DRAG2BUILD: INTERACTIVE POINT-BASED MANIPULATION OF 3D ARCHITECTURAL POINT CLOUDS GENERATED FROM A SINGLE IMAGE

JUN YIN[1], PENGJIAN XU[2], WEN GAO[3], PENGYU ZENG[4] and SHUAI LU[5]
[1,4,5]*Tsinghua University,* [2]*Zoomtech Engineering Co., Ltd,* [3]*Beijing University of Technology*
[1]*yinj22@mails.tsinghua.edu.cn, 0009-0005-7530-533X*

Abstract. At present, 3D reconstruction from images has made notable advancements in simple, small-scale scenes, but faces significant challenges in intricate, expansive architectural scenes. Focusing on the early stage of design stage, we present Drag2Build, a tool for converting images into point clouds for 3D reconstruction and modification in detailed architectural contexts. Our first step involved the creation of ArchiNet, a specialized 3D reconstruction dataset dedicated to elaborate architectural scenes. Next, we developed a 3D reconstruction approach using a conditional denoising diffusion model, enhanced by incorporating a model for segmenting objects, thereby improving segmentation and identification in complex scenes. Additionally, our system features an interactive component that allows for immediate modification of 2D images via an easy drag-and-drop action, synchronously updating 3D architectural point clouds. The performance of Drag2Build in 3D reconstruction precision was assessed and benchmarked against mainstream methods using ArchiNet. The experiments showed that our approach is capable of producing high-quality 3D point clouds, facilitating swift editing and efficient handling of intricate backgrounds.

Keywords. 3D building Generation, Diffusion Model, Single Image Reconstruction, DragDiffusion.

1. Introduction

The recent progress in transforming 2D images into 3D models significantly benefits spatial data capture and is crucial in fields like GIS, and City Modelling(Melas-Kyriazi et al., 2015). It's important in Architecture, Engineering, and Construction (AEC), with uses from Building Modeling to design and interior projects. However,this field needs further research, particularly during AEC's design stage.

Image-based 3D reconstruction provides an efficient design feedback method in early AEC stages. Yet, current studies often use uniform 3D objects such as hydrants, lacking diverse architectural scenes(Chang et al., 2015). This limits precision and adaptability

ACCELERATED DESIGN, Proceedings of the 29th International Conference of the Association for Computer-Aided Architectural Design Research in Asia (CAADRIA) 2024, Volume 1, 169-178.

in reconstructing complex architectural scenes.

Architects need a precise, affordable, fast 3D reconstruction method for early-stage complex architectural designs. It should allow easy 3D model edits for preview. We propose Drag2Build, a point cloud diffusion model for single-view architectural images, utilizing a unique single-image reconstruction method and an advanced segmentation model. It updates 3D models in real-time with user inputs, enhancing accuracy and editability.Our contributions are outlined as follows:

- We have collected ArchiNet, a specialized architectural dataset featuring comprehensive images (architectural line drawings, shaded diagrams, and renderings) alongside corresponding 3D architectural point cloud data, complete with precise camera parameters.
- We propose Drag2Build, a novel interactive 3D reconstruction framework, based on a 3D point cloud diffusion model, capable of editing through drag-and-drop actions. This method generates sparse yet precise 3D point clouds from single images lacking depth information, improving segmentation and recognition with the Segment-Anything object segmentation model. Architects can conveniently adjust and refine the generated 3D models.
- Our extensive experimental evaluations and comparisons with existing baselines establish that our algorithm excels in producing high-quality 3D point clouds, enabling rapid editing and effective handling of complex background scenes.

