ARCHITECTURAL GENERATIVE MODEL EVALUATION METHODS: IMAGE QUALITY ASSESSMENT METRICS AND EXPERT-BASED APPROACH

Taking the Chinese Campus Layout as an Examples

YING LIN[1] and FEI YE[2]
[1,2] Xi'an University of Architecture and Technology.
[1]lynn813910@gmail.com, 0009-0008-0132-381X
[2]feiye@xauat.edu.cn, 0009-0002-1444-2419

Abstract. The feasibility of using machine learning methods to generative architectural design solutions has been widely recognized as an effective in enhancing innovation, diversity, and efficiency of solutions. However, in generative design methods, the accuracy and quality of design results often rely on empirical evaluation of expert, which is challenging to evaluate and quantify by unified standards. This paper proposes a comprehensive method for evaluating model performance in architectural design tasks. The evaluation is based on computational criteria (i.e., FID, IS, SIMM indicators) and expert system criteria. The computational metrics will measure the distance, diversity, and similarity between the feature vectors of the real image and the generated image. In contrast, the expert criteria will measure the accuracy, intentionality, and rationality of the layout scheme. This study applies this framework to evaluate three widely used generative models in architectural design: GANs, Diffusion Models, and VAE. The framework also guides the optimization of generative models in architectural applications and assists architects in validating generative outcomes with more efficient workflows.

Keywords. Deep Learning, Generate Design, Evaluation Metrics, Campus Planning.

1. Introduction

Recently, with the development of computing power and technological breakthroughs, artificial intelligence and its application scenarios have experienced significant growth, especially in the fields of NLP (Natural Language Processing) and CV (Computer Vision). The widespread use of deep learning technology is driving architecture towards a new paradigm of digital design, playing an important role in the field of architecture and urban planning. This interdisciplinary approach uses complex neural networks to analysis, predict, and improve aspects of building and urban environment. Key application areas include design generation and optimization, urban data analysis,

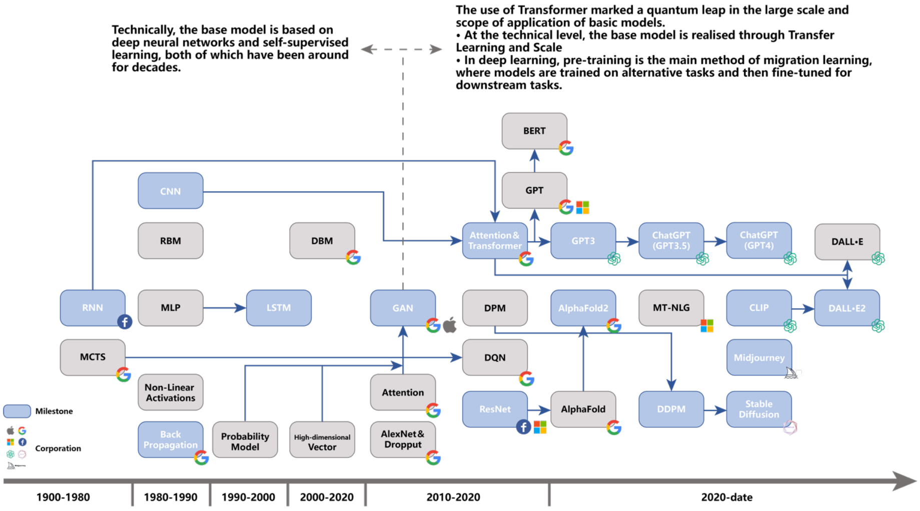
ACCELERATED DESIGN, Proceedings of the 29th International Conference of the Association for Computer-Aided Architectural Design Research in Asia (CAADRIA) 2024, Volume 1, 79-88. © 2024 and published by the Association for Computer-Aided Architectural Design Research in Asia (CAADRIA), Hong Kong.

prediction model, building performance simulation, historic preservation and reconstruction, construction management, and smart cities.

