Augmenting Minecraft with Fine-Tuned Stable Diffusion Models

by Jenny Zhang

Abstract

This project explores the use of Img2Img Stable Diffusion to create more realistic architecture representations based on input images from Minecraft. The technique is refined using Fast Dreambooth fine-tuning to capture the style of a specific architect for style transfer. Architecture renderings require a lot of time, expertise, and resources to be created from scratch, so this process uses Minecraft as a simplified 3D environment to quickly create and iterate on rough designs and augments these designs with user-defined prompts for stable diffusion.

Methodology

Figure 1. Prompt: “A realistic photo of a residential street, surrounded by grass and flowers in the style of 4 architects”, 1.0 image strength

Testing the original stable diffusion 2.1 model with prompt “a realistic photo of a residential street, surrounded by grass and flowers, 4k, featured on artstation, sunny, octane render, cinematic in the style of X” for various architects X, it was apparent that the model was capable of generating generic buildings for a residential street but was not trained on specific architects or architectural styles.

Figure 2. Prompt: “a realistic photo of a residential street, surrounded by grass and flowers, 4k, featured on artstation, sunny, octane render, cinematic, Zaha Hadid”, 0.5 image strength

Figure 3. Prompt: “a realistic photo of a building, surrounded by grass, 4k, featured on artstation, sunny, octane render, cinematic, Le Corbusier”, 0.5 image strength

To get outputs closer to a distinct architectural style, additional testing was done using dvArch - Multi-Prompt Architecture Tuned Model from CivitAI. This model was fine tuned to produce images with modern, gothic, and victorian architecture, and generated much more realistic buildings with a certain architectural style, but could not be steered towards a specific architect’s style.

The dvArch model and other stable diffusion models are fine-tuned using the Dreambooth technique developed by researchers from Google Research and Boston University in 2022. The process uses 3-5 images of a target subject and associates it with a specific keyword to expand the vocabulary of stable diffusion and allows stable diffusion to generate images of the subject in different contexts.

The implementation of Dreambooth that this project uses to fine-tune stable diffusion for specific architect’s styles is Fast Dreambooth by TheLastBen on github which can be run using Google Colab. After collecting 15-20 images for each of the architects (Zaha Hadid, Le Corbusier, Frank Gehry, and Antoni Gaudi) and preprocessing them to be of size 512x512 (with padding if necessary), Fast Dreambooth was used to fine-tune the stable diffusion 2.1 model with 4 new keywords (sdZahaHadid, sdLeCorbusier, sdFrankGehry, sdAntoniGaudi). The model was

trained with 1800 U-Net training steps with a learning rate of 2e-6 and 1800 text encoder steps with a learning rate of 1e-6 and took around 30 minutes per architect.

Results

The fine tuned model was tested on screenshots of Minecraft recreations of Villa Savoye by Le Corbusier and the Stata Center by Frank Gehry from the Minecraft replica of MIT Campus created by MIT undergrads in 2020. Each image was first evaluated with the original architect to confirm that the original architect's style was being successfully emulated, and then tested with the other architect keywords to envision what the building could potentially look like in a different architect's style. Each image uses the same prompt + architect keyword and the different outputs are generated due to the random seeds that were used.

Figure 4. Prompt: "sdLeCorbusier a photo of a building surrounded by green grass", 0.5 image strength. Top Left: original image + 5 different output images with the same prompt