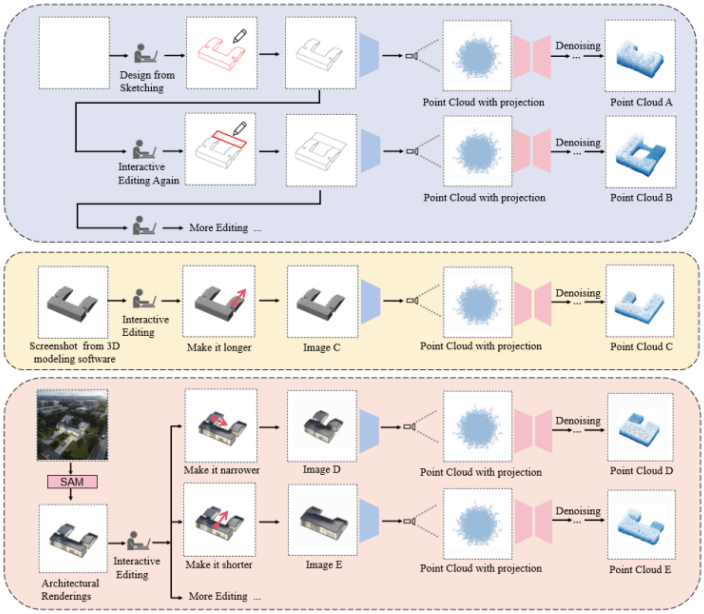

Figure 1. Utilizing Drag2Build, architects are now able to accomplish single-view 3D reconstructions within the design workflow by transforming sketches into three-dimensional models. This platform allows for immediate alterations to 2D illustrations, with concurrent updates being mirrored in the 3D architectural point cloud via straightforward drag-and-drop action.

2. Related Work

3D BUILDING GENERATION

The integration of computer technology and architecture has advanced significantly with the automated generation of 3D models(Wei et al., 2023). Traditional methods relied on fixed rules, limiting variety and increasing architects' workload (Li et al., 2021). Deep learning has brought new opportunities, merging 3D model generation with advanced algorithms (Alidoost. et al., 2019).However, these methods often face challenges in complex scenes and lack flexibility in modifications. Prior research, like Wei's diffusion point cloud model, had limitations in datasets and camera parameters, leading to less accurate reconstructions(Wei et al., 2023).Our approach innovates in 3D architectural reconstruction using a single-view, conditioned projection point-cloud diffusion model. This method enhances point cloud accuracy and allows real-time image-based modifications through user-friendly drag-and-drop interactions, contributing meaningfully to the field of 3D architectural modeling.

SINGLE IMAGE RECONSTRUTION

3D reconstruction from Single-image, blending computer vision and graphics, aims to create 3D scenes from a single 2D image. Traditional methods used 2D and 3D convolutional networks for this transformation, but often fell short in results. With deep learning, the Neural Radiance Field (NeRF) approach, requiring multiple images from different angles, has gained traction (Wang et al., 202) . However, its single-view performance remains subpar.

The rise of diffusion models has led to new single-image 3D reconstruction techniques. 3DFuse enhances diffusion models for better 3D consistency, while Zero123 uses them to render new perspectives based on camera position. OpenAI's Point-E(Nichol et al., 2022) and Shap-E (Jun et al., 2023) use internal models and latent function parameters for point clouds and textured meshes. However, these methods largely rely on common object datasets like Shapenet, with limited resource in specialized architectural datasets, and struggle with complex backgrounds.

POINT-BASED EDITING

To enhance detailed editing, a variety of point-based editing techniques have been introduced. Among these, DragGAN showcased remarkable capabilities for drag-based adjustments, utilizing two key elements: (1) the optimization of latent variable codes to guide the processing point towards its intended position, and (2) a tracking system to monitor the processing point's movement(Pan et al., 2023). Expanding upon this, Mou et al. adapted DragGAN's editing approach for diffusion models, proving its adaptability in various fields. However, despite these methods offering flexible editing between 2D images, they fall short in directly converting 2D edits into 3D structures, a crucial need in architectural design workflows.

3. METHOD

In this part, we explore the methodologies utilized in our innovative framework, Drag2Build. Our platform effectively combines SAM(Kirillovet al., 2022) with

methods for converting single images to point clouds, relying on 3D diffusion models. The primary objective is to tackle the intricate task of creating editable 3D architectural reconstructions from single image.Our approach introduces an intuitive, interactive editing feature, capitalizing on the distinct advantages of each integrated component. Through straightforward drag-and-drop actions, users can edit 2D images, and these modifications are immediately reflected in the corresponding 3D point cloud. This integration significantly enhances both the convenience and efficiency of the architectural design.

