Text Semantics to Image Generation: A Method of Building Facades Design Base on Stable Diffusion Model

Haoran Ma[1] and Hao Zheng[2](✉)

[1] School of Design, Jiangnan University, Wuxi, China
6210307146@stu.jiangnan.edu.cn
[2] Department of Architecture and Civil Engineering, City University of Hong Kong, HKSAR, China
hazheng@cityu.edu.hk

Abstract. Stable Diffusion model has been extensively employed in the study of architectural image generation, but there is still an opportunity to enhance in terms of the controllability of the generated image content. A multi-network combined text-to-building facade image generating method is proposed in this work. We first fine-tuned the Stable Diffusion model on the CMP Facades dataset using the LoRA (Low-Rank Adaptation) approach, then we apply the ControlNet model to further control the output. Finally, we contrasted the facade generating outcomes under various architectural style text contents and control strategies. The results demonstrate that the LoRA training approach significantly decreases the possibility of fine-tuning the Stable Diffusion large model, and the addition of the ControlNet model increases the controllability of the creation of text to building facade images. This provides a foundation for subsequent studies on the generation of architectural images.

Keywords: Architecture generation · Stable diffusion · LoRA · ControlNet

1 Introduction

Artificial intelligence has entered a new period of integration as of the twenty-first century. Machine learning, the foundational technology of artificial intelligence, is also the attention of architects. 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. These studies demonstrate that GAN trained on labeled samples is very adept at learning the shape of architectural features and their position arranged in building faces and planes. Supervised GANs like Pix2Pix HD (Park et al. 2019), and cycleGAN (Zhu et al. 2017) require conditional input during both training and inference. However, it is challenging to transition between tasks using this GAN model, which was trained using specific samples. For instance, creating structures with various architectural styles necessitates training several models. Moreover, the sample size is the key impediment. Even while unsupervised models like DCGAN

C. Yan et al. (Eds.): CDRF 2023, *Phygital Intelligence*, pp. 24–34, 2024.
https://doi.org/10.1007/978-981-99-8405-3_3

(Radford et al. 2015) can train a lot of samples, handling downstream tasks is still challenging.

Multi-modal task processing has become a hot area of research in recent years, including text-to-image generation, as the drawbacks of training samples have been greatly reduced. The Stable Diffusion model (Rombach et al. 2021) is a model for text-to-image generating tasks and creates detailed images from text descriptions. Midjourney (Borji 2022) and DALLE 2 (Ramesh et al. 2022) are models that are comparable to this one. These techniques for creating images to text have been applied broadly in disciplines like architectural design. In the AI Spring series of courses co-organized by DigitalFUTURES & FIU DDes in 2022, the application of text-to-image advanced technology in the field of architecture will be discussed.

However, 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). Hence, 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.

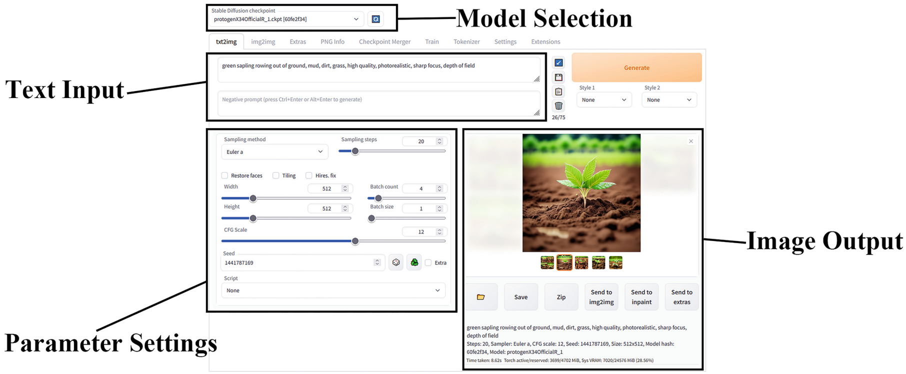
2 Methodology

2.1 Network Architecture

Stable Diffusion is a Text-to-Image generation technique based on Latent Diffusion Models (LDMs) (Rombach et al. 2021). It can generate better outcomes for image generation than the GAN model. Such as Unconditional image synthesis, image restoration (Inpainting), Super-resolution, Text-to-image generation, etc., random Gaussian noise can be gradually denoised after training. There are three essential parts to the stable diffusion model: (1) Variational autoencoders (VAE), which include both an encoder and a decoder. The first preserves significant deep picture features and transforms the image into a low-dimensional latent space representation for U-Net. The latter creates images from representations in the latent space. (2) 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. (3) 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. In order to facilitate model loading and image generation, this paper uses Stable Diffusion Web-UI as the control system (Fig. 1). The Web-UI enables Stable Diffusion to have a more intuitive user interface and integrates Text-to-Images, Super-Resolution and model training function.

