

# MAGRITTE: MANIPULATIVE AND GENERATIVE 3D REALIZATION FROM IMAGE, TOPVIEW AND TEXT

**Takayuki Hara**

The University of Tokyo

hara@mi.t.u-tokyo.ac.jp

**Tatsuya Harada**

The University of Tokyo / RIKEN

harada@mi.t.u-tokyo.ac.jp

## ABSTRACT

The generation of 3D scenes from user-specified conditions offers a promising avenue for alleviating the production burden in 3D applications. Previous studies required significant effort to realize the desired scene, owing to limited control conditions. We propose a method for controlling and generating 3D scenes under multimodal conditions using partial images, layout information represented in the top view, and text prompts. Combining these conditions to generate a 3D scene involves the following significant difficulties: (1) the creation of large datasets, (2) reflection on the interaction of multimodal conditions, and (3) domain dependence of the layout conditions. We decompose the process of 3D scene generation into 2D image generation from the given conditions and 3D scene generation from 2D images. 2D image generation is achieved by fine-tuning a pretrained text-to-image model with a small artificial dataset of partial images and layouts, and 3D scene generation is achieved by layout-conditioned depth estimation and neural radiance fields (NeRF), thereby avoiding the creation of large datasets. The use of a common representation of spatial information using 360-degree images allows for the consideration of multimodal condition interactions and reduces the domain dependence of the layout control. The experimental results qualitatively and quantitatively demonstrated that the proposed method can generate 3D scenes in diverse domains, from indoor to outdoor, according to multimodal conditions. A project website with supplementary video is here <https://hara012.github.io/MaGRITTe-project>.

**Keywords** 3D scene generation · 360-degree image generation · image outpainting · text-to-3D · layout-to-3D

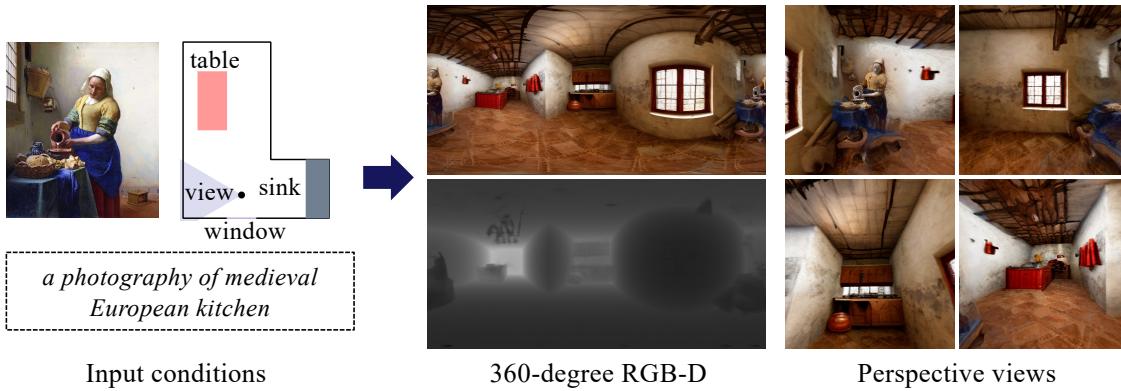


Figure 1: From a given partial image, layout information represented in top view, and text prompts, our method generates a 3D scene represented by the 360-degree RGB-D, and NeRF. Free perspective views can be rendered from the NeRF model.

## 1 Introduction

3D scene generation under user-specified conditions is a fundamental task in the fields of computer vision and graphics. In particular, the generation of 3D scenes extending in all directions from the observer’s viewpoint is a promising technology that reduces the burden and time of creators and provides them with new ideas for creation in 3D applications such as VR/AR, digital twins, and the metaverse.

In recent years, 3D scene generation under user-specified conditions using generative models [30, 44, 19, 57, 50, 25] has been extensively studied. A wide range of methods exist for generating 3D scenes from partial images [14, 6, 15, 12], layout information such as floor plans and bird’s-eye views [58, 5, 28, 69, 10, 48], and text prompts [63, 49, 26, 54]. However, these methods are limited by the conditions they can take as input, making it difficult to generate the 3D scene intended by the user. This is due to the fact that each condition has its own advantages and disadvantages. For example, when partial images are given, it is possible to present a detailed appearance; however, it is difficult to create information outside the image; when a layout is given, it is possible to accurately describe object alignment but not to specify a detailed appearance; when text is given as a condition, it is suitable for specifying the overall context; however, it is difficult to determine the exact shape and appearance of objects.

Considering these problems, we propose a method for generating 3D scenes by simultaneously providing a combination of three conditions: partial images, layout information represented in the top view, and text prompts (fig. 1). This approach aims to compensate for the shortcomings of each condition in a complementary manner, making it easier to create the 3D scenes intended by the creator. That is, details of appearance from partial images, shape and object placement from layout information, and overall context can be controlled using text prompts.

Integrating partial images, layouts, and texts to control a 3D scene involves the following significant difficulties that cannot be addressed by a simple combination of existing methods: (1) creation of large datasets, (2) reflection of the interaction of multimodal conditions, and (3) domain dependence of the layout representations. To overcome these difficulties, we initially decomposed the process of 3D scene generation into two steps: 2D image generation from the given conditions and 3D generation from 2D images. For 2D image generation, our approach is to create small artificial datasets for partial images and layout conditions and fine-tune the text-to-image model trained on a large dataset. We then generated a 3D scene from a 2D image using layout-conditioned monocular depth estimation and training NeRF [39]. This approach eliminates the need to create large datasets of 3D scenes. This study aimed to improve scene consistency and reduce computational costs using 360-degree images for 2D image generation. To address the second issue, which reflects the interaction of multimodal conditions, we encoded the input conditions into a common latent space in the form of equirectangular projection (ERP) for 360-degree images. To address the third issue of domain dependence of layout representations, we present a framework for incorporating domain-specific top-view representations with less effort by converting them into more generic intermediate representations of depth and semantic maps in ERP format. This allows for generating various scenes from indoor to outdoor by simply replacing the converter.

The contributions of this study are as follows:

- We introduce a method to control and generate 3D scenes from partial images, layouts, and texts, complementing the advantages of each condition.
- We present a method that avoids the need for creating large datasets by fine-tuning a pre-trained large-scale text-to-image model with a small artificial dataset of partial images and layouts for 2D image generation, and by generating 3D scenes from 2D images through layout-conditioned depth estimation and training NeRF.
- We address the integration of different modalities by converting the input information into ERP format, passing it through an encoder, and embedding the information in the same latent space.
- We present a framework for generating various scenes from indoor to outdoor with a module for converting top view layout representations into depth maps and semantic maps in ERP format.
- Experimental results validate that the proposed method can generate 3D scenes with controlled appearance, geometry, and overall context based on input information, even beyond the dataset used for fine-tuning.

## 2 Related Work

### 2.1 3D Scene Generation

3D scene generation involves the creation of a model of a 3D space that includes objects and backgrounds, based on user-specified conditions. In recent years, the use of generative models, such as VAE [30, 44], GAN [19], autoregressive

models [57], and diffusion models [50, 25], has made rapid progress. There are methods to generate a 3D scene from random variables [37, 8], from one or a few images [14, 6, 35, 15, 12], from layout information such as floor plans [58, 5], bird’s-eye views (semantic maps in top view) [28, 69], terrain maps [10] and 3D proxies [48], and as well as from text prompts [63, 49, 26, 54, 17]. However, each method has its own advantages and disadvantages in terms of scene control characteristics, and it is difficult to generate a 3D scene that appropriately reflects the intentions. We propose a method to address these challenges by integrating partial images, layout information, and text prompts as input conditions in a complementary manner. Furthermore, layout conditions were designed for each domain. Furthermore, while layout conditions need to be designed for each domain, the proposed method switches between converters for layout representations, enabling the generation of a variety of scenes from indoor to outdoor.

## 2.2 Scene Generation Using 360-Degree Image

Image generation methods have been studied for 360-degree images that record the field of view in all directions from a single observer’s viewpoint. Methods to generate 360-degree images from one or a few normal images [18, 51, 3, 2, 21, 4, 22, 64] and text prompts [11, 62, 56] have been reported. Methods for panoramic three-dimensional structure prediction were also proposed [52, 53].

Studies have also extended the observer space to generate 3D scenes with six degrees of freedom (DoF) from 360-degree RGB-D. In [27, 20, 31, 61], methods were proposed for constructing a 6-DoF 3D scene by training the NeRF from 360-degree RGB-D. LDM3D [54] shows a series of pipelines that add channels of depth to the latent diffusion model (LDM) [45], generate 360-degree RGB-D from the text, and mesh it. Generating 3D scenes via 360-degree images is advantageous in terms of guaranteeing scene consistency and reducing computation. Our research attempts to generate 360-degree images from multiple conditions and 6-DoF 3D scenes by layout-conditioned depth estimation and training the NeRF.

## 2.3 Monocular Depth Estimation

Monocular depth estimation involves estimating the depth of each pixel in a single RGB image. In recent years, deep learning-based methods have progressed significantly, and methods based on convolutional neural networks [47, 34, 32, 66, 67, 70, 38] and transformers [7, 13, 68, 55, 42] have been proposed. Monocular depth estimation for 360-degree images was also investigated [73, 33, 16, 59, 74, 60, 40, 43, 1]. In this study, we incorporated monocular depth estimation into a 360-degree RGB-D generation pipeline.

