TEXT-TO-CITY

Controllable 3D Urban Block Generation with Latent Diffusion Model

JUNLING ZHUANG[1*], GUANHONG LI[2*], HANG XU[3], JINTU XU[4], RUNJIA TIAN[5]
[1]*Columbia University in the City of New York,* [2]*Architectural Association School of Architecture,* [3]*The University of Sheffield,* [4]*Southeast University,* [5]*Harvard University,* [*]*indicates equal contribution to work*
[1]*junling.zhuang@columbia.edu, 0000-0002-2895-5445*
[2]*wafo210715@gmail.com, 0009-0009-9989-5513*
[3]*815191192@qq.com, 0000-0002-4063-6295*
[4]*tayhsu@outlook.com, 0009-0000-9519-161X*
[5]*runjiatian@gmail.com, 0000-0002-5983-9754*

Abstract. The rise of deep learning has introduced novel computational tools for urban block design. Many researchers have explored generative urban block design using either rule-based or deep learning methods. However, these methods often fall short in adequately capturing morphological features and essential design indicators like building density. Latent diffusion models, particularly in the context of urban design, offer a groundbreaking solution. These models can generate cityscapes directly from text descriptions, incorporating a wide array of design indicators. This paper introduces a novel workflow that utilizes Stable Diffusion, a state-of-the-art latent diffusion model, to generate 3D urban environments. The process involves reconstructing 3D urban block models from generated depth images, employing a systematic depth-to-height mapping technique. Additionally, the paper explores the extrapolation between various urban morphological characteristics, aiming to generate novel urban forms that transcend existing city models. This innovative approach not only facilitates the accurate generation of urban blocks with specific morphological characteristics and design metrics, such as building density, but also demonstrates its versatility through application to three distinct cities. This methodology, tested on select cities, holds potential for broader range of urban environments and more design indicators, setting the stage for future computational urban design research.

Keywords. deep learning, generative design, latent diffusion model, urban block morphology, artificial intelligence.

1. Introduction

Urban block design, a critical facet of urban planning, has been transformed by novel

ACCELERATED DESIGN, Proceedings of the 29th International Conference of the Association for Computer-Aided Architectural Design Research in Asia (CAADRIA) 2024, Volume 2, 169-178.

generative design methods, introducing advanced computational tools. Despite significant research employing rule-based and deep learning methods for generative urban block design, these approaches often inadequately capture essential morphological features and design indicators like building density.

Latent diffusion models present innovative solutions for urban block design. These models can generate cityscapes directly from text descriptions, which encompass various design indicators. These models can also be customized for different cities, enhancing their adaptability in capturing diverse urban characteristics. They can even be conditioned based on input masks to better represent the surrounding urban context.

This study introduces a cutting-edge workflow for creating 3D urban models using the advanced latent diffusion model, Stable Diffusion (Rombach et al., 2021). Our methodology includes: (1) dataset development, (2) model fine-tuning, and (3) model evaluation. We initially generated paired depth images and text data from 3D urban models, utilizing OpenStreetMap to source data from cities like Berlin, Hamburg, and Cambridge, chosen for their unique traits. Depth mapping and segmentation into patches were conducted to calculate crucial design metrics, such as building density. These metrics, combined with image data, formed our dataset. Subsequently, we fine-tuned the Stable Diffusion model using Dreambooth (Ruiz et al., 2023). The model's performance was assessed by comparing generated quantitative metrics with actual data.

In the post-processing step, we input urban morphology keywords and building density into the model to produce depth maps, which were then used to reconstruct 3D urban blocks using the depth-to-height mapping obtained from the dataset creation process. Our parametric generation process enables adjustment of building heights to meet exact building density input constraints. We also experimented interpolating between different urban morphological characteristics to extrapolate novel urban forms beyond existing cities. This workflow holds potential for broader application across more cities and diverse design indicators.