However, the rapid iteration of Artificial Intelligence Generated Content (AIGC) technology has significantly improved the performance of generative models, leading to the emergence of many variants with superior capabilities. These advances have come from typical generative models widely used in architecture research, such as GANs (Generative Adversarial Networks) (Goodfellow et al., 2014), Diffusion models (Ho et al., 2020), VAEs (Variational Auto-Encoders) (Kingma and Welling. et al., 2013), and Transformers (Vaswani et al., 2017). They often employ specific evaluation methods that are unique to computer science. Expert systems must empirically evaluate the accuracy and quality of results produced by generative design models. However, there is currently a lack of standardized evaluation frameworks in architectural research, which remains unavailable for objective measurement and quantification through consistent criteria. This research gap highlights the need to develop a comprehensive, interdisciplinary evaluation framework that can effectively combine computer science methods with the practical and aesthetic factors inherent in architectural design. Such a framework would facilitate not only the rigorous evaluation of generative models but also their practical application in architecture and urban planning.

1.1. GENERATIVE MODELS LITERATURE REVIEW

Generating new image from qualifying conditions is one of the challenging tasks in CV. So it have received a lot of attention in machine learning for their ability to generate new data instances that mimic real-world data distributions.



Technically, the base model is based on deep neural networks and self-supervised learning, both of which have been around for decades.

The use of Transformer marked a quantum leap in the large scale and scope of application of basic models.
- At the technical level, the base model is realised through Transfer Learning and Scale
- In deep learning, pre-training is the main method of migration learning, where models are trained on alternative tasks and then fine-tuned for downstream tasks.

Figure 1. Generation model development process

With the proliferation of generative model technology (Figure 1), architectural generative design is currently experiencing a significant era of opportunity, particularly in the realms of production tools and innovative methodologies. The application of artificial neural networks to the field of architectural design generation can be traced back to GANs, with several scholars undertaking studies on the transference of architectural facade styles and the creation of architectural plans. In June 2020, diffusion models garnered widespread attention. Subsequently, OpenAI's research, for the first time, demonstrated superiority over GANs (Dhariwal and Nichol, 2021). In 2022, OpenAI released DALL-E 2, Google introduced Imagen, and Stability AI made Stable Diffusion open source, marking the formal advent of a new era in the field of image generation. In 2023, research on diffusion models continues to optimize model training methods and multiple-modality. According to George Guidai, the latest advances in NLP and Diffusion Models will lead to significant changes in the way of architectural design (Guida, 2023).

1.2. RIDING ARTIFICIAL INTELLIGENCE WAVE IN ARCHITECTURE

In recent years, architectural image generation research has advanced significantly with the aid of cutting-edge computer technology, enabling architects to explore design possibilities, despite not achieving full autonomy.In the early stages of schematic design, Huang and Zheng (Huang and Zheng, 2018) and Chaillou (Chaillou, 2020) used pix2pix or other modified GANs to generated floor plans of apartments, gradually learning the locations of doors and windows. Sun (Sun et al., 2022) and Ali (Ali and Lee, 2023) explored architectural facade generation. Technological advances have also facilitated architectural rendering and representation studies, for example Wang (Wang et al., 2023) and Meng (Meng, 2022). Due to the complexity of the physical space of buildings, 3D generation has been the focus of research in the field of architecture, Zheng (Zheng, 2019) and Pang (Pang and Biljecki, 2022) have explored the methods of generating high-rise and street building 3D models respectively. YOUSIF (YOUSIF and BOLOJAN, 2021) studied the application of pix2pix model in automated building performance simulation, and SSIM index was introduced to evaluate the results, achieving a score of 0.94. The average similarity between Jia's (Jia, 2021) predicted daylight autonomy maps and the simulation results is as high as 91.51%.

1.3 EVALUATE THE GENERATION MODEL

When using generative models for architectural image generation tasks, the aim is to obtain high-quality generated images, mainly considering the following quantitative and qualitative aspects:

- The quality of the image itself, such as whether it is clear, whether it is realistic, whether it is diverse;
- The expert opinion that is similar to the ground truth, such as whether the layout is reasonable and whether the requirements of the specification are taken into account, among others.

In order to make a fair comparison between the three main categories and five sub-categories of models, we use pre-trained classifiers for computational purposes. Specifically, there is a class of automatic evaluation criteria within the computational metrics, that can be used to quickly measure the quality of the generated images, including IS for measuring the diversity and quality of the generated images; FID for comparing the similarity between the distribution of the generated images and the distribution of the real images. Furthermore, we measure the structural similarity between generated and real images using the SSIM.

2. Methodology

With reference to previous research experience, the evaluation of generative models in this paper revolves around four parts: 1. Model Training, 2. Image Generation, 3. Evaluation, 4. Analysis of Results (Figure 2).

