Architectural DigitalFUTURES (ADF) 2023
Computation and Formation (AI Design)

Text Semantics to Image Generation: A method of building facades design base on Stable Diffusion model

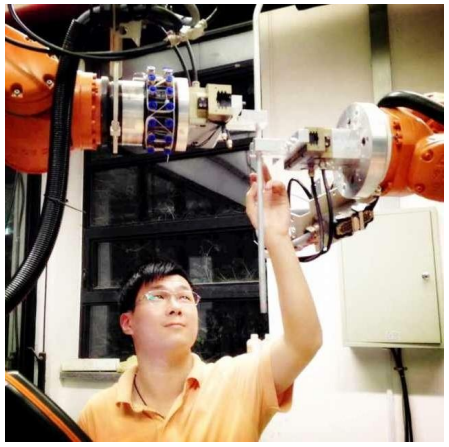
Haoran Ma / *School of Design, Jiangnan University, Wuxi, China*
Hao Zheng / *Department of Architecture and Civil Engineering, City University of Hong Kong, HKSAR, China*

The 5th International Conference on Computational Design and Robotic Fabricaion (CDRF 2023)

Author Introduction

Haoran Ma ***Master Student***
School of Design, Jiangnan University, Wuxi, China

6210307146@stu.jiangnan.edu.cn



Hao Zheng ***Professor (Assistant)***
Department of Architecture and Civil Engineering, City University of Hong Kong, HKSAR, China

hazheng@cityu.edu.hk

CONTENT

1 Introduction

- **Unimodal Image Generation**

Most research use generative adversarial networks (GAN), which produce building facades (Isola et al., 2016) and layouts (Huang and Zheng, 2018), to apply machine learning to generative design (Pix2Pix HD, cycleGAN, DCGAN).

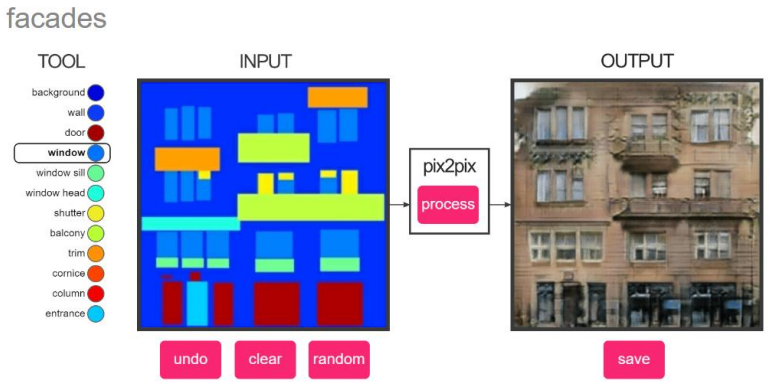


Fig. 1. Generates building facade. (Phillip et al., 2016)

Advantages

- Lightweight, and the model design is based on the task, which is more targeted.

Disadvantages

- A large number of samples are required for training.
- Weak migration ability, pictures outside the dataset cannot be generated.
- Single-mode control conditions. For example, pictures are generated from pictures.

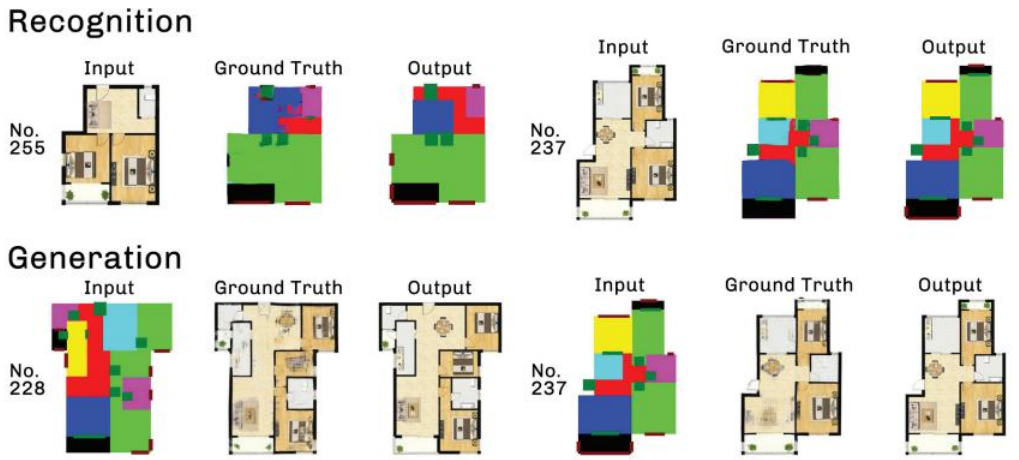


Fig. 2. Generates building plans. (Huang and Zheng, 2018)

1 Introduction

- **Cross-modal image generation**

Multi-modal task processing has become a hot area of research in recent years, including text-to-image generation. These techniques for creating images to text have been applied broadly in architectural design.