## 3 Proposed Method

This section describes the proposed method called *MaGRITTe*, that generates 3D scenes under multiple conditions. fig. 2 illustrates the overview of our method. Three input conditions are considered: a partial image, layout information represented in the top view, text prompts, and outputs from a 360-degree RGB-D and NeRF model. The proposed method comprises four steps: (a) ERP conversion of partial images and layouts, (b) 360-degree RGB image generation, (c) layout-conditioned depth estimation, and (d) NeRF training. The following sections describe each step.

### 3.1 Conversion of Partial Image and Layout

First, we describe the conversion of the partial image and layout in (a) of fig. 2. This study uses two layout representations, floor plans and terrain maps, for indoor and outdoor scenes, respectively.

#### 3.1.1 Floor Plans

A floor plan is a top-view representation of the room shape and the position/size/class of objects. The room shape comprises the two-dimensional coordinates of the corners and the height positions of the floor and ceiling, based on the assumption that the walls stand vertically. The objects are specified by 2D bounding box, height from the floor at the top and bottom, and class, such as chair or table.

#### 3.1.2 Terrain Maps

As shown in fig. 3, a terrain map describes the height of the terrain relative to the horizontal plane. This is a set  $\mathbb{R}^{H_{ter} \times W_{ter}}$  that constitutes a  $H_{ter} \times W_{ter}$  grid with the height of the ground surface at each grid point.

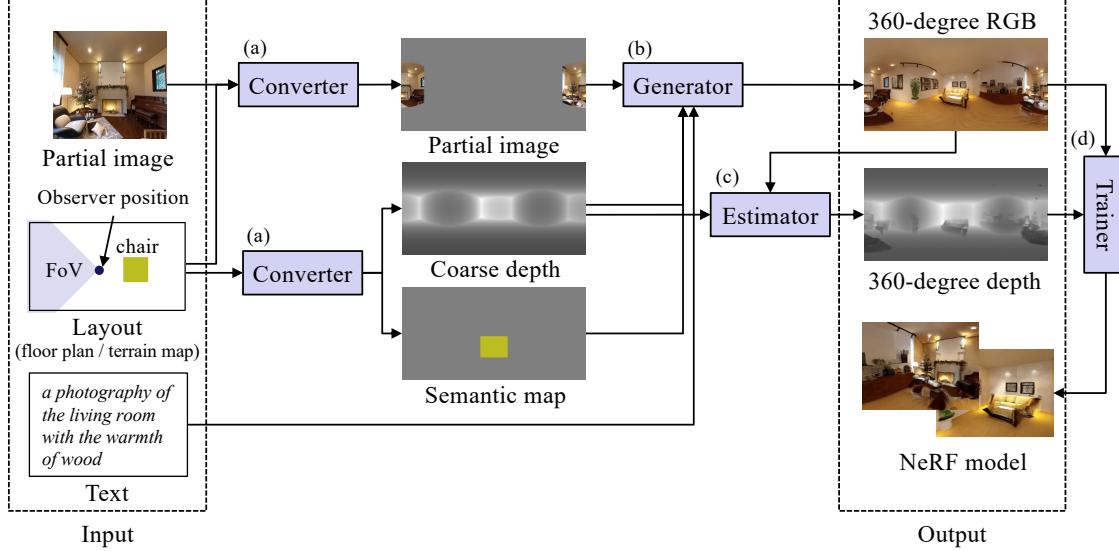


Figure 2: Overview of the proposed method to generate 360-degree RGB-D and NeRF models from a partial image, layouts and text prompts. (a) The partial image is converted to an ERP image from the observer position with the specified direction and field-of-view (FoV). The layout represented in top view is converted to a coarse depth and a semantic map in ERP format with the observer position as the projection center. (b) These ERP images and texts are combined to generate a 360-degree RGB. (c) The generated RGB is combined with the coarse depth to estimate the fine depth. (d) a NeRF model is trained from 360-degree RGB-D.

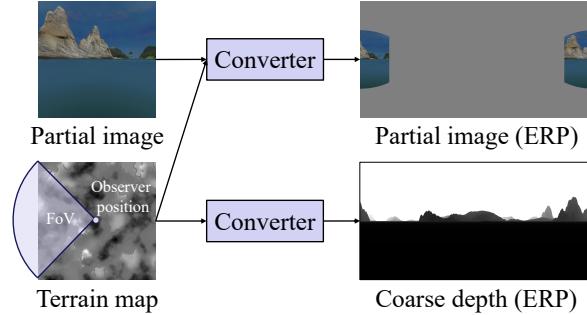


Figure 3: The case of using a terrain map for the layout format. The partial image and the terrain map are converted into ERP images from the observer’s viewpoint, respectively.

### 3.1.3 ERP Conversion

The observer position and field of view (FoV) of the partial image are provided in the layout. Based on this information, a partial RGB  $\mathcal{P} \in \mathbb{R}^{H_{\text{ERP}} \times W_{\text{ERP}} \times 3}$ , coarse depth  $\mathcal{D} \in \mathbb{R}^{H_{\text{ERP}} \times W_{\text{ERP}}}$ , and semantic map  $\mathcal{S} \in \{0, 1\}^{H_{\text{ERP}} \times W_{\text{ERP}} \times C}$  are created in the ERP format, as shown in fig. 2 (a), where  $H_{\text{ERP}}$  and  $W_{\text{ERP}}$  are the height and width of the ERP image, respectively, and  $C$  denotes the number of classes. The semantic map takes  $\mathcal{S}_{ijc} = 1$  when an object of class  $c$  exists at position  $(i, j)$  and  $\mathcal{S}_{ijc} = 0$  otherwise. For floor plans, the distance from the observer’s viewpoint to the room wall is recorded, and for terrain maps, the distance from the observer’s viewpoint to the terrain surface is recorded in ERP format and used as the coarse depth. A semantic map is created for a floor plan; the regions specifying the objects are projected onto the ERP image with the observer position of the partial image as the projection center, and object classes are assigned to the locations of their presence.

## 3.2 360-Degree RGB Generation

We combine partial images, coarse depths, and semantic maps represented in the ERP format and integrate them with text prompts to generate a 360-degree RGB image. Using the ERP format for the input and output allows the use

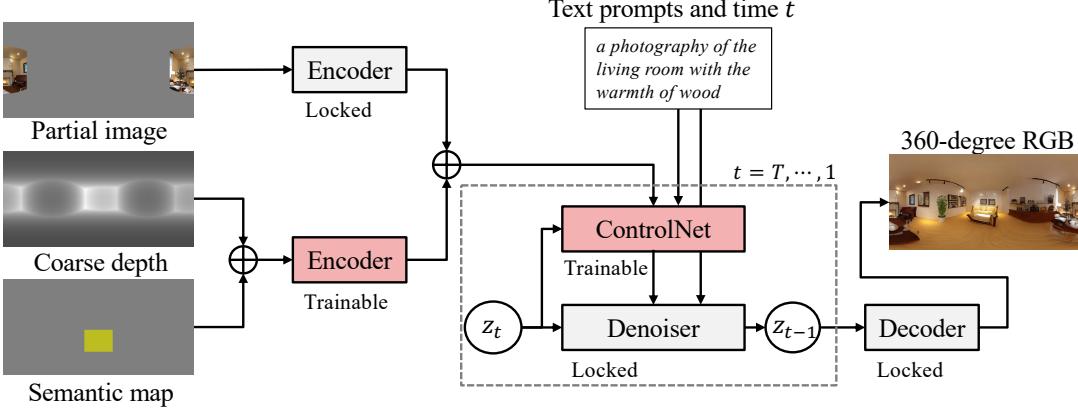


Figure 4: The pipeline of generating 360-degree RGB from a partial image, coarse depth map, semantic map, and text prompts. The partial image, coarse depth, and semantic map are embedded in the common latent space using encoders, and their channels are concatenated. Along with the text prompt, 360-degree RGB is generated in the framework of ControlNet[71] based on LDM[45].

of text-to-image models trained on large datasets. In this study, we employ StableDiffusion (SD) [45], a pre-trained diffusion model with an encoder and decoder, as the base text-to-image model. We fine-tune the model for our purposes using ControlNet [71], which controls the diffusion model with an additional network of conditional inputs. fig. 4 shows the pipeline to generate 360-degree RGB. A partial image, coarse depth, and semantic maps are embedded in the latent space, channel merged, and provided as conditional inputs to ControlNet along with text prompts. This is an improvement on PanoDiff [62], which generates 360-degree images from partial images, and our method embeds layout information into a common latent space in ERP format as well, allowing for interaction between conditions while preserving spatial information. The encoder for partial images is from SD, and the encoder for layout information is a network with the same structure as that used in ControlNet. The weights of the network derived from SD are fixed, and only the weights of the network derived from ControlNet are updated during training.

Fine-tuning of the base model degrades image-to-text performance. To mitigate this phenomenon, we additionally use a dataset with text annotations only for fine-tuning. If one model is trained for different combinations of conditions, the learning may not be generalized to other combinations of conditions. We introduce condition dropout (CD), in which training is performed by randomly changing the combination of conditions. Each condition is dropped with a probability of 50%, with the ERP image conditions being replaced by pixel values of 0 and text replaced by an empty string.

### 3.3 Layout-Conditioned Depth Estimation

Next, a fine depth is estimated from the coarse depth and the generated 360-degree RGB. In this study, we propose and compare two methods: end-to-end estimation and depth integration.

#### 3.3.1 End-to-End Estimation

In the end-to-end approach, the depth is estimated using U-Net [46] with a self-attention mechanism [57] with four channels of RGB-D as the input, and one channel of depth as the output. The network is trained to minimize the L1 loss between the network outputs and ground truth. Details of the network configuration are provided in appendix A.

