Vol.2024-CG-194 No.3
2024/6/29

Retrieval-Augmented Multi-Floor Building Image Generation

Zhengyang Wang[1,a)] Hao Jin[1] Haoran Xie[1]

Abstract: Demand for generating building images from text prompts grows, despite recent advances in diffusion models greatly enhancing image quality. The current generative models struggle with controlling the number of floors. To this end, we propose a retrieval-augmented framework for generating building images with provided floor count using a diffusion model. Initially, the text prompts with the provided floor count to retrieve the most suitable image from a building image database. Then, we adopted a multi-level structure detection algorithm to obtain a sketch from the matched image to ensure structural consistency. Finally, the building image with the desired floor count and style is generated by diffusion model, guided by the detected building sketch. Our proposed framework enables accurate control over the floor count in building image synthesis. We demonstrate the robustness and scalability of generating building images with a specific floor count from text prompts.

Keywords: Image generation, Retrieval-augmented generation, Diffusion model

1. Introduction

Image generation is currently undergoing significant improvements, particularly in the area of image personalization. ControlNet [14] and T2I-Adapter [6] incorporate conditions such as sketches and depth maps, enhancing control over the generated results to better meet specific requirements. However, the generation of images with precise, user-defined attributes remains a challenge, particularly in the domain of building image generation. The state-of-the-art models fail to control the number of floors in building images, as shown in Figure 1.

To address this challenge, we propose a retrieval-augmented generation method that utilizes text-to-image diffusion models to generate building images with precise control over the number of floors. Initially, A building image is generated with a specified style using latent diffusion model conditioned by text. Simultaneously, the generated image is utilized to retrieve the most suitable reference image from a building images database, ensuring structural consistency with the user-specified number of floors. Finally, we combine the generated styled building image with the structural information from the reference images to generate the building image with the provided floor count. Our framework offers the advantage of generating building images with the provided floor count and style.

The main contributions of this work are listed as follows:

- We proposed an innovative retrieval-augmented approach that effectively leverages data from a building image database to improve the structural consistency of generated images.

Fig. 1 Building image generation. These results generated using stable diffusion 1.5, all fail to accurately reflect the specified number of floors.

- We can achieve precise control in generating building images with the provided number of floors, overcoming the limitations of existing generation models. This provides a robust and scalable solution for generating detailed and structurally consistent building images from text prompts.

2. Related Works

2.1 Retrieval-augmented Generation

Retrieval-Augmented Generation (RAG) [5] provides a method for leveraging external database to perform retrieval, using the retrieved relevant examples to more effectively guide and enhance the generation process. For example, in various natural language processing tasks, relevant documents can be retrieved to improve the content of the generated text [1, 3]. Similarly, in image generation, RAG improves the quality and

[1] Japan Advanced Institute of Science and Technology, Nomi, Ishikawa 923–1211, Japan
[a)] s2310021@jaist.ac.jp