Advantages

- Generate content across modalities with strong migration capabilities.
- The adjustment of output results is more flexible and efficient.

Disadvantages

- Large model training is difficult and requires a lot of computing resources.
- The generation result of the extended mode is strong in randomness and poor in control ability.

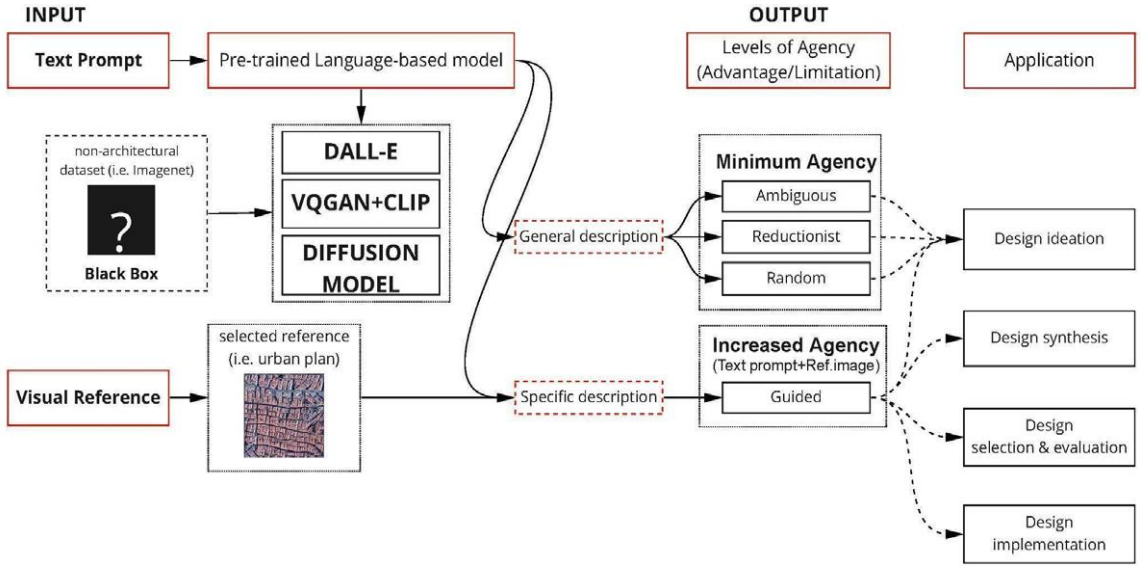


Fig. 2. Extended modal urban texture generation. (Bolojan et al., 2022)

STABLE DIFFUSION

Midjourney

DALL·E 2

1 Introduction

Large sizable diffusion model have **poor adaptation to tasks** requiring the creation of building facades, and it is typically challenging to regulate the training and generation results (Ruiz et al., 2022).

- **This research starts with the Stable Diffusion model, utilizes the LoRA approach to refine the model, trains on the building facade dataset, and then integrates the ControlNet model to control the generated results to accomplish the accuracy and controllability of the generated results.**

- **This will provide an easier creative tool for architects to generate a large number of controllable building facade design results by changing a few prompt words.**

2.1 Network Architecture

- **Variational autoencoders (VAE)**: Encoder-preserves significant deep picture features and transforms the image into a low-dimensional latent space representation for U-Net. Decode-Creating images from representations in the latent space.
- **U-Net**: U-Net is a residual module-based encoder and decoder that decodes low-resolution images into high-resolution images after the encoder compresses the images as the predicted noise in latent sapce.
- **Text-Encoder**: which translates the tagged sequence to a potential text embedding sequence, transforms the input text into a meaning that U-Net can comprehend and uses to direct the model as it denoises the embedding.