2.2 LoRA and ControlNet

Microsoft researchers have developed a new technology called Low-Rank Adaptation of Big Linguistic Models (LoRA), which is primarily used to address the issue of large



Stable Diffusion checkpoint: protogenX34OfficialR_1.ckpt [60fe2f34]

green sapling rowing out of ground, mud, dirt, grass, high quality, photorealistic, sharp focus, depth of field

Negative prompt (press Ctrl+Enter or Alt+Enter to generate)

Style 1: None
Style 2: None

Sampling method: Euler a
Sampling steps: 20
□ Restore faces □ Tiling □ Hires. fix
Width: 512
Batch count: 4
Height: 512
Batch size: 1
CFG Scale: 12
Seed: 1441787169 □ Extra
Script: None

green sapling rowing out of ground, mud, dirt, grass, high quality, photorealistic, sharp focus, depth of field
Steps: 20, Sampler: Euler a, CFG scale: 12, Seed: 1441787169, Size: 512x512, Model hash: 60fe2f34, Model: protogenX34OfficialR_1
Time taken: 8.62s Torch active/reserved: 3699/4702 MiB, Sys VRAM: 7020/24576 MiB (28.56%)

Fig. 1. Stable Diffusion Web-UI interface

model fine-tuning (Hu et al. 2021). The entire Stable Diffusion model used to be slow and challenging to adjust. Although techniques for lightweight large-scale model fine-tuning, like Textual Inversion or Dreambooth, are growing in popularity, the graphics card's computational power is still quite demanding. Because the model weight is not necessary to calculate the gradient, 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.

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. The recently released ControlNet model addresses this issue by controlling the picture production outcomes by adding more conditions to the Stable Diffusion model (Zhang and Agrawala 2023). As a result, 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.

In this study, we use LoRA to optimize the Stable Diffusion big model and train it on the CMP Facades dataset, and then we apply various ControlNet model conditions to regulate the development of building facades.

2.3 Training Process

In this study, 200 images from the CMP Facades dataset (Tylecek 2012) are initially chosen at random to serve as training samples (LoRA fine-tuning training requires very few samples, and the results are excellent). These 200 images are then resized into 512 × 512 pixel. Next, use the text data from each image as the training set for the trigger words. 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 h to complete. The model eventually produced a size of 144MB (the file has been opened in the Civitai community, https://civitai.com/models/11661/buildingfacade). Second, we utilized the model supplied by (Zhang and Agrawala 2023) for the ControlNet model (Fig. 2).