#### 3.3.2 Depth Integration

In the depth integration approach, depth estimates are obtained from 360-degree RGB using the monocular depth estimation method, LeRes [70] is employed in this study, and the final depth is obtained so as to minimize the weighted squared error for the coarse depth and depth estimates. Since LeRes is designed for normal field-of-view images, the 360-degree image is projected onto  $N$  tangent images, and depth estimation and integration are performed on each tangent image. Let  $\hat{d}_n \in \mathbb{R}^{H_d W_d}$  ( $n = 1, 2, \dots, N$ ) be the monocular depth estimate for  $n$ -th tangent image in ERP format, where  $H_d$  and  $W_d$  are the height and width of the depth map, respectively. Since the estimated depth  $\hat{d}_n$  has unknown scale and offset, it is transformed using the affine transformation coefficient  $s_n \in \mathbb{R}^2$  as  $\tilde{d}_n s_n$ , where

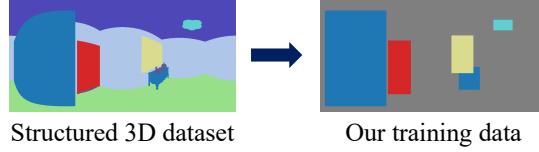


Figure 5: Semantic map. Regions related to objects are extracted, excluding regions derived from the shape of the room, such as walls, floor, and ceiling, which are enclosed in a bounding box to form a semantic map in the proposed method.

$\tilde{d}_n = (\hat{d}_n - 1) \in \mathbb{R}^{H_d W_d \times 2}$ . We consider the following evaluation function  $\mathcal{L}_{\text{depth}}$ , where  $d_0 \in \mathbb{R}^{H_d W_d}$  is the coarse depth,  $\Phi_n \in \mathbb{R}^{H_d W_d \times H_d W_d}$  ( $n = 0, 1, \dots, N$ ) is the weight matrix, and  $x \in \mathbb{R}^{H_d W_d}$  is the integrated depth.

$$\mathcal{L}_{\text{depth}} = \|x - d_0\|_{\Phi_0}^2 + \sum_{n=1}^N \|x - \tilde{d}_n s_n\|_{\Phi_n}^2, \quad (1)$$

where quadratic form  $\|v\|_Q^2 = v^\top Q v$ . The fine depth  $x$  and coefficients  $s_n$  ( $n = 1, 2, \dots, N$ ) that minimize  $\mathcal{L}_{\text{depth}}$  can be obtained in closed form from the extreme value conditions as follows:

$$x = \left( \sum_{n=0}^N \Phi_n \right)^{-1} \left( \Phi_0 d_0 + \sum_{n=1}^N \Phi_n \tilde{d}_n s_n \right), \quad (2)$$

$$\begin{bmatrix} s_1 \\ s_2 \\ \vdots \\ s_N \end{bmatrix} = \begin{bmatrix} D_1 & U_{1,2} & \cdots & U_{1,N} \\ U_{2,1} & D_2 & \cdots & U_{2,N} \\ \vdots & \vdots & \ddots & \vdots \\ U_{N,1} & U_{N,2} & \cdots & D_N \end{bmatrix}^{-1} \begin{bmatrix} b_1 \\ b_2 \\ \vdots \\ b_N \end{bmatrix}, \quad (3)$$

where,  $D_k = \tilde{d}_k^\top \{\Phi_k^{-1} + (\sum_{n=0 \setminus k}^N \Phi_n)^{-1}\}^{-1} \tilde{d}_k$ ,  $U_{k,l} = -\tilde{d}_k^\top \Phi_k (\sum_{n=0}^N \Phi_n)^{-1} \Phi_l \tilde{d}_l$ ,  $b_k = \tilde{d}_k^\top \Phi_k (\sum_{n=0}^N \Phi_n)^{-1} \Phi_0 d_0$ . The derivation of the equation and setting of weights  $\{\Phi_n\}_{n=0}^N$  are described in appendix A.

### 3.4 Training NeRF

Finally, we train the NeRF model using the generated 360-degree RGB-D. In this study, we employ a method from [20] that can train NeRF by inpainting the occluded regions from a single image.

## 4 Dataset

We build fine-tuning datasets for indoor and outdoor scenes, respectively. We create artificial datasets with layout annotations using computer graphics as the *base dataset*, whereas datasets without layout annotations are created using actual captured datasets as the *auxiliary dataset*.

### 4.1 Indoor Scene

For the base dataset, we modified and used a structured 3D dataset [72] containing 3500 synthetic departments (scenes) with 185,985 panoramic renderings for RGB, depth, and semantic maps. The same room had both furnished and unfurnished patterns, and the depth of the unfurnished room was used as the coarse depth. For consistency with the ERP conversion in section 3.1, the semantic map was transformed, as shown in (fig. 5). Each image was annotated with text using BLIP [36] and partial images were created using a perspective projection transformation of 360-degree RGB with random camera parameters. The data were divided into 161,126 samples for training, 2048 samples for validation, and 2048 samples for testing.

For the auxiliary dataset, we used the Matterport 3D dataset [9], which is an indoor real-world 360° dataset captured by Matterport's Pro 3D camera including 10,800 RGB-D panoramic images. Similar to the structured 3D dataset, partial images and text were annotated. The depth and semantic maps included in the dataset were not used, and zero was assigned as the default value for the coarse depth and semantic map during training. The data were divided into 7675 samples for training and 2174 samples for testing.

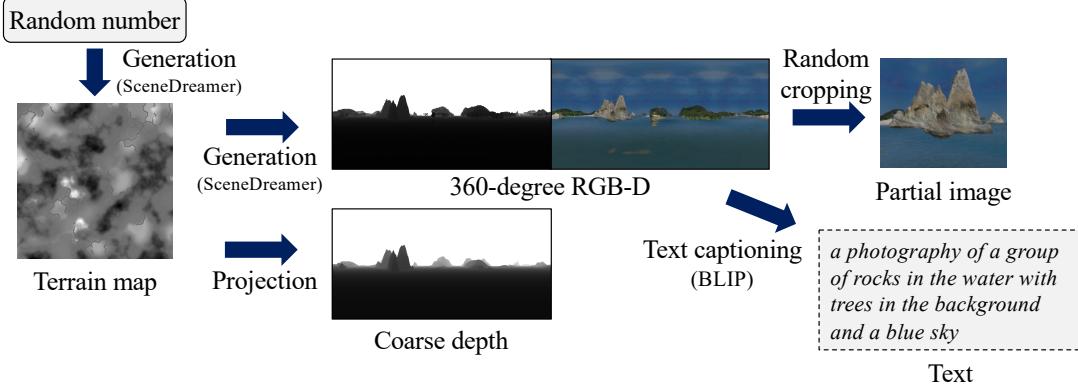


Figure 6: Dataset creation for outdoor scene. SceneDreamer [10] generates a terrain map from a random number, and renders 360-degree RGB-D. The generated RGB image is annotated with text using BLIP [36], and partial images are created by a perspective projection transformation of 360-degree RGB with random camera parameters. A coarse depth is converted from the terrain maps

## 4.2 Outdoor Scene

As the base dataset, we created the *SceneDreamer dataset* using SceneDreamer [10], which is a model for generating 3D scenes. As shown in fig. 6, a 360-degree RGB-D image was generated from random numbers via a terrain map to annotate the partial images and texts. A semantic map was not used in this study because of limited object classes. The data were divided into 12,600 samples for training, 2,052 samples for validation, and 2052 samples for testing.

For the auxiliary dataset, we used the SUN360 dataset [65] which includes various real captured 360-degree RGB images. We extracted only outdoor scenes from the dataset, and partial images and text were annotated. The distance to the horizontal plane was set as the default value for the coarse depth during training. The data were divided into 39,174 training samples and 2048 testing samples.

## 5 Experimental Results

Quantitative and qualitative experiments were conducted to verify the effectiveness of the proposed method for generating 3D scenes under multiple conditions.

### 5.1 Implementation Details

The partial images, coarse depths, and semantic maps were in ERP format with a resolution of  $512 \times 512$ , and the shape of the latent variable in the LDM was  $64 \times 64 \times 4$ . We trained the 360-degree RGB generation model based on the pretrained SD v2.1 using the Adam optimizer [29] with a learning rate of  $1.0 \times 10^{-5}$  and batch size of 16. We trained the end-to-end depth estimation model from scratch using the Adam optimizer with a learning rate of  $4.5 \times 10^{-6}$  and batch size of 6. The convolutional layers in the networks use circular padding [22] to resolve the left-right discontinuity in ERP.

