Evolving Evocative 2D Views of Generated 3D Objects

Eric Chu
MIT Media Lab
echu@mit.edu

Abstract

We present a method for jointly generating 3D models of objects and 2D renders at different viewing angles, with the process guided by ImageNet and CLIP -based models. Our results indicate that it can generate anamorphic objects, with renders that both evoke the target caption and look visually appealing.

1 Introduction

While there has been significant effort by both the research and artistic communities to use image models such as BigGAN [1] and CLIP [6] to produce artwork, using deep generative models to create 3D works of art is comparatively underexplored. In working towards that goal, we are also motivated to model *anamorphic art* [7], which change or only become recognizable upon certain viewing angles [1] [2] [3] . This work also loosely parallels computer vision research aiming for more *interpretable* and *controllable* image generation by jointly learning the 3D generative model and the 2D rendering process [4].

Figure 1: Rendered views of generated 3D objects, under different optimization objectives. Row 1: optimizing for the given class under an ImageNet-based model. Row 2: optimizing for similarity to the text "A [photo|painting] of ..." under the CLIP model.

[1] Skull when viewed from bottom left: `https://en.wikipedia.org/wiki/The_Ambassadors_(Holbein)`

[2] Portrait of Ferdinand Cheval in junk heap: `https://www.fastcompany.com/3032876/perspective-is-everything-this-anamorphic-sculpture-is-and-isnt-what-it-appears`

[3] Ambigram on cover of Gödel, Escher, Bach: `https://i.stack.imgur.com/OKBvZ.jpg`

Preprint. Under review.

2 Method and results

The pipeline for generating 3D structures, and ultimately images of the generated objects, consists of three components: (1) a generator, (2) a scorer, and (3) an optimization loop using a genetic algorithm as shown in Figure 2.

The **first component of our generator** is the 3D version of the "superformula", a generalization of the superellipse originally introduced as a simple, "universal" equation to model a wide range of natural and abstract shapes [2]. This serves as a powerful and computationally tractable 3D model, but could eventually be replaced by deep generative models [8, 5] as they improve. The 2D equation, with 6 parameters m, a, b, n_1, n_2, n_3, is given as:

$$r(\varphi) = \left(\left| \frac{\cos\left(\frac{m\varphi}{4}\right)}{a} \right|^{n_2} + \left| \frac{\sin\left(\frac{m\varphi}{4}\right)}{b} \right|^{n_3} \right)^{-\frac{1}{n_1}}$$

This can be generalized to three dimensions using two instances of the superformula r_1 and r_2 by:

$$\begin{aligned} x &= r_1(\theta)\cos\theta \cdot r_2(\phi)\cos\phi \\ y &= r_1(\theta)\sin\theta \cdot r_2(\phi)\cos\phi \\ z &= r_2(\phi)\sin\phi \end{aligned}$$

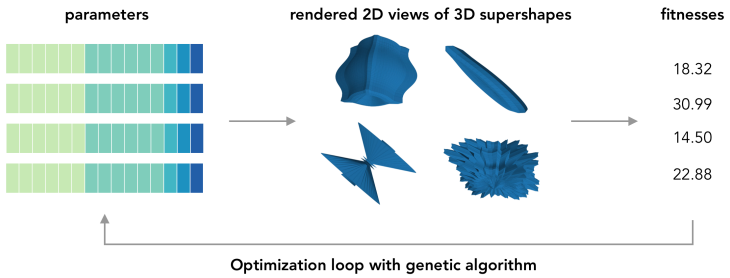

Figure 2: Pipeline for generating 3D objects and views.

In order to render the 3D surface to a 2D image, the **second component of our generator** controls the viewing angle by setting three additional parameters – the elevation, azimuth, and rotation. This can introduce considerable variability and asymmetry into the final product.

These images are then assessed by a **scorer**. Specifically, the fitness of an image is calculated by either (a) the loss against a specific ImageNet class under a trained MobileNetV3 model [3], or (b) the similarity to a caption under the CLIP model [6]. As the rendering proecss is non-differentiable, we optimize the 15 parameters in a genetic algorithm -based **optimization loop**. The parameters are mutated at a rate of 0.1 and a selection rate of 0.5 while using the roulette wheel selection strategy. We keep a population size of 40 at every iteration.

As shown in Figures 1 and 3, our method is able to effectively simulatneously generate surfaces and control the viewing angle in order to evoke the desired object. Often times, the target is only clear at highly specific angles. Finally, we present some examples in the Appendix of (1) explorations of generating views of Richard Serra -style sculptures using a purely 2D-based generative model, and (2) cherry-picked views evolved using novelty search that the author found personally visually appealing as abstract art.

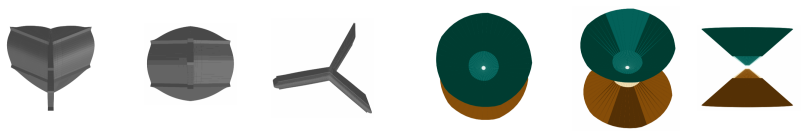

Figure 3: Different views of the generated surfaces for "elephant's head" and "record player".

