Research on Image-to-Image Generation and Optimization Methods Based on Diffusion Model Compared with Traditional Methods: Taking Façade as the Optimization Object

Zexi Lyu[1(✉)], Zao Li[2], and Zijing Wu[3]

[1] Hefei University of Technology, Hefei, China
1079036667@qq.com
[2] Anhui Jianzhu University, Hefei, China
[3] University of Waterloo, Waterloo, ON, Canada
z256wu@uwaterloo.ca

Abstract. The intersection of technology and culture has become a topic of great interest worldwide, with China's development embracing this integration as an essential direction. One critical area where these two fields converge is in the inheritance, translation, and creative design of cultural heritage. In line with this trend, our study explores the potential of stable diffusion to produce highly detailed and visually stunning building façades. We start by providing an overall survey and algorithm fundamentals of the generative deep learning models used so far, namely, GAN and Diffusion models. Then, we present our methodology for using Diffusion Model to generate architecture façades. We then demonstrate how the fine-tuning is done for Stable Diffusion is done to yield optimal performance and then evaluate four different training methods of SD. We also compare existing GAN based façade generation method with our Diffusion based method. Our results show that our Diffusion-based approach outperforms existing methods in terms of detail and quality, highlighting the potential of stable diffusion in generating visually appealing building façades. This research contributes to the growing body of work on the integration of technology and culture in architecture and offers insights into the potential of stable diffusion for creative design applications.

Keywords: Façade · Diffusion · GAN · Image-to-image · Image generation · Fine-tuning

1 Introduction

Rapid urbanization in China has incited a conundrum of architectural style disarray, necessitating urgent preservation of vanishing features. Façade enhancement, a vital aspect of architectural style, demands collecting, organizing, analyzing, evaluating, and redesigning extant styles. Traditionally, this labor-intensive process yielded subjective outcomes. This study focuses on generating building façades via stable diffusion, initially

C. Yan et al. (Eds.): CDRF 2023, *Phygital Intelligence*, pp. 35–50, 2024.
https://doi.org/10.1007/978-981-99-8405-3_4

establishing a dataset of neo-Chinese architectural façades based on component types and distribution patterns. Subsequently, this dataset evaluates the performance of four stable diffusion methods for façade images and tests existing labeled façade datasets.

Related work. Over the past decade, generative image synthesis has been extensively researched and applied, particularly in architectural design. GANs [1], which have dominated the field, consist of a generator producing data samples and a discriminator identifying samples as real or generated. Both components, typically U-Nets, iteratively improve until the generator successfully deceives the discriminator. The generator initiates with random noise sampled from a distribution (e.g., Gaussian), while the discriminator, trained on ground truth datasets, outputs the probability of a sample's authenticity. The process minimizes the loss function:

$$\min maxV(D, G) = E_{x\sim\rho_{data}(x)}\big[logD(x)\big] + E_{z\sim\rho_z(z)}[\log(1 - D(G(z)))]$$

Original GAN has limited performance on conditional outputs, so Conditional GAN [2] was proposed by computing the $D(x|y)$ and $G(z|y)$. Pix2Pix [5] further improved CGAN by improving generator and discriminator with U-Net and PatchGAN as well as optimizing loss-function using *L1 loss* as below.

$$G^* = \arg\min maxL_{cGAN}(G, D) + \lrcorner L_{L1}(G)$$

Further work on Pix2Pix by Yu et al. [7] in their paper on architectural façade generation suggest that Pix2Pix perform well in façade generation and façade style conversion after 100 epochs of training (Fig. 1).

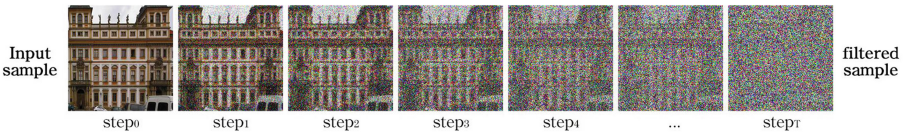


Fig. 1. The diffusion process for an input image. Going from left to right is the forward process where Gaussian noises are added step by step until the image is completely Gaussian. The goal of the model is to learn the function that best approximates the reverse process, going from step t to step 0.