### 5.2 360-Degree RGB Generation

First, we evaluate 360-degree RGB generation. We used the peak-signal-to-noise-ratio (PSNR) as the evaluation metric: PSNR (whole) for the entire image between the ground truth and generated images, PSNR (partial) for the region of the partial image given by the input. We also employ the FID [24], which is a measure of the divergence of feature distributions between the ground truth and generated images, and the CLIP score (CS) [41, 23], which promptly quantifies the similarity with the input text. PSNR (whole) measures the reproducibility of the ground truth at the pixel value level, PSNR (partial) measures the degree to which the partial image given as the input is reflected in the generated result, FID measures the plausibility of the generated image, and CS measures the fidelity of the generated result to text prompts. Because there is no comparison method that uses partial images, layouts, and text prompts as inputs to generate a 360-degree image, we compared our method with PanoDiff [62], which is a state-of-the-art 360-degree RGB image generation model that uses partial images and texts. We implemented it and used PanoDiff with the encoder of

Table 1: Evaluation results of 360-degree RGB generation on the Modified Structured 3D dataset and the SceneDreamer dataset.

method	Structured3D dataset				SceneDreamer dataset			
	PSNR↑ (whole)	PSNR↑ (partial)	FID↓	CS↑	PSNR↑ (whole)	PSNR↑ (partial)	FID↓	CS↑
PanoDiff [62]	11.59	<b>36.00</b>	21.23	<b>30.75</b>	<u>12.91</u>	<b>37.19</b>	30.94	<u>29.86</u>
Ours (w/o CD)	<b>12.56</b>	<u>35.39</u>	<u>18.87</u>	<u>30.72</u>	12.46	34.68	<u>29.54</u>	29.71
Ours (w/ CD)	<u>12.42</u>	33.29	<b>18.84</b>	30.71	<b>13.29</b>	<u>34.81</u>	<b>29.05</b>	<b>29.93</b>

Table 2: CS evaluation results for base model forgetting

Trained on base dataset	Trained on auxiliary dataset	Condition dropout	Indoor	Outdoor
✓			29.48	24.75
✓	✓		29.34	26.24
✓	✓	✓	<b>30.23</b>	<b>29.26</b>

the layout information removed in the proposed method for a fair comparison using the same network configurations and pretrained models.

table 1 shows the quantitative evaluation results of 360-degree RGB generation on the Structured 3D dataset and the SceneDreamer dataset. The results of the proposed method are shown with and without CD. PanoDiff is superior in terms of PSNR (partial) and CS, which is a reasonable result since PanoDiff is a method that takes only partial images and text prompts as conditions for image generation. However, the proposed method is superior to PSNR (whole) and FID, which indicates that the reproducibility and plausibility of the generated images can be enhanced by considering layout information as a condition as well. Comparing the proposed method with and without CD, FID tended to be slightly better when CD was present, whereas PSNR (whole), PSNR (partial), and CS were superior or inferior depending on the two datasets. The better performance of CD on the SceneDreamer dataset can be attributed to the larger number of samples in the auxiliary dataset.

fig. 7 shows the examples of generating a 360-degree RGB image for the test set of the Structured 3D dataset and the SceneDreamer dataset. PanoDiff, which does not use the layout information as a condition, generates images that differ significantly from the ground truth. This may have led to the degradation of PSNR (whole) and FID. Although the image generated by the proposed method differs from the ground-truth image at the pixel level, it can generate images with room geometry, terrain, and object placement in accordance with the given conditions. When comparing methods with and without CD, it is difficult to determine the perceived superiority in most cases. A closer look at fig. 7 (d) shows that additional objects, such as trees, are more likely to appear in Ours (w/o CD) on the SceneDreamer dataset, which may be a factor in the PSNR degradation.

Next, we present the results of the evaluation of the experiment in a setting in which the conditions were crossed between datasets. table 2 shows the results of the CS for generated results with the text prompt of the auxiliary dataset for the depth of the base dataset. This indicates that CS can be improved by using the auxiliary dataset and CD. fig. 8 shows the difference with and without CD. These results show that the use of CD better reflects text prompts, and the generalization of text prompts in combination with depth is possible.

Table 3: Evaluation results of 360-degree depth generation on the Modified Structured 3D dataset and the SceneDreamer dataset

Method	Structured3D dataset		SceneDreamer dataset	
	RMSE↓	AbsRel↓	RMSE↓	AbsRel↓
Coarse depth	8.858	0.0117	15.30	0.0200
360MonoDepth [43]	21.67	0.0138	15.30	0.0202
LeRes (ERP) [70]	19.03	0.0149	<u>15.24</u>	<u>0.0187</u>
LeRes (multi views)	21.90	0.0147	15.25	0.0188
Ours (end-to-end)	<b>6.649</b>	<b>0.0056</b>	15.29	0.0196
Ours (depth integration)	7.432	0.0119	<b>15.20</b>	<b>0.0185</b>

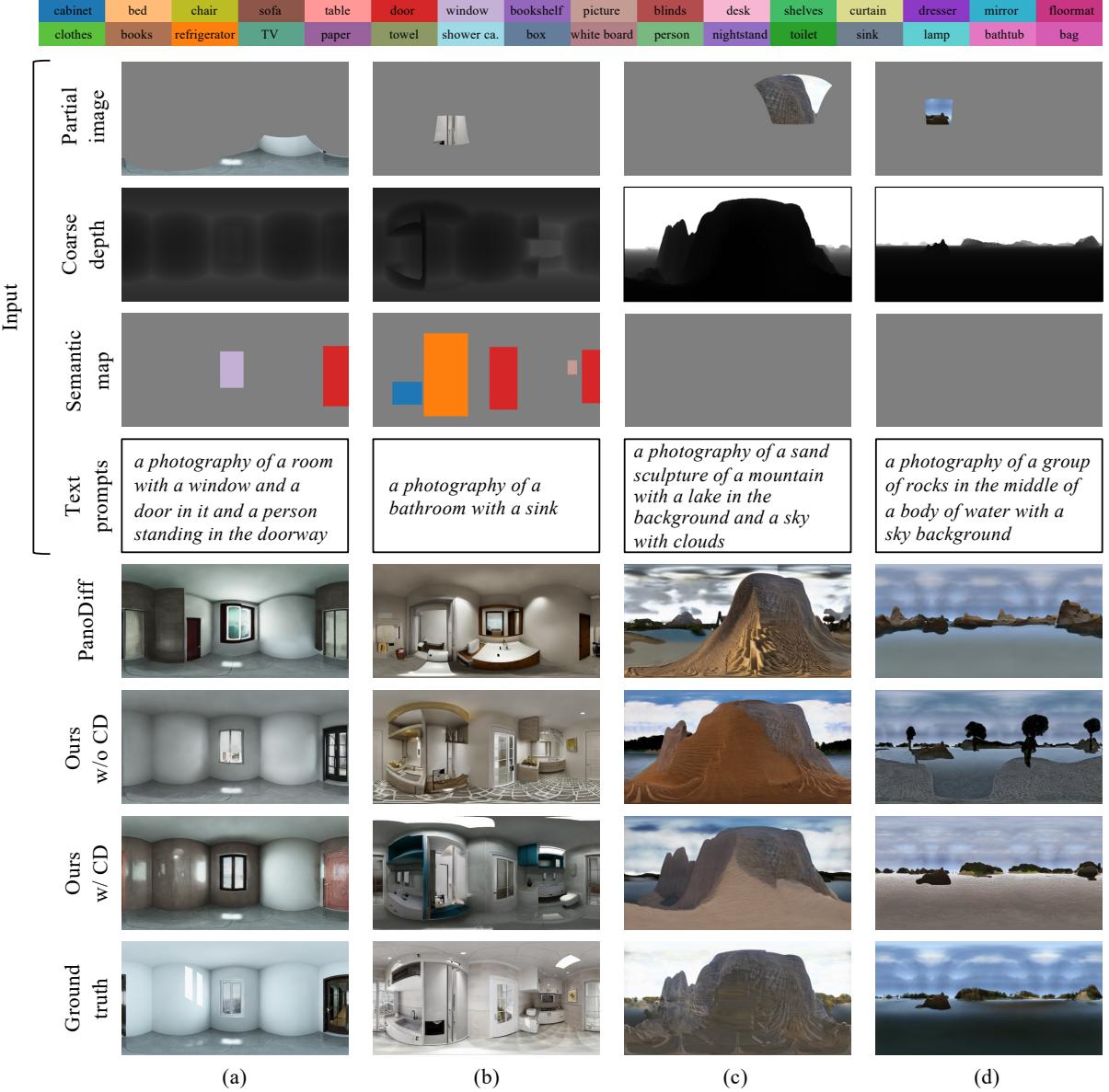


Figure 7: The results of generating a 3D scene for the test set of (a)(b) the Stuructured 3D dataset and (c)(d) the SceneDreamer dataset.

### 5.3 360-Degree Depth Generation

Next, we evaluate the depth of the generated 360-degree image. Because the estimated depth has scale and offset degrees of freedom, its value was determined to minimize the squared error with the ground-truth depth, similar to the method presented in [42]. We used the root mean squared error (RMSE) and mean absolute value of the relative error,  $\text{AbsRel} = \frac{1}{M} \sum_{i=1}^M \frac{|z_i - z_i^*|}{z_i^*}$ , where  $M$  is the number of pixels,  $z_i$  is the estimated depth of the  $i$ th pixel, and  $z_i^*$  is the ground-truth depth of the  $i$ th pixel. Pixels at infinity were excluded from evaluation. table 3 shows the results of the quantitative evaluation of depth generation on the Structured 3D dataset and the SceneDreamer dataset. For comparison, the results of 360MonoDepth which is a 360° monocular depth estimation [43] method; LeRes (ERP), which is LeRes [70] directly applied to ERP; and LeRes (multi views), which applies LeRes to multiple tangent images of a 360-degree image and integrates the estimated depths in a section 3.3 manner without using coarse depth, are also shown. In terms of RMSE and AbsRel, our method (end-to-end) was the best for the structured 3D dataset, and our method (depth integration) was the best for the SceneDreamer dataset. It was also shown that combining LeRes with



Figure 8: The difference with and without CD. In this example, "piano" in the text prompt is reflected only for the method with CD.

coarse depth increased accuracy compared to using LeRes alone. The end-to-end method is relatively ineffective for the SceneDreamer dataset. This may be because the number of samples in the dataset was small and the depth was estimated to be close to the coarse depth.