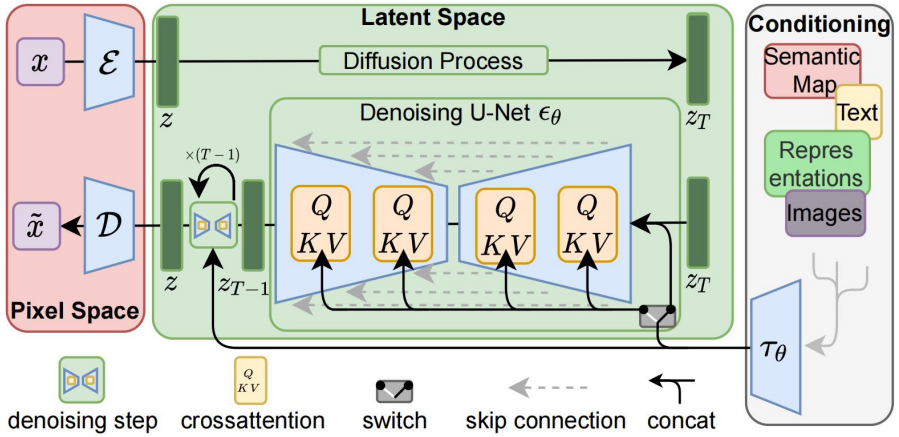


Fig. 1. Schematic diagram of Stable Diffusion

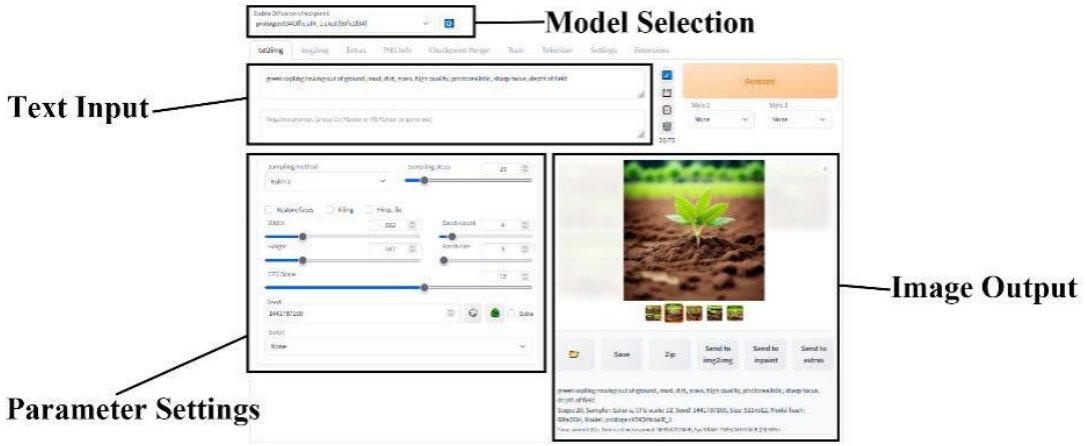


Fig. 2. Stable Diffusion Web-UI interface

2.2 LoRA and ControlNet

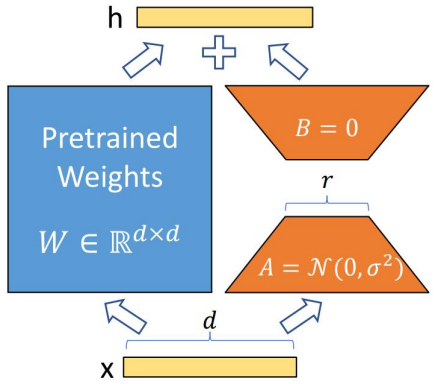


Fig. 3. LoRA schematic diagram

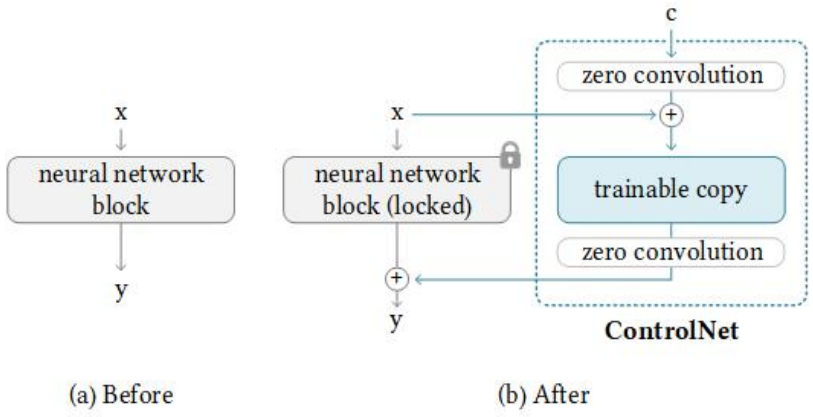


Fig. 4. ControlNet schematic diagram

- **LoRA (Low-Rank Adaptation)**

Traditional methods require high computing power of the graphics card, such as Textual Inversion or Dreambooth. The LoRA method injects the trainable layer instead of the pre-trained model weight in each Transformer block, drastically reducing the number of training parameters.

LoRA fine-tuning is quicker and less computationally intensive while maintaining the same level of quality as full-model fine-tuning.

- **ControlNet Model**

On the other hand, the diffusion model generates text and images in a highly random manner, making it challenging to manage the outcome. Furthermore, it can be challenging to precisely regulate the final generated content given the information provided in the text.

It is now much easier to regulate the diffusion model's strong randomness generation results. Many control conditions are included in ControlNet, including Canny Edge, and Segmentation Map, etc.

