

DreamControl-3D: ControlNet for Text-to-3D

Eileen Li (chenyil), Sanjan Das (sanjand)
16-825: Learning for 3D
Carnegie Mellon University

1 Introduction

Creating professional 3D models is a task that requires significant time and effort. On the other hand, 3D content has been increasing in demand, permeating industries such as entertainment, architecture, and robotics. With diffusion models, 2D image generation has become easier and more accessible. There are several approaches for conditional generation as well. However, the same for 3D model generation is yet to catch up. How can we make this easier?

Describing what you envision in natural language is an intuitive way to generate content. Moreover, being able to provide a rough sketch as guidance would be incredibly powerful. With user text and a rough sketch as input, we want to output a high fidelity 3D model that well caters to the user’s requirement.



Figure 1: Example of what we would like to do

This codebase for our project is available at: <https://github.com/eileenforwhat/dream-control-3d>.

2 Related Work

To accomplish our goal, we build on several recent advancements: StableDiffusion [Rom+22] for text-to-image, ControlNet [ZA23] for conditional 2D control, DreamFusion [Poo+22] for text-to-3D, and Zero-1-to-3 [Liu+23] for novel view synthesis.

2.1 StableDiffusion: High Resolution Image Synthesis with Diffusion Models

Diffusion models are becoming increasingly popular in recent years as they have surpassed other image synthesis methods in producing high fidelity results. In a nutshell, these models learn to incrementally denoise an image and can be conditioned using text and/or another image. StableDiffusion [Rom+22] is a text-to-image model released by Stability AI and is ubiquitous in followup works involving generative AI, since it is the only open-source large-scale diffusion model (unlike Midjourney, OpenAI’s Dalle-2, or Google’s Imagen).

2.2 ControlNet: Adding Conditional Control to Text-to-Image Diffusion Models

ControlNet [ZA23] is a neural network architecture that can control image diffusion models (like StableDiffusion) to learn task-specific input conditions. The ControlNet clones the weights of a large diffusion model into a “trainable copy” and a “locked copy”. The locked copy preserves the network capability learned from billions of images, while

the trainable copy is trained on task-specific datasets to learn the conditional control. The trainable and locked neural network blocks are connected with a unique type of convolution layer called “zero convolution”, where the convolution weights progressively grow from zeros to optimized parameters in a learned manner. Since the zero convolution does not add new noise to deep features, the training is as fast as fine-tuning a diffusion model, compared to training new layers from scratch.

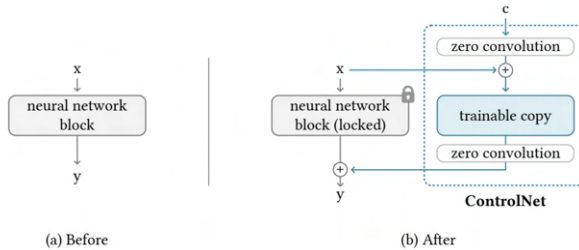


Figure 2: ControlNet architecture.