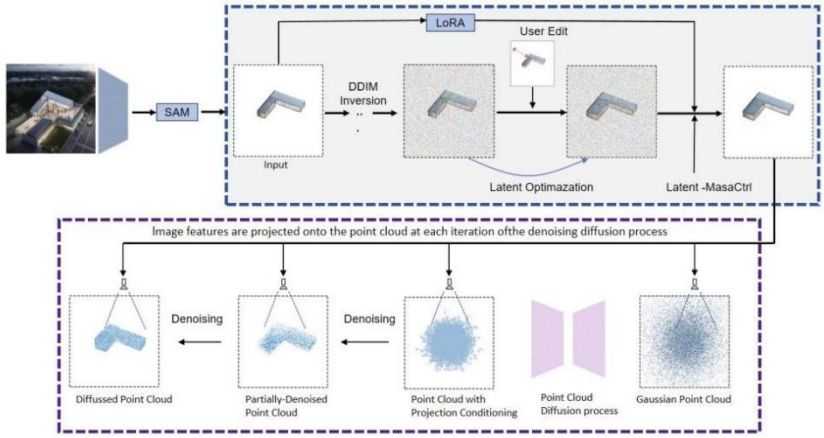

Figure 2. Drag2Build reconstructs architectural point clouds from a single input image and its camera pose in three steps. First, the SAM precisely identifies and segments buildings. Second, the point-based image editing model uses the UNet portion of the LoRA fine-tuned diffusion model to make interactive modifications. Third, the 3D point-cloud diffusion model incrementally generates a point cloud that matches the input image.

3.1. SAM

The SAM is introduced to extract buildings within the given image. To be specific, SAM processes an image, denoted as I, along with a prompt p. an image I and prompt p. SAM then produces a accurate segmentation mask M that effectively masks the background. The process involves the application of an affine transformation to produce localized image segments I' that encompass the bounding box, where the symbol $\odot$ denotes element-wise multiplication using the image mask as follows:

$$\mathrm{M} = \mathrm{SAM}(\mathrm{I}, \mathrm{p});\ \mathrm{I}' = \ \mathrm{Affine}(\mathrm{M} \odot \mathrm{I}; \mathrm{b}) \#(1)$$

SAM is a state-of-the-art object segmentation model capable of effectively handling multiple object categories. It addresses the challenge of extracting the principal architectural elements from images with complex scenes.

3.2. POINT-BASED EDITING

Inspired by the impressive success of DragGAN and DragDiffusion, we integrated a

similar approach into our model to address the challenges of interactive 2D and 3D modifications in the architectural design process. Initially, we fine-tune the UNet of the diffusion model using LoRA (Augustin et al., 2016) to enhance its accuracy in encoding the features of the input image. Next, we optimize the latent of the diffusion based on user instructions, such as the positions of the handles and target points, along with an optional mask to specify the editable area. This step is crucial to achieve the desired point-based interactive editing. Considering the Markov chain of the diffusion model's latent representation, this step focuses on optimizing the latent representation of single-step diffusion to enhance efficiency and efficacy during the editing process. After completing the optimization of the latent representations, the final edited result is obtained through a denoising step. Similar to DragDiffusion, we employ the Latent-MasaCtrl mechanism to improve the consistency between the original image and edited outcome.As for the differences from DragDiffusion, we have changed the base model and employed ProbSparse Attention, which reduces interaction time and prevents overfitting.