Diffusion model [8] is another family of latent variable model that had been researched extensively for image synthesis purposes. The main idea behind DMs is to construct a Markov Chain that adds random Gaussian noise to sample image gradually until it is no longer visually meaningful and that learns how to reverse this process. The forward process is defined as below:

$$q(x_t|x_{t-1}) = \aleph\Big(x_t; \sqrt{1-\beta_t}x_{t-1}, \beta_t \mathrm{I}\Big) q(x_{1:T}|x_0) = \prod_{t=1}^{T} q(x_t|x_{t-1})$$

where t denotes the timestep of each operation and beta denotes the variance sched-ule or noise schedule such that (variance schedule).

$$\{\beta_t \in (0, 1)\}_{t=1}^{T}$$

This is done by finding the estimating $q(x_{t-1})$ conditioned on original data, that is, $q(x_{t-1}|x_t, x_0)$. Hence rewriting the conditional probability using Bayes rule gives:

$$Q(x_t|x_t, x_0) \sim G(\mu, \beta)$$

$$\tilde{\beta}_t = \frac{1-\overline{\alpha}_{t-1}}{1-\overline{\alpha}_t} \cdot \beta_t$$

$$\tilde{\mu}_t(x_t, x_0) = \frac{\sqrt{\alpha_t}(1-\overline{\alpha}_{t-1})}{1-\overline{\alpha}_t} x_t + \frac{\sqrt{\overline{\alpha}_{t-1}}\beta_t}{1-\overline{\alpha}_t} x_0$$

where $\alpha = 1 - \beta$, a simplification trick used in forward diffusion process that makes $q(x_{t-1})$ can be conditioned on x_0 alone. This way with the reverse process defined, the loss function could be modeled as following:

$$E\big[-\log \rho_\theta(x_0)\big] \le E_q\left[-\log \frac{\rho_\theta(x_{0:T})}{q(x_{1:T}|x_0)}\right] = E_q\left[-\log \rho(x_T) - \sum_{t\ge 1} \log \frac{\rho_\theta(x_{t-1}|x_t)}{q(x_t|x_{t-1})}\right]$$

By optimizing ρ_θ, the reverse process, the model's loss function can be modeled by taking the negative log-likelihood function to get to the variational lower bound of the loss. Ho et al. in his paper on DDPM [3] further simplified the loss function and improved the training efficiency by ignoring the weights in the original function and keeping the variance fixed while train only the mean of the normal distribution.

Rombach et al. in the paper Latent Diffusion Model [4], which is the model we will be using for this paper, further improved the training efficiency for generating high resolution images by first encoding the input into latent variable using an encoder network and then feed the lower dimension latent variables into a DDPM-like U-Net architecture for image generation.

2 Methodology

We propose in this paper to use fine-tuned Stable Diffusion, an implementation of Latent Diffusion Model to conduct façade generation and compare the effect of various diffusion model training methods and parameter sets have on the final generated façades. We also compare the quality of generated façades with previous work on generative architectural façade using earlier methods such as cGAN (Figs. 2 and 3).

2.1 Introduction to Diffusion Training Methods

2.1.1 Textual Inversion

Textual Inversion is a feature in the Stable Diffusion model, which allows for personalizing the model by training a small part of the neural network with custom images. This way, the model can be guided to generate new images based on the concepts taught through Textual Inversion. The Textual Inversion process involves feeding a set of images into the model, which then outputs a vector that represents a specific concept. This vector can then be used in the text-to-image generation process to generate new images based on the taught concepts.

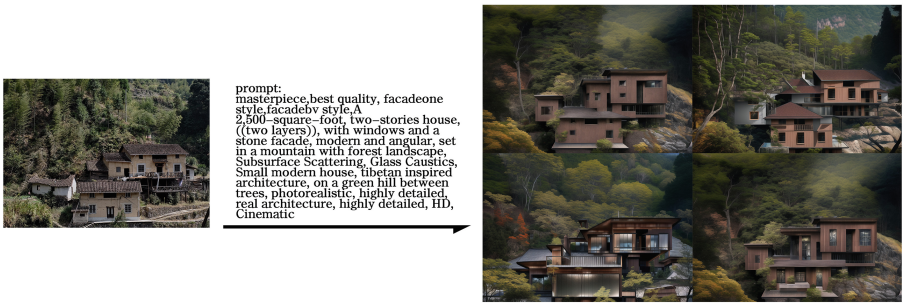


Fig. 2. An illustration of img-to-img generation. To the left is the original architecture image[1] taken at Song Yang country, Zhejiang Province of China, and to the right are four img-to-img images generated with respect to the prompts listed in the middle.