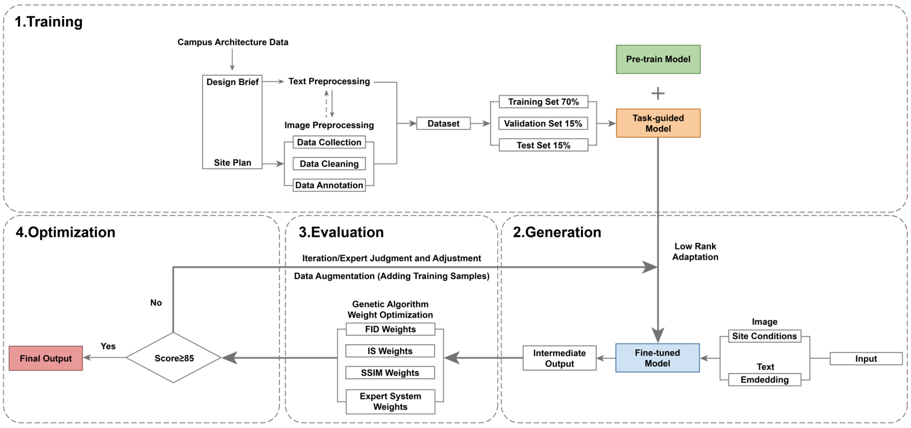


Figure 2. Workflow of the methodology

2.1. DATASET CONSTRUCTION

The dataset of this study, focusing on Chinese Campus Layout research, has three parts as data sources: Mapbox, public cases from ArchDaily, and design projects from our subject group. The publicly available master plans were collected manually and through Python crawling, and the valid cases were filtered. Specific screening rules are: 1. The zoning of the building master plan is clearly visible, and the compass is clear. 2. There is an independent standard playground runway (except those stacked vertically with other functional zones). 3. Classrooms are arranged unilaterally, and the spacing of the teaching buildings meets the requirements of the specification. 4. The neighbouring roads and land classification are clear. We screened a total of 155 valid samples this time.

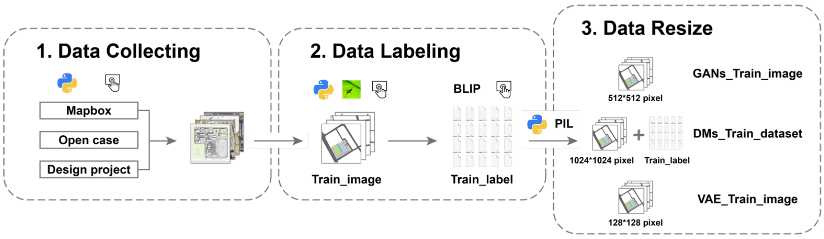
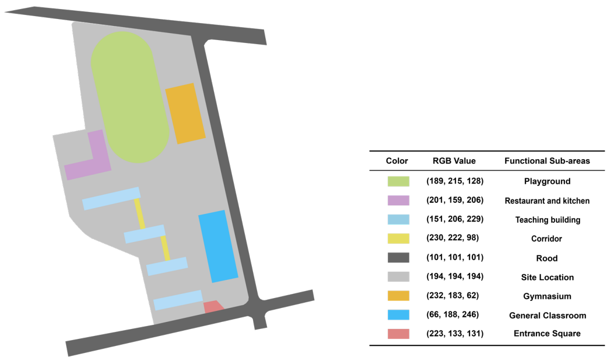


Figure 3. Dataset construction process

Next, we marked and labelled the samples (Figure 3). Here we use the Python Imaging Library (PIL) to process the image, extract the outline of the teaching building through OpenCV threshold segmentation and extraction of land boundary information, and store the coordinate information of the outline points, read the document information in Grasshopper, and use Polyline to connect lines to form a closed polygon, so that the vectorization of the site boundary and building profile is completed. Threshold segmentation of the elements in the general plane is labelled with RGB values (Figure 4), completing the image sample processing. Finally, using the general label model BLIP, ensuring one-to-one correspondence of the labelled image to generated TXT files. The sample label processing is finished,