#### 5.4 Results in the Wild

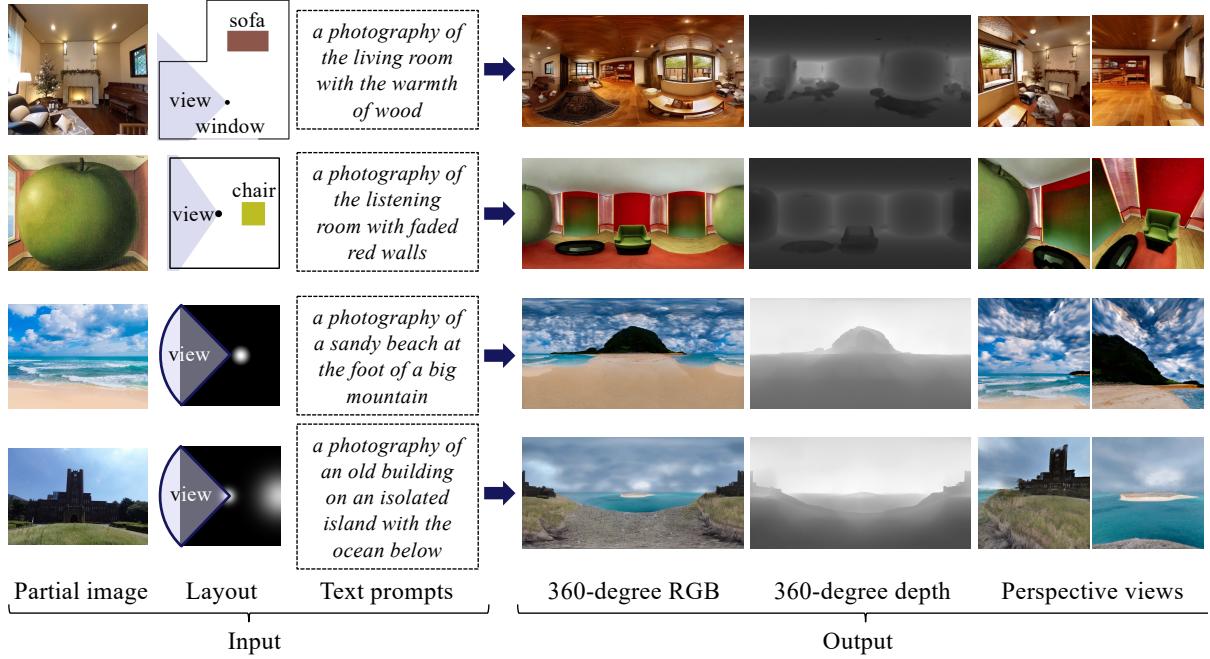


Figure 9: Samples of the 3D scene generation based on user-generated conditions. Perspective views are rendered using the learned NeRF model. The first and fourth partial images are taken by the author using a camera, the second is a painting entitled "The Listening Room" by René Magritte and the third was downloaded from the web (<https://www.photo-ac.com/>).

Finally, we evaluated the results of 3D scene generation based on user-generated conditions outside the dataset used for fine-tuning. Examples of 3D scenes generated by the proposed method, conditioned on partial images, layouts, and text, are shown in figs. 1 and 9. These conditions were created freely by the authors. It can be seen that the generated scene contains the given partial image and conforms to the instructions of the text prompt according to the given layout. These results show that the proposed method can generate 3D scenes with the appearance, geometry, and overall context controlled according to the input information, even outside the dataset used for fine-tuning. Details of the experimental setup, additional samples, ablation studies, and limitations are described in appendices B and C.

## 6 Conclusions

We proposed a method for generating and controlling 3D scenes using partial images, layout information, and text prompts. We confirmed that fine-tuning a large-scale text-to-image model with small artificial datasets can generate

360-degree images from multiple conditions, and free perspective views can be generated by layout-conditioned depth estimation and training NeRF. This enables 3D scene generation from multimodal conditions without creating a new large dataset. It is also indicated that the interaction of multiple spatial conditions can be performed using a common ERP latent space, and that both indoor and outdoor scenes can be handled by replacing the conversions.

Future studies will include the detection of inconsistent input conditions and suggestions for users on how to resolve these inconsistencies. Creating conditions under which the layout and partial images match perfectly is difficult, and a method that aligns with the approximate settings is desirable.

## Acknowledgements

This work was partially supported by JST Moonshot R&D Grant Number JPMJPS2011, CREST Grant Number JPMJCR2015 and Basic Research Grant (Super AI) of Institute for AI and Beyond of the University of Tokyo. We would like to thank Yusuke Kurose, Jingen Chou, Haruo Fujiwara, and Sota Oizumi for helpful discussions.

## References

- [1] Ai, H., Cao, Z., pei Cao, Y., Shan, Y., Wang, L.: Hrdfuse: Monocular 360° depth estimation by collaboratively learning holistic-with-regional depth distributions. In: IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) (2023)
- [2] Akimoto, N., Aoki, Y.: Image completion of 360-degree images by cgan with residual multi-scale dilated convolution. IIEEJ Transactions on Image Electronics and Visual Computing **8**(1), 35–43 (2020)
- [3] Akimoto, N., Kasai, S., Hayashi, M., Aoki, Y.: 360-degree image completion by two-stage conditionalgans. In: IEEE International Conference on Image Processing (ICIP) (2019)
- [4] Akimoto, N., Matsuo, Y., Aoki, Y.: Diverse plausible 360-degree image outpainting for efficient 3dcg background creation. In: IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) (2022)
- [5] Bahmani, S., Park, J.J., Paschalidou, D., Yan, X., Wetzstein, G., Guibas, L., Tagliasacchi, A.: Cc3d: Layout-conditioned generation of compositional 3d scenes. arXiv:2303.12074 (2023)
- [6] Bautista, M.A., Guo, P., Abnar, S., Talbott, W., Toshev, A., Chen, Z., Dinh, L., Zhai, S., Goh, H., Ulbricht, D., Dehghan, A., Susskind, J.: Gaudi: A neural architect for immersive 3d scene generation. In: Advances in Neural Information Processing Systems (NeurIPS) (2022)
- [7] Bhat, S.F., Alhashim, I., Wonka, P.: AdaBins: Depth estimation using adaptive bins. In: IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) (2021)
- [8] Chai, L., Tucker, R., Li, Z., Isola, P., Snavely, N.: Persistent nature: A generative model of unbounded 3d worlds. In: IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) (2023)
- [9] Chang, A., Dai, A., Funkhouser, T., Halber, M., Nießner, M., Savva, M., Song, S., Zeng, A., Zhang, Y.: Matterport3d: Learning from rgb-d data in indoor environments. In: International Conference on 3D Imaging, Modeling, Processing, Visualization and Transmission (3DIMPVT) (2017)
- [10] Chen, Z., Wang, G., Liu, Z.: Scenedreamer: Unbounded 3d scene generation from 2d image collections. IEEE Transactions on Pattern Analysis and Machine Intelligence (TPAMI) **45**(12), 15562–15576 (2023)
- [11] Chen, Z., Wang, G., Liu, Z.: Text2light: Zero-shot text-driven hdr panorama generation. ACM Transactions on Graphics (TOG) **41**(6), 1–16 (2022)
- [12] Cheng, W., Cao, Y.P., Shan, Y.: Sparsegnv: Generating novel views of indoor scenes with sparse input views. arXiv:2305.07024 (2023)
- [13] Cheng, Z., Zhang, Y., Tang, C.: Swin-depth: Using transformers and multi-scale fusion for monocular-based depth estimation. IEEE Sensors Journal (2021)
- [14] DeVries, T., Bautista, M.A., Srivastava, N., Taylor, G.W., Susskind, J.M.: Unconstrained scene generation with locally conditioned radiance fields. In: IEEE/CVF International Conference on Computer Vision (ICCV) (2021)
- [15] Du, Y., Smith, C., Tewari, A., Sitzmann, V.: Learning to render novel views from wide-baseline stereo pairs. In: IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) (2023)
- [16] Eder, M., Moulon, P., Guan, L.: Pano popups: Indoor 3d reconstruction with a plane-aware network. In: International Conference on 3D Vision (3DV) (2019)