2. Related Work

2.1. RULE-BASED URBAN BLOCK GENERATION

Early development in computational urban block generation usually involves use of rule-based algorithms. Rahimian, M. proposed the use of shape grammar for generative urban design in San Diego (Rahimian, M. et al. 2019). Specifically, Kelly (2021) developed a rule-based urban modelling platform to transform 2D data into 3D urban models. This method enables users to promptly tailor their models to urban design objectives. However, finding a rule writer for this approach is tough. It needs urban design expertise, some math understanding, and rule-to-algorithm skills. Also, it's complex for non-computer experts in early urban planning.

2.2. GAN-BASED GENERATION

With the development of deep generative models such as Generative Adversarial Networks (GANs), significant strides have been made in urban block generation. Fedorova (2021) employed GANs (Goodfellow et al. 2014) for predicting urban block

layouts, learning from the surrounding urban context. Tian et al. (2021) utilized Pix2Pix for generating urban layouts based on boundary shapes, while Boim et al. (2022) applied Pix2Pix GANs for creating non-planned settlements from existing urban features. Researchers have increasingly used additional conditions to enhance GAN outputs. Shen et al. (2020) integrated constraints like roads and landscapes, Liu et al. (2021) included factors such as boundaries and water features, and Zhong et al. (2022) explored a GAN Data Label Setting (DLS) for customizable urban designs.

2.3. LATENT DIFFUSION MODEL: TEXT-TO-IMAGE GENERATION

Text-to-image models, integrating natural language processing and computer vision, bridge textual and visual domains. Mansimov et al. (2015) pioneered image generation from natural language descriptions using deep neural networks. Building on GANs (Goodfellow et al. 2014), Wang et al. (2018) developed a method for creating images from semantic label maps using conditional GANs. Ramesh et al. (2021) explored a text-to-image generation method using an autoregressive transformer. Dhariwal (2021) suggested diffusion models surpass GANs in image quality, leading Rombach (2022) to create latent diffusion models (LDMs) that accept multiple conditions like text, image, and semantic maps in latent space. Recently, Ruiz et al. (2023) introduced Dream Booth, a fine-tuning approach for LDMs for specific applications. These advancements underscore the feasibility of generating urban designs from textual representations and hold promise for large-scale urban generation with flexible control.

3. Methodology

Our steps of model training and conditional generation is shown in Figure 1

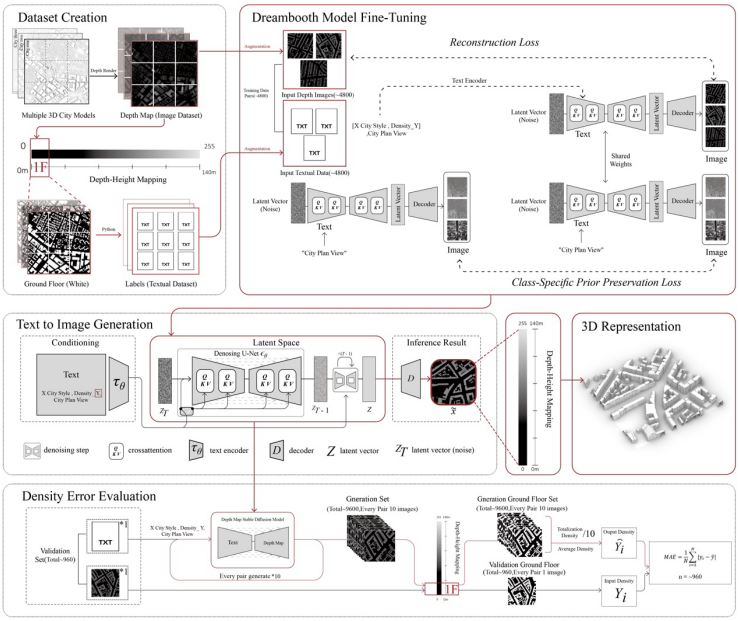


Figure 1. Research Methodology Diagram

3.1. DATASET CREATION AND AUGMENTATION

Depth Image from 3D City Model Data. From OpenStreetMap (OSM), we extracted 3D city models of several cities with unique morphology at Level of Detail2(LoD2) which allow various shapes of building geometry including pyramids, domes and gabled, skillion roofs, enriching the diversity of urban context. The city models are then rendered with z-depth information. We normalized the depth values by placing blocks with the same maximum height to ensure consistent value range mapping for urban blocks with various density conditions. The city depth map was then sliced into 300*300-meter segments representing individual urban blocks.