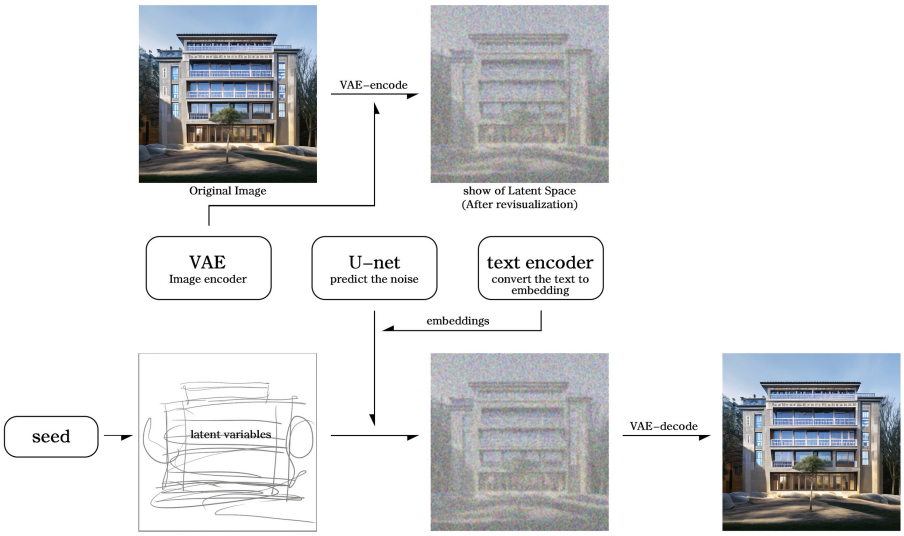


Fig. 3. The architecture for training and tuning LDM to perform façade design tasks. Random seed is also included to add more variety in generated contents.

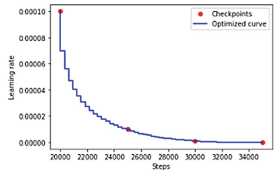
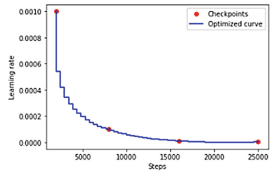
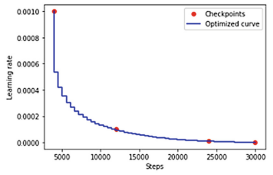
2.1.2 Hypernetwork

Hypernetwork is a novel concept used to fine-tune models without touching any weights. This technology is widely used in style transfer and has better generalization performance compared to textual inversion. In Stable Diffusion refers to an additional layer that is processed after an image has been rendered through the model. It tends to skew all results from the model towards the training data, essentially changing the model.

The learning rate for the Hypernetwork may be different than the learning rate for the embedding, with a lower value for the Hypernetwork (Table 1).

[1] Original architecture images are from CRCV· The second National Architectural Design Competition of Songyang Rural Revitalization.

Table 1. Comparison of three experiments on Hypernetwork structure[2]

First experiment: Turn on the LN, Dropout, Layer 1,2,1	Second experiment: Turn on the LN, Dropout, Layer 1, 2, 1, activate the function Swish	Third experiment: Turn on the LN, Dropout, Layer 1, 2, 2, 1, activate the function Swish
		

For the training set we selected, the learning rate of the third experiment achieved a good effect, about 70% of the performance can be restored. LN makes sense for training to be more stable by preventing overfitting. Enabling Dropout can prevent hypernetwork overfitting. Custom dropout ratio is not currently supported, with a default of 0.3. Although the extended layer structure can obtain good training effect, the pt file with layer structure 1, 2, 1 occupies about 83.8 MB of real-time memory, while the PT file with layer structure 1, 2, 2, 1 occupies about 167 MB.

2.1.3 DreamBooth

DreamBooth [9] is an innovative tool for refining text-to-image diffusion models, such as Stable Diffusion, enabling subject-driven generation. The fine-tuning process entails retraining the model with minimal subject-specific images and identifiers, resulting in a model adept at discerning subjects, isolating them from existing contexts, and accurately synthesizing them within new desired settings. Described as a photo booth by its Google research team creators, DreamBooth facilitates the customization of personalized diffusion models with limited training data. Utilizing Imagen as its foundation, the model can be exported as.ckpt, easily integrated into various UIs. While heralded as the preeminent image generation model, it demands a mid-tier gaming GPU and restricts simultaneous usage with other models.

2.1.4 LoRA: Low-Rank Adaptation for Fast Diffusion Fine-Tuning

LoRA [10] is a technique for adapting pre-trained language models to new tasks by freezing the original model's weights and adding trainable rank decomposition matrices to each Transformer layer. This approach significantly reduces storage requirements while maintaining input and output dimensions. Implemented as a Python package called loralib, it integrates with PyTorch models like HuggingFace. LoRA introduces minimal inference latency and capitalizes on the inherent low-rank characteristics of large models

[2] Test code from https://colab.research.google.com/drive/1qzweYEMIFkG6jPa04tD1MhWWOzgSnDvP?usp=sharing.

by adding a bypass matrix, simulating full fine-tuning. This method presents a simple, effective solution for lightweight fine-tuning.