3.3. PROJECTION CONDITONAL POINT CLOUD DIFFUSION MODEL-ING

In machine learning, diffusion denoising probabilistic models are famous for their ability to produce images and data of high quality .These models utilize a progressive approach of noise integration, incorporating disturbances into a sample, referred to as , rooted from the targeted data distribution .Each stage's noise intensity is regulated according to the variance values in , as described below:

$$q(X_t|X_{t-1}) = \mathcal{N}\left(X_t;\ \sqrt{1-\beta_t}X_{t-1}, \beta_t I\right)\#(2)$$

Every $q(X_t|X_{t-1})$ follows a Gaussian distribution, determined through the widely employed method of reparameterization, as follows:

$$q(X_t|X_0) = \sqrt{\overline{\alpha}_t}X_0 + \epsilon\sqrt{1-\overline{\alpha}_t}\#(3),$$

where $\alpha_t = 1-\beta_t$, $\overline{\alpha}_t = \prod_{s=0}^{t}\alpha_s$, and $\epsilon \sim \mathcal{N}(0, I)$.

In the development of a generative model, the reverse process is important. It begins with a noisy distribution $q(X_T)$, and reduces noise in the data points, ultimately producing samples that closely mimic the intended distribution $q(X_0)$ .In a 3D point cloud having N points, which is considered as an entity with 3N- dimensions, we train the diffusion model $s_\theta : \mathbb{R}^{3N} \to \mathbb{R}^{3N}$. This framework is to enhance the accuracy of point positions within a spherical Gaussian domain by converting them into recognizable shapes. The model, at each phase, estimates the positional variance of each point relative to its current location, and repeatedly applies this methodology to generate a sample aligned with the target distribution$q(X_0)$.

This network is specifically trained to anticipate the noise $\epsilon \in \mathbb{R}^{3N}$ injected in the preceding step. It is made by employing the L_2 loss, which calculates the disparity between the groundtruths and those noise values as below:

$$\mathcal{L} = E_{\epsilon\sim\mathcal{N}(0,I)}[\|\epsilon - s_\theta(X_t, t)\|_2^2]\#(4).$$

In the prediction stage, a arbitrary point cloud $X_T \sim N(0, I_{3N})$ is extracted from a

Gaussian distribution encompassing 3N dimensions. A backward process is then initiated to generate create sample instances X_0. In each stage of this sequence, the average value of the approximation of $q(X_{t-1}|X_t)$ is derived from predictions$_\theta(X_t, t)$, which is subsequently utilized as a foundational element to sample $q(X_{t-1}|X_t)$.

In this framework, 3D reconstruction is viewed as a generative process with condition, where the distribution $q(X_0|I, V)$ is dependent on the provided input image I and its camera view V. To tackle the challenge of geometric consistency, our method involves mapping the image to partially denoised points during each phase of the diffusion sequence. At first, the image undergoes transformation into a dense feature volume, accomplished through the application of a standard 2D image model, such as a convolutional neural network (CNN) or a Vision Transformer(ViT). Before every diffusion step, we map the aforementioned features onto a point cloud.

4. EXPERIMENTS AND RESULTS

4.1. DATASET

Collecting a 3D architectural model database is complex, time-consuming, and resource-intensive, involving data collection, modeling adjustments, and annotations. Traditional 3D point-cloud datasets often face incompleteness, inaccuracies, and limited styles (Li et al., 2021). To evaluate our method, we created ArchiNet, a diverse dataset with 4,688 SketchUp models from 132 architecture students (with a mix of 95 undergraduates and 37 postgraduates) and 675,180 images, each with detailed camera settings. We initially processed these models in Grasshopper to transform them into standardized point-cloud files (PLY format), and subsequently utilized Matplotlib to export images of each PLY file from 36 different angles, ensuring a comprehensive view. Our collection includes three visual types: architectural line drawings (AL), shading diagrams (AS), and renderings (AR), each with point cloud data and individual masks. The dataset is divided into training, testing, and validation sets with a ratio of 92:21:21, facilitating a balanced evaluation of our models across diverse architectural styles and complexities.