2.3 Training Process

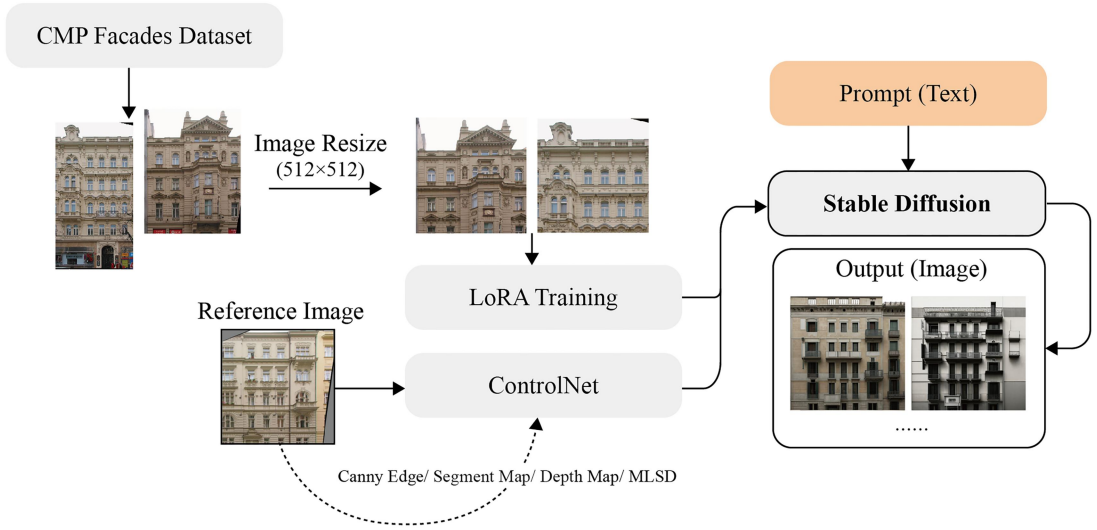
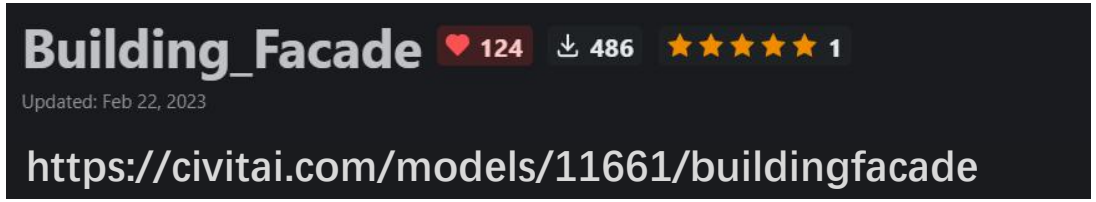


Fig. 5. Experimental Workflow



- **Prepare Dataset**

200 images from the CMP Facades dataset (Tylecek, 2012) are initially chosen at random to serve as training samples. And then resized into 512×512 pixel. We use the CLIP model for textual reasoning on building facades. These texts serve as trigger words.

- **LoRA Training**

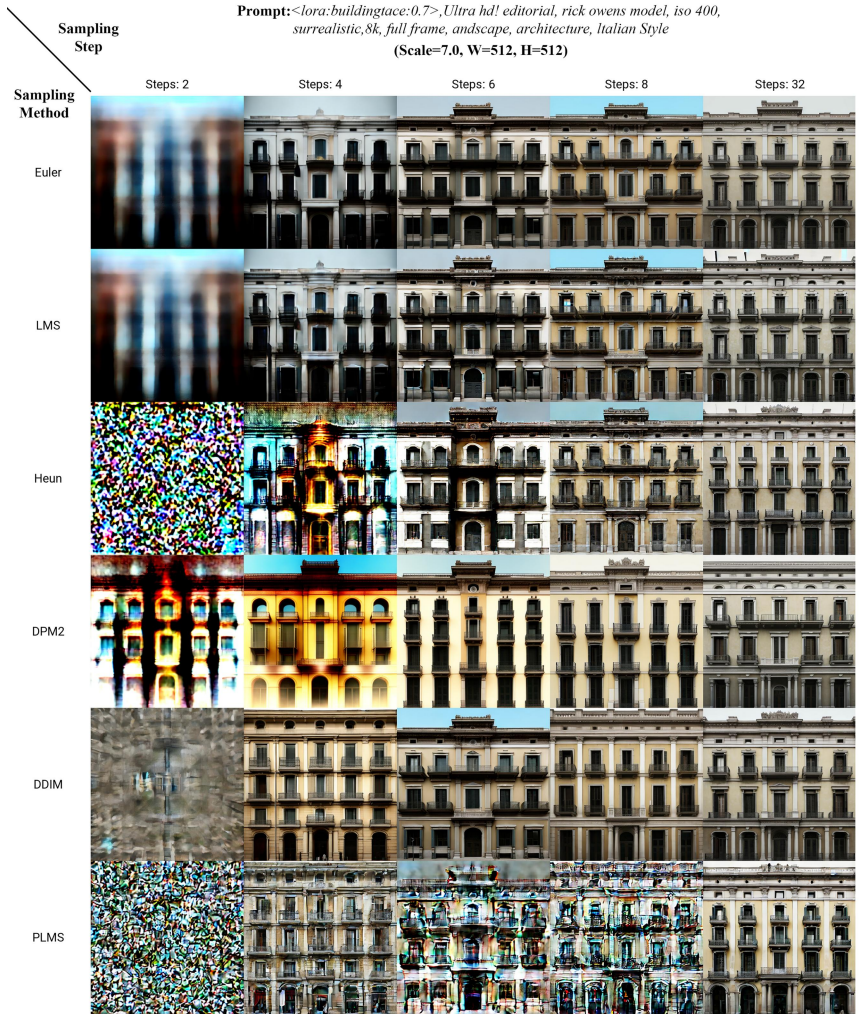
Stable Diffusion v1-4 (Rombach et al., 2021) was selected as the base model, and it was adjusted on an NVIDIA RTX 2060 with 6GB of memory (*Epoch=1, Batch Size=20000, Learning Rate=0.00001*), taking more than 2 hours to complete (*Output model size=144MB*).

