

Top2Pano: Learning to Generate Indoor Panoramas from Top-Down View

Zitong Zhang Suranjan Gautam Rui Yu

University of Louisville

{zitong.zhang, suranjan.gautam, rui.yu}@louisville.edu

<https://top2pano.github.io/>



Figure 1. **Top:** We present *Top2Pano*, a method for synthesizing high-quality indoor panoramas from a top-down view. Given a camera position, *Top2Pano* generates panoramas that are both visually compelling and geometrically accurate. **Bottom:** Our model demonstrates strong generalization capabilities. When provided with schematic floor plans as input, *Top2Pano* produces photorealistic and structurally coherent panoramas. Additionally, our approach can be easily adapted for stylized synthesis, enabling diverse design variations. Note: The original dataset (Matterport3D [3]) contains blurry regions near the upper and lower edges of the panoramic image.

Abstract

Generating immersive 360° indoor panoramas from 2D top-down views has applications in virtual reality, interior design, real estate, and robotics. This task is challenging due to the lack of explicit 3D structure and the need for geometric consistency and photorealism. We propose **Top2Pano**, an end-to-end model for synthesizing realistic indoor panoramas from top-down views. Our method estimates volumetric occupancy to infer 3D structures, then uses volumetric rendering to generate coarse color and depth panoramas. These guide a diffusion-based refinement stage using ControlNet, enhancing realism and structural fidelity. Evaluations on two datasets show *Top2Pano* outperforms baselines, effectively reconstructing geometry, occlusions, and spatial arrangements. It also generalizes well, producing high-quality panoramas from schematic floorplans. Our results highlight *Top2Pano*’s potential in bridging top-down views with immersive indoor synthesis.

1. Introduction

Understanding and synthesizing immersive indoor scenes from minimal structural information is a fundamental chal-

lenge in computer vision and graphics [16, 20, 21, 32, 40, 44]. The ability to generate realistic indoor panorama images from a 2D top-down view holds immense potential for a wide range of applications, including virtual reality (VR) [23], interior design [27], real estate visualization [15], and robotics [11]. For instance, real estate platforms can leverage this technology to offer potential buyers photorealistic virtual walkthroughs generated directly from architectural floorplans, enhancing the property viewing experience. Similarly, VR applications can benefit from automatically synthesized environments that create more engaging and immersive user experiences. Additionally, robots operating in indoor environments can utilize synthesized panoramas to improve their spatial understanding and navigation capabilities, enabling more efficient and accurate movement in complex spaces. Despite its broad applicability, the task of generating high-quality indoor panoramas from top-down views remains surprisingly underexplored in the literature. Recent advancements in large multimodal models have enabled the synthesis of panoramas directly from text input [10, 39, 41]; however, these approaches often overlook critical geometric and textural constraints. Other studies have focused on generating 3D models from semantic layouts [2, 4, 8, 9, 30, 38];

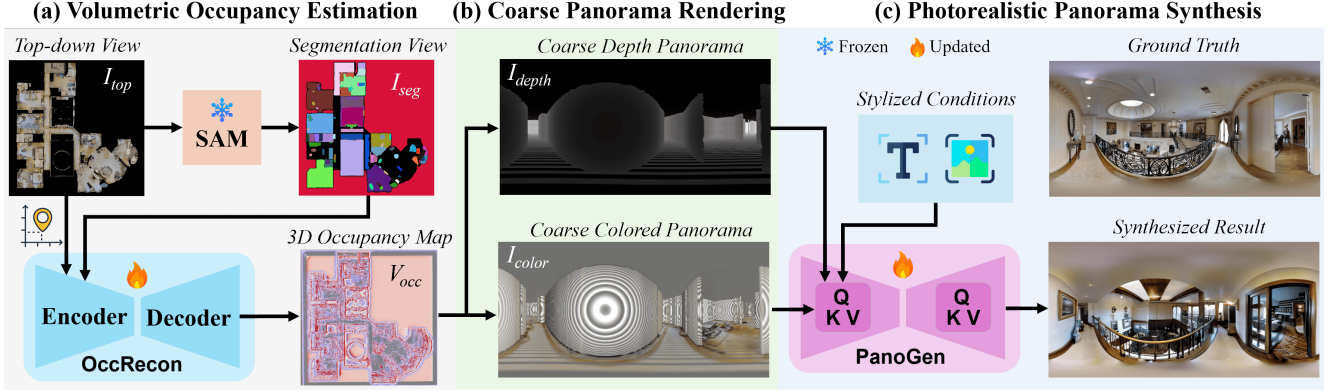


Figure 2. **Overview of the proposed Top2Pano pipeline.** The pipeline begins by segmenting the top-down view using SAM [19]. Both the segmented top-down image and the original top-down view are then processed by the OccRecon module to estimate the scene’s 3D volumetric occupancy. Next, given the camera position, the system employs volumetric rendering to generate coarse depth and color panoramas. These coarse images are subsequently refined by the PanoGen module to produce the final photorealistic panorama. The PanoGen module also supports stylized panorama generation based on textual or visual conditions.

however, these methods are often limited by the quality of the resulting 3D meshes, making them unsuitable for rendering high-quality panoramas. Furthermore, semantic information is often unavailable in top-down views or floorplans, posing an additional challenge for existing approaches.

Addressing this gap requires overcoming several significant technical challenges. First, a 2D top-down view provides only limited visual cues about the actual appearance and layout of the scene, making it difficult to infer occluded structures and fine texture details. Second, generating geometrically consistent indoor scenes demands accurate reasoning about 3D spatial occupancy from a 2D input, which is inherently ambiguous. Third, achieving photorealism while maintaining structural coherence necessitates a synthesis approach that effectively balances fidelity and realism, ensuring that the generated scenes are both visually appealing and functionally accurate.