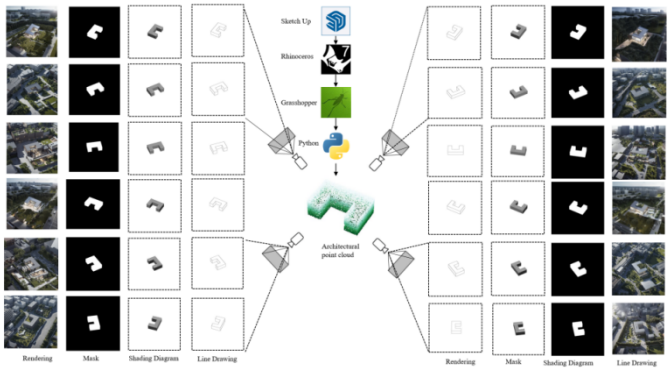

Figure 3. Dataset Collection Process Diagram

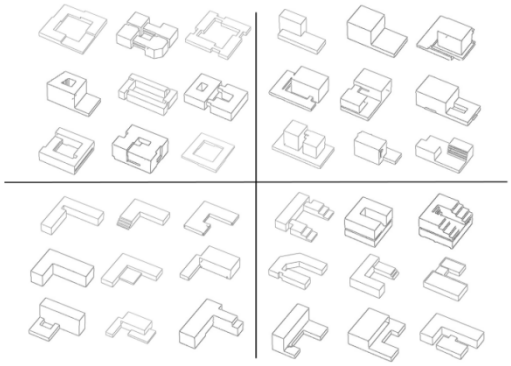

Figure 4. Example instances from our ArchiNet dataset.ArchiNet 3D assets are semantically diverse, highquality, and paired with camera parameters

4.2. EXPERIMENTS DETAILS

In our research, we utilized a high-performance workstation equipped with two NVIDIA A6000 GPUs, each boasting 48 GB of RAM, and powered by an Intel(R) Xeon(R) Gold 6326 CPU. The training phase leveraged our ArchiNet dataset, employing an Adam optimizer. The parameters set for this phase included a learning rate of 0.0001, a batch size of 24, and a cap of 100,000 steps for training. For uniformity and consistency in our analysis, each 3D point-cloud model was standardized into a OBJ file format, containing precisely 8192 points.

4.3. BENCHMARKS AND METRICS

In our study, we compared our model's point clouds with others, focusing on intricate background handling and point-cloud fidelity due to our task's interactive nature. We retrained 3D-R2N2 using our dataset, sticking to its original settings, to create voxel files from images, later converted to point clouds. For Point-E(Nichol et al., 2022), we used official code and models, following their single-image 3D reconstruction examples. These predictions required no extra scaling. With Shap-E(Jun et al., 2023), we maintained hyperparameters and followed their single-image examples, converting meshes to point clouds for consistency in distribution and count.Our evaluation was based on two key metrics:

Fréchet Inception Distance (FID): This metric assesses the disparity between the generated and actual point clouds by comparing the covariance of point cloud features. Lower FID values suggest a closer resemblance to real point clouds.

Chamfer Distance (CD): CD is a prevalent metric for gauging the likeness between two point cloud sets, calculating the mean nearest-point distance between them. Smaller CD values indicate higher similarity.

4.4. QUANTITATIVE RESULTS

Tables 1-2 provide a comprehensive summary of the performance evaluations

conducted on the ArchiNet dataset. The data conclusively shows that our model surpasses competing models in all four evaluated metrics. When comparing the mean scores across the three datasets (AL, AS, and AR), our model shows a substantial reduction in Chamfer Distance (CD) results by 42.3%, 43.2%, and 64.2% compared to Shap-E(Jun et al., 2023), Point-E(Nichol et al., 2022), and 3D-R2N2. Concurrently, it demonstrates a remarkable increase in F-Scores by 25%, 104%, and 112%. This superiority of our model in terms of reconstruction accuracy and detail capture is evident. Additionally, the results highlight the variability in the adaptability of different models to distinct types of images. For example, Shap-E exhibits subpar performance with line drawing datasets as opposed to other types, while Point-E struggles more with rendered image data. In contrast, our methodology maintains consistent and stable performance across all three data categories.

Table 1.CD results on ArchiNet. The best results are indicated in bold.

Method	AL	AS	AR
Ours	**0.0828**	**0.0800**	**0.0799**
Shap-E	0.1703	0.1240	0.1263
Point-E	0.1512	0.1505	0.1262
3D-R2N2	0.2276	0.2241	0.2240

Table 2. F-Score results on ArchiNet. The best results are indicated in bold .