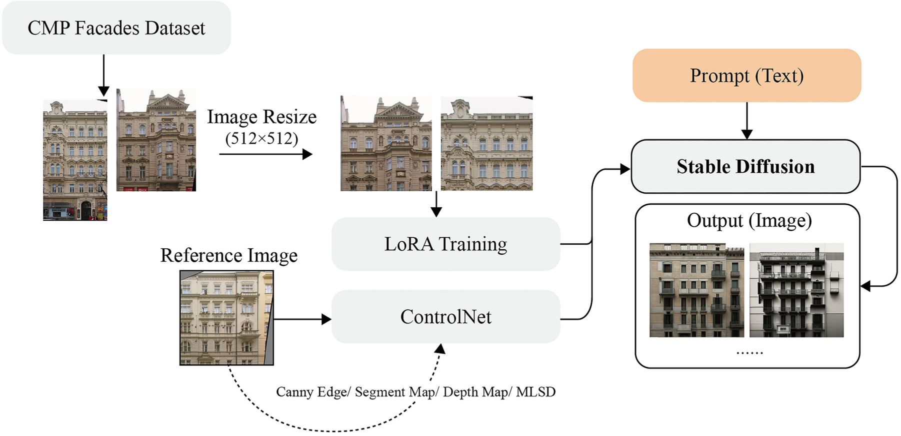


Fig. 2. Experimental Workflow

3 Results

3.1 Generation with Different Style Semantic Base on LoRA

As shown in Fig. 3, we enter a prompt to compare and analyze different image sampling methods and sampling steps (<lora: buildingface:0.7>, Ultra hd! editorial, rick owens model, iso 400, surrealistic, 8k, full frame, landscape, architecture, Italian Style). The prompt's representation of the LoRA model is <lora: buildingface:0.7>, where 0.7 stands for the model's weight. The CFG Scale was set to 7. The degree of influence the text has on the results generated increases with decreasing CFG Scale value, while unpredictability increases with decreasing CFG Scale value. The number of sampling steps is also set at 2, 4, 6, 8, and 32.

We discovered that while the Euler and LMS approaches produce similar content at each sampling step, each sampling method creates distinct content at various sample steps. Heun method is similar to the PLMS method. Until the content of the fourth step starts to emerge, the noise content in the second sampling step is random. It should be noted that stages 6 through 8 of the PLMS method alter at random. Also, the results are remarkably similar despite the fact that DPM2 and DDIM use distinct sampling methods. 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).

Then we tried the generation effect of different style semantics in the fine-tuned Stable Diffusion model, and the parameter settings were the same as those in Sect. 3.1. Figure 4 shows how the created building facade style adapts to the features of the prompt when the architectural style is changed (for example, to Italianate). In the case of the