Textual Data Generation. Figure2 shows a threshold-based algorithm was employed to count building density by dividing the number of pixels in the ground floor of each depth map by the total number of pixels and output as textual data that are associated with depth maps then paired to form a training dataset. For each city morphology characteristic, we prepared 800 data pairs, culminating in a total of 4800 pairs.

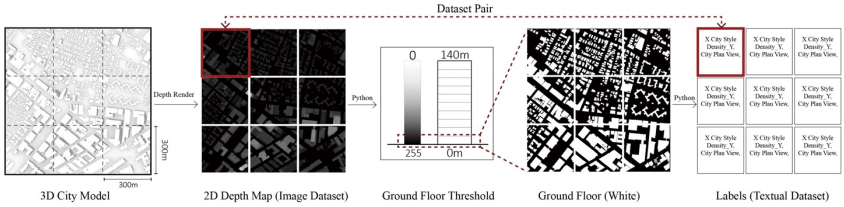


Figure 2. Depth map render and automatically building density labelling

3.2. TEXT-TO-DEPTH-IMAGE GENERATION WITH FINE-TUNED LATENT DIFFUSION MODEL

Text-to-Depth-Image Generation with Latent Diffusion. We leverage the cross-modality characteristics of latent diffusion models to learn the mapping from text to depth map. We chose Stable Diffusion version 1.5, the State-of-the-Art latent diffusion model as our base model, which is capable of generating high-quality images from text inputs. Stable Diffusion models can also be fine-tuned using methods such as DreamBooth (Ruiz et al., 2023) or LoRA to implant novel a-prior knowledge to adapt to more use cases.

Fine-Tuning Latent Diffusion Model. Our objective is to "implant" new (key, value) pairs into the "dictionary" of the diffusion model so that given the key for our subject, the production of depth maps of particular city and density by a text prompt using DreamBooth (Ruiz et al., 2023). The first step is to insert the depth image subject instance into the model and connect the depth map subject to a distinctive identifier. So we input 4800 depth images with three different city which paired with a text prompt containing two unique identifier which reflect the certain city style and accurate building density value, also the name of the class the subject belongs to (e.g., "[X City Style] , [Density_Y] ,City Plan View "). In this way our depth images with the same city style and building density values share the same two unique identifiers, allows pairing the corresponding semantics and pictures in the process of reconstructing U-Net's Loss values, so that our trained model can generate depth maps corresponding to

the key design metrics based on different text prompts. In parallel, we apply a class-specific prior preservation loss:

$$\mathbb{E}_{x,c,\epsilon,\epsilon',t}\left[w_t \parallel \hat{X}_\theta(\alpha_t x + \sigma_t \epsilon, c) - x \parallel_2^2 + \lambda w_{t'} \parallel \hat{X}_\theta(\alpha_{t'} x_{pr} + \sigma_{t'} \epsilon', c_{pr}) - x_{pr} \parallel_2^2\right],$$

which leverages the semantic prior that the model has on the class and encourages it to generate diverse instances belonging to the subject's class using the class name in a text prompt (e.g., "City Plan View").

3.3. TESTING AND EVALUATION

Text-to-image tasks face challenges in quantitatively evaluating design indicator accuracy. We addressed this by combining quantitative analysis with image generation, dividing our dataset into training and validation sets. For model fine-tuning, the validation set's density values 'y' were used to generate and evaluate depth maps, comparing predicted and actual 'y' values. The L1 loss across these sets gauged model accuracy. This process, crucial for model selection, does not affect DreamBooth training convergence but aids in identifying the model with the best metrics.

3.4. 3D POST-PROCESSING

Our approach incorporates a Grasshopper script to translate depth map grey values into accurate building heights using a depth-to-height mapping obtained from the training data. As shown in Figure 3, users can adjust the restored building heights by setting an average height value. The script recalculates the height-to-grey scale mapping, modifying the grayscale in the building areas to maintain roof slopes and relationships during 3D extrusion. This ensures that luminance values do not exceed the grayscale maximum of 255, preserving original geometric features of roofs in the 3D urban landscape.