To tackle these challenges, we introduce **Top2Pano**, a novel framework for generating photorealistic indoor panorama images from 2D top-down views. Our approach consists of three main stages. First, we learn the volumetric occupancy of the indoor scene, enabling the model to infer plausible spatial structures and layout configurations. Next, we employ volumetric rendering to generate coarse depth and colored panorama images, providing an initial estimate of the scene’s appearance and geometry. Finally, we refine the synthesized panoramas using a diffusion-based model [42] conditioned on the coarse representations, enhancing both realism and structural consistency. By incorporating learned occupancy priors and diffusion-based refinement, our model effectively bridges the gap between schematic top-down views and immersive indoor panoramas, producing results that are both visually compelling and geometrically accurate.

We evaluate Top2Pano on two indoor datasets and demonstrate its effectiveness compared to baseline methods. Our model not only generates higher-quality images with improved geometric consistency but also exhibits strong generalization capabilities. Even when provided with schematic floorplans as input, Top2Pano can produce photorealistic and structurally coherent panoramas. Moreover, we show that our method can be easily adapted for stylized synthesis, allowing for diverse design variations and enabling users to explore different interior aesthetics with ease.

Our key contributions are as follows:

- We introduce Top2Pano, a novel framework for generating indoor panoramas from 2D top-down views, integrating volumetric occupancy learning, coarse synthesis, and diffusion-based refinement to achieve high-quality results.
- We conduct extensive experiments on two indoor datasets, demonstrating that our model surpasses baseline methods in both image quality and structural consistency, setting a new benchmark for this task.
- We show that Top2Pano generalizes well to schematic floorplans, producing high-quality, geometry-consistent panoramas. Furthermore, our approach supports stylish synthesis, enabling the generation of panoramas with diverse interior design aesthetics, making it a versatile tool for various applications.

2. Related Work

2.1. Panorama Generation

Traditionally, panoramas were generated using image stitching and feature matching methods. With the recent advancements in generative machine learning, text-driven panorama generation techniques have gained popularity. These methods have utilized GANs, VAEs or a combina-

tion of GANs and VAEs [5] and more recently, diffusion models [10, 39, 41] to synthesize panoramic images from textual descriptions. Another popular field of research is panorama synthesis from narrow-FoV images using image out-painting. Some methods rely solely on narrow-FOV images [1], while others incorporate textual descriptions alongside the images [6, 17, 34]. Cross-view panorama generation is another well-explored area, particularly challenging due to large shifts in camera perspective. Most techniques in this domain have focused on generating ground-view panoramas from aerial images. Some approaches directly use top-down images as input [35], while others [22, 28, 31, 37] extract geometric and segmentation information from top-down images to enhance quality. To the best of our knowledge, there is no prior work that explores the generation of indoor panoramic images from floor plans or top-down views of indoor spaces.

2.2. Layout-Guided 3D Scene Generation

Recent approaches to 3D scene generation leverage layouts for semantic and physical plausibility. Plan2Scene [33] reconstructs 3D meshes from floor plans, while ATISS [26] employs Transformers conditioned on scene layout. CC3D [2] follows a 3D GAN-based approach using 2D semantic layouts. Diffusion-based methods such as SceneCraft [38], Layout2Scene [4], and Prim2Room [9] have further improved synthesis quality. ControlRoom3D [30] and Ctrl-Room [8] are closely related to our work, as both generate panoramas during reconstruction process. Unlike prior methods that rely on *explicit semantic layouts* detailing object classes, positions, orientations, and sizes, our approach only requires a top-down view – an easily obtainable and lightweight input format. Instead of generating full 3D scenes, we produce panoramas, offering a more efficient and realistic solution for AR/VR, and autonomous robotic navigation by enabling immersive experiences like virtual tours and supporting real-time robotic navigation.

3. Method

The pipeline of our Top2Pano method is illustrated in Figure 2. Given an input top-down view $I_{\text{top}} \in \mathbb{R}^{H \times W \times 3}$, we first generate its segmentation map $I_{\text{seg}} \in \mathbb{R}^{H \times W \times 3}$ using a pretrained model. Both the top-down view and its segmentation, along with the specified camera position, are then fed into an encoder-decoder occupancy estimation module, *OccRecon*, which reconstructs a 3D volumetric occupancy map $V_{\text{occ}} \in \mathbb{R}^{H \times W \times N}$, where N represents the number of vertical voxels. From this occupancy map, we render a coarse depth panorama I_{depth} via volumetric rendering and project colors from the top-down view to obtain a coarse color panorama I_{color} . To ensure geometric consistency, we enforce structural constraints on walls and floor, preserving occlusion relationships and realistic spatial structure. Finally,

both coarse depth and color panoramas serve as conditions for a diffusion-based synthesis module, *PanoGen*, which generates photorealistic panoramic images $I_{\text{pano}} \in \mathbb{R}^{H \times W \times 3}$. These images faithfully capture the scene’s spatial layout, furniture, and fine color details. Additionally, PanoGen module supports stylized synthesis with optional textual or imagery-based controls.

3.1. Volumetric Occupancy Estimation

The 2D top-down views lack 3D structural information about objects and furniture. To render panoramas that accurately reflect the geometric spatial relationships of objects, we propose training an *OccRecon* module to estimate the scene’s 3D occupancy or density.

Input Representations. Unlike the layout-guided 3D scene generation setting, our top-down input lacks semantic information. In our preliminary study, we found that current semantic segmentation models struggle to generalize to indoor top-down views, making it challenging to estimate the 3D structure without semantic guidance. To address this, we propose leveraging a pretrained segmentation model SAM [19] to extract the 2D structure of the scene. Both the top-down image and the segmentation view are then fed into the encoder of the *OccRecon* module. The segmentation provides valuable details, such as room boundaries, furniture positions, and shapes, which significantly enhance the *OccRecon* module’s ability to learn the overall 3D structure of the rooms. Moreover, this semantic-free input design enables our model to generalize effectively to more abstract inputs, such as semantic floorplans.