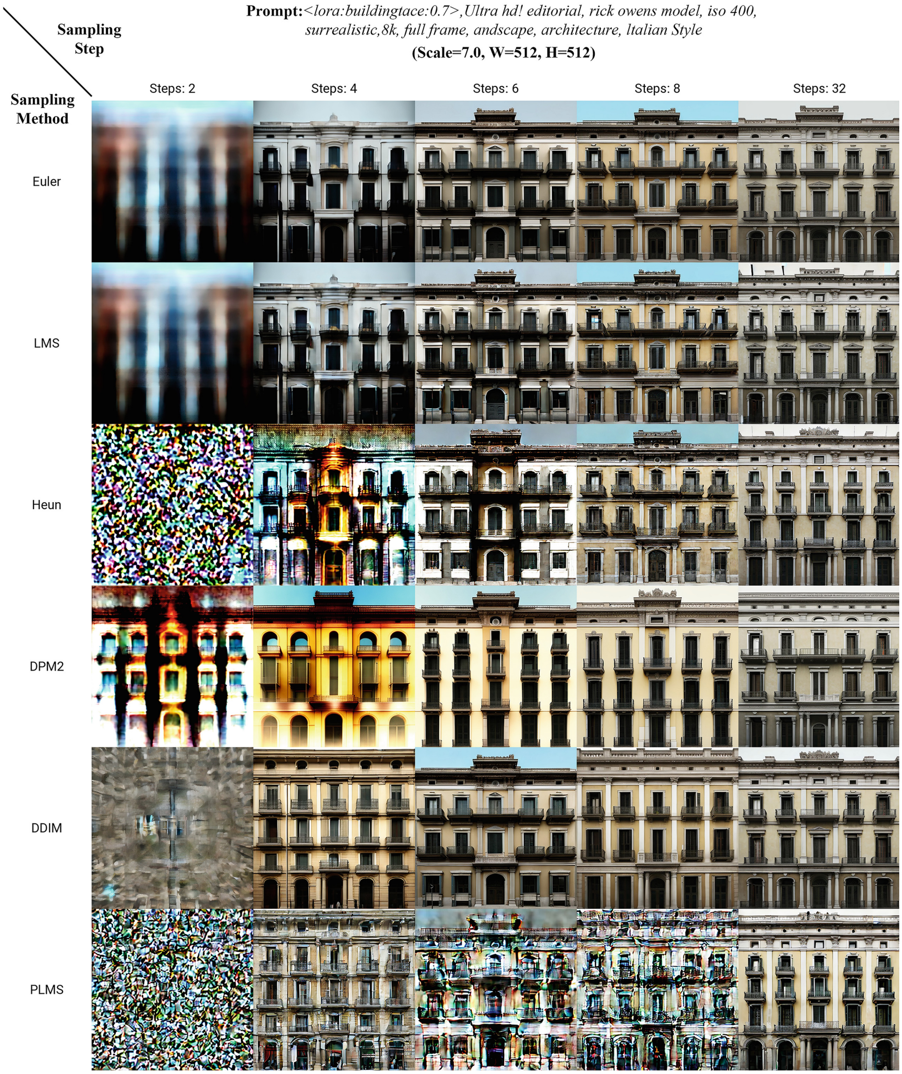

Prompt:<lora:buildingtace:0.7>,Ultra hd! editorial, rick owens model, iso 400, surrealistic,8k, full frame, andscape, architecture, ltalian Style
(Scale=7.0, W=512, H=512)
Sampling Step
Sampling Method
Steps: 2
Steps: 4
Steps: 6
Steps: 8
Steps: 32
Euler
LMS
Heun
DPM2
DDIM
PLMS

Fig. 3. Generation results of different sampling methods and sampling steps

new Chinese style, the model generates beautiful Chinese stone railings in the base and captures the features of the large eaves of Chinese architecture. In the treatment of the windows and doors, the complicated decorative lines are eliminated and a supporting scene with pine trees is created in the front. Despite their strong similarities, the Italian, French and Rococo styles each have their own distinctive features. The French style facade has a classical form with carvings and lines in the details, while the Italian style facade has a window frame with elaborate carvings. The Rococo facade has elaborate ornamentation. With the absence of intricate carving and multiple layers of decorative lines, the Modern style facade is the most understated.

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.

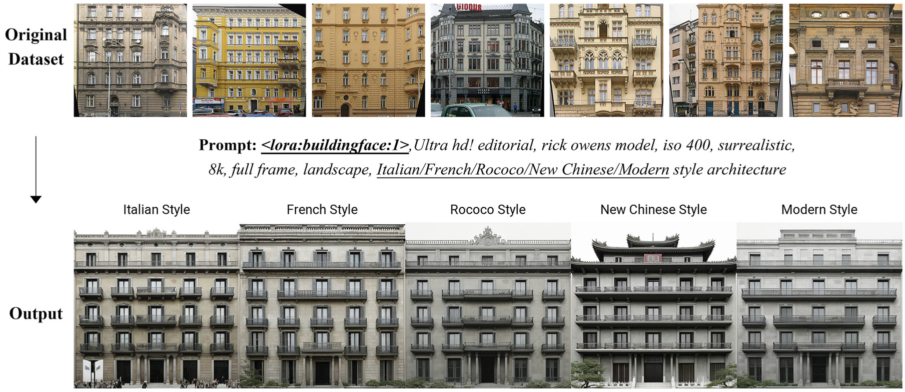

Original Dataset
Prompt: <lora:buildingface:1>,Ultra hd! editorial, rick owens model, iso 400, surrealistic, 8k, full frame, landscape, Italian/French/Rococo/New Chinese/Modern style architecture
Italian Style
French Style
Rococo Style
New Chinese Style
Modern Style
Output

Fig. 4. Generation results of different style semantic

3.2 Generation of Different Control Model Base on ControlNet

The diffusion model from text to image has a significant degree of unpredictability when no control constraints are added. In this part, we add generative conditions to the diffusion model using the ControlNet model. Several control models are offered by ControlNet. In this article, we primarily employ the Canny Edge, Segment Map, Depth Map, and MLSD models to provide control conditions for the production of building facades, and we compare the generation of different ControlNet weights (0.2, 0.4, 0.6, 0.8, and 1.0) results.

As can be seen in Fig. 4, the reference base image of the ControlNet model is a building facade with 512 x 512 pixels. The prompt input and parameter settings are the same as in Sect. 4.1. The results show that the canny edge model produces the best results when the ControlNet weight is set to its maximum ($W = 1.0$), preserving the facade layout of the reference image while also taking into account the requirement for prompts. The elevation layout of the reference image was less similar to the results generated by the other models in the same conditions. For example, neither the layout pattern of the reference image nor the artistic requirements of the prompt were maintained in the results produced by the depth map model. The degree of similarity between the generated outputs is particularly high when the weights of the ControlNet model are relatively low ($W = 0.2$), even though the two models ultimately produce different results. However, when the weights were close to 0.4, the model outputs showed markedly different results.

In general, (1) 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. (2) 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.