- **Imgae Generation**

we utilized the model supplied by (Zhang and Agrawala, 2023) for the ControlNet model as the control condition to generate building facades through Stabel Diffusion model.

3 Results

3.1 Generation with Different Style Semantic Base on LoRA



Prompt: *(<lora: buildingface:0.7>, Ultra hd! editorial, rick owens model, iso 400, surrealistic, 8k, full frame, landscape, architecture, Italian Style)*

- **Euler** and **LMS** approaches produce similar content at each sampling step.
- **Heun** method is similar to the **PLMS** method. Until the content of the fourth step starts to emerge.
- Also, the results are remarkably similar despite the fact that **DPM2** and **DDIM** use distinct algorithm. The DPM2 sampling method, which has the highest tag use rate at over 80%, serves as the foundation for the follow-up study (Rombach et al., 2021).

Fig. 6. Generation results of different sampling methods and sampling steps

3.1 Generation with Different Style Semantic Base on LoRA

- **Generation with Different Style Semantic**

We tried the generation effect of different style semantics in the fine-tuned Stable Diffusion model. (Italian Style/French Style/Rococo Style/New Chinese Style/Modern Style)

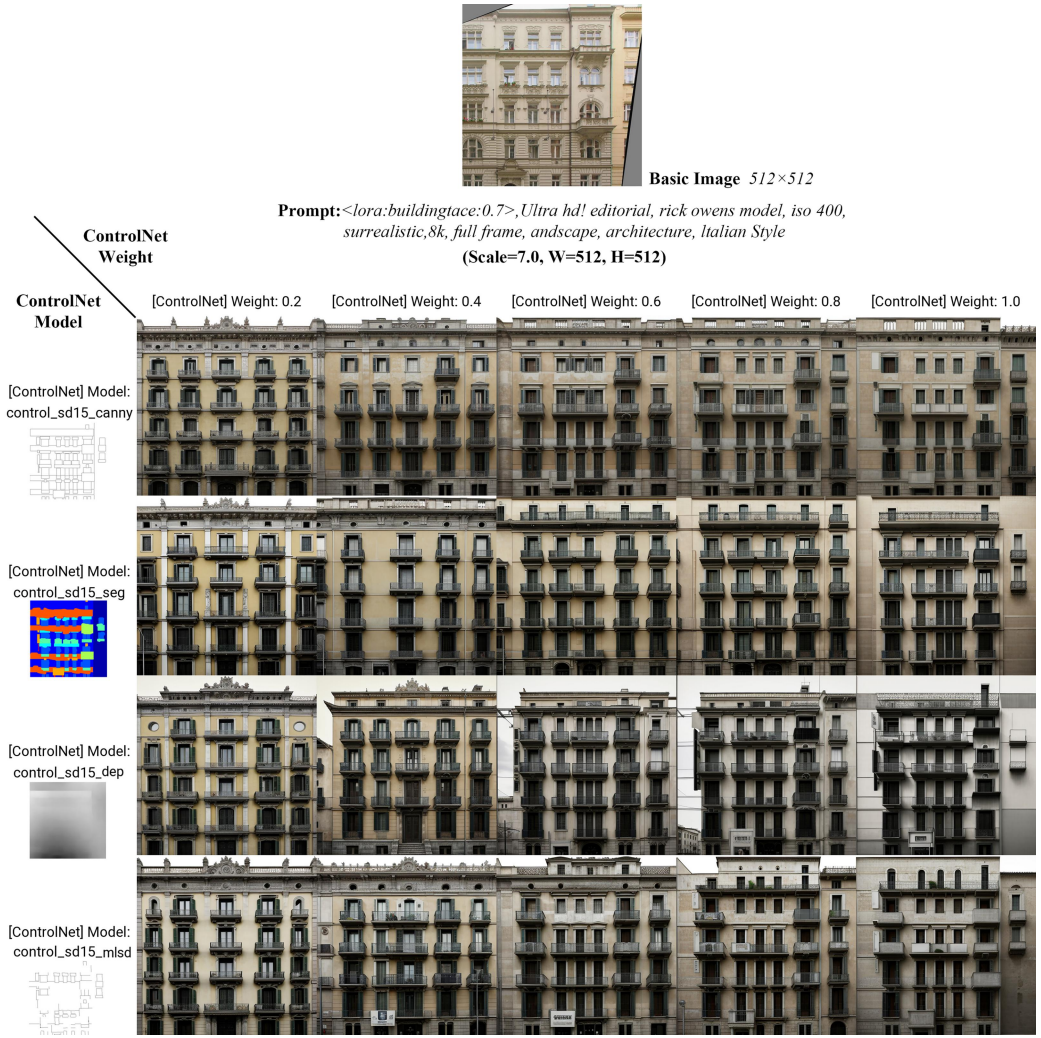
- **Differences from Original Dataset**