OccRecon Module. We propose a diffusion-based encoder-decoder framework that efficiently extracts and reconstructs the 3D spatial structure of a scene. Instead of using computationally intensive 3D diffusion models, our approach leverages a 2D diffusion model [42], significantly reducing resource demands. The *OccRecon* module processes 2D images and segmentations to extract spatial information and reconstruct room structures, including wall and furniture heights. In the final step, a 3D convolutional layer integrates the learned spatial features to generate a comprehensive 3D occupancy map. By relying on 2D inputs for most of the process, our method drastically lowers computational costs while still capturing the full scene layout. This balance of efficiency and accuracy makes it a practical solution for 3D structural modeling without semantic inputs. The *OccRecon* module outputs a 3D volumetric occupancy map V_{occ} , which explicitly represents the scene’s 3D geometry.

$$V_{\text{occ}} = \text{OccRecon}(I_{\text{top}}, I_{\text{seg}})_{\text{condition}} \quad (1)$$

We normalize the occupancy values to the range $[0, 1]$. Given the scale information of the scene, we transform the real-world coordinates of a 3D point into the estimated volumetric

space. The point’s density is then obtained by querying the learned occupancy map V_{occ} using tri-linear interpolation.

Structural Reinforcement. Top-down views offer a comprehensive overview of the entire floor, but when observed from a first-person perspective, the scene typically reveals only the details of the current space, obscuring rooms beyond the walls. After the model learns the overall geometric structure of the rooms through the OccRecon module, we apply structural reinforcement to refine the wall geometry. This provides precise depth information, enhancing the reconstruction of the room’s geometric layout. We solidify the wall voxels by applying the maximum value (1 after normalization). Additionally, top-down priors can constrain the floor’s geometry and help infer the texture of the furniture, which is essential for generating accurate indoor panoramas with correctly placed and colored furniture. Since the OccNet modules are optimized end-to-end with the subsequent modules, the height of the furniture can be inferred from the 3D occupancy map. By encoding structural constraints from the walls and floor, we achieve a more accurate representation of the scene’s geometry, including the positions, colors, and other attributes of the furniture.

3.2. Coarse Panorama Rendering

Given the 3D occupancy map of the scene, we render coarse depth and color images based on the specified camera position. To ensure accurate mapping, we first compute the ratio between pixel resolution and physical dimensions. With the room height known, we apply this ratio along with the pixel coordinates in the top-down image to determine the corresponding position in the occupancy map V_{occ} . To generate the panoramic image from the 3D occupancy map, we employ equirectangular projection along with a spherical coordinate system.

The coarse depth panorama I_{depth} is obtained through volumetric rendering [24] of the occupancy map. The depth of a projected ray at pixel (u, v) is computed as follows:

$$I_{\text{depth}}^{(u,v)} = \sum_{i=1}^S T_i \alpha_i d_i, \quad T_i = \prod_{j=1}^{i-1} (1 - \alpha_j), \quad (2)$$

where α_i represents the transparency level, and d_i denotes the distance from the sampled position to the camera.

The coarse color panorama I_{color} is obtained by directly projecting the indoor top-down view along the camera rays [28], assigning colors based on the corresponding intersections and bilinear interpolation. Specifically, the color at pixel (u, v) is determined by sampling the top-down image without learning a radiance field.

$$I_{\text{color}}^{(u,v)} = \sum_i T_i \alpha_i c_i \quad (3)$$

where c_i is the color copied from the top-down image. This approach directly maps color from the top-down view, offering a more straightforward and computationally efficient solution compared to NeRF [25], which reconstructs scenes by learning a radiance field for view synthesis.

We employed a uniform voxel sampling strategy, where voxels were sampled along a fixed-length ray for both coarse color and coarse depth. However, this approach introduced banding artifacts, particularly noticeable on the floor directly beneath the camera in the coarse color image. Since the top-down view served as the floor texture, maintaining high-quality details was essential for achieving realistic rendering. To mitigate these artifacts, we reduced the ray length by half for coarse color sampling. This adjustment increased the density of sample points within the same spatial region, enhancing sampling accuracy and producing smoother color and texture transitions. As a result, banding artifacts were significantly diminished, improving overall rendering quality. For coarse depth, preserving scene structure was paramount, so we retained the original sampling strategy.

3.3. Photorealistic Synthesis

Generating photorealistic panoramic images directly from top-down views is challenging. To address this, we propose a two-stage pipeline. In the second stage, the PanoGen module synthesizes photorealistic indoor panoramic images from coarse color and depth inputs. We implement PanoGen using a diffusion-based ControlNet [42], enabling the restoration of fine details. This two-stage approach not only reconstructs the house’s structural layout, including precise wall positions, but also restores elements such as windows, lighting, and furniture. The final panoramic image I_{pano} is generated based on the two coarse inputs from the previous stage:

$$I_{\text{pano}} = \text{PanoGen}(I_{\text{color}}, I_{\text{depth}})_{\text{condition}} \quad (4)$$

The PanoGen module modifies the conditioning mechanism of ControlNet by treating coarse color and depth images as separate inputs before combining them. This design enables PanoGen to capture both the scene’s geometric structure and the furniture’s color and position, resulting in more accurate and high-quality panoramic images. To further enhance consistency, we incorporate an alignment loss to prevent structural distortions when the viewpoint changes and a color loss to ensure accurate color reproduction in the synthesized output.

3.4. Training and Optimization

The Top2Pano model employs denoising MSE loss, alignment loss, and color loss functions for optimization. The denoising MSE loss function [42] is defined as:

$$\mathcal{L}_{\text{diff}} = \mathbb{E}_{z_0, t, c_t, c_c, c_d, \epsilon \sim \mathcal{N}(0,1)} [\|\epsilon - \epsilon_{\theta}(z_t, t, c_t, c_c, c_d)\|_2^2], \quad (5)$$

where t denotes the number of noise addition steps, c_c represents the corresponding coarse color image, and c_d represents the coarse depth image. The variables t and c_t correspond to the time step and simple text prompts, respectively. The diffusion algorithm trains a neural network ϵ_θ to predict the noise added to the noisy image z_t . The alignment loss is formulated as:

$$\mathcal{L}_{\text{alignment}} = \left[\|I_D - \hat{I}_D\|_2^2 \right], \quad (6)$$

where I_D represents the depth images generated by the model and \hat{I}_D represents the ground truth depth images. This loss function is used to address distortions in the furniture.