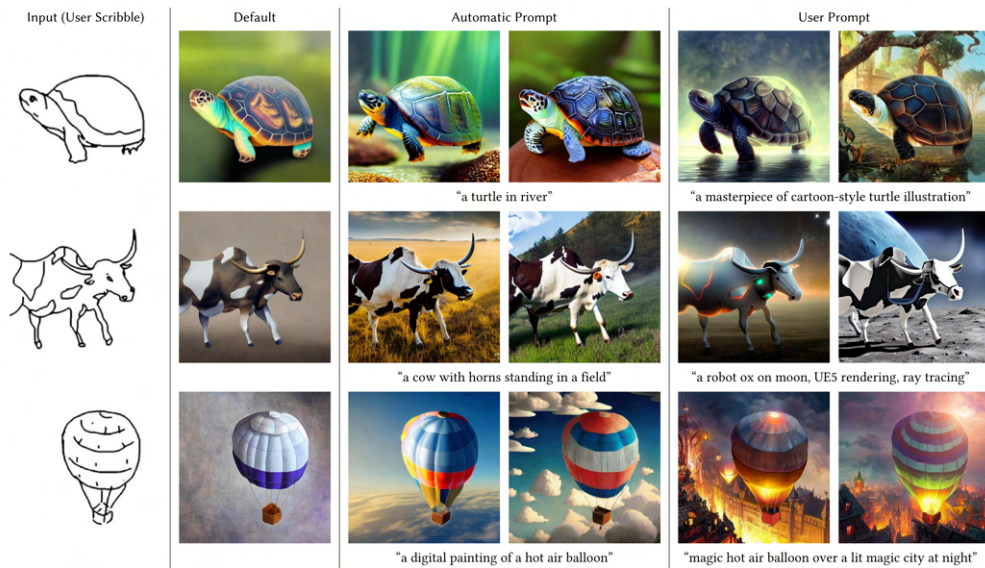


Figure 3: Example results from ControlNet, with scribble conditioning.

While ControlNet has shown amazing results in the 2D domain, it has not been applied to 3D yet. For our project, we explore this direction of work and its effectiveness.

2.3 DreamFusion: Text-To-3D Using 2D Diffusion

In this work, the authors use 2D text-to-image diffusion model to perform text-to-3D synthesis. They use a pretrained Imagen (text-to-image) model and optimize a randomly initialized NeRF such that its 2D renderings from random angles achieve a low loss. More specifically, at each training iteration, a random camera and light setting is used to render a NeRF model. They then add some sampled noise, ϵ , at timestep t , and feed the noisy image into a frozen diffusion model, which tries to predict the added noise. The delta between added and predicted noise is backpropagated onto the NeRF parameters as gradient update.

The results look promising, but the exact setup and quality has not been reproduced by the community since the underlying models used are not open-sourced. Moreover, there hasn't been a clear way to establish control over the 3D generation process.



Figure 4: Example results from DreamFusion

For our project, we build on an open source version of DreamFusion that uses StableDiffusion instead of Imagen as the diffusion model backend [Tan22].

2.4 Zero-1-to-3: Novel View Synthesis

Zero-1-to-3 is a framework for changing the camera viewpoint of an object given just a single RGB image. The authors train a conditional diffusion model using a synthetic dataset to learn controls of the relative camera viewpoint, which allow new images to be generated of the same object under a specified camera transformation.

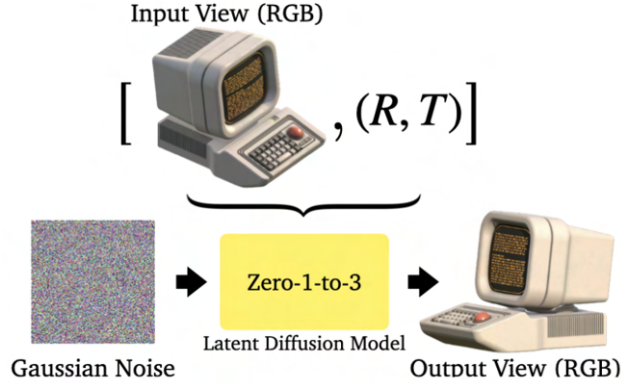


Figure 5: Zero-1-to-3 takes a novel viewpoint as input and outputs an image of the object from that viewpoint.

We use a pretrained Zero-1-to-3 model that generates 2D images from different viewpoints for our StableDiffusion output.

3 Method

3.1 ControlNet on StableDiffusion

Figure 6 shows how we combine ControlNet and DreamFusion to achieve user scribble control of 3D generation. We add ControlNet guidance by replacing the DreamFusion diffusion pipeline with one that contains a ControlNet module. Furthermore, we use a pretrained model that has been conditioned on user scribbles (ControlNet-Scribble). This module’s weights are locked during 3D NeRF optimization.