3 Experiments

We conduct three types of experiments. First one is a comparison of diffusion model with GAN, pix2pix in particular; second one is a comparison of different parameter tunings among LDM, including sampling methods, steps, CFG Scales, img2img redraw etc.; the last one is a comparison of different training methods, Textual Inversion, Hypernetwork, DreamBooth and LoRA on our own generated dataset. We aim to find an efficient, high quality parameter and training methods that can fulfill the exact needs of architects.

3.1 Comparison of Façades Generated by Pix2Pix and Latent Diffusion Model

We first compare Conditional GAN Pix2Pix with the LDM model used by Stable Diffusion. Pix2Pix is one of the most used generative GAN models in many different fields and it has yielded decent quality and accuracy in the area of architectural façade design. Qiu et al. experimented with *Pix2Pix* on façade design and trained their network on CMP[3] Façade dataset by Tylecek et al. for 100 epochs. We use the same dataset and train our LDM and presents a comparison of generated façades as in Figure. As can be seen in the comparison, LDM can achieve better quality and se-mantic understanding in the generated façades then those of the Pix2Pix models (Figs. 4, 5 and 6).

Fig. 4. CMP Façade dataset

[3] Dataset from https://cmp.felk.cvut.cz/~tylecr1/facade/, hereby declare.

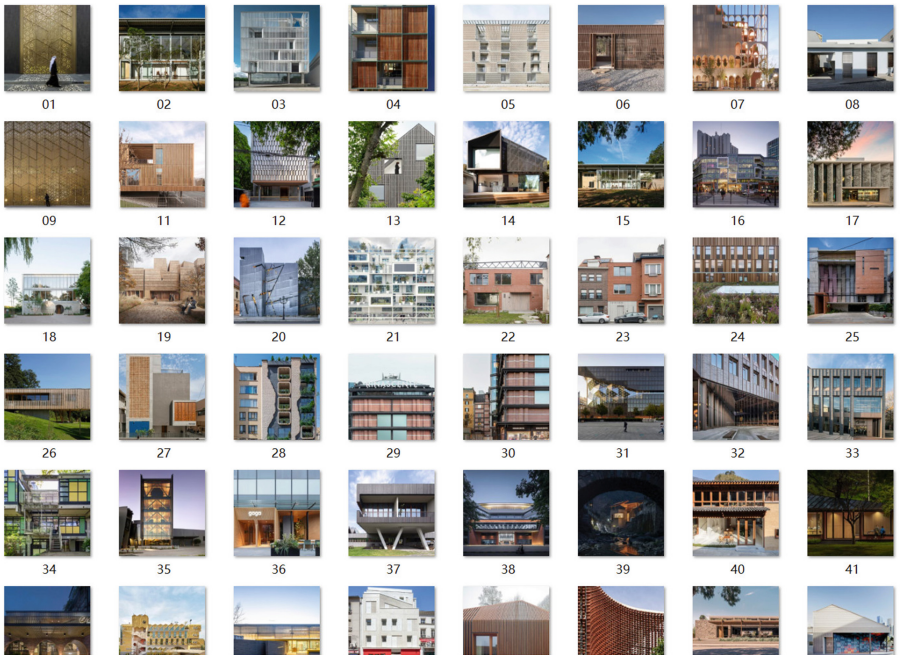


Fig. 5. Homemade Façade dataset

Fig. 6. Comparison of architecture façades generated from img-to-img translation using Pix2Pix from Qiu et al.'s work and Stable Diffusion from our tuning.

Another advantage of LDM over Pix2Pix is that LDM is an unsupervised model that does not require any labeling on data for training. We used only the original images in CMP Façade dataset for training while Pix2Pix network also used the label images to assist in training to yield optimal results.

3.2 Comparison of Images Generated by Different Prompts

Stable Diffusion, a prompt-based text-to-image model, comprises two key components: Contrastive Language-Image Pre-Training (CLIP) [17] and the generative Diffusion Model. CLIP, a multimodal model, is trained on text and image data to generate textual summaries from images. It transforms input text prompts into embeddings fed into the reverse diffusion process, conditioning generation. Prompt words stem from the model's natural language processing (NLP) scheme and tagged words in initial training materials. These prompts directly influence the final image elements, with accuracy being vital for effective AI-generated images. Thus, prompt selection and design require meticulous attention for optimal results (Fig. 7).