Let I be the rendered image and G the ground truth image, both with $C = 3$ color channels. For each channel $c \in \{1, \dots, C\}$, we define the normalized histogram as

$$H^c(I) = (h_1^c(I), h_2^c(I), \dots, h_{\text{bins}}^c(I)), \quad (7)$$

where bins is the number of histogram bins (e.g., 256), and each $h_k^c(I)$ represents the normalized frequency (probability) of pixel intensities falling into bin k . Likewise, for the ground truth image G , we have

$$H^c(G) = (h_1^c(G), h_2^c(G), \dots, h_{\text{bins}}^c(G)). \quad (8)$$

The color histogram loss, measured using the L_1 norm, is given by

$$\begin{aligned} \mathcal{L}_{\text{color}}(I, G) &= \sum_{c=1}^C \|H^c(I) - H^c(G)\|_1 \\ &= \sum_{c=1}^C \sum_{k=1}^{\text{bins}} |h_k^c(I) - h_k^c(G)|. \end{aligned} \quad (9)$$

The final loss function combines three loss terms as follows:

$$\mathcal{L} = \mathcal{L}_{\text{diff}} + \mathcal{L}_{\text{alignment}} + \mathcal{L}_{\text{color}}. \quad (10)$$

3.5. Generalization and Stylized Synthesis

Generalization to Floorplan. We aim to generate first-person view panoramas that faithfully represent a scene. While top-down views lack vertical details like walls and windows, this omission grants flexibility in generating panoramas. This flexibility is especially useful in interior design, where inputs are often simple floorplans. To improve generalization, particularly to schematic floorplans, we train on *orthographic* rather than perspective views, as they better match floorplans. Empirical results show that Top2Pano generalizes well to schematic and even hand-drawn floorplans while maintaining photorealism. Additionally, our model enables stylized synthesis guided by text or images, supporting diverse design needs.

Text-Guided Stylization. The PanoGen module, built upon the text-driven Stable Diffusion model, inherently supports

text-conditioned image generation. As illustrated in Figure 2, PanoGen synthesizes panoramas using three conditions: coarse depth, coarse colored image, and stylized textual guidance. When the input is a textureless floorplan, the stylized textual condition effectively guides the style of the synthesized result, as demonstrated in Figure 1. However, when a colored top-down view is provided, the influence of text-guided stylization becomes less pronounced. This occurs because the rendered coarse colored panorama constrains the final output to closely align with the input view, thereby diminishing the impact of the textual stylization.

To enhance text-guided stylization, the weight of the coarse colored panorama in the PanoGen conditions can be reduced. However, this introduces a tradeoff: prioritizing stylization may come at the expense of fidelity to the top-down view. Regardless of this tradeoff, the coarse depth condition consistently ensures that the synthesized panorama adheres to the underlying scene geometry. Notably, this tradeoff does not apply to the text-to-panorama generation task, where textual guidance plays a more dominant role.

Image-Guided Stylization. Given several scene images (not necessarily panoramas), we can fine-tune the PanoGen module using low-rank adaptation (LoRA) [14] to generate panoramas that align with the visual styles present in the provided images. To guide this process, we introduce structured textural prompts augmented with style tags (e.g., [Japanese]) at the beginning of the input prompts. These tags act as conditional modifiers, steering the model toward synthesizing images that follow specific aesthetic themes, such as regional design styles. The framework is applied solely to the PanoGen module, ensuring both computational and parameter efficiency. Notably, our method requires fewer than five in-the-wild images per target style, substantially reducing data demands. This approach achieves its efficiency by decomposing weight updates into low-rank matrices. For a pretrained weight matrix $\mathbf{W}_0 \in \mathbb{R}^{d \times d}$, the update $\Delta \mathbf{W}$ is constrained as follows:

$$\Delta \mathbf{W} = \mathbf{B}\mathbf{A}, \quad \text{where } \mathbf{A} \in \mathbb{R}^{d \times r}, \mathbf{B} \in \mathbb{R}^{r \times d} \quad (11)$$

Here, $r \ll d$ represents the intrinsic rank (we use $r = 8$). During fine-tuning, only the matrices \mathbf{A} and \mathbf{B} are updated, while the original weights \mathbf{W}_0 remain fixed. The forward pass is modified as follows:

$$\mathbf{h}_{\text{out}} = \mathbf{W}_0 \mathbf{h}_{\text{in}} + \alpha \cdot \mathbf{B}\mathbf{A}\mathbf{h}_{\text{in}} \quad (12)$$

where α is a scaling coefficient. This lightweight adaptation (0.8% new parameters) mitigates catastrophic forgetting and maintains the model’s baseline generation quality for generic prompts while enabling precise style control through our [style] textual conditioning.

	Training			Testing		
	Scenes	Floors	Panoramas	Scenes	Floors	Panoramas
Matterport3D	61	127	6177	14	29	1405
Gibson	152	203	5379	39	76	1672

Table 1. The numbers of scenes, floors, and panorama images in the training and testing sets of the two datasets

	Method	PSNR \uparrow	SSIM \uparrow	FID \downarrow	LPIPS \downarrow
Matterport3D	Sat2Density[28]+LDM[29]	11.27	0.4221	100.06	0.6237
	Sat2Density[28]+ControlNet[42]	11.42	0.4222	85.78	0.6163
	PanFusion[41]	11.45	0.4372	85.74	0.6153
	Top2Pano (Ours)	11.72	0.4409	30.84	0.6029
	Sat2Density[28]+LDM[29]	10.54	0.4480	84.33	0.6462
Gibson	Sat2Density[28]+ControlNet[42]	10.97	0.4582	85.21	0.6645
	PanFusion[41]	11.36	0.4744	79.53	0.6634
	Top2Pano (Ours)	11.58	0.4851	28.68	0.6282

Table 2. Quantitative comparison with existing methods on the Matterport3D [3] and Gibson [36] datasets.