Color	RGB Value	Functional Sub-areas
🟩	(189, 215, 128)	Playground
🟪	(201, 159, 206)	Restaurant and kitchen
🟦	(151, 206, 229)	Teaching building
🟨	(230, 222, 98)	Corridor
⬛	(101, 101, 101)	Rood
⬜	(194, 194, 194)	Site Location
🟧	(232, 183, 62)	Gymnasium
🟦	(66, 188, 246)	General Classroom
🟥	(223, 133, 131)	Entrance Square

Figure 4. Labelling rule

2.2. MODEL CONSTRUCTION

The idea of GAN confrontation is inspired by Game Theory, where the generator is trained, at the same time, a discriminator is applied to determine whether the input is a real image or a generated image, the two are getting stronger by playing each other in a Zero-sum Game, and a large number of realistic images can be generated by sampling (Figure 5a). The Diffusion Models defines two processes, forward and backward, both of each sample from the real data distribution and gradually add Gaussian noise to the samples, and generate a series of noisy samples, and the noise addition process can be controlled by the variance parameter (Figure 5b). VAE is a variant of auto-encoder, its purpose is to train neural networks in an unsupervised way, including Encoder and Decoder: Encoder process is to compress the original data into low-dimensional vectors, and Decoder is to restore the low-dimensional vectors to the original data (Figure 5c).

In recent years, GAN and VAE have shown great potential in the task of sampling a given data distribution to generate a new one. GAN learns the sampling procedure of complex distributions in an adversarial manner to learned them, while VAE seeks a model that is high likelihood to the distribution of data samples. Although the satisfactory performance of these models in producing high-quality images, they have some limitations of their own. Due to the adversarial of training, GANs tend to suffer from training mode collapse and less distribution coverage, so therefore inferior to SOTA model likelihood-based VAE models in terms of diversity. VAEs can capture more diversity and are often easier to scale and train than GANs, but still fall short in terms of visual sample quality and sampling efficiency (Chai et al., 2023). Recently, diffusion models such as DDPM have emerged as another powerful class of generative models capable of producing high-quality images comparable to GANs (Dhariwal and

Nichol, 2021), with desirable properties such as strong sample diversity, realistic probability distributions, adaptability to different training goals, and ease of scaling. This means that Diffusion Models are well suited for learning complex and diverse data, which motivates us to explore the potential of Diffusion-based generative models for architectural image.

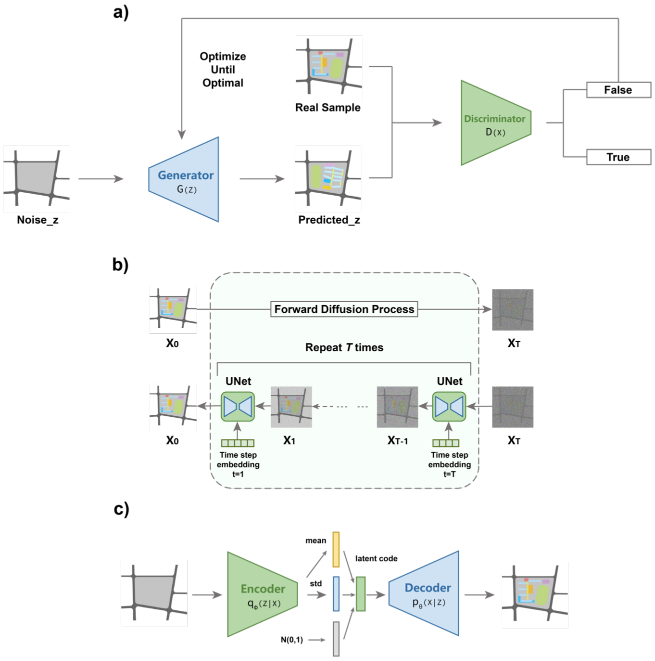


Figure 5. a) Workflow of GAN; b) Workflow of Diffusion Model; c) Workflow of VAE

In this article, three major categories and five subcategories of generative models are selected for evaluation: GANs (pix2pix and cyclegan), DMs (Stable Diffusion, SDXL), and VAE (VQ-VAE), input the real site boundaries and neighbouring road conditions to output the corresponding campus layout images. The models are trained to learn the layout rules from the real campus, and the general plan is automatically generated. All experiments are performed on NVIDIA GeForce RTX 4060 GPU (8 GB).