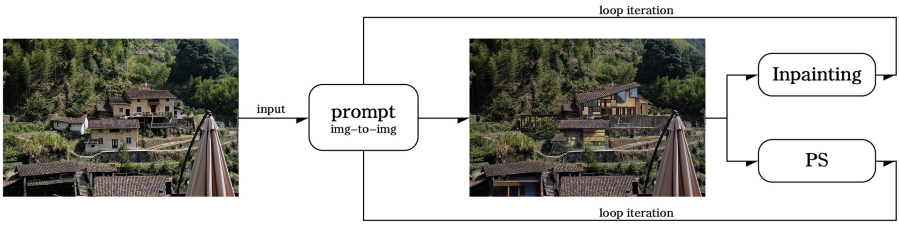


Fig. 7. Prompt + PS/Inpainting img-to-img loop iteration

The above figure illustrates the iterative process of img-to-img used in this research. The current workflow involves the use of prompt and post-processing techniques, such as Photoshop (PS) or inpainting. Using the figure as an example, the forward prompt used by the author is "(masterpiece), (best quality), ((façade-one style)), three 2000-square-foot, two-stories small modern houses, ((two layers)), with windows and a stone façade, modern and angular, set in a mountain with forest landscape, Subsurface Scattering, Glass Caustics, Small modern house, photorealistic, highly detailed, real architecture, ((low saturation)), highly detailed, HD, Cinematic". "façade-one style" is the label/trigger word trained by the author's model, and using this label for image generation can achieve desirable results. () adds emphasis to a term, [] decreases emphasis, both by a factor of 1.1. You can either stack ()/[] for increasing/decreasing emphasis or use the new syntax which takes a number directly-it looks like this:

(word: 1.1) = (word)

(word: 1.21) = ((word))

(word: 0.91) = [word]

The negative prompt used by the author is "lowers, text, error, extra digit, low quality, jpeg artifacts, signature, blurry, normal quality, cropped, worst quality".

When keeping the seed (the starting point of the random number generator) unchanged, different image effects can be generated by changing the prompt or modifying the match degree between the prompt and the generated image, as shown in Fig. 8.

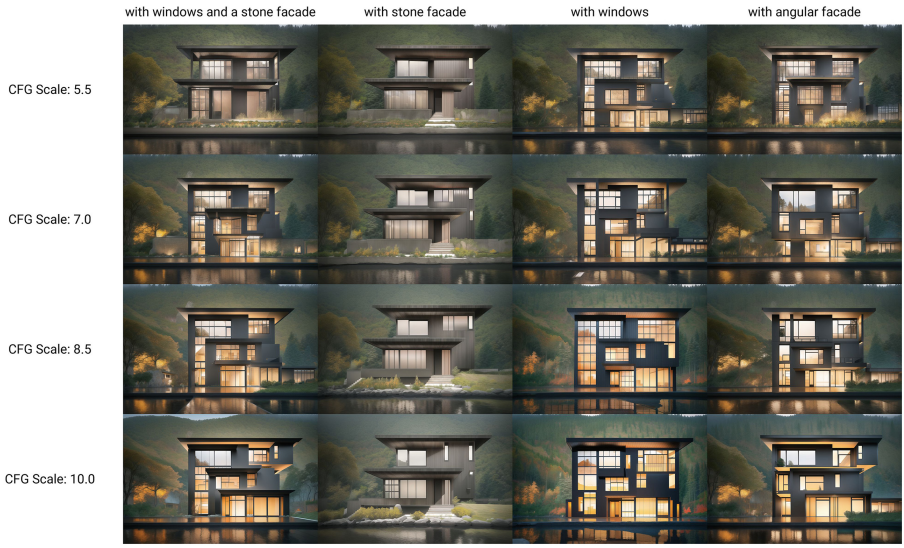


Fig. 8. Prompt replacement—CFG Scale X-Y graphs

3.3 Comparison of Images Generated by Sampling Method, Sampling Steps, Classifier Free Guidance Scale, Img-to-Img Redraw Amplitude

The diffusion model generates clear images from noisy counterparts via a forward noise-adding process and a backward denoising process. This sampling method, crucial for image generation, affects denoising, quantization, and operational speed. This study compares popular methods, including Euler a, DDIM, and the DPM series. Non-linear iterative methods like DPM a and Euler a exhibit declining quality beyond a certain iteration count, while linear iterative methods, such as DDIM/Euler, display an opposing tendency, with quality relying on iteration count. However, marginal effects limit significant improvements beyond a certain point (Fig. 9).

As shown in the figure, the image generation performance is better with the *Euler a* sampling method and Sampling Steps between 50 and 60.

The Classifier Free Guidance Scale (CFG Scale) balances sample quality and diversity by jointly training conditional and unconditional diffusion models without using a sampler. Higher prompt relevance yields increased prompt frequency and reduced object-environment fusion, while lower relevance allows greater AI creativity and enhanced fusion.