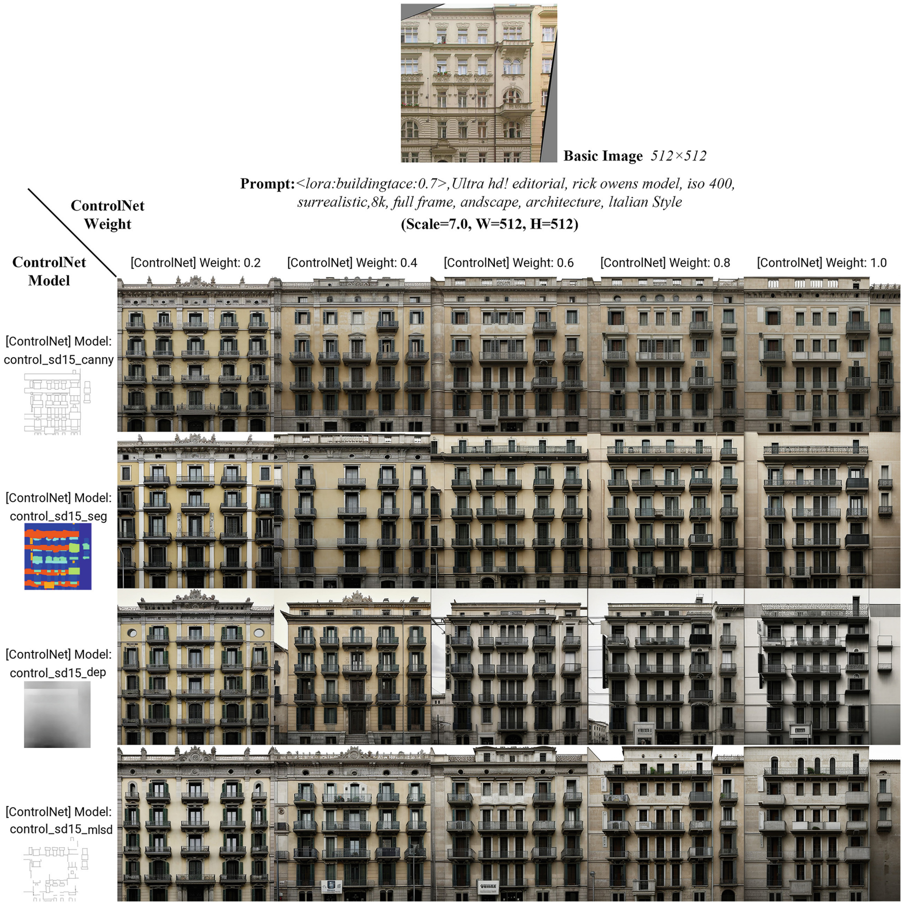

Basic Image 512×512
Prompt:<lora:buildingtace:0.7>,Ultra hd! editorial, rick owens model, iso 400, surrealistic,8k, full frame, andscape, architecture, ltalian Style
(Scale=7.0, W=512, H=512)
ControlNet Weight
ControlNet Model
[ControlNet] Weight: 0.2
[ControlNet] Weight: 0.4
[ControlNet] Weight: 0.6
[ControlNet] Weight: 0.8
[ControlNet] Weight: 1.0
[ControlNet] Model: control_sd15_canny
[ControlNet] Model: control_sd15_seg
[ControlNet] Model: control_sd15_dep
[ControlNet] Model: control_sd15_mlsd

Fig. 5. Generation results of different ControlNet Model and Weight

We also generated various building facade styles using ControlNet's Canny Edge model, comparing the effects of various weight values on the outcomes. As seen in Fig. 5, when the weight value of ControlNet increases, the architectural style gradually unifies and the building facade's elements become more condensed. For instance, in the New Chinese style, when the weight value is more than 0.4, some elements are still present but the huge eaves' characteristics progressively fade. 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).