2.3. EXPERIMENTAL SETTINGS

	Model	sample number	resolution	batch_size	learning rate	epoch
GANs	pix2pix	134	512*512	8	0.0002	220
	cyclegan	134	256*256	8	0.0002	180
DMs	SD	48	512*512	12	0.0001	100
	SDXL	48	1024*1024	10	0.0001	60
VAE	VQ-VAE	155	128*128	32	0.0002	150

Table 1. Implementing details of model training.

Initially, we produce corresponding datasets for each of the five different models based on their requirements for images and text label. For instance, the GANs model necessitates paired input and real images, the diffusion model demands one-to-one correspondence between the input image and text label, and the VAE model's training duration is directly proportional to the pixel count of the image. Consequently, the VAE dataset is divided to optimize computational costs. Upon reviewing the training data from prior studies and conducting numerous experiments, we selected the optimal batch size, learning rate, and number of epochs to enhance training (Table 1).

2.4. GENERATING RESULTS

After training, the neural network can be used to generate a new layout based on the input site boundaries and road conditions (Figure 6). Based on the results, the generated images are compared with ground truth to assess similarity.

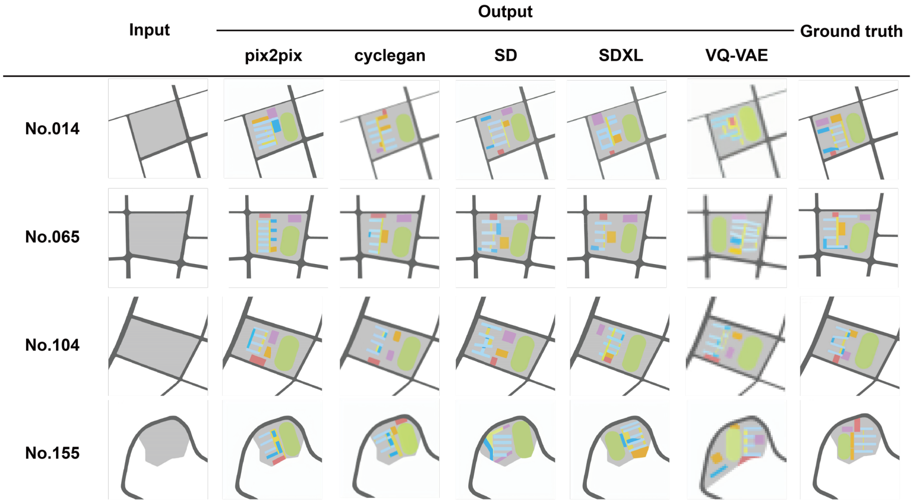


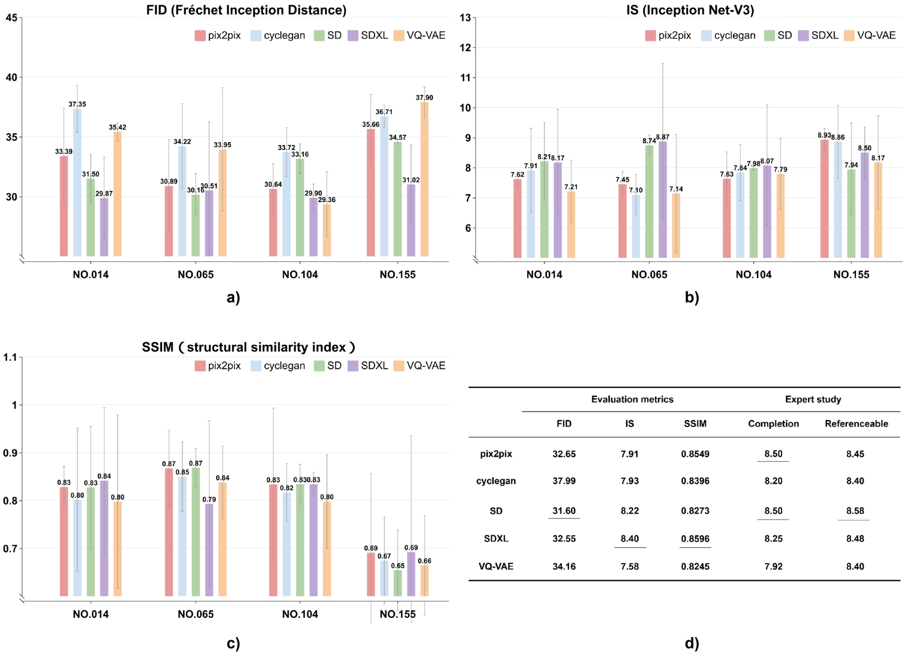
Figure 6. Generated results buy different Generative *Models*