Fig. 9. Sampling Steps–Sampling Methods X-Y graphs

When the Denoising strength is less than 0.5, local modifications will be made directly on the original image. When the Denoising strength is greater than 0.6, elements that match the original image will be rarely seen (Fig. 10).

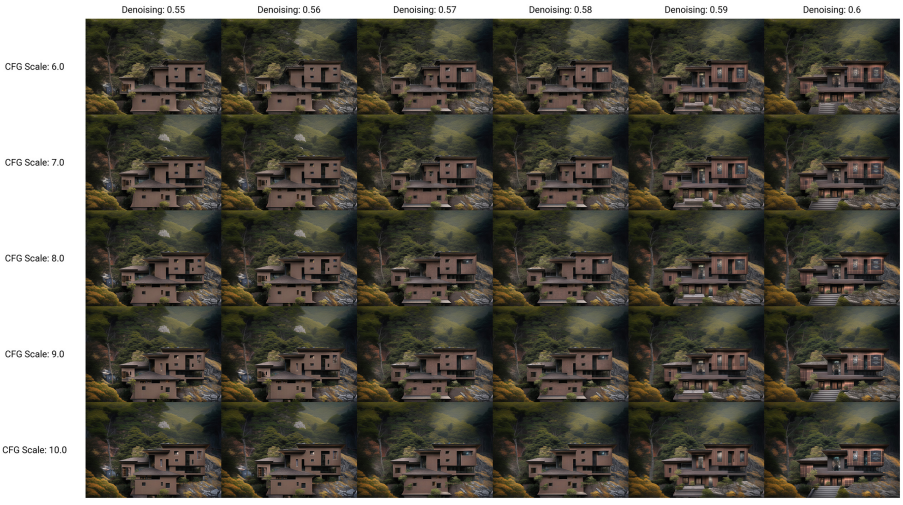


Fig. 10. Denoising strength—CFG Scale X-Y graphs

As shown in the figure, the image generation performance is better with the CFG Scale is between 7 and 10, and the Denoising is 0.59.

3.4 Comparison of Images Generated by the Training Methods: Textual Inversion, Hypernetwork, DreamBooth, LoRA

After the training models are completed, the variables are strictly controlled and the tags of the generated embedding and DB model are tested (Fig. 11).

Fig. 11. Training models—Hypernet Strength X-Y graphs

Hypernetworks differ from Textual Inversion as they fine-tune the model, leading to better generalization and better training aesthetics. DreamBooth can generate good results with just a few input images of a specific object and its corresponding class name (e.g., dog), along with a unique identifier implanted in different textual descriptions. DB is better than Textual Inversion as it inserts training data into the output, leading to high similarity and great results.

LoRA approximates full fine-tuning expressiveness by setting rank r equal to pre-trained weight matrices' rank, with increasing trainable parameters. Consequently, LoRA converges to the original model, whereas adapter-based methods converge to an MLP and prefix-based methods to a model restricted by input sequence length (Fig. 12).

Fig. 12. LoRA's datasets composition schematic

With the assistance of textual prompts, the training dataset for LoRA can be more guided, resulting in more directed and desirable style transfer outcomes.

LoRA offers a lightweight, efficient alternative to full model fine-tuning of Stable Diffusion, outperforming DreamBooth in speed and adaptability. Low-rank adaptation yields compact results (1-6MB) for easy sharing and compatibility with diffusers and inpainting. In some cases, LoRA surpasses full fine-tuning, with potential for checkpoint merging, recipe creation, and enhanced fine-tuning via CLIP, Unet, and tokens. Offering multi-vector pivotal tuning inversion, LoRA models are smaller than 2GB + DB counterparts, enabling rapid training, art style replication, and DB training with minimal VRAM requirements.

3.5 Using Loopback Method to Optimize Images

Fig. 13. Using Loopback method to improve image quality

Loopback is a method by Stable Diffusion to use generated image output, in our case, generated façades, as input for the next round of generation. The process is similar to a cycle of repeating image-to-image translation. We set the iteration steps to 2 steps and Fig. 13 is the yielded result. It can be seen that Loopback can provide better details in generated façades.

3.6 Using ControlNet to Guide the Façade Generation Process

ControlNet is a method proposed by Zhang [17] to control the output of a pretrained Diffusion model to achieve better accuracy. It is achieved by having a locked neural network(the original pretrained model) and trainable copy of the original network at the same time and feed the control conditions(i.e., a edge map or line sketch) to the trainable copy and then connect the copy with the locked model layer-wise.