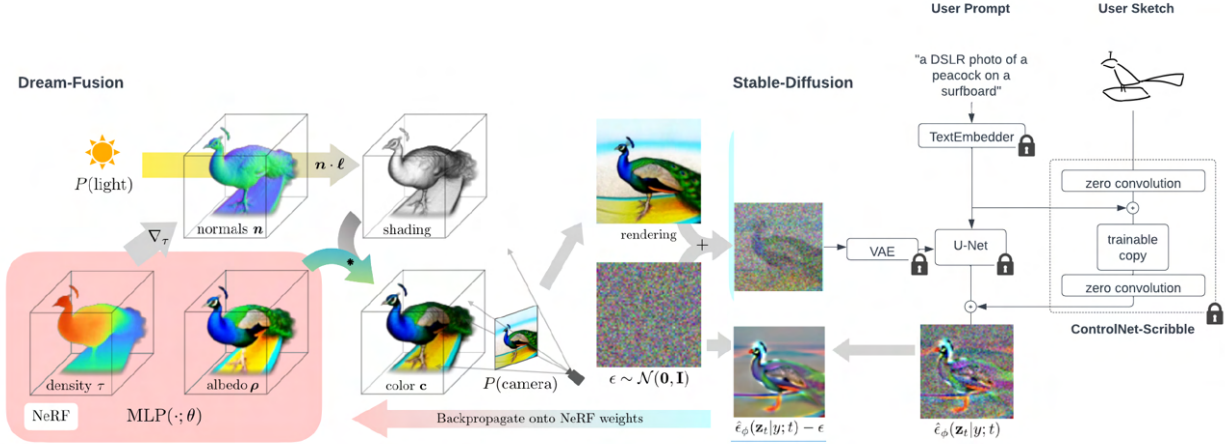


Figure 6: Adding ControlNet guidance to DreamFusion

3.2 Multi-viewpoint Optimization

There is an inherent issue in the approach described previously that may impact the quality of results. The user scribble is provided from a single viewpoint which means that ControlNet only helps to condition the StableDiffusion output from a certain viewpoint only. NeRF optimization actually needs images sampled from hundreds of different viewpoints for optimization. In the original StableDiffusion guidance, this is accomplished in a somewhat hacky way by appending phrases such as "from the {right, left, back, front} view" to the original input text conditioning string. In other words, we are relying on the text conditioning to provide images from different viewpoints.

To tackle this problem, we experimented with two different methods:

1. Viewpoint emphasis: add a weighted loss such that we place more emphasis on the primary viewpoint
2. ControlNet with Zero-1-to-3 guidance: use pretrained zero-1-to-3 model to generate object images from different camera viewpoints

3.2.1 Viewpoint emphasis

This method weighs training loss differently depending on the viewpoint of the camera during NeRF optimization. In addition to the user scribble input, we also require from them the "view" from which they've drawn their object (whether it's from the side, front or back). When a random camera position is sampled during NeRF training, if the camera position matches the scribble's intended view, then the loss is weighted by a factor. For this experiment, we simply multiply the loss term by a factor of 2.0.

3.2.2 ControlNet with Zero-1-to-3 guidance

Using 'controlnet-zero1to3' guidance, we first run ControlNet upstream to generate RGB images from user scribble. We then replace the StableDiffusion model of DreamFusion with a pretrained Zero-1-to-3 model that takes as inputs the rendered image and the sampled viewpoint that was used to produce the rendering. In this way, the Zero-1-to-3 module takes the viewpoint into account when calculating the estimated noise and the loss propagation is much more accurate during NeRF optimization. This method is shown in Figure 7.

3.3 Constraints in Implementation

The official DreamFusion uses Imagen for their diffusion model and is not publically released. We use an open-source implementation of DreamFusion that uses StableDiffusion instead. This implementation is known for not producing results at the same quality level as the original DreamFusion paper.