On the other hand, utilizing a fine-tuned LoRA model based on Stable Diffusion can produce content that is entirely different from the original dataset and offers a wide range of adjusting options.

By quickly generating numerous designs in various styles and types for building facades with just text input, this technique to facade design for buildings is more effective.

Fig. 7. Generation results of different style semantic

3.2 Generation of Different Control Model Base on ControlNet



- **Comparison of Control Conditions**

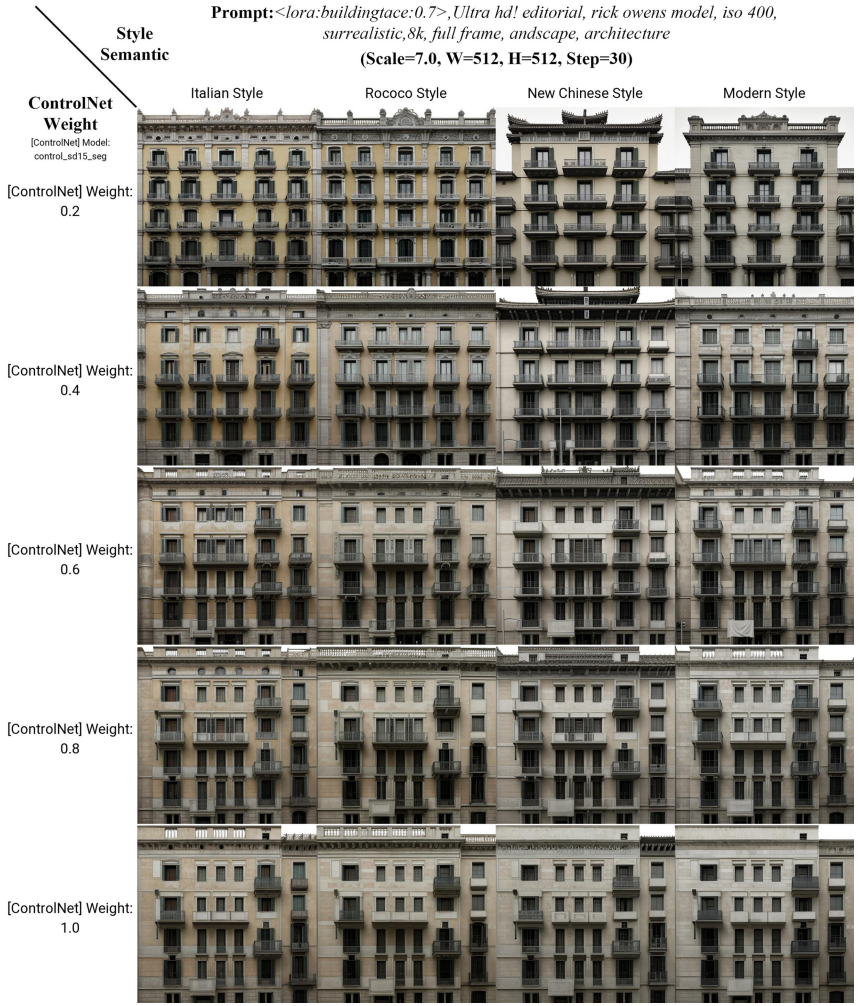
The generated results are affected differently by the various ControlNet control models, with the **Canny Edge model** producing more results than the Segment Map and MLSD models. The depth map model's output has a better sense of spatial orientation.

- **Comparison of ControlNet Weights**

Fewer ControlNet weight values produce more varied results under the same conditions. The building structure gets more similar to the reference object as the weight value rises, while the building facade has less detail. Increasing the weight value, in other words, restricts the machine's reasoning.

Fig. 6. Generation results of different ControlNet Model and Weight

3.2 Generation of Different Control Model Base on ControlNet



We also generated various building facade styles using **ControlNet's Canny Edge model**, comparing the effects of various weight values on the outcomes.