- [17] Fridman, R., Abecasis, A., Kasten, Y., Dekel, T.: Scenescape: Text-driven consistent scene generation. arXiv:2302.01133 (2023)
- [18] Gardner, M.A., Sunkavalli, K., Yumer, E., Shen, X., Gambaretto, E., Christian, G., Lalonde, J.F.: Learning to predict indoor illumination from a single image. ACM Transactions on Graphics (TOG) **9**(4) (2017)
- [19] Goodfellow, I., Pouget-Abadie, J., Mirza, M., Xu, B., Warde-Farley, D., Ozair, S., Courville, A., Bengio, Y.: Generative adversarial nets. In: Advances in Neural Information Processing Systems (NeurIPS) (2014)
- [20] Hara, T., Harada, T.: Enhancement of novel view synthesis using omnidirectional image completion. arXiv:2203.09957 (2022)
- [21] Hara, T., Mukuta, Y., Harada, T.: Spherical image generation from a single image by considering scene symmetry. In: AAAI Conference on Artificial Intelligence (AAAI) (2021)
- [22] Hara, T., Mukuta, Y., Harada, T.: Spherical image generation from a few normal-field-of-view images by considering scene symmetry. IEEE Transactions on Pattern Analysis and Machine Intelligence (TPAMI) **45**(5), 6339–6353 (2022)
- [23] Hessel, J., Holtzman, A., Forbes, M., Le Bras, R., Choi, Y.: CLIPScore: A reference-free evaluation metric for image captioning. In: Moens, M.F., Huang, X., Specia, L., Yih, S.W.t. (eds.) Conference on Empirical Methods in Natural Language Processing (EMNLP) (2021)
- [24] Heusel, M., Ramsauer, H., Unterthiner, T., Nessler, B.: Gans trained by a two time-scale update rule converge to a local nash equilibrium. In: Advances in Neural Information Processing Systems (NeurIPS) (2017)
- [25] Ho, J., Jain, A., Abbeel, P.: Denoising diffusion probabilistic models. In: Advances in Neural Information Processing Systems (NeurIPS) (2020)
- [26] Höllerin, L., Cao, A., Owens, A., Johnson, J., Nießner, M.: Text2room: Extracting textured 3d meshes from 2d text-to-image models. In: IEEE/CVF International Conference on Computer Vision (ICCV) (2023)
- [27] Hsu, C.Y., Sun, C., Chen, H.T.: Moving in a 360 world: Synthesizing panoramic parallaxes from a single panorama. arXiv:2106.10859 (2021)
- [28] Kim, S.W., Brown, B., Yin, K., Kreis, K., Schwarz, K., Li, D., Rombach, R., Torralba, A., Fidler, S.: Neuralfield-ldm: Scene generation with hierarchical latent diffusion models. In: IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) (2023)
- [29] Kingma, D.P., Ba, J.L.: Adam: A method for stochastic optimization. arXiv:1412.6980 (2014)
- [30] Kingma, D.P., Welling, M.: Auto-encoding variational bayes. arXiv:1312.6114 (2013)
- [31] Kulkarni, S., Yin, P., Scherer, S.: 360fusionnerf: Panoramic neural radiance fields with joint guidance. arXiv:2209.14265 (2022)
- [32] Kuznetsov, Y., Stuckler, J., Leibe, B.: Semi-supervised deep learning for monocular depth map prediction. In: The IEEE Conference on Computer Vision and Pattern Recognition (CVPR) (2017)
- [33] de La Garanderie, G.P., Atapour-Abarghouei, A., Breckon, T.: Eliminating the blind spot: Adapting 3d object detection and monocular depth estimation to 360° panoramic imagery. In: European Conference on Computer Vision (ECCV) (2018)
- [34] Laina, I., Rupprecht, C., Belagiannis, V., Tombari, F., Navab, N.: Deeper depth prediction with fully convolutional residual networks. In: International Conference on 3D Vision (3DV). IEEE (2016)
- [35] Lei, J., Tang, J., Jia, K.: Rgbd2: Generative scene synthesis via incremental view inpainting using rgbd diffusion models. In: IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) (2023)
- [36] Li, J., Li, D., Xiong, C., Hoi, S.: Blip: Bootstrapping language-image pre-training for unified vision-language understanding and generation. In: International Conference on Machine Learning (ICML) (2022)
- [37] Lin, C.H., Lee, H.Y., Menapace, W., Chai, M., Siarohin, A., Yang, M.H., Tulyakov, S.: InfiniCity: Infinite-scale city synthesis. In: IEEE/CVF International Conference on Computer Vision (ICCV) (2023)
- [38] Masoumian, A., Rashwan, H.A., Abdulwahab, S., Cristiano, J., Asif, M.S., Puig, D.: Gcndepth: Self-supervised monocular depth estimation based on graph convolutional network. Neurocomputing (2022)
- [39] Mildenhall, B., Srinivasan, P.P., Tancik, M., Barron, J.T., Ramamoorthi, R., Ng, R.: Nerf: Representing scenes as neural radiance fields for view synthesis. In: European Conference on Computer Vision (ECCV) (2020)
- [40] Pintore, G., Agus, M., Almansa, E., Schneider, J., Gobbetti, E.: SliceNet: deep dense depth estimation from a single indoor panorama using a slice-based representation. In: IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) (2021)

- [41] Radford, A., Kim, J.W., Hallacy, C., Ramesh, A., Goh, G., Agarwal, S., Sastry, G., Askell, A., Mishkin, P., Clark, J., Krueger, G., Sutskever, I.: Learning transferable visual models from natural language supervision. In: International Conference on Machine Learning (ICML) (2021)
- [42] Ranftl, R., Lasinger, K., Hafner, D., Schindler, K., Koltun, V.: Towards robust monocular depth estimation: Mixing datasets for zero-shot cross-dataset transfer. IEEE Transactions on Pattern Analysis and Machine Intelligence (TPAMI) **44**(3), 1623–1637 (2022)
- [43] Rey-Area, M., Yuan, M., Richardt, C.: 360MonoDepth: High-resolution 360deg monocular depth estimation. In: IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) (2022)
- [44] Rezende, D.J., Mohamed, S., Wierstra, D.: Stochastic backpropagation and approximate inference in deep generative models. In: International Conference on Machine Learning (ICML) (2014)
- [45] Rombach, R., Blattmann, A., Lorenz, D., Esser, P., Ommer, B.: High-resolution image synthesis with latent diffusion models. In: IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) (2022)
- [46] Ronneberger, O., Fischer, P., Brox, T.: U-net: Convolutional networks for biomedical image segmentation. In: Navab, N., Hornegger, J., Wells, W.M., Frangi, A.F. (eds.) Medical Image Computing and Computer-Assisted Intervention (MICCAI) (2015)
- [47] Roy, A., Todorovic, S.: Monocular depth estimation using neural regression forest. In: IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) (2016)
- [48] Schult, J., Tsai, S., Höllerin, L., Wu, B., Wang, J., Ma, C.Y., Li, K., Wang, X., Wimbauer, F., He, Z., et al.: Controlroom3d: Room generation using semantic proxy rooms. arXiv:2312.05208 (2023)
- [49] Shi, Y., Wang, P., Ye, J., Long, M., Li, K., Yang, X.: Mvdream: Multi-view diffusion for 3d generation. arXiv:2308.16512 (2023)
- [50] Sohl-Dickstein, J., Weiss, E.A., Maheswaranathan, N., Ganguli, S.: Deep unsupervised learning using nonequilibrium thermodynamics. In: International Conference on Machine Learning (ICML) (2015)
- [51] Song, S., Funkhouser, T.: Neural illumination: Lighting prediction for indoor environments. In: IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) (2019)
- [52] Song, S., Zeng, A., Chang, A.X., Savva, M., Savarese, S., Funkhouser, T.: Im2pano3d: Extrapolating 360° structure and semantics beyond the field of view. In: IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) (2018)
- [53] Srinivasan, P.P., Mildenhall, B., Tancik, M., Barron, J.T., Tucker, R., Snavely, N.: Lighthouse: Predicting lighting volumes for spatially-coherent illumination. In: IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) (2020)
- [54] Stan, G.B.M., Wofk, D., Fox, S., Redden, A., Saxton, W., Yu, J., Aflalo, E., Tseng, S.Y., Nonato, F., Müller, M., Lal, V.: Ldm3d: Latent diffusion model for 3d. arXiv:2305.10853 (2023)
- [55] Sun, C., Sun, M., Chen, H.: Hohonet: 360 indoor holistic understanding with latent horizontal features. In: IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) (2021)
- [56] Tang, S., Zhang, F., Chen, J., Wang, P., Furukawa, Y.: Mvdiffusion: Enabling holistic multi-view image generation with correspondence-aware diffusion. arXiv (2023)
- [57] Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A.N., Kaiser, L.u., Polosukhin, I.: Attention is all you need. In: Advances in Neural Information Processing Systems (NeurIPS). vol. 30 (2017)
- [58] Vidanapathirana, M., Wu, Q., Furukawa, Y., Chang, A.X., Savva, M.: Plan2scene: Converting floorplans to 3d scenes. In: IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) (2021)
- [59] Wang, F.E., Hu, H.N., Cheng, H.T., Lin, J.T., Yang, S.T., Shih, M.L., Chu, H.K., Sun, M.: Self-supervised learning of depth and camera motion from 360°*circ* videos. In: Jawahar, C., Li, H., Mori, G., Schindler, K. (eds.) Asian Conference on Computer Vision (ACCV) (2019)
- [60] Wang, F.E., Yeh, Y.H., Sun, M., Chiu, W.C., Tsai, Y.H.: Bifuse: Monocular 360 depth estimation via bi-projection fusion. In: The IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) (2020)
- [61] Wang, G., Wang, P., Chen, Z., Wang, W., Loy, C.C., Liu, Z.: Perf: Panoramic neural radiance field from a single panorama. arXiv:2310.16831 (2023)
- [62] Wang, J., Chen, Z., Ling, J., Xie, R., Song, L.: 360-degree panorama generation from few unregistered nfov images. In: ACM International Conference on Multimedia (2023)
- [63] Wang, Z., Lu, C., Wang, Y., Bao, F., Li, C., Su, H., Zhu, J.: Prolificdreamer: High-fidelity and diverse text-to-3d generation with variational score distillation. arXiv:2305.16213 (2023)