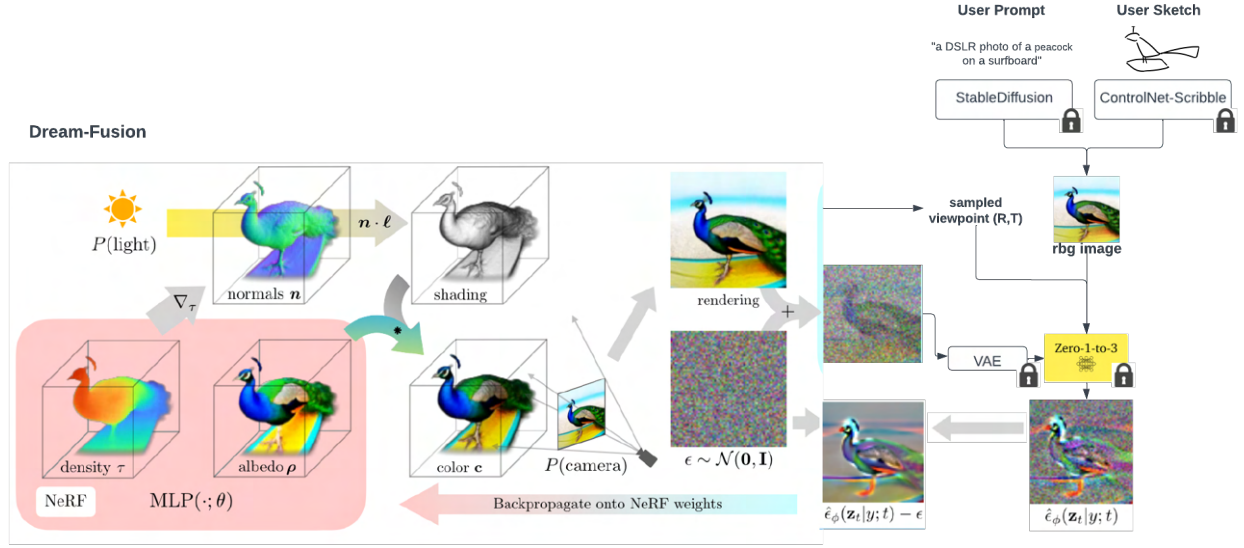


Figure 7: Adding ControlNet + Zero-1-to-3 guidance to DreamFusion

Even with ‘instant-ngp’ [Mül+22] enabled, NeRF optimization is quite slow with 3D model generation taking around 1.5 hours for each object. This bottleneck makes a live demo (which was our original plan) incredibly difficult and limits the number of experiments we can run.

4 Experiments

4.1 Training Details

For ‘controlnet’ guidance, we use controlnet_conditioning_scale = 0.5 and StableDiffusion guidance_scale = 100. For ‘controlnet-zero1to3’ guidance, we use controlnet_conditioning_scale = 0.5 and StableDiffusion guidance_scale = 3. For all experiments, we train with learning rate = 1e-3, instant-ngp enabled, and for 10000 iterations.

4.2 Experiment Setup

We ran the following experiments for comparison:

1. text-only stable-diffusion guidance (labeled as **baseline**)
2. controlnet guidance on user sketch (labeled as **exp1-cn**)
3. controlnet guidance with view emphasis (labeled as **exp2-cn+v**)
4. controlnet-zero1to3 guidance (labeled as **exp3-cn+z**)

We also ran some ablations with varying controlnet_conditioning_scale and StableDiffusion guidance_scale, but did not find any meaningful patterns that outperformed the default values.

5 Results and Discussion

Figure 8 shows results from experiment groups as described in the previous section.



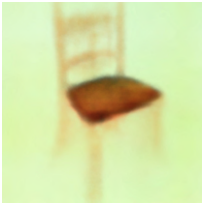


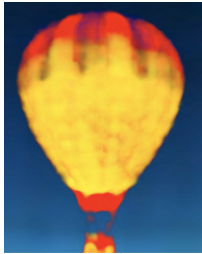
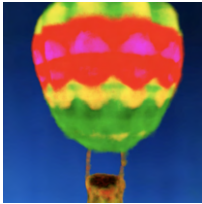