References

[1] A. Brock, J. Donahue, and K. Simonyan. Large scale gan training for high fidelity natural image synthesis. In *International Conference on Learning Representations*, 2018.

[2] J. Gielis. A generic geometric transformation that unifies a wide range of natural and abstract shapes. *American journal of botany*, 90(3):333–338, 2003.

[3] A. Howard, M. Sandler, G. Chu, L.-C. Chen, B. Chen, M. Tan, W. Wang, Y. Zhu, R. Pang, V. Vasudevan, et al. Searching for mobilenetv3. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 1314–1324, 2019.

[4] Y. Liao, K. Schwarz, L. Mescheder, and A. Geiger. Towards unsupervised learning of generative models for 3d controllable image synthesis. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 5871–5880, 2020.

[5] C. Nash, Y. Ganin, S. A. Eslami, and P. Battaglia. Polygen: An autoregressive generative model of 3d meshes. In *International Conference on Machine Learning*, pages 7220–7229. PMLR, 2020.

[6] A. Radford, J. W. Kim, C. Hallacy, A. Ramesh, G. Goh, S. Agarwal, G. Sastry, A. Askell, P. Mishkin, J. Clark, et al. Learning transferable visual models from natural language supervision. *arXiv preprint arXiv:2103.00020*, 2021.

[7] D. Topper. On anamorphosis: Setting some things straight. *Leonardo*, 33(2):115–124, 2000.

[8] Y. Zhou and O. Tuzel. Voxelnet: End-to-end learning for point cloud based 3d object detection. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 4490–4499, 2018.

Appendix A Views of Richard Serra -style sculptures using VQGAN+CLIP

The artist Richard Serra has a series of sculptures, such as *To Lift*[4], that were created by "applying a verb" [5] on a material such as fiberglass, neon, vulcanized rubber, or lead.

"It struck me that instead of thinking what a sculpture is going to be and how you're going to do it compositionally, what if you just enacted those verbs in relation to a material, and didn't worry about the results?"

We viewed these kind of sculptures as an interesting test bed for material modeling and compositionality (material + verb) for existing image generative models. While one might prefer a joint 3D-2D model as we propose, we investigate the popular VQGAN+CLIP pipeline's ability to replicate and produce new Serra -style art. Shown below in Figure 4, we see that the model is able to produce a limited approximation of the target captions.

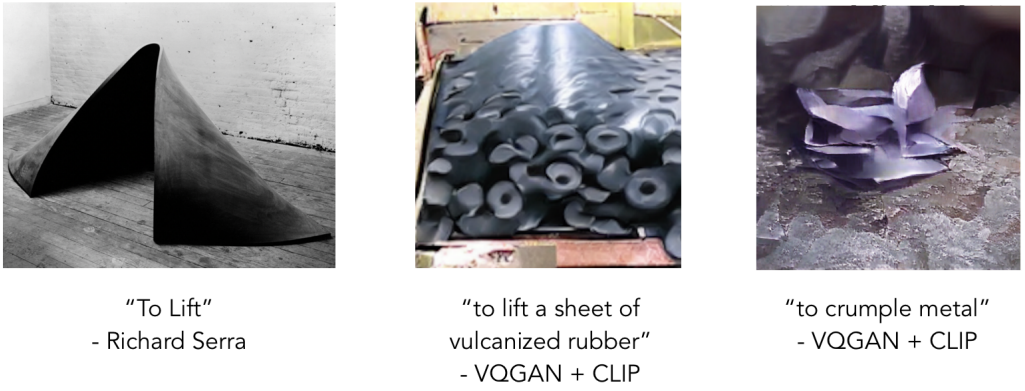

Figure 4: Richard Serra's verb sculpture vs. VQGAN+CLIP generations.

Appendix B Abstract art discovered through novelty search

We also exlored the use of novelty search (instead of optimization against CLIP or an ImageNet model) to generate views. Several examples are shown below:

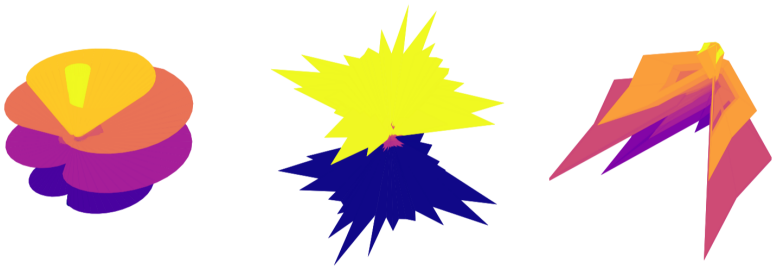

Figure 5: Views discovered during novelty search.

[4]To Lift: `https://www.moma.org/collection/works/101902`

[5]Verb list: `https://www.moma.org/collection/works/152793?artist_id=5349&page=1&sov_referrer=artist`