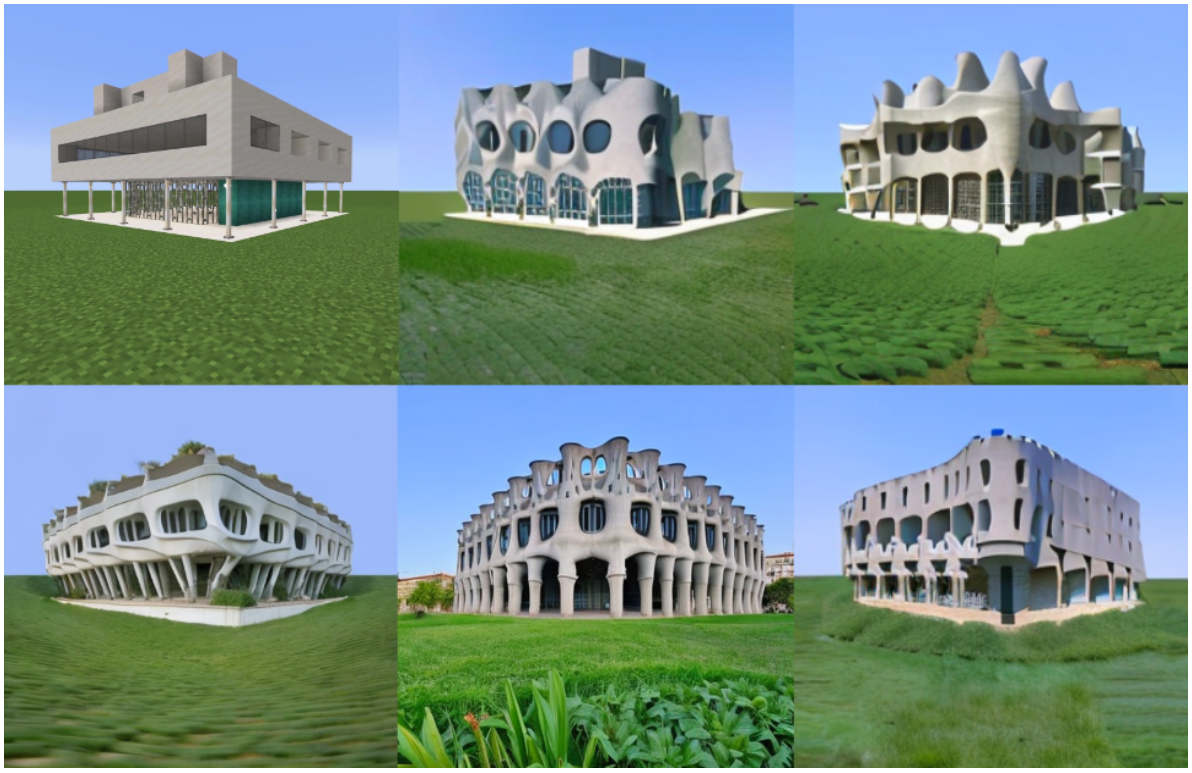

Figure 5. Prompt: “sdAntoniGaudi a photo of a building surrounded by green grass”, 0.5 image strength. Top Left: original image + 5 different output images with the same prompt

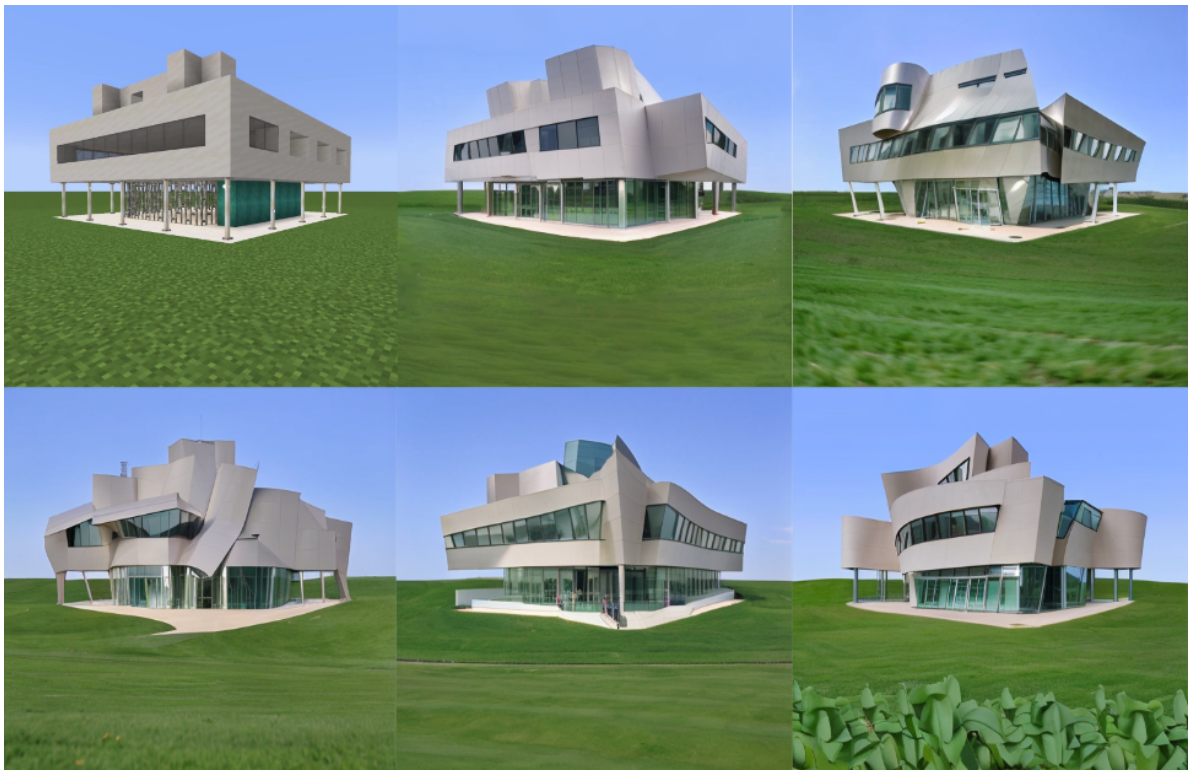

Figure 6. Prompt: “sdFrankGehry a photo of a building surrounded by green grass”, 0.5 image strength. Top Left: original image + 5 different output images with the same prompt, 0.5 image strength