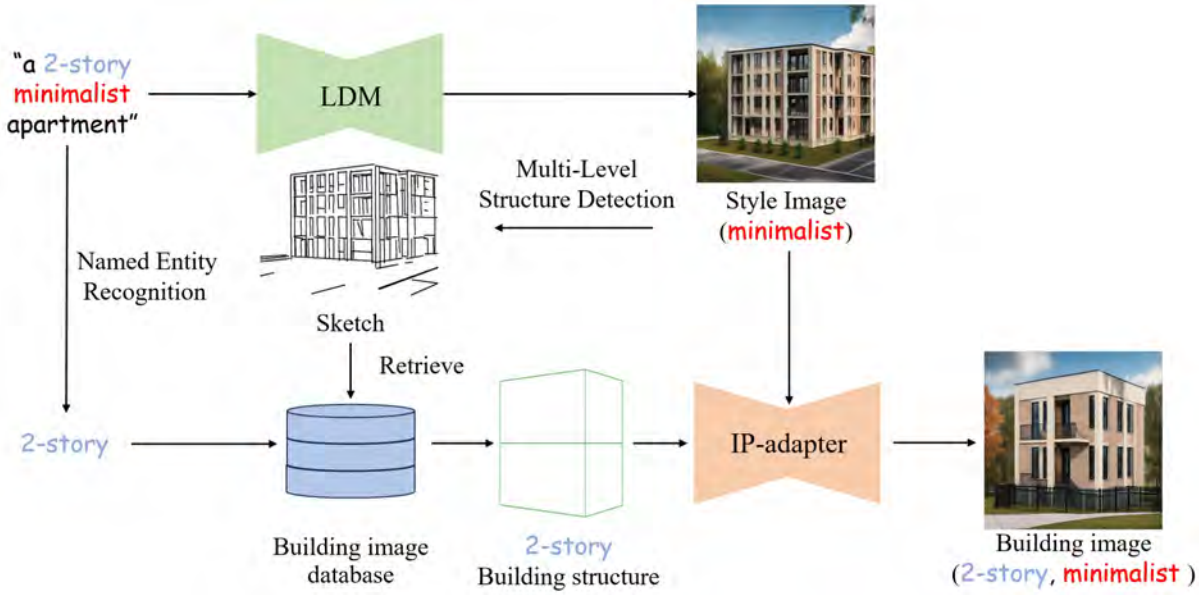

Fig. 2 The workflow of retrival-augmented multi-floor building image generation

accuracy of image synthesis in generative models [10].

2.2 Diffusion Model

Diffusion models were first proposed by Sohl-Dickstein et al. [11], and later improved by Song et al. [12] and Ho et al. [4]. Diffusion models are a class of probabilistic likelihood-based models. Diffusion models consist of two processes: (1) adding noise to the data through a forward diffusion process, and (2) gradually recovering the data through a reverse diffusion process. Diffusion models perform exceptionally well in terms of image fidelity and diversity. However, diffusion models have high computational costs and long training times. Additionally, while the generated images are realistic, they exhibit a high degree of randomness.

Compared to traditional diffusion models, conditional diffusion models enhance control and style guidance by incorporating additional conditions, such as ControlNet [14] and T2I-Adapter [6], further improving the model's ability to generate detailed and realistic images. However, conditional diffusion models still face challenges, such as dependence on complex image attributes and difficulty in precisely controlling all details in certain complex scenarios. The conditional diffusion models have been extensively explored in image generation with the controls of spatial relationship [15], semantic parts [8], and material design [16]. In this work, we especially focus on the control of counting numbers in building image generation.

3. Method

In this paper, we propose a retrieval-augmented framework for generating building images with the provided floor count using a diffusion model. Figure 2 illustrates the workflow of our framework. Firstly, we adopt the Latent Diffusion Model (LDM) to generate styled building images (Section 3.1) and retrieve building image structures through Retrieval-Augmented Generation (RAG) (Section 3.2). Next, we adopt the styled building images and the retrieved building image structures to generate building images with accurate floor counts using the IP-Adapter (Section 3.3).

3.1 Prelimiaries

Latent Diffusion Model (LDM) [9] is a variant of the diffusion model. The diffusion process of LDM does not directly operate in the image space. Instead, it maps image data away from the high-dimensional space to a latent space with strong representational capabilities, where the diffusion process is executed. It includes an encoder $\mathcal{E}$ and a decoder D. The image x is compressed into latent representation $z = \mathcal{E}(x)$ by the encoder $\mathcal{E}$, and is reconstructed as $x \approx D(z)$ using the decoder D. Given a latent sample z_0, it undergoes a forward diffusion process, where Gaussian noise is gradually added at each timestep t, producing latent samples z_t:

$$q(z_t \mid z_{t-1}) = \mathcal{N}\left(z_t;\ \sqrt{1-\beta_t} z_{t-1}, \beta_t I\right), \tag{1}$$

where the noise scale at each step is controlled by the hyperparameter β, and I represents the variance. The reverse diffusion process is the inference process of denoising, approximated by p_θ to predict the previous step z_{t-1}:

$$p_\theta(z_{t-1} \mid z_t) = \mathcal{N}\left(z_{t-1}; \mu_\theta(z_t, t), \Sigma_\theta(z_t, t)\right), \tag{2}$$

where μ_θ and Σ_θ can be achieved using a denoising model ϵ_θ with learnable paraments θ. To minimize the gap between the true posterior distribution of the forward diffusion process and the model distribution of the reverse diffusion process, by the objective function L:

$$\mathcal{L} := \mathbb{E}_{\mathcal{E}(z),\epsilon\sim\mathcal{N}(0,1),t}\left[\ \|\epsilon - \epsilon_\theta(z_t, t)\|_2^2\ \right]. \tag{3}$$