4. Experiments

4.1. Data Preparation

For evaluation, we use the Matterport3D [3] and Gibson [36] datasets. Since no existing dataset provides both top-down views and high-quality panoramic images, we generate top-down views from 3D models in these datasets using Blender. Specifically, we import textured 3D meshes into Blender and render top-down views with an orthographic camera. The top-down view we render closely resembles a floorplan, unlike the perspective-rendered views used in embodied dialog localization [12]. This similarity enhances our model’s ability to generalize to floorplan inputs. To determine the number of floors in each scene, we apply DBSCAN [7] clustering to the camera positions within the datasets. We exclude certain scenes, such as airports and large supermarkets, as well as panoramic images depicting outdoor environments to ensure alignment with our task. After processing, the final dataset sizes are summarized in Table 1.

4.2. Evaluation Metrics

To assess the quality of the generated panoramas, we employ both pixel-based and perceptual evaluation metrics. For pixel-level assessment, we utilize peak signal-to-noise ratio (PSNR) and structural similarity index measure (SSIM) to quantify image fidelity. Additionally, we incorporate perceptual metrics such as Fréchet Inception Distance (FID) [13] and Learned Perceptual Image Patch Similarity (LPIPS) [43] to capture higher-level visual realism.

4.3. Implement Details

Our code runs on an NVIDIA RTX A6000 GPU with 48GB of memory. The model has 3.3 billion parameters and is trained with a batch size of 21 for 100 epochs. On average, each experiment takes approximately two days to complete on both the Matterport3D and Gibson datasets. We optimize

Modules					Matterport3D				Gibson			
seg	floor	wall	depth	color	PSNR \uparrow	SSIM \uparrow	FID \downarrow	LPIPS \downarrow	PSNR \uparrow	SSIM \uparrow	FID \downarrow	LPIPS \downarrow
			✓	✓	11.08	0.4122	104.03	0.6463	11.07	0.4551	117.08	0.6574
✓			✓	✓	11.26	0.4206	102.84	0.6471	11.10	0.4564	115.88	0.6568
✓	✓			✓	10.78	0.4091	55.73	0.6161	11.19	0.4668	32.28	0.6310
✓		✓		✓	11.22	0.4207	54.44	0.6193	11.28	0.4613	83.31	0.6429
✓		✓	✓	✓	11.57	0.4378	43.99	0.6067	11.16	0.4733	81.87	0.6549
✓	✓		✓	✓	10.98	0.4196	53.34	0.6115	11.33	0.4601	33.89	0.6307
	✓	✓	✓	✓	11.21	0.4199	35.08	0.6131	11.38	0.4641	36.29	0.6384
✓	✓	✓		✓	11.26	0.4299	34.61	0.6060	11.49	0.4673	34.36	0.6336
✓	✓	✓	✓	✓	11.59	0.4381	67.02	0.6192	11.38	0.4761	65.66	0.6557
✓	✓	✓	✓	✓	11.72	0.4409	30.84	0.6029	11.58	0.4851	28.68	0.6282

Table 3. Ablation study on five designs in our Top2Pano model (top-down view **segmentation**, **floor** reinforcement, **wall** reinforcement, coarse **depth** panorama, and coarse **colored** panorama).

our model using the Adam optimizer [18] with default parameters ($\beta_1 = 0.9$, $\beta_2 = 0.999$, $\epsilon = 10^{-8}$) and a learning rate of 10^{-5} .

4.4. Comparison with Previous Methods

As shown in Table 2, we compare our method against three baseline approaches across four evaluation metrics. Sat2Density [28] is a satellite-to-ground panorama synthesis method, which we adapt for indoor panorama generation using latent diffusion model (LDM) [29] and ControlNet [42]. PanFusion [41] is a text-to-panorama generation framework that also incorporates layout-conditioned generation via ControlNet. Table 2 shows that our method outperforms all baselines across all four metrics on both datasets, demonstrating its effectiveness. Furthermore, qualitative comparisons in Figures 3 and 4 highlight that our approach generates more realistic and structurally accurate house reconstructions, including furniture placement.

4.5. Ablation Study

We conducted comprehensive ablation studies to analyze and validate the contribution of each component in our model. Specifically, we performed experiments comparing several model variants against the original. These experiments involved removing key elements such as the structural reinforcement of the floor and walls, the segmentation input to the OccRecon module, and the coarse depth and colored panoramas as conditional inputs to the PanoGen module.

As shown in Table 3, our original model achieves the highest overall scores across all four metrics, with any modification leading to some degree of performance degradation. Removing coarse colored panoramas or the embedded floor significantly disrupts furniture placement and color accuracy. While the room structure remains mostly intact, furniture positions become unreliable, and color representations appear distorted. Conversely, excluding coarse depth panoramas or embedded walls maintains color and furniture accuracy but compromises the spatial understanding and overall quality of room structure reconstruction. These effects are further illustrated in the qualitative results in the supp. materials.



Figure 3. Qualitative comparisons on the Matterport3D dataset.



Figure 4. Qualitative comparisons on the Gibson dataset.

	PSNR \uparrow	SSIM \uparrow	FID \downarrow	LPIPS \downarrow
$G \rightarrow M$	11.08	0.4397	40.74	0.6285
$M \rightarrow G$	11.03	0.4366	45.87	0.6353

Table 4. Cross-dataset evaluation. “G” represents Gibson; “M” represents Matterport3D.



Figure 5. Our Top2Pano model generalizes to both textured floorplans (first row) and hand-drawn sketched floorplans (second row), incorporating stylized [Japanese] control.

4.6. Generalization, Stylization, Manipulation

We employed cross-dataset evaluation to assess our model’s generalization capability. As shown in Table 4, although there is a slight decline in metric scores, our model maintains strong performance. Additionally, the trained parameters tend to generate photorealistic walls that reflect characteristics of the training dataset.