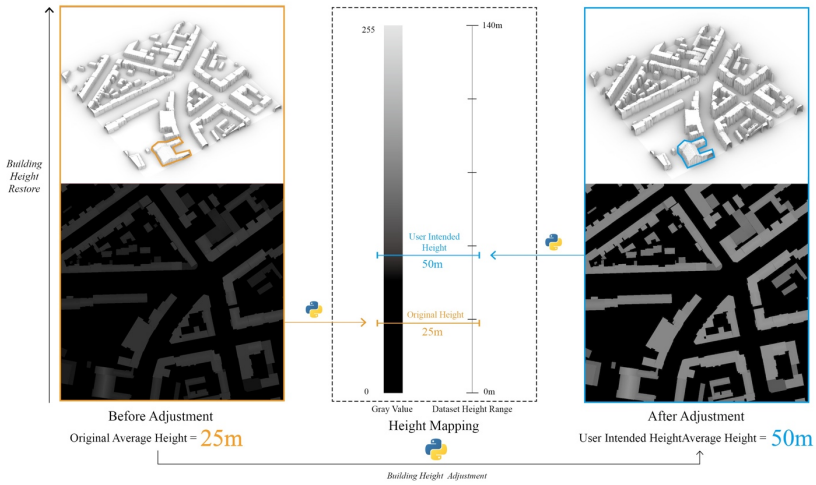


Figure 3. Depth map render and automatically building density labelling

4. Experiment

4.1. DATASET FORMAT, AUGMENTATION AND VALIDATION

Depth Image Data Collection. The study analysed Berlin, Hamburg, and Cambridge (USA), each with unique street patterns—Berlin's linear, Hamburg's enclosed, and Cambridge's grid-like. The dataset featured diverse roof styles and used a standardized 140-meter height mapping for depth images. Building density metrics were calculated for balanced distribution within each city, resulting in a dataset of 100 images per city.

Textual Data Format. In our study, we tested eight caption formats for image-text pairing, with Figure 6 highlighting the top three. Format 3, such as "Baroque town texture, Density_16, City Plan View" for Berlin, was the most effective. It aligns with Dreambooth's identifier strategy, ensuring accurate density labels and reducing ambiguity compared to natural language formats.

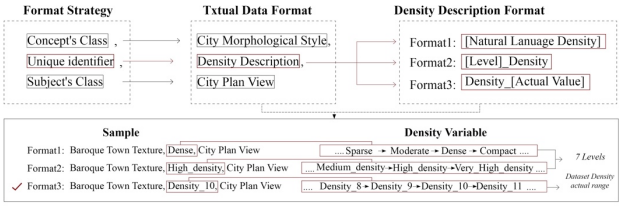


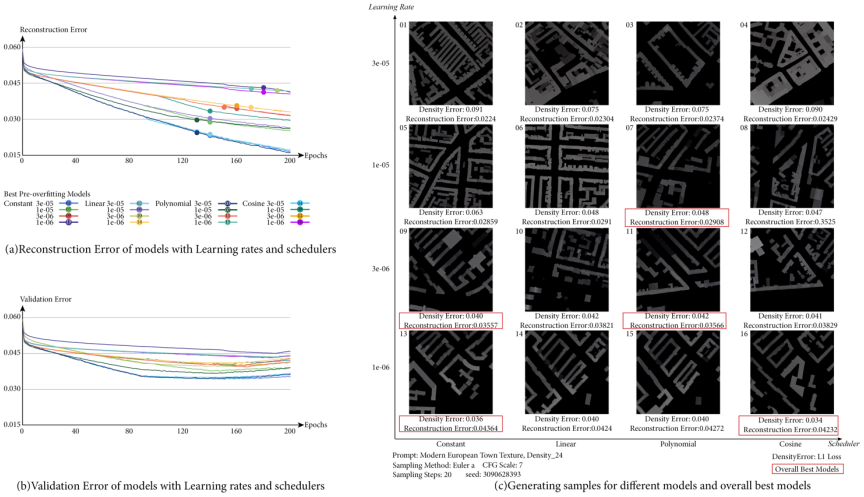
Figure 4. Textual data format strategy and experiment

Data Augmentation. We expanded our dataset sixteen fold using rotation and flip transformations, creating three datasets for different cities. Each contains 1660 pairs of depth maps and text descriptions, totalling 4800 training pairs. Figure 7 shows these pairs, highlighting city-wise building density variations.