- When the **weight value of ControlNet increases**, the architectural style gradually unifies and the building facade's elements become more condensed.
- When the **weight value is set to 1.0**, the large eaves feature nearly completely vanishes, although the upper right corner still contains some content.

In general, ControlNet may be used to successfully manage the consistency of the results that are created and the reference images, but more building facade features are sacrificed.

The ideal range for ControlNet's weight value is between 0.6 and 0.8.

Fig. 6. Generation results of different ControlNet Weight and style semantic

3.3 Final Generation Experiments

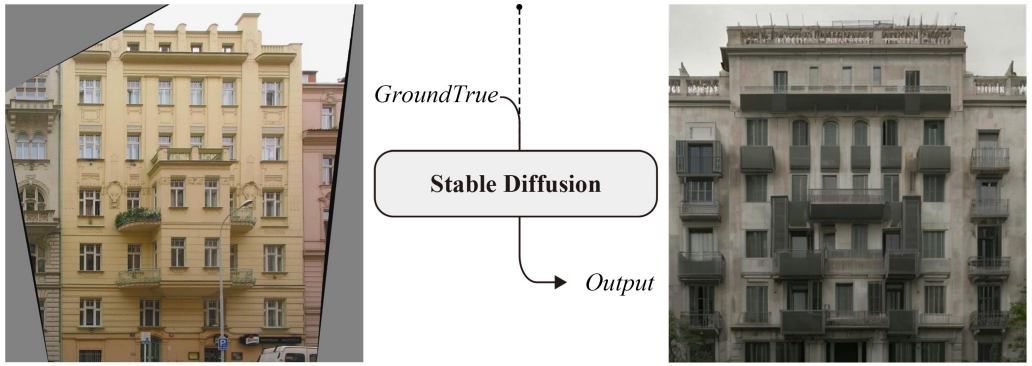
We used the best parameters in our migration experiments.

We tried to get the model to generate a **Modern Building Facade Style**, and the stable diffusion model fine-tuned using LoRA understood exactly what we wanted to get. Not only that, but the building facade remained consistent with the architecture of the reference image under the control of the CotrolNet model, and this process took only 0.2 seconds. We then added the words "white and chrome" to the prompt and the model outputted a white facade based on the text.

By simply adding text, it was possible to quickly obtain a different output. This will provide architects with a more efficient concept output.

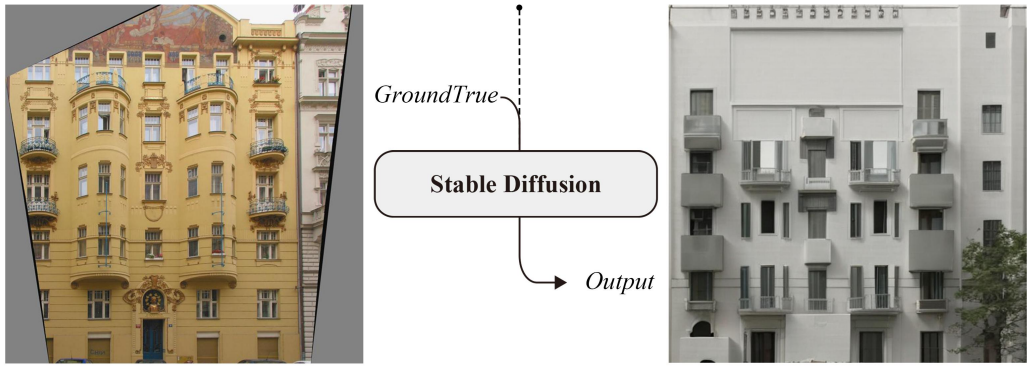
Prompt: *<lora:buildingface:1>,Ultra hd! editorial, rick owens model, iso 400, surrealistic, 8k, full frame, landscape, modern architecture*

(Scale=7.0, W=512, H=512, Step=20, ControlNet Weight=0.75)



Prompt: *<lora:buildingface:1>,Ultra hd! editorial, rick owens model, **white and chrome**, iso 400, surrealistic, 8k, full frame, landscape, modern architecture*

(Scale=7.0, W=512, H=512, Step=20, ControlNet Weight=0.75)