Our model also generalizes well to floorplans. When given a specific floorplan and camera positions, the model assists users in generating indoor panoramic images. By using different text prompts, users can create various interior design styles, explore the house from a first-person perspective, and modify its style according to their preferences. We tested our model with three types of floorplans. The first type is a colored floorplan (Figure 5, first row), which provides detailed information about room colors and furniture hues, making it highly informative. The second type is a plain floorplan (Figure 1, bottom), which lacks color information and shows only the room structure and furniture layout. The third type is a hand-drawn floorplan sketch (Figure 5, first row), which offers a rough visual representation of the room. Our model successfully generates accurate panoramic images from these floorplans and adapts the visual style based on the provided textual descriptions. Moreover, our model enables panorama manipulation by editing objects in the floorplan, such as adding new items, as illustrated in Figure 6.

4.7. Limitations

Failure cases. Figure 7 shows representative failure cases. We annotate different types of failures with numbered labels:

- *Ceiling*: ① missing fan, ② vaulted ceiling, ③ false light;
- *Wall*: ④ height error, ⑤ false or missing decorations;
- *Window*: ⑥ false window;



Figure 6. Top2Pano enables panorama manipulation via compositional floorplan editing. In the second row, adding a rectangular object to the floorplan (compared to the first row) leads the model to generate a washstand with a mirror in the panorama.



Top-down

Our Prediction

Ground Truth

Figure 7. Failure cases (zoom in to view error types).

- *Furniture*: ⑦ height error;
- *Thin object*: ⑧ missing flat-screen TV;
- *Stairs*: ⑨ false direction.

The failure cases stem from the ambiguity of the 2D input, leading to hallucinated objects that are not observable from the top-down view. The limitations in handling vertical structural details are largely due to the inherent ambiguity of the task.

Limited Vertical FoV. The generated panoramas exhibit a limited vertical field of view (FoV), reflecting the constraints of the training data. We expect improved performance with future datasets that include full vertical FoV panoramas.

5. Conclusions

We present **Top2Pano**, a novel method for generating high-quality 360° indoor panoramas from 2D top-down views. The model first estimates volumetric occupancy to infer 3D structure, then applies volumetric rendering to produce coarse color and depth panoramas. These guide a diffusion-based refinement stage via ControlNet. To our knowledge, this is the first approach to generate panoramas from top-down views. Experiments on two datasets show that Top2Pano outperforms baselines in reconstructing room layouts and realistic furniture.

References

- [1] Naofumi Akimoto, Yuhi Matsuo, and Yoshimitsu Aoki. Diverse plausible 360-degree image outpainting for efficient 3deg background creation. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR*, pages 11431–11440, 2022. 3
- [2] Sherwin Bahmani, Jeong Joon Park, Despoina Paschalidou, Xingguang Yan, Gordon Wetzstein, Leonidas J. Guibas, and Andrea Tagliasacchi. CC3D: layout-conditioned generation of compositional 3d scenes. In *IEEE/CVF International Conference on Computer Vision, ICCV*, pages 7137–7147, 2023. 1, 3
- [3] Angel Chang, Angela Dai, Thomas Funkhouser, Maciej Halber, Matthias Niessner, Manolis Savva, Shuran Song, Andy Zeng, and Yinda Zhang. Matterport3d: Learning from rgb-d data in indoor environments. *International Conference on 3D Vision (3DV)*, 2017. 1, 6
- [4] Minglin Chen, Longguang Wang, Sheng Ao, Ye Zhang, Kai Xu, and Yulan Guo. Layout2scene: 3d semantic layout guided scene generation via geometry and appearance diffusion priors. *arXiv preprint arXiv:2501.02519*, 2025. 1, 3
- [5] Zhaoxi Chen, Guangcong Wang, and Ziwei Liu. Text2light: Zero-shot text-driven hdr panorama generation. *ACM Transactions on Graphics (TOG)*, 41(6):1–16, 2022. 3
- [6] Mohammad Reza Karimi Dastjerdi, Yannick Hold-Geoffroy, Jonathan Eisenmann, Siavash Khodadadeh, and Jean-François Lalonde. Guided co-modulated gan for 360° field of view extrapolation. In *2022 International Conference on 3D Vision (3DV)*, pages 475–485, 2022. 3
- [7] Martin Ester, Hans-Peter Kriegel, Jörg Sander, and Xiaowei Xu. A density-based algorithm for discovering clusters in large spatial databases with noise. In *Proceedings of the Second International Conference on Knowledge Discovery and Data Mining (KDD)*, pages 226–231, 1996. 6
- [8] Chuan Fang, Xiaotao Hu, Kunming Luo, and Ping Tan. Ctrlroom: Controllable text-to-3d room meshes generation with layout constraints. *arXiv preprint arXiv:2310.03602*, 2023. 1, 3
- [9] Chengzeng Feng, Jiacheng Wei, Cheng Chen, Yang Li, Pan Ji, Fayao Liu, Hongdong Li, and Guosheng Lin. Prim2room: Layout-controllable room mesh generation from primitives. *arXiv preprint arXiv:2409.05380*, 2024. 1, 3
- [10] Mengyang Feng, Jinlin Liu, Miaomiao Cui, and Xuansong Xie. Diffusion360: Seamless 360 degree panoramic image generation based on diffusion models. *arXiv preprint arXiv:2311.13141*, 2023. 1, 3
- [11] Julia Guerrero-Viu, Clara Fernandez-Labrador, Cédric Demonceaux, and José Jesús Guerrero. What’s in my room? object recognition on indoor panoramic images. In *IEEE International Conference on Robotics and Automation, ICRA*, pages 567–573, 2020. 1
- [12] Meera Hahn, Jacob Krantz, Dhruv Batra, Devi Parikh, James M Rehg, Stefan Lee, and Peter Anderson. Where are you? localization from embodied dialog. *arXiv preprint arXiv:2011.08277*, 2020. 6
- [13] Martin Heusel, Hubert Ramsauer, Thomas Unterthiner, Bernhard Nessler, and Sepp Hochreiter. Gans trained by a two time-scale update rule converge to a local nash equilibrium. In *Annual Conference on Neural Information Processing Systems*, pages 6626–6637, 2017. 6
- [14] Edward J Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. LoRA: Low-rank adaptation of large language models. In *International Conference on Learning Representations, ICLR*, 2022. 5
- [15] Zilong Huang, Jun He, Junyan Ye, Lihan Jiang, Weijia Li, Yiping Chen, and Ting Han. Scene4u: Hierarchical layered 3d scene reconstruction from single panoramic image for your immerse exploration. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR*, pages 26723–26733, 2025. 1
- [16] Zhe Huang, Yizhe Zhao, Hao Xiao, Chenyan Wu, and Lingting Ge. Duospacenet: Leveraging both bird’s-eye-view and perspective view representations for 3d object detection. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops, CVPR Workshops*, pages 2560–2570, 2025. 1
- [17] Nikolai Kalischek, Michael Oechsle, Fabian Manhardt, Philipp Henzler, Konrad Schindler, and Federico Tombari. Cubediff: Repurposing diffusion-based image models for panorama generation. In *International Conference on Learning Representations, ICLR*, 2025. 3
- [18] Diederik P. Kingma and Jimmy Ba. Adam: A method for stochastic optimization. In *International Conference on Learning Representations, ICLR*, 2015. 6
- [19] Alexander Kirillov, Eric Mintun, Nikhila Ravi, Hanzi Mao, Chloé Rolland, Laura Gustafson, Tete Xiao, Spencer Whitehead, Alexander C. Berg, Wan-Yen Lo, Piotr Dollár, and Ross B. Girshick. Segment anything. In *IEEE/CVF International Conference on Computer Vision, ICCV*, pages 3992–4003, 2023. 2, 3
- [20] Jiachen Liu, Yuan Xue, Haomiao Ni, Rui Yu, Zihan Zhou, and Sharon X Huang. Computer-aided layout generation for building design: A review. *arXiv preprint arXiv:2504.09694*, 2025. 1
- [21] Jiachen Liu, Rui Yu, Sili Chen, Sharon X. Huang, and Hengkai Guo. Towards in-the-wild 3d plane reconstruction from a single image. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR*, pages 27027–27037, 2025. 1
- [22] Xiaohu Lu, Zuoyue Li, Zhaopeng Cui, Martin R. Oswald, Marc Pollefeys, and Rongjun Qin. Geometry-aware satellite-to-ground image synthesis for urban areas. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 856–864, 2020. 3
- [23] Wentao Lyu, Peng Ding, Yingliang Zhang, Anpei Chen, Minye Wu, Shu Yin, and Jingyi Yu. Refocusable gigapixel panoramas for immersive VR experiences. *IEEE Trans. Vis. Comput. Graph.*, 27(3):2028–2040, 2021. 1
- [24] Nelson L. Max. Optical models for direct volume rendering. *IEEE Trans. Vis. Comput. Graph.*, 1(2):99–108, 1995. 4
- [25] Ben Mildenhall, Pratul P. Srinivasan, Matthew Tancik, Jonathan T. Barron, Ravi Ramamoorthi, and Ren Ng. Nerf: Representing scenes as neural radiance fields for view syn-