The experimental results show that our trained model somewhat learns the layout rules of the middle school campus. The selection of test samples focuses on rectangular site, with irregular terrain also chosen as an example. In No. 014, 065 and 155, GANs and DMs learned the layout rules of the teaching buildings and the playground well. Regarding functional layout, the results generated mainly by SD (Stable Diffusion) and SDXL are highly consistent with the original plan. The relationship between corridors, general classrooms, and teaching building, as well as between the gymnasium and playground, and the entrance square and the road, are more reasonable. Almost all test samples successfully output the orientation of the classroom buildings and the long-axis direction of the playground. The final result is as expected. In sample No. 155, the overall orientation of the layout is skewed, but it's adapted to terrain. After a preliminary analysis, it appears that the dataset is insufficient, particularly regarding the number of samples of irregularly curved terrain, resulting in the model not learning the relationship between the road, the terrain, and the layout very well.

3. Evaluating discussion

3.1. QUANTITATIVE EVALUATION

First, we evaluated the FID metrics on images generated by each generative model. A lower FID corresponds that generated images are more similar to the real images. Figure 7a shows that the DMs (SD and SDXL) consistently show lower FID scores, in tests Nos. 014, 065, and 155, indicating they often produce images closest to the real image distribution, SDXL in particular performed better. Pix2pix also has lower FID scores in tasks with similar delicate structures, due to the model properties applied to paired images,it has the second-highest generation quality after the DMs in all four tests, shows great potential in architectural image generation.



	Evaluation metrics			Expert study	
	FID	IS	SSIM	Completion	Referenceable
pix2pix	32.65	7.91	0.8549	8.50	8.45
cyclegan	37.99	7.93	0.8396	8.20	8.40
SD	31.60	8.22	0.8273	8.50	8.58
SDXL	32.55	8.40	0.8596	8.25	8.48
VQ-VAE	34.16	7.58	0.8245	7.92	8.40

d)

Figure 7. Evaluation results; a) FID; b) IS; c) SSIM; d) Evaluation metrics and Expert study

In this paper, the Inception V3 model is used to evaluate IS indicators, which has better classification performance, widely used in image recognition and classification tasks (Figure 7b). The high scores indicate that the model generates diverse and reasonably clear images. The SD consistently scores the highest across all samples, indicating it generally produces clearer and more diverse images compared to the other models. The VQ-VAE is generally good at accurately reconstructing or interpolating between existing images, thus, which might be due to less diversity in the images it generates.

The SSIM measures the similarity between two images and takes values in the range of [0,1]. The larger the value, the closer in structure to ground truth. Figure 7c shows the images generated by each model and the ground truth input function. Except for test NO. 155, pix2pix consistently performed the best across all samples in other datasets. The DMs showed strong performance but were not always superior to other models, possibly due to the samples' specificity or the model's training.

3.2. QUALITATIVE COMPARISONS

We distributed questionnaires through online and offline methods, targeting mainly architects, architectural faculty and students, designers, and other related practitioners. A total of 67 pieces of valid data were returned, each subject compared four pairs of layouts, sampled from five generative models that generate images, with the ground truth.and the statistical results are shown in Figure 7d. Expert Study with a range of [0,10], calculating the mean of each score and tallying. The metrics 'Completion' and 'Referenceable' are subjective measures evaluated by professionals proficient in the relevant field for the generated images.

Both pix2pix and SD scored 8.50 in Completion, the highest in the category. This indicates that they were able to generate fully realised images for the assigned tasks. Pix2pix is particularly good at processing detailed structural information due to its paired training data, making it well suited for targeted explicit tasks. SD's high score indicates its ability to generate complete images by understanding and reproducing complex image features. SD had the highest reference score of 8.58, indicating that experts believed it generated the most accurate and reliable images relative to real ones. SDXL followed closely at 8.48, indicating that it also generated high-quality, reference-worthy images.