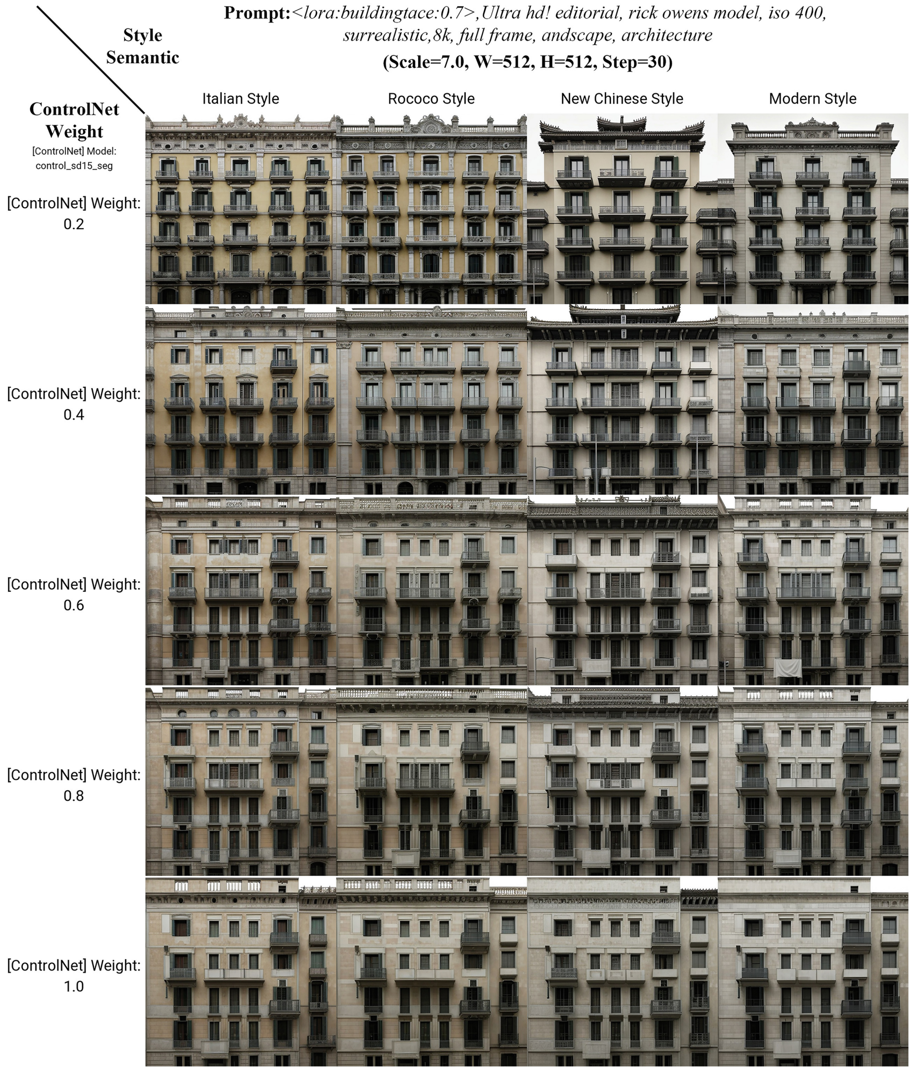


Fig. 6. Generation results of different ControlNet Weight and style semantic

3.3 Final Generation Experiments

We used the best parameters in our migration experiments. As shown in Fig. 7, 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 s. 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. The results of our experiments have been presented in the Civitai community.[1]