- thesis. In *European Conference on Computer Vision, ECCV*, pages 405–421, 2020. [4](#)
- [26] Despoina Paschalidou, Amlan Kar, Maria Shugrina, Karsten Kreis, Andreas Geiger, and Sanja Fidler. Atiss: Autoregressive transformers for indoor scene synthesis. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2021. [3](#)
- [27] Giovanni Pintore, Fabio Bettio, Marco Agus, and Enrico Gobbetti. Deep scene synthesis of atlanta-world interiors from a single omnidirectional image. *IEEE Trans. Vis. Comput. Graph.*, 29(11):4708–4718, 2023. [1](#)
- [28] Ming Qian, Jincheng Xiong, Gui-Song Xia, and Nan Xue. Sat2density: Faithful density learning from satellite-ground image pairs. In *IEEE/CVF International Conference on Computer Vision, ICCV*, pages 3660–3669, 2023. [3](#), [4](#), [6](#), [7](#)
- [29] Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. High-resolution image synthesis with latent diffusion models. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR*, pages 10674–10685, 2022. [6](#)
- [30] Jonas Schult, Sam Tsai, Lukas Höllein, Bichen Wu, Jialiang Wang, Chih-Yao Ma, Kunpeng Li, Xiaofang Wang, Felix Wimbauer, Zijian He, Peizhao Zhang, Bastian Leibe, Peter Vajda, and Ji Hou. Controlroom3d: Room generation using semantic proxy rooms. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2024. [1](#), [3](#)
- [31] Yujiao Shi, Dylan Campbell, Xin Yu, and Hongdong Li. Geometry-guided street-view panorama synthesis from satellite imagery. *IEEE Trans. Pattern Anal. Mach. Intell.*, 44(12):10009–10022, 2022. [3](#)
- [32] Bin Tan, Rui Yu, Yujun Shen, and Nan Xue. Planarsplatting: Accurate planar surface reconstruction in 3 minutes. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR*, pages 1190–1199, 2025. [1](#)
- [33] Madhawa Vidanapathirana, Qirui Wu, Yasutaka Furukawa, Angel X. Chang, and Manolis Savva. Plan2scene: Converting floorplans to 3d scenes. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR*, pages 10733–10742, 2021. [3](#)
- [34] Jionghao Wang, Ziyu Chen, Jun Ling, Rong Xie, and Li Song. 360-degree panorama generation from few unregistered nfov images. In *Proceedings of the 31st ACM International Conference on Multimedia, MM*, pages 6811–6821, 2023. [3](#)
- [35] Songsong Wu, Hao Tang, Xiao-Yuan Jing, Haifeng Zhao, Jianjun Qian, Nicu Sebe, and Yan Yan. Cross-view panorama image synthesis. *IEEE Trans. Multimed.*, 25:3546–3559, 2023. [3](#)
- [36] Fei Xia, Amir R. Zamir, Zhi-Yang He, Alexander Sax, Jitendra Malik, and Silvio Savarese. Gibson env: real-world perception for embodied agents. In *IEEE Conference on Computer Vision and Pattern Recognition, CVPR*, 2018. [6](#)
- [37] Ningli Xu and Rongjun Qin. Geospecific view generation geometry-context aware high-resolution ground view inference from satellite views. In *European Conference on Computer Vision, ECCV*, pages 349–366, 2024. [3](#)
- [38] Xiuyu Yang, Yunze Man, Jun-Kun Chen, and Yu-Xiong Wang. Scenecraft: Layout-guided 3d scene generation. In *Advances in Neural Information Processing Systems*, 2024. [1](#), [3](#)
- [39] Weicai Ye, Chenhao Ji, Zheng Chen, Junyao Gao, Xiaoshui Huang, Song-Hai Zhang, Wanli Ouyang, Tong He, Cairong Zhao, and Guofeng Zhang. Diffpano: Scalable and consistent text to panorama generation with spherical epipolar-aware diffusion. In *Annual Conference on Neural Information Processing Systems, NeurIPS*, 2024. [1](#), [3](#)
- [40] Rui Yu, Jiachen Liu, Zihan Zhou, and Sharon X. Huang. Nerf-enhanced outpainting for faithful field-of-view extrapolation. In *IEEE International Conference on Robotics and Automation, ICRA*, pages 16826–16833, 2024. [1](#)
- [41] Cheng Zhang, Qianyi Wu, Camilo Cruz Gambardella, Xiaoshui Huang, Dinh Phung, Wanli Ouyang, and Jianfei Cai. Taming stable diffusion for text to 360° panorama image generation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2024. [1](#), [3](#), [6](#), [7](#)
- [42] Lvmin Zhang, Anyi Rao, and Maneesh Agrawala. Adding conditional control to text-to-image diffusion models. In *IEEE/CVF International Conference on Computer Vision, ICCV*, pages 3813–3824, 2023. [2](#), [3](#), [4](#), [6](#)
- [43] Richard Zhang, Phillip Isola, Alexei A. Efros, Eli Shechtman, and Oliver Wang. The unreasonable effectiveness of deep features as a perceptual metric. In *IEEE Conference on Computer Vision and Pattern Recognition, CVPR*, pages 586–595, 2018. [6](#)
- [44] Dian Zheng, Cheng Zhang, Xiao-Ming Wu, Cao Li, Chengfei Lv, Jian-Fang Hu, and Wei-Shi Zheng. Panorama generation from nfov image done right. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR*, pages 21610–21619, 2025. [1](#)