Method	AL	AS	AR
Ours	**0.6688**	**0.6894**	**0.6848**
Shap-E	0.4900	0.5771	0.5781
Point-E	0.3159	0.3066	0.3685
3D-R2N2	0.3377	0.3144	0.3152

4.5. QUALITATIVE RESULTS

Beyond the quantitative evaluation, qualitative results are also shown for the ArchiNet dataset (Figures 5-8). These visual examples illustrate how our technique not only facilitates immediate, user-friendly interaction with 2D images via straightforward drag-and-drop actions, directly reflecting changes in 3D architectural point clouds, but also precisely captures intricate architectural elements like platforms, extended volumes, and exterior staircases.

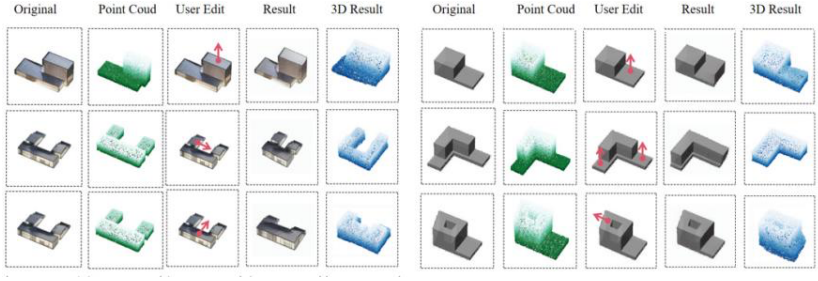

Figure 5: Example results from different methods on the ArchiNet dataset. The image demonstrates our model's ability to modify 3d shapes by editing images

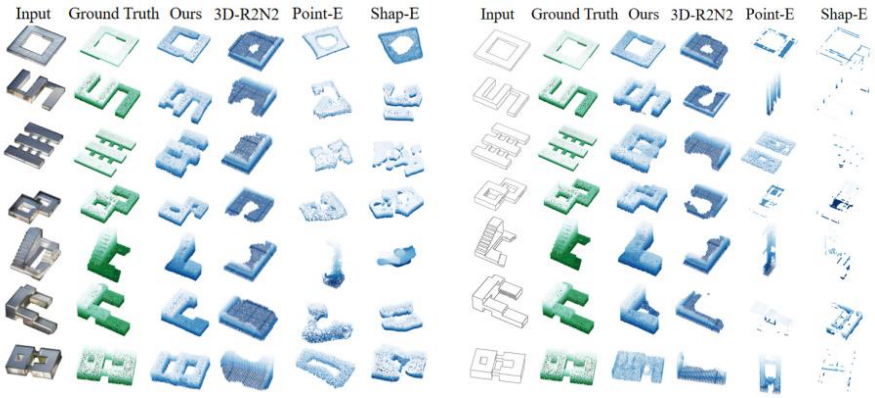

Figure 6.Example results from different methods on the ArchiNet AR dataset and the ArchiNet AL dataset. The first column shows the input image, and the second column displays the ground truth. The subsequent four columns each showcase the point cloud reconstruction results of different models.

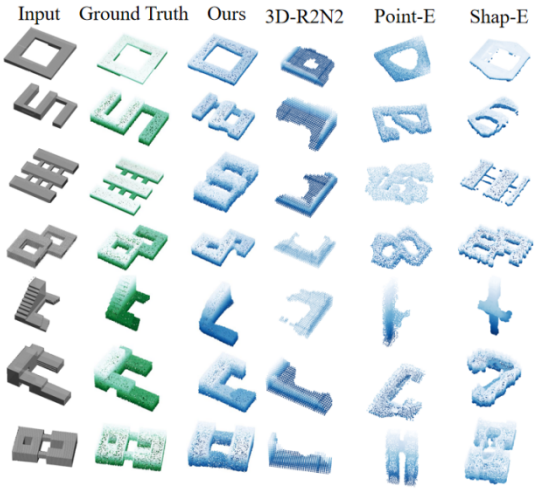

Figure 7. Example results from diverse methodologies applied to the ArchiNet AS dataset are shown.