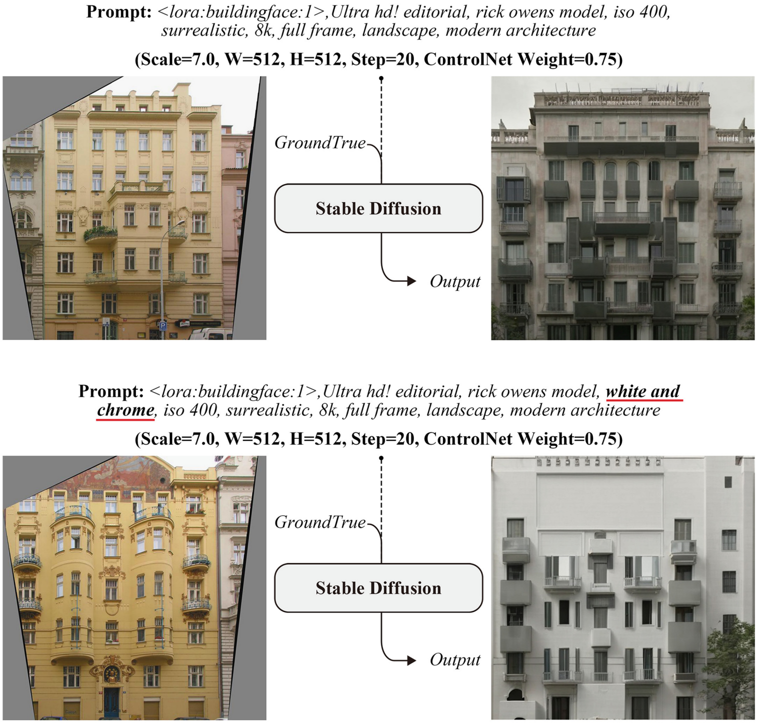

Prompt: *<lora:buildingface:1>,Ultra hd! editorial, rick owens model, iso 400, surrealistic, 8k, full frame, landscape, modern architecture*
(Scale=7.0, W=512, H=512, Step=20, ControlNet Weight=0.75)
GroundTrue
Stable Diffusion
Output
Prompt: *<lora:buildingface:1>,Ultra hd! editorial, rick owens model, **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)
GroundTrue
Stable Diffusion
Output

Fig. 7. Examples of migration experiments

4 Conclusion and Discussion

This research proposed a method for generating building facades based on the Stable Diffusion model. The LoRA method is used to fine-tuned the huge model, which was trained on 200 building facades. Also, use ControlNet to regulate the generation outcomes during the future generation process. The controllable operational research from text semantics to building facade generation is completed in this work. The findings

[1] https://civitai.com/gallery/133518?modelId=11661&modelVersionId=13784&infinite=false&returnUrl=%2Fmodels%2F11661%2Fbuildingfacade.

demonstrate that: (1) 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. (2) 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. (3) 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. Overall, this makes it easier to design building facades, simply by changing the prompt words and adjusting the model weights to obtain a large number of quality results. Future research could combine morphological generative algorithms with AI to produce more accurate and richer results.

This study still has some restrictions, though. The amount of data is insufficient in the first place since training with additional data necessitates more digital memory space. 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 may result in the production of more high-quality building facade content.

References

Borji, A.: Generated faces in the wild: Quantitative comparison of stable diffusion, midjourney and dall-e 2 (2022). arXiv:2210.00586

Hu, E.J., Shen, Y., Wallis, P., Allen-Zhu, Z., Li, Y., Wang, S., Chen, W.: LoRA: Low-Rank Adaptation of Large Language Models (2021). abs/2106.09685

Huang, W., Zheng, H.: Architectural drawings recognition and generation through machine learning. In: Proceedings of the 38th Annual Conference of the Association for Computer Aided Design in Architecture, Mexico City, Mexico, pp. 18–20 (2018)

Isola, P., Zhu, J.-Y., Zhou, T., Efros, A.A.: Image-to-image translation with conditional adversarial networks. IEEE Conf. Comput. Vision and Pattern Recogn. (CVPR) **2017**, 5967–5976 (2016)

Park, T., Liu, M.-Y., Wang, T.-C., Zhu, J.-Y.: Semantic image synthesis with spatially-adaptive normalization. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition, pp. 2337–2346 (2019)

Radford, A., Metz, L., Chintala, S.: Unsupervised Representation Learning with Deep Convolutional Generative Adversarial Networks (2015). CoRR, abs/1511.06434

Ramesh, A., Dhariwal, P., Nichol, A., Chu, C., Chen, M.: Hierarchical Text-Conditional Image Generation with CLIP Latents (2022). ArXiv, abs/2204.06125

Rombach, R., Blattmann, A., Lorenz, D., Esser, P., Ommer, B.: High-resolution image synthesis with latent diffusion models. IEEE/CVF Conf. Comput. Vision Pattern Recogn. (CVPR) **2022**, 10674–10685 (2021)

Ruiz, N., Li, Y., Jampani, V., Pritch, Y., Rubinstein, M., Aberman, K.: DreamBooth: Fine Tuning Text-to-Image Diffusion Models for Subject-Driven Generation (2022). ArXiv, abs/2208.12242

TYLECEK, R. 2012. The cmp facade database. Tech. rep., CTU–CMP–2012–24, Czech Technical University

Zhang, L., Agrawala, M.: Adding Conditional Control to Text-to-Image Diffusion Models (2023). ArXiv, abs/2302.05543

Zhu, J.-Y., Park, T., Isola, P., Efros, A.A.: Unpaired image-to-image translation using cycle-consistent adversarial networks. In: Proceedings of the IEEE international conference on computer vision, pp. 2223–2232 (2017)