Figure 5. Berlin, Cambridge and Hamburg partial training dataset pairs samples

4.2. MODEL FINE-TUNING

Fitting Model on Dataset. For model validation, we used 20% of our dataset as a validation set, focusing on learning rate and scheduler during hyperparameter optimization. The best setup (No. 09, Figure 6) achieved a 0.03557 reconstruction error and 0.040 density loss over 160 epochs on an NVIDIA 4090 GPU, using a 3e-06 learning rate with a constant scheduler. Extensive testing over 200 epochs revealed that neural network loss is more affected by learning rate than schedulers. A balance between fast convergence and low density error was found with lower learning rates and a cosine scheduler. Our top models (Nos. 07, 09, 11, and 16), selected for their balanced performance, are adaptable to different use cases, as depicted in Figure 6-C.



(a)Reconstruction Error of models with Learning rates and schedulers

(b)Validation Error of models with Learning rates and schedulers

(c)Generating samples for different models and overall best models

Figure 6. Models' evaluation with different learning rates and schedulers

4.3. RESULTS

Density Error Evaluation. Figure 7 illustrates the performance of our top model, No. 9, using Berlin as an example. It shows higher accuracy in generating textual prompts within the dataset's density range and increased error values outside this range.

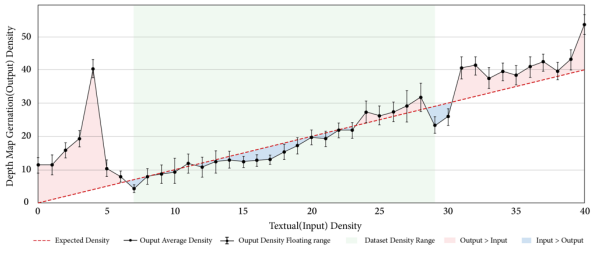


Figure 7. Models' evaluation with different learning rates and schedulers

2D-to-3D Post-Processing. We ran batch inference generation of grayscale urban morphology patterns and extrude these images using our Grasshopper from section 3.4 to reconstruct 3D urban models as shown in Figure 8.

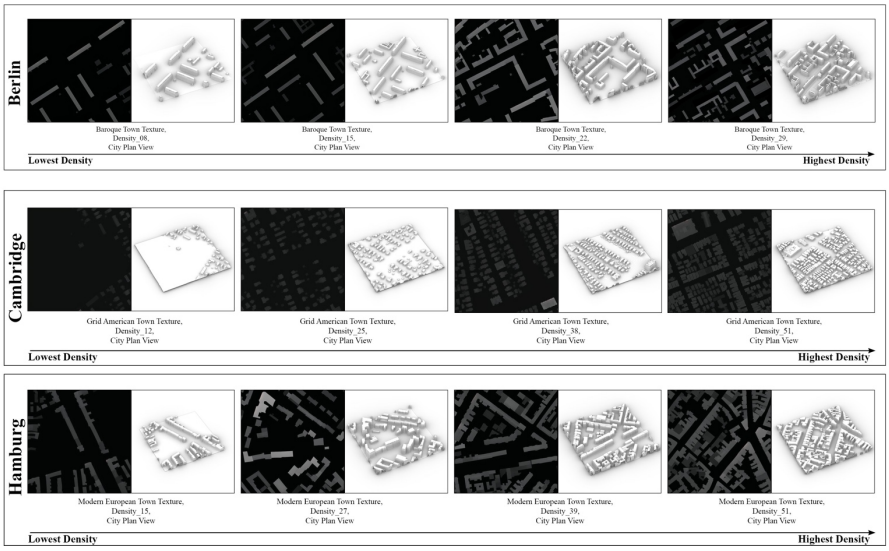
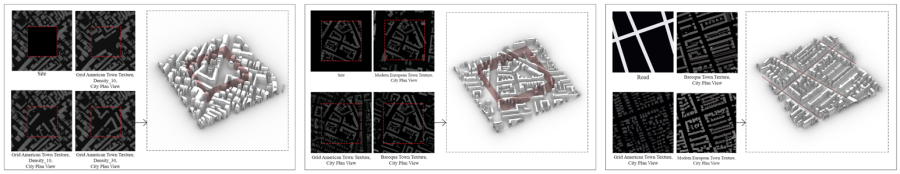