- [64] Wu, T., Zheng, C., Cham, T.J.: Ipo-ldm: Depth-aided 360-degree indoor rgb panorama outpainting via latent diffusion model. arXiv:2307.03177 (2023)
- [65] Xiao, J., Ehinger, K.A., Oliva, A., Torralba, A.: Recognizing scene viewpoint using panoramic place representation. In: IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) (2012)
- [66] Xu, D., Ricci, E., Ouyang, W., Wang, X., Sebe, N.: Multi-scale continuous crfs as sequential deep networks for monocular depth estimation. In: IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) (2017)
- [67] Xu, D., Wang, W., Tang, H., Liu, H., Sebe, N., Ricci, E.: Structured attention guided convolutional neural fields for monocular depth estimation. In: IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) (2018)
- [68] Yang, G., Tang, H., Ding, M., Sebe, N., Ricci, E.: Transformer-based attention networks for continuous pixel-wise prediction. In: IEEE/CVF International Conference on Computer Vision (ICCV) (2021)
- [69] Yang, K., Ma, E., Peng, J., Guo, Q., Lin, D., Yu, K.: Bevcontrol: Accurately controlling street-view elements with multi-perspective consistency via bev sketch layout. arXiv:2308.01661 (2023)
- [70] Yin, W., Zhang, J., Wang, O., Niklaus, S., Mai, L., Chen, S., Shen, C.: Learning to recover 3d scene shape from a single image. In: IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) (2021)
- [71] Zhang, L., Rao, A., Agrawala, M.: Adding conditional control to text-to-image diffusion models. In: IEEE International Conference on Computer Vision (ICCV) (2023)
- [72] Zheng, J., Zhang, J., Li, J., Tang, R., Gao, S., Zhou, Z.: Structured3d: A large photo-realistic dataset for structured 3d modeling. In: European Conference on Computer Vision (ECCV) (2020)
- [73] Zioulis, N., Karakottas, A., Zarpalas, D., Daras, P.: Omnidepth: Dense depth estimation for indoors spherical panoramas. In: Ferrari, V., Hebert, M., Sminchisescu, C., Weiss, Y. (eds.) European Conference on Computer Vision (ECCV) (2018)
- [74] Zioulis, N., Karakottas, A., Zarpalas, D., Alvarez, F., Daras, P.: Spherical view synthesis for self-supervised 360° depth estimation. In: International Conference on 3D Vision (3DV) (2019)

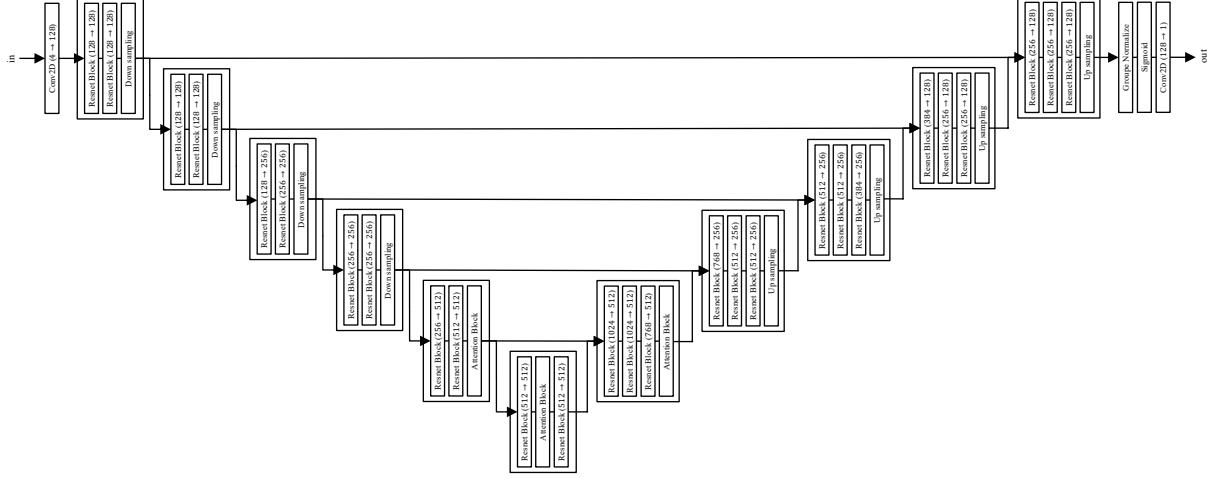


Figure 10: The structure of the layout-conditioned depth estimation network. Conv2D ( $N \rightarrow M$ ) is a two-dimensional convolutional layer with  $N$  input channels,  $M$  output channels, and a kernel size of  $3 \times 3$ . The Resnet Block shown in fig. 11 is combined into a U-Net structure. Downsampling and upsampling are performed using a factor of 2. In the Attention Block, self-attention [57] in the form of a query, key, and value is applied in pixels.

## A Details of Layout-Conditioned Depth Estimation

In this section, we describe the details of the layout-conditioned depth estimation, which generates a fine depth from the coarse depth and generated RGB.

### A.1 End-toEnd Network Configuration

The structure of the network that generates a fine depth from a coarse depth and the generated RGB end-to-end is shown in figs. 10 and 11. The network consists of a combination of U-Net [46] and self-attention [57], with four channels of RGB-D as the input and one channel of depth as the output. The network was trained to minimize the L1 loss between the depth output from the network and the depth of the ground truth. The model was trained from scratch using the Adam optimizer with a learning rate of  $4.5 \times 10^{-6}$  and a batch size of six.

### A.2 Equation Derivation for Depth Integration

Let  $\hat{d}_n \in \mathbb{R}^{H_d W_d}$  ( $n = 1, 2, \dots, N$ ) be the monocular depth estimate for  $n$ -th tangent image in ERP format, where  $H_d$  and  $W_d$  are the height and width of the depth map, respectively. Since the estimated depth  $\hat{d}_n$  has unknown scale and offset, it is transformed using the affine transformation coefficient  $s_n \in \mathbb{R}^2$  as  $\tilde{d}_n s_n$ , where  $\tilde{d}_n = (\hat{d}_n - 1) \in \mathbb{R}^{H_d W_d \times 2}$ . We consider the following evaluation function  $\mathcal{L}_{\text{depth}}$ , where  $d_0 \in \mathbb{R}^{H_d W_d}$  is the coarse depth,  $\Phi_n \in \mathbb{R}^{H_d W_d \times H_d W_d}$  ( $n = 0, 1, \dots, N$ ) is the weight matrix, and  $x \in \mathbb{R}^{H_d W_d}$  is the integrated depth.

$$\mathcal{L}_{\text{depth}} = \|x - d_0\|_{\Phi_0}^2 + \sum_{n=1}^N \|x - \tilde{d}_n s_n\|_{\Phi_n}^2, \quad (4)$$

where the quadratic form  $\|v\|_Q^2 = v^\top Q v$ . We find the affine transformation coefficient  $s_n$  ( $n = 1, 2, \dots, N$ ) and fine depth  $x$  from the extreme-value conditions to minimize  $\mathcal{L}_{\text{depth}}$ . The partial differentiation of eq. (4) with  $x$  yields:

$$\begin{aligned} \frac{\partial \mathcal{L}_{\text{depth}}}{\partial x} &= 2\Phi_0(x - d_0) + 2 \sum_{n=1}^N \Phi_n(x - \tilde{d}_n s_n) \\ &= 2 \sum_{n=0}^N \Phi_n x - 2 \left( \Phi_0 d_0 + \sum_{n=1}^N \Phi_n \tilde{d}_n s_n \right), \end{aligned} \quad (5)$$

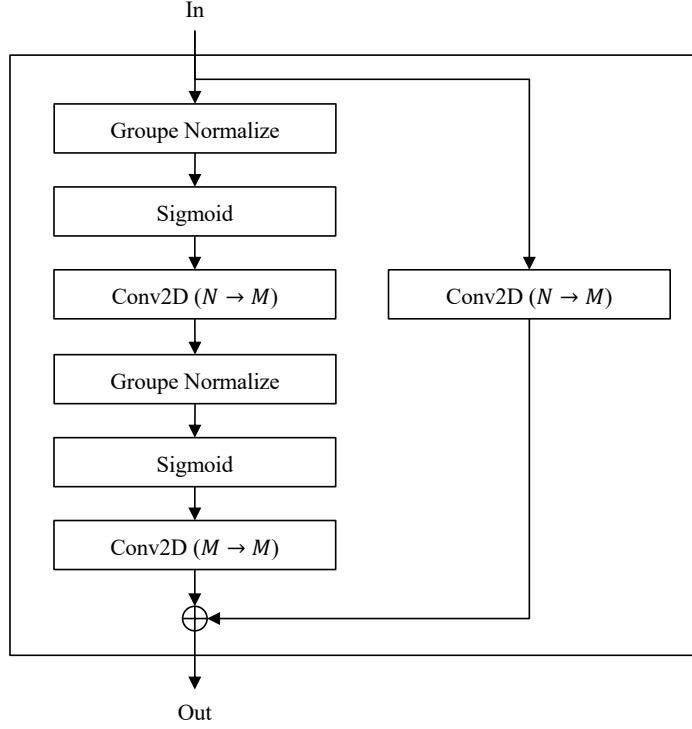


Figure 11: The structure of a Resnet Block ( $N \rightarrow M$ ).  $N$  is the number of input channels, and  $M$  is the number of output channels. In the groupe normalize, the number of split channels is fixed at 32. Conv2D refers to a two-dimensional convolutional layer, and the numbers in parentheses indicate the conversion of the number of channels.

and  $x$  satisfying the extreme-value conditions are as follows:

$$x = \left( \sum_{n=0}^N \Phi_n \right)^{-1} \left( \Phi_0 d_0 + \sum_{n=1}^N \Phi_n \tilde{d}_n s_n \right). \quad (6)$$

Next, the partial differentiation of eq. (4) with  $s_k$  yields:

$$\frac{\partial \mathcal{L}_{\text{depth}}}{\partial s_k} = -2\tilde{d}_k^\top \Phi_k (x - \tilde{d}_k s_k), \quad (7)$$

and  $s_k$  satisfying the extreme-value conditions are as follows:

$$\tilde{d}_k^\top \Phi_k \tilde{d}_k s_k = \tilde{d}_k^\top \Phi_k x. \quad (8)$$