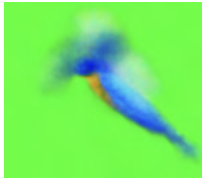


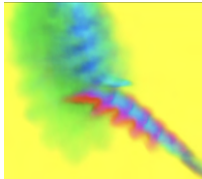
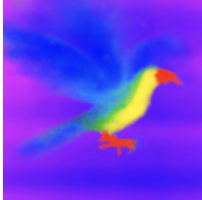
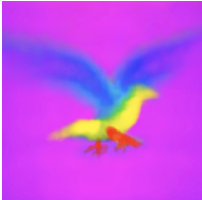


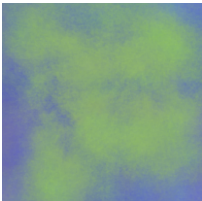

user input	baseline (text-only)	exp1-cn	exp2-cn+v	exp3-cn+z
 "a wooden chair with cushions"				
 "a hot air balloon"				
 "a bird with blue feathers"				
 "a bird with rainbow feathers"				
 "a black horse"				

Figure 8: Experiment results.

baseline

Baseline DreamFusion works well on objects that are more common (chair) or are rotation invariant (balloon). For more complex objects (bird, horse), it tends to have much lower quality. Importantly, it does not take user scribble as guidance and produces its output solely on text.

exp1-cn

ControlNet guidance works well enough without much tweaking, but adding scribble conditioning tends to break down with more complex objects. We see this when the results take longer to converge than usual or doesn't converge at all (horse). However, with some inputs (bird), adding a sketch for conditioning seem to help with result quality.

In all examples, we see that the results does take guidance from the user sketch. In the chair example, we observe the back of the chair add a horizontal bar, and in the balloon example, we observe the horizontal strips of the balloon that did not exist in the baseline.

exp2-cn+v

Adding weighted loss seem to lead to slight improvements, though it is not significantly better, perhaps due to the fact that the primary viewpoint is not sampled often enough. A further exploration could be to force sample the primary viewpoint every N iteration.

exp3-cn+z

The experiments for 'controlnet-zero1to3' guidance produced the highest quality results and was able to converge on objects that previous experiments failed on (horse). However, this method does have the drawback that it relies heavily on the single RGB image that was produced by ControlNet. On the chance that this initial step output a lower quality image, the downstream Zero-1-to-3 model cannot self-correct and is doomed (as shown in the chair example).

6 Conclusion

Controllability of 3d generation is still an open area of research. In our work, we explore one promising approach using ControlNet, which has shown incredibly impressive results in the 2d domain. We have implemented a proof-of-concept conditioning on user sketch, but our findings could easily extend to other flavors of image conditioning, such as depth map or segmentation map. More and more people are using AI as a tool to unlock creativity and productivity. We hope generating 3D models will soon be as accessible as scribbling a pencil drawing on the back of a napkin.

References

- [Mül+22] Thomas Müller et al. "Instant neural graphics primitives with a multiresolution hash encoding". In: *ACM Transactions on Graphics* 41.4 (July 2022), pp. 1–15. DOI: [10.1145/3528223.3530127](https://doi.org/10.1145/3528223.3530127). URL: <https://doi.org/10.1145/3528223.3530127>.
- [Poo+22] Ben Poole et al. *DreamFusion: Text-to-3D using 2D Diffusion*. 2022. arXiv: [2209.14988](https://arxiv.org/abs/2209.14988) [cs.CV].
- [Rom+22] Robin Rombach et al. *High-Resolution Image Synthesis with Latent Diffusion Models*. 2022. arXiv: [2212.10752](https://arxiv.org/abs/2212.10752) [cs.CV].
- [Tan22] Jiaxiang Tang. *Stable-dreamfusion: Text-to-3D with Stable-diffusion*. <https://github.com/ashawkey/stable-dreamfusion>. 2022.
- [Liu+23] Ruoshi Liu et al. *Zero-1-to-3: Zero-shot One Image to 3D Object*. 2023. arXiv: [2303.11328](https://arxiv.org/abs/2303.11328) [cs.CV].
- [ZA23] Lvmin Zhang and Maneesh Agrawala. "Adding conditional control to text-to-image diffusion models". In: *arXiv preprint arXiv:2302.05543* (2023).