3.2 Retrieval Augmentation

This work efficiently retrieves building image with different number of floors from our database. Retrieval augmentation involves the following three stages: (1) extracting the floor count

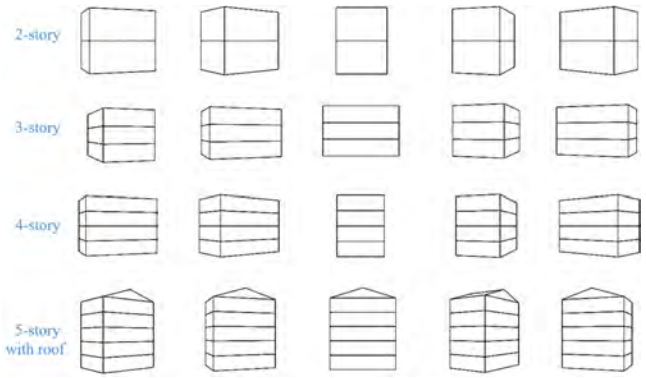

Fig. 3 Example images from the building database.

from the user's text prompt, (2) using a multi-level structure detection algorithm to extract a sketch during the LDM process, and (3) retrieving the building image structure from the dataset that best matches the sketch based on the obtained floor count. We elaborate on the details of these three stages.

Firstly, we obtain the floor count from the provided text prompt using Named Entity Recognition (NER) [7] method. NER is widely used as a fundamental method in natural language processing (NLP). NER is a method for extracting information from text, capable of identifying entities such as dates, locations, and quantities, and classifying this information into predefined categories. Therefore, NER serves as a preprocessing method for various downstream tasks such as information retrieval and translation. Therefore, we parsed the textual prompts through NER to extract information about the number of floors explicitly mentioned by the user. For example, if the text prompt contains a number followed by the word like "story" or "floors", we identify that number as the floor count.

Subsequently, the Multi-Level Structure Detection (MLSD) method efficiently extracts linear structures from images through feature extraction. After generating the building images with the LDM's inference process, we use the MLSD preprocessing method to extract the sketches.

Finally, the obtained floor count information and sketch are served to retrieve the most structurally consistent building structure from the database. First, we use the floor count information as an index to search in the relevant floor database. Then, we use Open Sketch Search Engine (OpenSSE) [2] to calculate the similarity between the sketch and each image in the relevant floor database, retrieving the most structurally consistent building image structure.

In this work, we first utilized SketchUP[*1] to produce 3D building models containing different sizes and floors. Using OpenSSE, 102 different views of each model were generated to create a comprehensive database. Our work has established a database containing 2,040 building images. A portion of the data is shown in Figure 3.

3.3 Building Image Generation

In this stage, we introduced IP-Adapter [13]. IP-Adapter is a generative model implemented based on LDM. In order to effectively combine the features of the styled building images and the retrieved building image structures, decoupling cross-attention is used. Decoupling cross-attention is defined as follows,

$$\begin{aligned}\mathbf{Z}^{new} = \text{Softmax}(\frac{\mathbf{Q}(\mathbf{K}^{style})^\top}{\sqrt{d}})\mathbf{V}^{style} + \\ \text{Softmax}(\frac{\mathbf{Q}(\mathbf{K}^{struct})^\top}{\sqrt{d}})\mathbf{V}^{struct},\end{aligned} \tag{4}$$

where $\mathbf{Q} = \mathbf{Z}\mathbf{W}_q, \mathbf{K}^{style} = c_t\mathbf{W}_k^{style}, \mathbf{V}^{style} = c_t\mathbf{W}_v^{style}, \mathbf{K}^{struct} = \boldsymbol{c}_i\mathbf{W}_k^{struct}, \mathbf{V}^{struct} = \boldsymbol{c}_i\mathbf{W}_v^{struct}$ are the query, key, and value matrix for the styled building images and the retrieved building image structures of attention operations respectively.