4. Conclusion

In this paper, we discuss the training and evaluation of architectural image generation models, different metrics lead to different trade-offs, and different evaluation scores will benefit different models. Therefore, it is important to train and evaluate according to the specific situation of the target application. In addition, when selecting models for architectural image generation tasks, we should be careful not to take good performance in one task as evidence of good performance in another application.

According to the evaluation results of the campus master plan layout generation experiment, in terms of computational criteria, pix2pix generates images closer to the ground truth, while the latest diffusion models SD and SDXL are more enlightening and diverse. Expert indicators show that the images generated by pix2pix and SD models are more complete, while SDXL is more informative and innovative. In summary, within the model evaluation framework proposed by this paper, the diffusion model excels, demonstrating superior computational performance and receiving high expert evaluations. This indicates that the diffusion model is adept at executing the building floor plan generation task, thereby informing future collective building layout planning. Although our proposed evaluation framework is impressive compared with existing architectural layout generation methods, it still has some limitations. For example, when architects want to apply the framework to filter the generated images as their main reference, they need to deploy the evaluation system manually. This greatly increases the learning cost and is not conducive to the promotion of the method.

References

Goodfellow, I., Pouget-Abadie, J., Mirza, M., Xu, B., Warde-Farley, D., Ozair, S., ... & Bengio, Y. (2014). Generative adversarial nets. *Advances in neural information processing systems, 27.*

Ho, J., Jain, A., & Abbeel, P. (2020). Denoising diffusion probabilistic models. *Advances in neural information processing systems, 33,* 6840-6851.

Kingma, D. P., & Welling, M. (2013). Auto-encoding variational bayes. *arXiv preprint arXiv:1312.6114.*

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., ... & Polosukhin, I. (2017). Attention is all you need. *Advances in neural information processing systems, 30.*

Dhariwal, P., & Nichol, A. (2021). Diffusion models beat gans on image synthesis. Advances in neural information processing systems, 34, 8780-8794.

Guida, G. E. O. R. G. E. (2023, March). Multimodal Architecture: Applications of Language in a Machine Learning Aided Design Process. In *HUMAN-CENTRIC-Proceedings of the 28th CAADRIA Conference. Ahmedabad* (pp. 18-24).

Huang, W., & Zheng, H. (2018, October). Architectural drawings recognition and generation through machine learning. In *Proceedings of the 38th annual conference of the*

Chaillou, S. (2020, September). Archigan: Artificial intelligence x architecture. In *Architectural Intelligence: Selected Papers from the 1st International Conference on Computational Design and Robotic Fabrication (CDRF 2019)* (pp. 117-127). Singapore: Springer Nature Singapore. *association for computer aided design in architecture, Mexico City, Mexico* (pp. 18-20).

Sun, C., Zhou, Y., & Han, Y. (2022). Automatic generation of architecture facade for historical urban renovation using generative adversarial network. *Building and Environment, 212,* 108781.

Ali, A. K., & Lee, O. J. (2023). Facade style mixing using artificial intelligence for urban infill. *Architecture, 3*(2), 258-269.

Wang, B., Zhang, S., Zhang, J., & Cai, Z. (2023). Architectural style classification based on CNN and channel–spatial attention. *Signal, Image and Video Processing, 17*(1), 99-107.

Meng, S. (2022). Exploring in the latent space of design: A method of plausible building facades images generation, properties control and model explanation base on stylegan2. In *Proceedings of the 2021 DigitalFUTURES: The 3rd International Conference on Computational Design and Robotic Fabrication (CDRF 2021) 3* (pp. 55-68). Springer Singapore.

Zhang, H. (2019). 3D model generation on architectural plan and section training through machine learning. *Technologies, 7*(4), 82.

Pang, H. E., & Biljecki, F. (2022). 3D building reconstruction from single street view images using deep learning. *International Journal of Applied Earth Observation and Geoinformation, 112,* 102859.

YOUSIF, S., & BOLOJAN, D. (2022). Deep learning-based surrogate modeling for performance-driven generative design systems. In *Proc of the 27th International Conference of the Association for Computer-Aided Architectural Design Research in Asia (CAADRIA), CAADRIA* (pp. 363-372).

Jia, M. (2021). Daylight prediction using Gan: General workflow, tool development and case study on Manhattan, New York.

Chai, S., Zhuang, L., & Yan, F. (2023). LayoutDM: Transformer-based Diffusion Model for Layout Generation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition* (pp. 18349-18358).