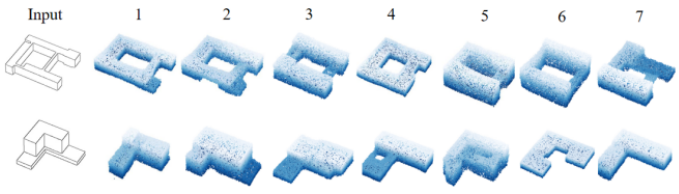

Figure 8.Examples of diversity in the results. The far left column presents a sample image selected from the architectural line drawing collection, notable for its notably ambiguous form. Subsequent images illustrate outputs generated by our model using seven distinct random seeds. Demonstrating

its proficiency, our model consistently generates considerable shape variations, all the while aligning accurately with the perspective of the input image from the reference vantage point.

5. CONCLUSION AND FUTURE WORK

The Drag2Build framework marks a meaningful progress in 3D building generation and interactive architectural design. Looking ahead, our objectives are set towards enhancing and expanding our model in several key areas:

Synergy with Digital Modeling Platforms: Our goal is to craft plugins or tools that seamlessly blend our framework with established digital modeling applications, harnessing the combined strengths of both systems.

Enhanced Precision in Editing: We are dedicated to refining the accuracy of our drag-and-drop features, focusing on the meticulous modification of small-scale elements and intricate details within 3D point clouds.

Advanced Point Cloud Processing: We plan to evolve point cloud processing into formats like meshes, making editing simpler for architects, thus enhancing practicality and utility in real-world architectural applications.

References

Alidoost, F., Arefi, H., & Tombari, F. (2019). 2D image-to-3D model: Knowledge-based 3D building reconstruction (3DBR) using single aerial images and convolutional neural networks (CNNs). Remote Sensing, 11(19), 2219.

Augustin, A., Yi, J., Clausen, T., & Townsley, W. M. (2016). A study of LoRa: Long range & low power networks for the internet of things. Sensors, 16(9), 1466.

Chang, A. X., Funkhouser, T., Guibas, L., Hanrahan, P., Huang, Q., Li, Z., ... & Yu, F. (2015). Shapenet: An information-rich 3d model repository. arXiv preprint arXiv:1512.03012.

Jun, H., & Nichol, A. (2023). Shap-e: Generating conditional 3d implicit functions. arXiv preprint arXiv:2305.02463.

Kirillov, A., Mintun, E., Ravi, N., Mao, H., Rolland, C., Gustafson, L., ... & Girshick, R. (2023). Segment anything. arXiv preprint arXiv:2304.02643.

Li, W., Meng, L., Wang, J., He, C., Xia, G. S., & Lin, D. (2021). 3D building reconstruction from monocular remote sensing images. In Proceedings of the IEEE/CVF International Conference on Computer Vision (pp. 12548-12557).

Melas-Kyriazi, L., Laina, I., Rupprecht, C., & Vedaldi, A. (2023). Realfusion: 360deg reconstruction of any object from a single image. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (pp. 8446-8455).

Nichol, A., Jun, H., Dhariwal, P., Mishkin, P., & Chen, M. (2022). Point-e: A system for generating 3d point clouds from complex prompts. arXiv preprint arXiv:2212.08751.

Pan, X., Tewari, A., Leimkühler, T., Liu, L., Meka, A., & Theobalt, C. (2023, July). Drag your gan: Interactive point-based manipulation on the generative image manifold. In ACM SIGGRAPH 2023 Conference Proceedings (pp. 1-11).

Wang, Z., Wu, S., Xie, W., Chen, M., & Prisacariu, V. A. (2021). NeRF--: Neural radiance fields without known camera parameters. arXiv preprint arXiv:2102.07064.

Wei, Y., Vosselman, G., & Yang, M. Y. (2023). BuilDiff: 3D Building Shape Generation using Single-Image Conditional Point Cloud Diffusion Models. In Proceedings of the IEEE/CVF International Conference on Computer Vision (pp. 2910-2919).