Figure 8. 3D Extrusion with different density text prompt and height adjustment

Conditional Generation. Figure 9 shows our model allows the user to control output according to site and road conditions using ControlNet (Zhang et al. 2023), including specific city morphology, which is seen that the generated results fit the text input and can well connect to the surrounding urban morphology.



(a)Application 1: Inpainting Site conditional generation (b)Application 2: OutPainting Site conditional generation (c)Application 3: Road Conditional Generation

Figure 9. Site and Road Conditional Generation

Morphological Extrapolation. Deforum Diffusion was utilized to merge city morphologies by interpolating between keyframes based on different text prompts. This approach generated hybrids of city styles, as shown in two examples: (1) Figure 10(a) depicts the blending of Berlin's large buildings with Cambridge's smaller ones at frame 120, and (2) Figure 10(b) shows the mix of Berlin's flat roofs with Hamburg's sloping roofs, also at frame 120. This shows that our model is capable of synthesizing novel urban forms.

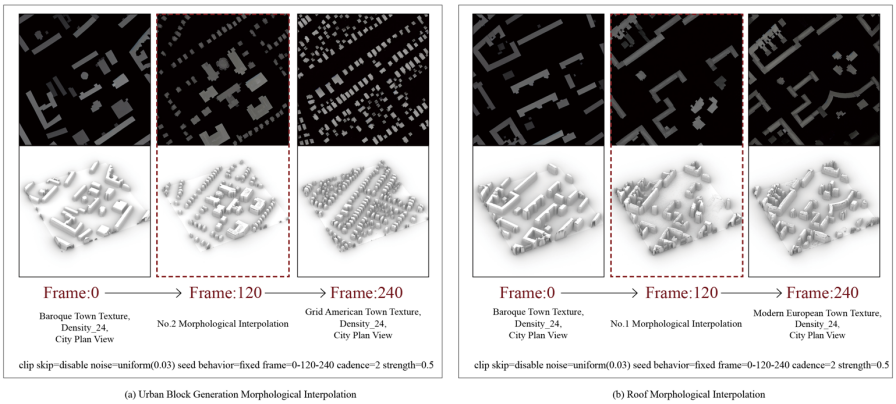


Figure 10. Urban blocks and roof morphological extrapolation

5. Conclusion and Future Work

Our research involved refining the Stable Diffusion Model for generating depth maps from text prompts detailing city style and density. These maps were then used to create 3D urban blocks, enabling the exploration of varied urban morphologies through conditional generation. Additionally, we developed a novel method for assessing building density error, applicable to various quantifiable design metrics.

Our model offers more controlled generation compared to Pix2Pix GANs, allowing precise prediction of urban layouts based on user input. However, it is limited to prompts within the training dataset's scope and struggles with multi-indicator prompts (e.g., combining Floor Area Ratio and Density), which compromise model quality.

Due to the multifarious factors impacting real-world systems, future research could benefit from expanding dataset scopes to encompass elements such as functionality, ecology, regional culture, and regulations, among others. Incorporating these components into text prompts could enable the tackling of more intricate urban planning challenges, paving the way for applications that are more aligned with real-world needs.

References

Boim, A., Dortheimer, J. & Sprecher, A. (2022). A machine-learning approach to urban design interventions in non-planned settlements. In 27th International Conference on Computer-Aided Architectural Design Research in Asia, CAADRIA 2022 (pp.223-232).The Association for Computer-Aided Architectural Design Research in Asia (CAADRIA).