4. Results

Figure 4 illustrates the generated building image of 3-story, 4-story and 5-story buildings after style image generation, sketch extraction, structure retrieval and building image generation. The results demonstrate that our framework can effectively generate building images with the provided number of floors while maintaining the style. To verify the effectiveness of our proposed framework, we designed an ablation experiment that uses only text prompts for LDM. By using the same text prompts, this experiment compares the generated results of our framework and LDM, as shown in Figure 5. The results indicate that our framework significantly outperforms LDM in generating building images with a specified number of floors.

5. Conclusion

In this work, we proposed the retrieval-augmented generation and conditional diffusion models to generate building images with a specified number of floors, without fine-tuning or training. By combining text prompts and retrieved building image structure, we can precisely control the style and structure of the buildings during the generation process, thereby producing images that meet user requirements.

However, this work has several limitations. First, the external database is relatively small, resulting in a limited variety of floor designs that can be generated. This constrains the model's performance in terms of diversity and detail, leading to building images that lack complexity and richness. Additionally, despite improvements in controlling the number of floors, there are still deficiencies in the style transfer of certain details, such as windows and balconies. The model's generation of these details does not fully meet expectations, affecting the overall consistency and accuracy of the final images.

[*1] https://www.sketchup.com/en

References

[1] Borgeaud, S., Mensch, A., Hoffmann, J., Cai, T., Rutherford, E., Millican, K., van den Driessche, G., Lespiau, J.-B., Damoc, B., Clark, A., de Las Casas, D., Guy, A., Menick, J., Ring, R., Hennigan, T., Huang, S., Maggiore, L., Jones, C., Cassirer, A., Brock, A., Paganini, M., Irving, G., Vinyals, O., Osindero, S., Simonyan, K., Rae, J. W., Elsen, E. and Sifre, L.: Improving language models by retrieving from trillions of tokens (2022).

[2] Eitz, M., Richter, R., Boubekeur, T., Hildebrand, K. and Alexa, M.: Sketch-based shape retrieval, *ACM Transactions on graphics (TOG)*, Vol. 31, No. 4, pp. 1–10 (2012).

[3] Guu, K., Lee, K., Tung, Z., Pasupat, P. and Chang, M.-W.: REALM: Retrieval-Augmented Language Model Pre-Training (2020).