Best practices for using ControlNet is to convert original image into an edge map. Edges or scratches can effectively control the output into desired results. Some edge detection methods we have tested and resulted decent output includes:

(i) Holistically-Nested Edge Detection Boundary (HED Boundary) [18], a convolutional neural network based edge detection model trained on labelled datasets that is capable of learning the hierarchical relations and other complicated spatial relations in image and combine these information when converting into edge maps;
(ii) Semantic Segmentation using Uniformer [19], a transformer based architecture that utilizes 3D convolution and spatiotemporal attention mechanism to achieve better compute efficiency and accuracy in various tasks, including segmentation on images.

Fig. 14. Holistically-Nested Edge Detection in img-to-img

Fig. 15. Semantic Segmentation in img-to-img

ControlNet along with edge detection and segmentation techniques enables architects to generate façades designs using a sketch drawing or an existing façade image with better accuracy and better alignment to the user's intentions. Edge detection technology plays a crucial role in controlling the creation of images in the Img-to-Img framework, allowing designers to achieve the desired rendering effects in the generated images, as shown in Fig. 14. The involvement of semantic segmentation allows for more accurate differentiation of the various elements in the original image, facilitating better subsequent translation: architectural elements are replaced with new architectural elements, and so on, resulting in better facade generation and better surrounding environment, as shown in Fig. 15.

To apply lighting to generated images, upload the light source image to the image generation area and place the original image in ControlNet, selecting the Depth model, as shown in Fig. 16. Depth [20], a valuable intermediate representation for actions in physical environments, facilitates realistic rendering in scenes by comparing pixel depth values and preventing distant objects from obscuring closer ones.

Due to the inherent principle of img-to-img, which generates images based on the original image with added Gaussian noise, color block distribution is generally similar, but controlling finer details is challenging. With ControlNet's intervention, the model, initially guided by text generation, can now comprehend information extracted from images. Combined with img-to-img, this yields more desirable control outcomes.

ControlNet also supports the combination of multiple models, enabling multi-condition control of images. For example, by setting up two ControlNets, the first one controls building façade contours using HED, while the second one manages background composition through Seg or Depth. Adjusting ControlNet weights, such as prioritizing HED over Depth, ensures accurate façade structure recognition, followed by content and style control through prompt words and style models.

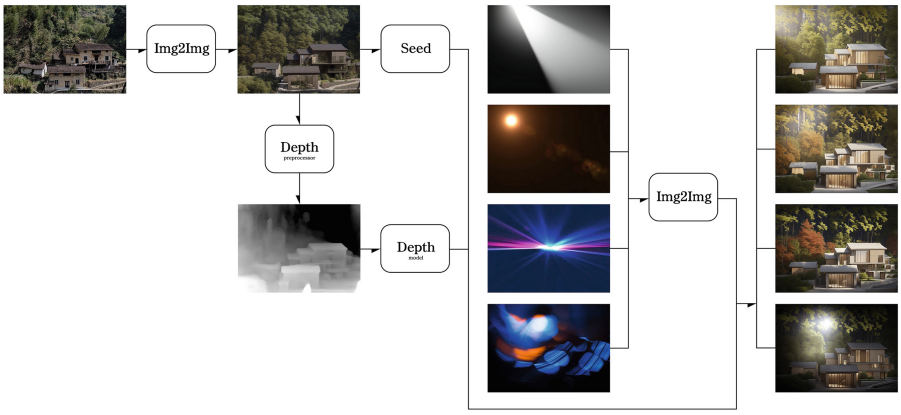


Fig. 16. Img-to-img combined with ControlNet--take Depth as an example

4 Conclusion and Discussion

Stable Diffusion outperforms earlier models like Pix2Pix in architecture façade generation, excelling in content quality and training efficiency. By adding a bypass matrix, based on the model's low-rank characteristics, LoRA achieves lightweight fine-tuning effectively.

This method offers potential in architectural style consistency and coherence. Despite some non-functionality, the generated images preserve the original photo's composition and color tone, with the structure well-extracted and translated, resulting in logical façade compositions. Utilizing this method during the sketch stage enables designers to evaluate color, form, and composition across multiple schemes.

However, Stable Diffusion has limitations, including potential inaccuracies in recognizing environmental factors, regulations, and engineering functionality. Thus, human experts should review and refine generated façades for feasibility.

Architectural AI's future is promising, providing assistance and inspiration for façade designs and allowing architects to focus on innovative tasks, elevating productivity. While serving as a valuable tool, it should not replace designers' emotional judgment and final decisions. The technology's success depends on the collaborative synergy between designers and AI tools, capitalizing on each other's strengths and weaknesses (Fig. 17).

Despite personal constraints in data collection and hardware configurations, this study addresses key issues in historical and cultural preservation. It targets challenges like updating historical core buildings, maintaining architectural style and quality, ensuring seamless style transitions in transitional zones, and integrating traditional design elements with modern urban functionality. Additionally, the research leverages digital technologies, including diffusion models, semantic ontology methods, and rough set screening, to develop innovative façade design strategies in preservation areas.