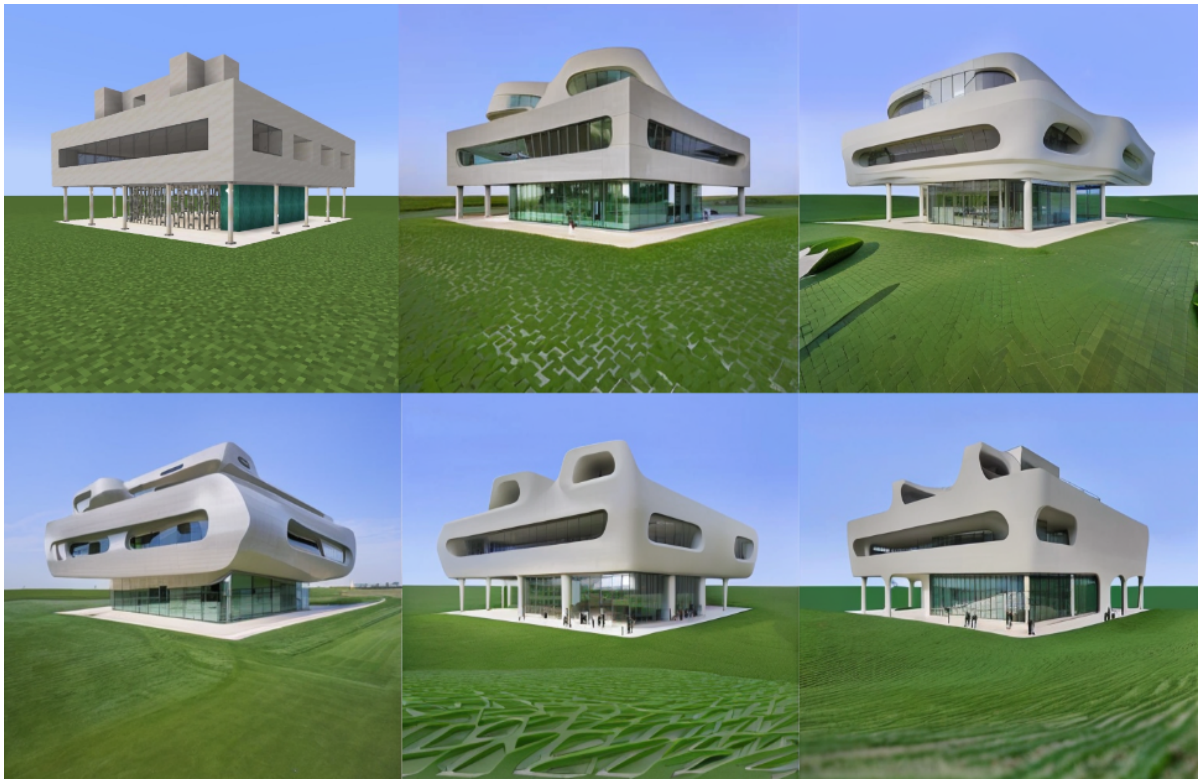

Figure 7. Prompt: “sdZahaHadid a photo of a building surrounded by green grass”, 0.5 image strength. Top Left: original image + 5 different output images with the same prompt

Figure 8. Prompt: “sdFrankGehry a photo of a building”, 0.5 image strength.
Top Left: original image + 5 different output images with the same prompt

Figure 9. Prompt: “sdZahaHadid a photo of a building”, 0.5 image strength
Top Left: original image + 5 different output images with the same prompt

Figure 10. Prompt: “sdLeCorbusier a photo of a building”, 0.5 image strength
Top Left: original image + 5 different output images with the same prompt

Figure 11. Prompt: “sdAntoniGaudi a photo of a building”, 0.5 image strength
Top Left: original image + 5 different output images with the same prompt

Conclusion

Stable Diffusion can augment architecture in Minecraft in a creative and meaningful way while keeping original image structure at image strength 0.5 or less. Fine-tuned models created using FastDreambooth were able to capture unique features of each of the four architects and apply them to buildings designed by other architects with 15-20 images and a relatively short training time. However, it is easy for this fine-tuned model to be overfit due to the small dataset, so it could be made more robust and consistent with more diverse images with different lighting conditions, different camera angles. The preliminary results are quite visually impressive and the fine-tuned models have the potential to be used as a brainstorming/inspiration tool as users can quickly iterate in 3D space in Minecraft Creative Mode.

Future Steps

1. More Dream Booth training for more architects/general architecture styles
2. Exploration of latent space to find intermediate architectural representations between architectural styles
3. Exploration of denoising process for better camera view consistency