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

Fedorova, S., (2021). Generative Adversarial Networks for Urban Block Design. In SimAUD 2021: Symposium on Simulation for Architecture and Urban Design.

Goodfellow, I., Pouget-Abadie, J., Mirza, M., Xu, B., Warde-Farley, D., Ozair, S., & Bengio, Y. (2020). Generative adversarial networks. Communications of the ACM, 63(11), 139-144. https://doi.org/10.1145/3422622

Kelly, T. (2021). CityEngine: An Introduction to Rule-Based Modeling. In: Shi, W., Goodchild, M.F., Batty, M., Kwan, MP., Zhang, A. (eds) Urban Informatics(pp.637–662).Springer

Liu, Y., Fang, C., Yang, Z., Wang, X., Zhou, Z., Deng, Q., & Liang, L. (2022). Exploration on machine learning layout generation of Chinese private garden in Southern Yangtze. In 3rd International Conference on Computational Design and Robotic Fabrication, CDRF 2021. (pp.35-44).

Rahimian, M., Beirão, J. N., Duarte, J. M. P., & Iulo, L. D. (2019). A Grammar-Based Generative Urban Design Tool Considering Topographic Constraints The Case for American Urban Planning. In 37th Conference on Education and Research in Computer Aided Architectural Design in Europe and 23rd Conference of the Iberoamerican Society Digital Graphics, eCAADe SIGraDi 2019 (pp. 267-276).

Mansimov, E., Parisotto, E., Ba, J. L., & Salakhutdinov, R. (2015). Generating images from captions with attention. arXiv preprint arXiv:1511.02793.

Ramesh, A., Pavlov, M., Goh, G., Gray, S., Voss, C., Radford, A., & Sutskever, I. (2021). Zero-shot text-to-image generation. In International Conference on Machine Learning, ICML 2021 (pp.8821-8831).

Rombach, R., Blattmann, A., Lorenz, D., Esser, P., & Ommer, B. (2022). High-resolution image synthesis with latent diffusion models. In Proceedings of the IEEE/CVF conference on computer vision and pattern recognition, CVPR 2022 (pp.10684-10695).

Ruiz, N., Li, Y., Jampani, V., Pritch, Y., Rubinstein, M., & Aberman, K. (2023). Dreambooth: Fine tuning text-to-image diffusion models for subject-driven generation. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2023 (pp.22500-22510).

Shen, J., Liu, C., Ren, Y., & Zheng, H. (2020). Machine learning assisted urban filling. In 25th International Conference on Computer-Aided Architectural Design Research in Asia, CAADRIA 2020 (pp.5-6). The Association for Computer-Aided Architectural Design Research in Asia (CAADRIA).

Tian, R. (2021). Suggestive Site Planning with Conditional GAN and Urban GIS Data. In The 2nd International Conference on Computational Design and Robotic Fabrication, CDRF 2020. (pp.103–113).

Vasanthakumar, S., Saha, N., Haymaker, J., & Bibil, S. D. (2017). A Performance-Based Framework to Determine Built Form Guidelines. In Proc. 37th Annual Proceeding of eCAADe., Cambridge (pp. 630-639). Education and research in Computer Aided Architectural Design in Europe

Wang, T. C., Liu, M. Y., Zhu, J. Y., Tao, A., Kautz, J., & Catanzaro, B. (2018). High-resolution image synthesis and semantic manipulation with conditional gans. In Proceedings of the IEEE conference on computer vision and pattern recognition, CVPR, 2018 (pp.8798-8807).

Zhang, L., Rao, A., & Agrawala, M. (2023). Adding conditional control to text-to-image diffusion models. In Proceedings of the IEEE/CVF International Conference on Computer Vision, ICCV 2023 (pp.3836-3847).

Zhong, X., Pia, F., Yu, F., Tan, C., & Pan, Y. (2022). A Discussion on an Urban Layout Workflow Utilizing Generative Adversarial Network (GAN): With a focus on automatized labeling and dataset acquisition. eCAADe 2022. (pp.583-592). Education and research in Computer Aided Architectural Design in Europe