The results of our experiments have been presented in the Civitai community

3.3 Final Generation Experiments

4 Conclusion and Discussion

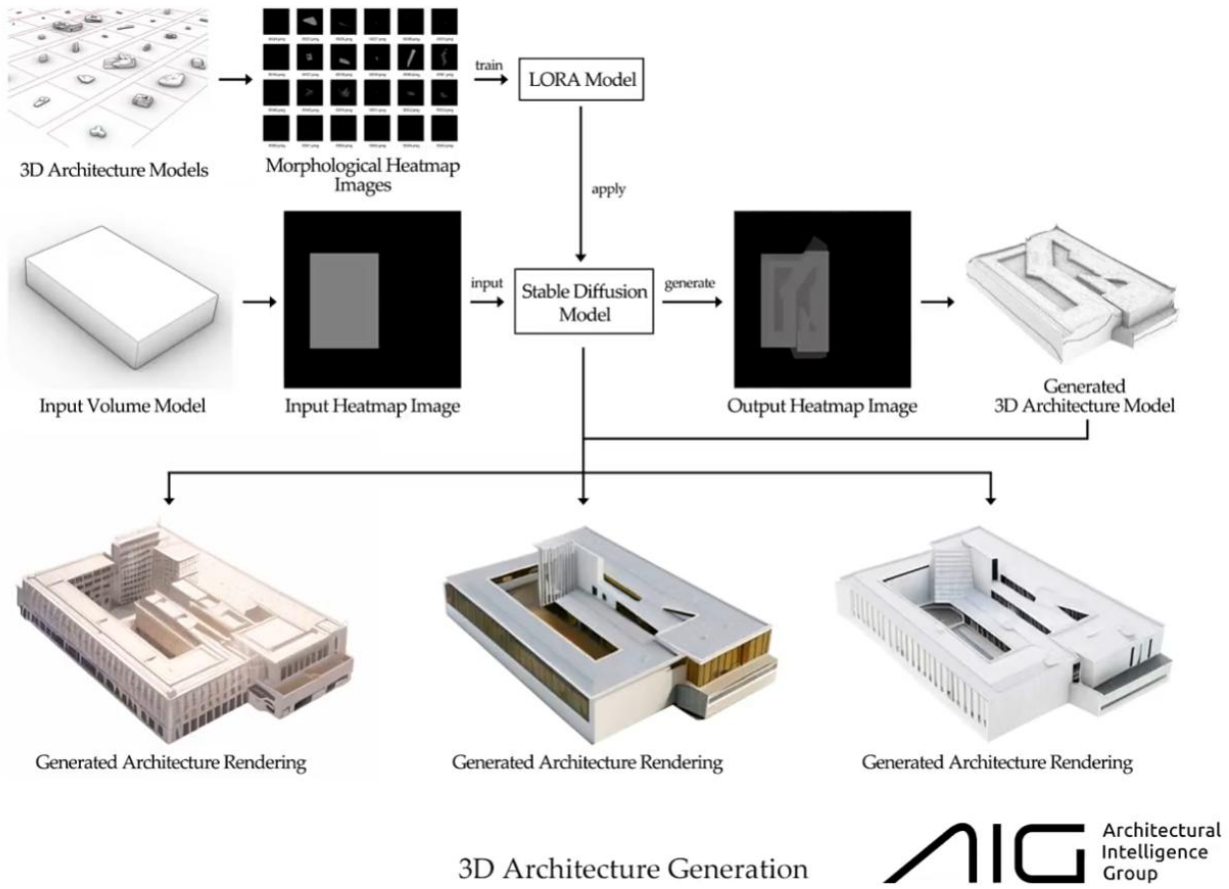
Conclusion

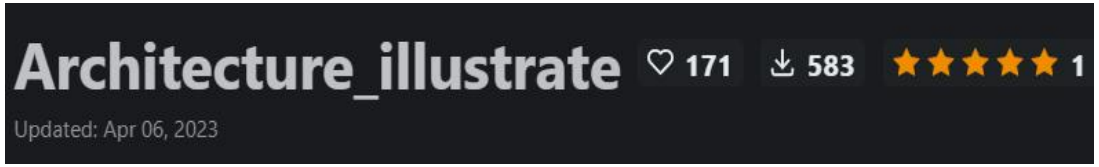
- The fine-tuning training of the Stable Diffusion model using the LoRA model reduces the computational power needs of the graphics card and saves a significant amount of time.
- The Stable Diffusion model that has been fine-tuned using LoRA is very flexible to tasks involving building facades, and the semantic characteristics of various styles can be effectively included into the outcomes produced.
- ControlNet can be used to effectively control the building facade generation results' consistency with the reference object structure, but too much model weight would reduce diversity of results.

Future Research

- And future research could **combine morphological generative algorithms** with AI to produce more accurate and richer results.
- According to further research, training can be done on **a cloud computing platform** with more powerful processing capacity.
- Second, the prompt's input can be improved further, **providing more details prompt** may result in the production of more high-quality building facade content.

5 Extended works





AI generated 3D building model.

Architectural DigitalFUTURES (ADF) 2023

Computation and Formation (AI Design)

Thanks.

Haoran Ma / *School of Design, Jiangnan University, Wuxi, China*
Hao Zheng / *Department of Architecture and Civil Engineering, City University of Hong Kong, HKSAR, China*

The 5th International Conference on Computational Design and Robotic Fabricaion (CDRF 2023)