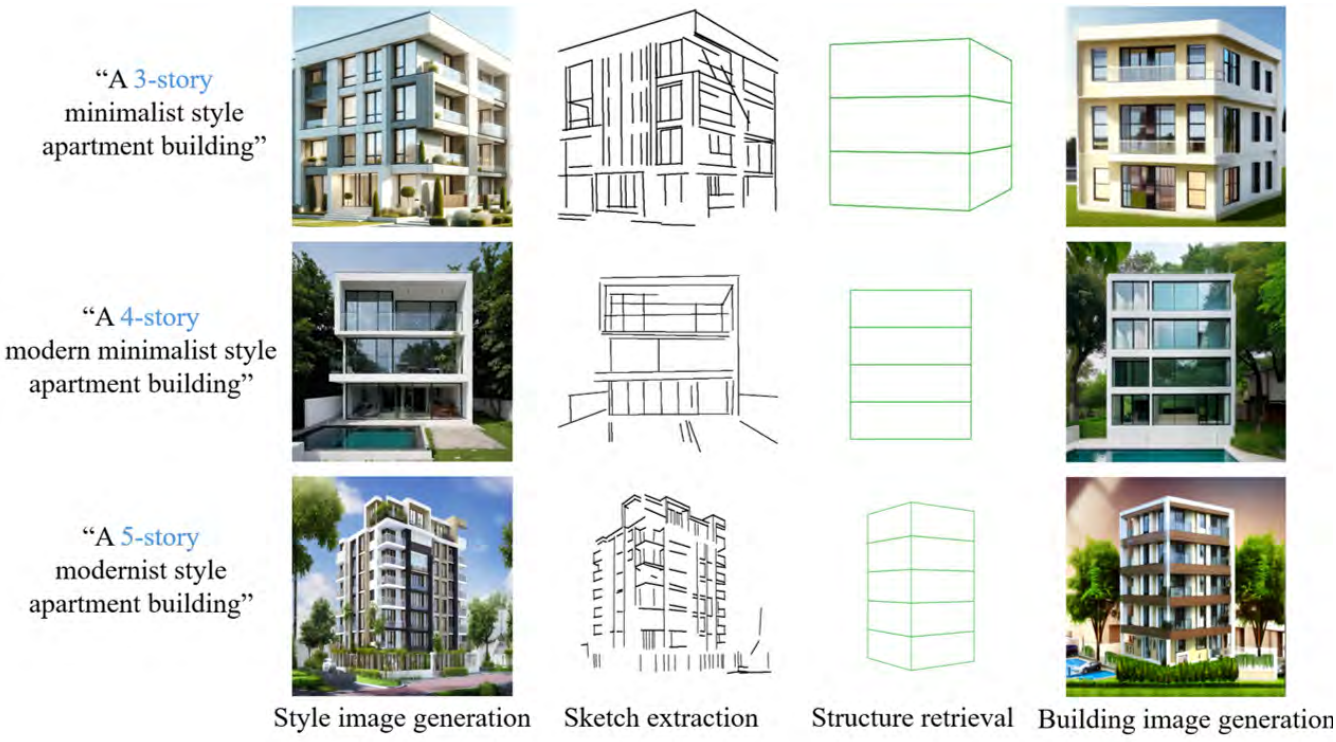

Fig. 4 The generation process of building image.

Fig. 5 The visual shows the LDM and our generated results respectively.

[4] Ho, J., Jain, A. and Abbeel, P.: Denoising diffusion probabilistic models, *Advances in Neural Information Processing Systems*, Vol. 33, pp. 6840–6851 (2020).
[5] Lewis, P., Perez, E., Piktus, A., Petroni, F., Karpukhin, V., Goyal, N., Küttler, H., Lewis, M., tau Yih, W., Rocktäschel, T., Riedel, S. and Kiela, D.: Retrieval-Augmented Generation for Knowledge-Intensive NLP Tasks (2021).
[6] Mou, C., Wang, X., Xie, L., Wu, Y., Zhang, J., Qi, Z., Shan, Y. and Qie, X.: T2I-Adapter: Learning Adapters to Dig out More Controllable Ability for Text-to-Image Diffusion Models (2023).
[7] Nadeau, D. and Sekine, S.: A survey of named entity recognition and classification, *Lingvisticae Investigationes*, Vol. 30, No. 1, pp. 3–26 (2007).
[8] Peng, Y., Zhao, C., Xie, H., Fukusato, T. and Miyata, K.: Sketch-Guided Latent Diffusion Model for High-Fidelity Face Image Synthesis, *IEEE Access*, Vol. 12, pp. 5770–5780 (online), DOI: 10.1109/ACCESS.2023.3346408 (2024).
[9] Rombach, R., Blattmann, A., Lorenz, D., Esser, P. and Ommer, B.: High-Resolution Image Synthesis with Latent Diffusion Models (2022).
[10] Sheynin, S., Ashual, O., Polyak, A., Singer, U., Gafni, O., Nachmani, E. and Taigman, Y.: KNN-Diffusion: Image Generation via Large-Scale Retrieval (2022).
[11] Sohl-Dickstein, J., Weiss, E., Maheswaranathan, N. and Ganguli, S.: Deep unsupervised learning using nonequilibrium thermodynamics, *International Conference on Machine Learning*, PMLR, pp. 2256–2265 (2015).
[12] Song, Y. and Ermon, S.: Generative modeling by estimating gradients of the data distribution, *Advances in neural information processing systems*, Vol. 32 (2019).
[13] Ye, H., Zhang, J., Liu, S., Han, X. and Yang, W.: IP-Adapter: Text Compatible Image Prompt Adapter for Text-to-Image Diffusion Models (2023).
[14] Zhang, L., Rao, A. and Agrawala, M.: Adding Conditional Control to Text-to-Image Diffusion Models (2023).
[15] Zhang, T. and Xie, H.: Sketch-Guided Text-to-Image Generation with Spatial Control, *2024 2nd International Conference on Computer Graphics and Image Processing (CGIP)*, pp. 153–159 (online), DOI: 10.1109/CGIP62525.2024.00035 (2024).
[16] Zhang, Y., Zhang, T. and Xie, H.: TexControl: Sketch-Based Two-Stage Fashion Image Generation Using Diffusion Model (2024).