A. Qualitative Results on Ablation Study

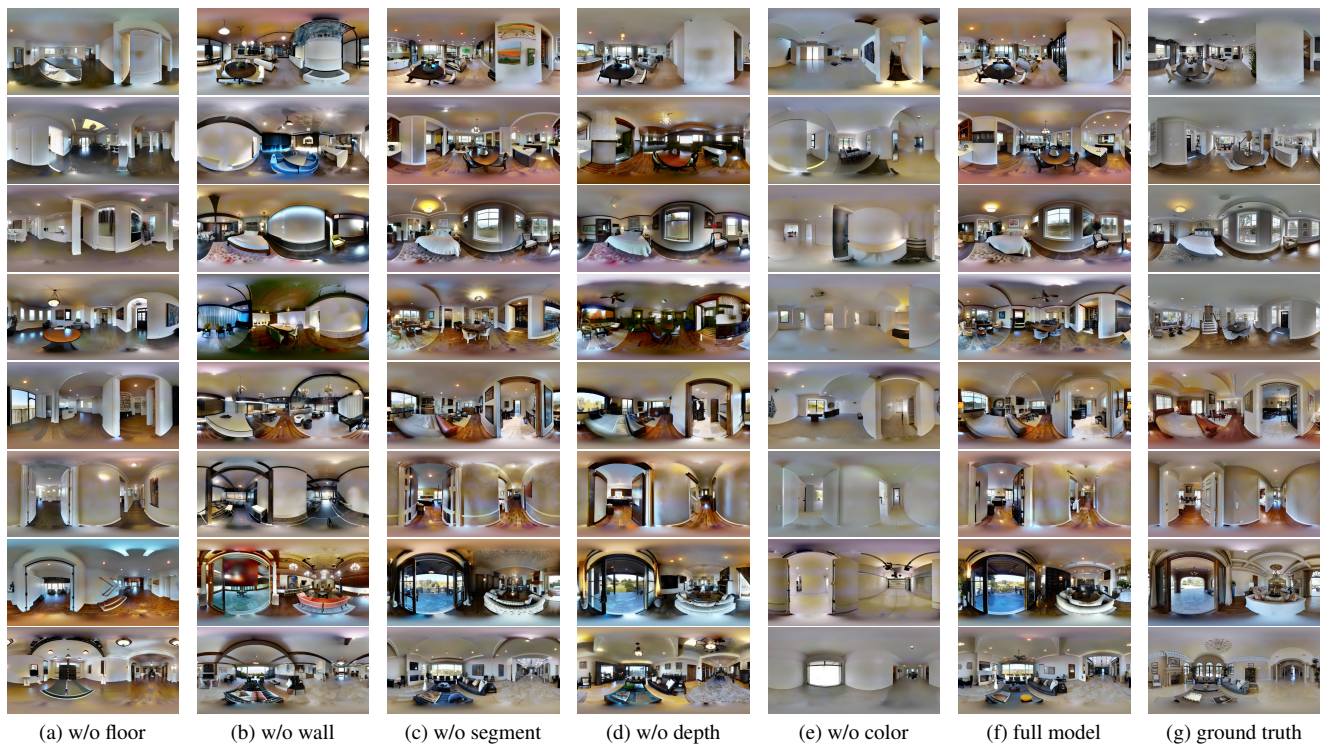


Figure A.1. Qualitative results of ablation experiments on the Matterport3D dataset.



Figure A.2. Qualitative results of ablation experiments on the Gibson dataset.