from forest fires and plastics, chocolate bars and soaps" #clip prompt for evolving a 3d model. the orangutan is seen, but also the choco bars look evil conceptually: a warning that if we do not protect nature, all we will have at the end are 3d models #aiart https://t.co/y2fushhpvs" / twitter. `https://twitter.com/NJetchev/status/1440722632697991181`. (Accessed on 09/24/2021).

[4] Kevin Frans, L. B. Soros, and Olaf Witkowski. Clipdraw: Exploring text-to-drawing synthesis through language-image encoders, 2021.

[5] Yue Jiang, Dantong Ji, Zhizhong Han, and Matthias Zwicker. Sdfdiff: Differentiable rendering of signed distance fields for 3d shape optimization, 2019.

[6] Or Patashnik, Zongze Wu, Eli Shechtman, Daniel Cohen-Or, and Dani Lischinski. Styleclip: Text-driven manipulation of stylegan imagery, 2021.

[7] Georgios Pavlakos, Vasileios Choutas, Nima Ghorbani, Timo Bolkart, Ahmed A. A. Osman, Dimitrios Tzionas, and Michael J. Black. Expressive body capture: 3D hands, face, and body from a single image. In *Proceedings IEEE Conf. on Computer Vision and Pattern Recognition (CVPR)*, pages 10975–10985, 2019.

[8] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, Gretchen Krueger, and Ilya Sutskever. Learning transferable visual models from natural language supervision, 2021.
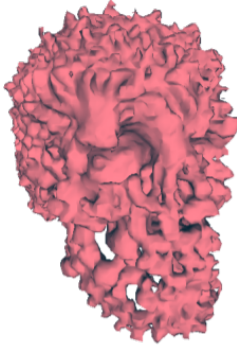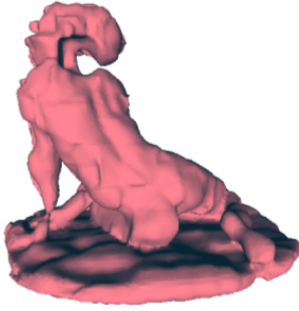
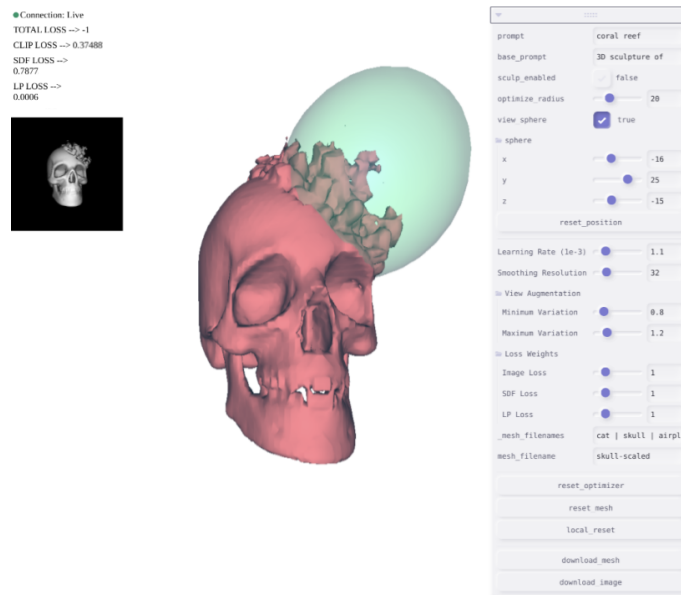| Prompt	Start Mesh	Final Mesh
"Coral reef"		
"Voxel Grid"		

Figure 3: Additional Creations

Figure 4: Screenshot of the sculpting process inside the client application