Future research will quantify image data for the training method, enhancing the generation of effective, realistic architectural images. Due to the extensive data required for optimal diffusion model training, subsequent work could explore data collection

Fig. 17. Extra effect display

and preprocessing collaborations with academic and commercial institutions, as well as employing automated tools for data identification and refinement. This research holds significant implications for urban design and preservation, with potential applications extending beyond the study's scope.

Funding. This research was funded by the National Natural Science Foundation of China, grant number 51978226, and the Anhui Province University Outstanding Scientific Research and Innovation Team, grant number 2022AH010021.

References

1. Goodfellow, I.J., et al.: Generative Adversarial Networks. arXiv, 10 June 2014. arXiv.org. https://doi.org/10.48550/arXiv.1406.2661
2. Mirza, M., Osindero, S.: Conditional Generative Adversarial Nets (2014). https://doi.org/10.48550/arXiv.1411.1784
3. Ho, J., et al.: Denoising Diffusion Probabilistic Models (2020). https://doi.org/10.48550/arXiv.2006.11239
4. Rombach, R., et al.: High-Resolution Image Synthesis with Latent Diffusion Models (2022). https://doi.org/10.48550/arXiv.2112.10752
5. Isola, P., et al.: Image-to-Image Translation with Conditional Adversarial Networks (2018). https://doi.org/10.48550/arXiv.1611.07004
6. Weng, L.: What Are Diffusion Models? (2021). https://lilianweng.github.io/posts/2021-07-11-diffusion-models/
7. Yu, Q., Malaeb, J., Ma, W.: Architectural facade recognition and generation through generative adversarial networks. In: 2020 International Conference on Big Data & Artificial Intelligence & Software Engineering (ICBASE), Bangkok, Thai-land, pp. 310–316 (2020). https://doi.org/10.1109/ICBASE51474.2020.00072
8. Sohl-Dickstein, J., et al.: Deep Unsupervised Learning Using Nonequilibrium Thermodynamics (2015). https://doi.org/10.48550/arXiv.1503.03585
9. Ruiz, N., et al.: DreamBooth: Fine Tuning Text-to-Image Diffusion Models for Subject-Driven Generation (2022). https://doi.org/10.48550/arXiv.2208.12242
10. Hu, E.J., et al.: LoRA: Low-Rank Adaptation of Large Language Models (2021). https://doi.org/10.48550/arXiv.2106.09685
11. Wu, Q., Ye, H., Gu, Y.: Guided diffusion model for adversarial purification from random noise (2022). https://doi.org/10.48550/arXiv.2206.10875
12. Dong, Z., Wei, P., Lin, L.: Dreamartist: Towards controllable one-shot text-to-image generation via contrastive prompt-tuning (2022). https://doi.org/10.48550/arXiv.2211.11337

13. Cao, H., Tan, C., Gao, Z., Chen, G., Heng, P.-A., Li, S.Z.: A survey on generative diffusion model (2022). https://doi.org/10.48550/arXiv.2209.02646
14. Bowen, R.S., Chang, H., Herrmann, C., Teterwak, P., Liu, C., Zabih, R.: OCONet: image extrapolation by object completion. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 2307–2317 (2021)
15. Saharia, C., Chan, W., Chang, H., Lee, C.A., Ho, J., Salimans, T., Fleet, D.J., Norouzi, M.: Palette: Image-to-image diffusion models (2022). https://doi.org/10.48550/arXiv.2111.05826
16. Barnes, C., Shechtman, E., Finkelstein, A., Goldman, D.B.: Patch-match: a randomized correspondence algorithm for structural image editing. ACM Trans. Graph. (Proc. SIGGRAPH) **28**, 3 (2009)
17. Zhang, L., Agrawala, M.: Adding Conditional Control to Text-to-Image Diffusion Models (2023). https://doi.org/10.48550/arXiv.2302.05543
18. Xie, S., Tu, Z.: Holistically-Nested Edge Detection (2015). https://doi.org/10.48550/arXiv.1504.06375
19. Li, K., et al.: UniFormer: Unified Transformer for Efficient Spatiotemporal Representation Learning (2022). https://doi.org/10.48550/arXiv.2201.04676
20. Ranftl, R., Lasinger, K., Hafner, D., Schindler, K., Koltun, V.: Towards robust monocular depth estimation: mixing datasets for zero-shot cross-dataset transfer. IEEE Trans. Pattern Anal. Mach. Intell.Intell. **44**(3), 1623–1637 (2020)