By substituting eq. (6) into eq. (10), we obtain

$$\tilde{d}_k^\top \Phi_k \tilde{d}_k s_k = \tilde{d}_k^\top \Phi_k \left( \sum_{n=0}^N \Phi_n \right)^{-1} \left( \Phi_0 d_0 + \sum_{n=1}^N \Phi_n \tilde{d}_n s_n \right). \quad (9)$$

Transposing  $s_n$  on the left-hand side yields

$$\tilde{d}_k^\top \Phi_k \tilde{d}_k s_k - \tilde{d}_k^\top \Phi_k \left( \sum_{n=0}^N \Phi_n \right)^{-1} \sum_{n=1}^N \Phi_n \tilde{d}_n s_n = \tilde{d}_k^\top \Phi_k \left( \sum_{n=0}^N \Phi_n \right)^{-1} \Phi_0 d_0. \quad (10)$$

Considering the coefficient of  $s_k$  as  $D_k \in \mathbb{R}^{2 \times 2}$ , we obtain

$$\begin{aligned}
 D_k &= \tilde{d}_k^\top \Phi_k \tilde{d}_k - \tilde{d}_k^\top \Phi_k \left( \sum_{n=0}^N \Phi_n \right)^{-1} \Phi_k \tilde{d}_k \\
 &= \tilde{d}_k^\top \Phi_k \left\{ I - \left( \sum_{n=0}^N \Phi_n \right)^{-1} \Phi_k \right\} \tilde{d}_k \\
 &= \tilde{d}_k^\top \Phi_k \left\{ I - \left( I + \Phi_k^{-1} \sum_{n=0}^{N \setminus k} \Phi_n \right)^{-1} \right\} \tilde{d}_k \\
 &= \tilde{d}_k^\top \Phi_k \left\{ I + \left( \sum_{n=0}^{N \setminus k} \Phi_n \right)^{-1} \Phi_k \right\}^{-1} \tilde{d}_k \\
 &= \tilde{d}_k^\top \left\{ \Phi_k^{-1} + \left( \sum_{n=0}^{N \setminus k} \Phi_n \right)^{-1} \right\}^{-1} \tilde{d}_k,
 \end{aligned} \tag{11}$$

where  $\sum_{n=0}^{N \setminus k} \Phi_n := \sum_{n=0}^N \Phi_n - \Phi_k$ . In addition, considering the coefficient of  $s_l (l \neq k)$  as  $U_{k,l} \in \mathbb{R}^{2 \times 2}$ , we obtain

$$U_{k,l} = -\tilde{d}_k^\top \Phi_k \left( \sum_{n=0}^N \Phi_n \right)^{-1} \Phi_l \tilde{d}_l. \tag{12}$$

The constant  $b_k \in \mathbb{R}^2$  is expressed as follows:

$$b_k = \tilde{d}_k^\top \Phi_k \left( \sum_{n=0}^N \Phi_n \right)^{-1} \Phi_0 d_0. \tag{13}$$

Therefore, when the conditions in eq. (10) are coupled for  $k = 1, 2, \dots, N$ , we obtain

$$\begin{bmatrix} D_1 & U_{1,2} & \cdots & U_{1,N} \\ U_{2,1} & D_2 & \cdots & U_{2,N} \\ \vdots & \vdots & \ddots & \vdots \\ U_{N,1} & U_{N,2} & \cdots & D_N \end{bmatrix} \begin{bmatrix} s_1 \\ s_2 \\ \vdots \\ s_N \end{bmatrix} = \begin{bmatrix} b_1 \\ b_2 \\ \vdots \\ b_N \end{bmatrix}. \tag{14}$$

We can then solve for  $s_n (n = 1, 2, \dots, N)$  as follows.

$$\begin{bmatrix} s_1 \\ s_2 \\ \vdots \\ s_N \end{bmatrix} = \begin{bmatrix} D_1 & U_{1,2} & \cdots & U_{1,N} \\ U_{2,1} & D_2 & \cdots & U_{2,N} \\ \vdots & \vdots & \ddots & \vdots \\ U_{N,1} & U_{N,2} & \cdots & D_N \end{bmatrix}^{-1} \begin{bmatrix} b_1 \\ b_2 \\ \vdots \\ b_N \end{bmatrix}. \tag{15}$$

From the above results, we can determine  $x$  that minimizes equation eq. (4) by first calculating  $s_n (n = 1, 2, \dots, N)$  using eq. (15) and then substituting the value into eq. (6).

### A.3 Weight Setting for Depth Integration

In this study, we set the weight matrix  $\Phi_n (n = 0, 1, \dots, N)$  to a diagonal matrix. By making it a diagonal matrix, the large matrix calculation in eqs. (11) to (13) can be avoided and can be attributed to element-by-element calculations. The diagonal components represent the reflected intensity at each location on each depth map. Since the weight matrices  $\Phi_n (n = 1, 2, \dots, N)$  are for depth maps that express the estimated depth for  $N$  tangent images in ERP format, the weights are increased for regions where tangent images are present, as shown in fig. 12 To smooth the boundary, we first set the following weights  $w_{ij}$  for pixel position  $(i, j)$  in the tangent image of height  $H_{\tan}$  and width  $W_{\tan}$ .

$$w_{ij} = \left\{ 1 - \left( \frac{2i}{H_{\tan}} - 1 \right)^2 \right\} \left\{ 1 - \left( \frac{2j}{W_{\tan}} - 1 \right)^2 \right\}. \tag{16}$$

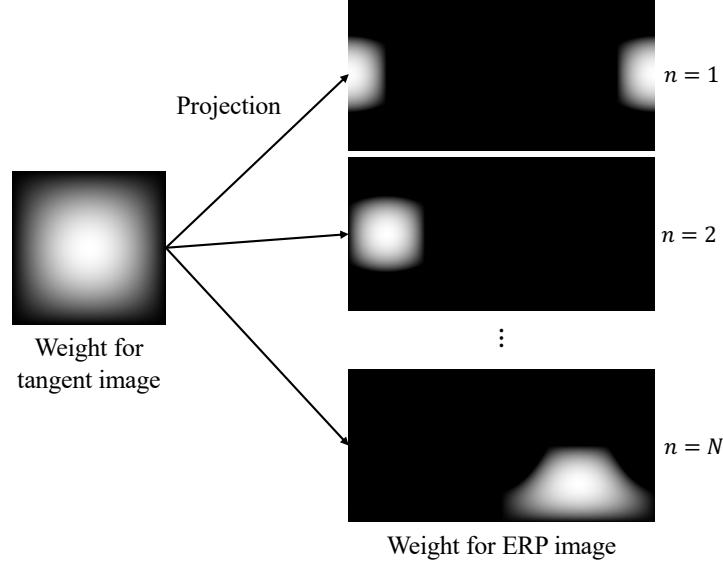


Figure 12: Weights for estimated depth maps. The weights are set such that the center of the tangent image is 1, the edges of the image are 0, and the weights are converted to ERP format for each depth map ( $n = 1, 2, \dots, N$ ).

This weight has a maximum value of 1 at the center of the tangent image and a minimum value of 0 at the edges of the image. The weights for the tangent image are converted to ERP format and set to the diagonal components of the weight matrix  $\Phi_n(n = 1, 2, \dots, N)$ . The weights of the outer regions of each tangential image are set to zero. Tangent images are created with a horizontal field of view of 90 degrees and resolution of  $512 \times 512$  pixels, and 16 images were created with the following latitude  $\theta_n$  and longitude  $\phi_n$  shooting directions.

$$\theta_n = \begin{cases} \frac{\pi}{4} & (1 \leq n \leq 4) \\ -\frac{\pi}{4} & (5 \leq n \leq 8) \\ 0 & (9 \leq n \leq 16) \end{cases} \quad (17)$$

$$\phi_n = \begin{cases} \frac{\pi n}{2} & (1 \leq n \leq 8) \\ \frac{\pi n}{4} & (9 \leq n \leq 16) \end{cases} \quad (18)$$

On the other hand, the weights for the coarse depth  $\Phi_0$  are set as follows. When using floor plans for the layout format, a low-weight  $\eta_L$  is set for areas in the partial image or layout condition where an object is specified, and a high-weight  $\eta_H (\geq \eta_L)$  for other areas. In this study, we set  $\eta_L = 0.0$ ,  $\eta_H = 2.0$ . When using the terrain map for the layout format, set the diagonal component of the weight matrix  $\Phi_0(i, j)$  according to the value of the coarse depth at each location  $(i, j)$  in the ERP as follows:

$$\Phi_0(i, j) = \frac{\alpha}{d_0(i, j)^2 + \epsilon}, \quad (19)$$

where  $\alpha$  and  $\epsilon$  are hyperparameters. In this study, the coarse depth is normalized to the interval  $[0, 1]$ , and we set  $\alpha = 1.0 \times 10^{-3}$  and  $\epsilon = 1.0 \times 10^{-8}$ . We set  $\Phi_0(i, j) = 0$  in the region where the coarse depth is infinite. The weights are inversely proportional to the square of the coarse depth to ensure that the squared error in eq. (4) assumes values of the same scale with respect to the coarse depth. This prevents the error from being overestimated when an object is generated in the foreground of a large-depth region, such as a tree in the foreground of the sky.

## B Additional Results

### B.1 360-Degree RGB Generation

figs. 13 and 14 show additional samples of 360-degree RGB image generation for the Structured 3D dataset and SceneDreamer dataset, respectively.

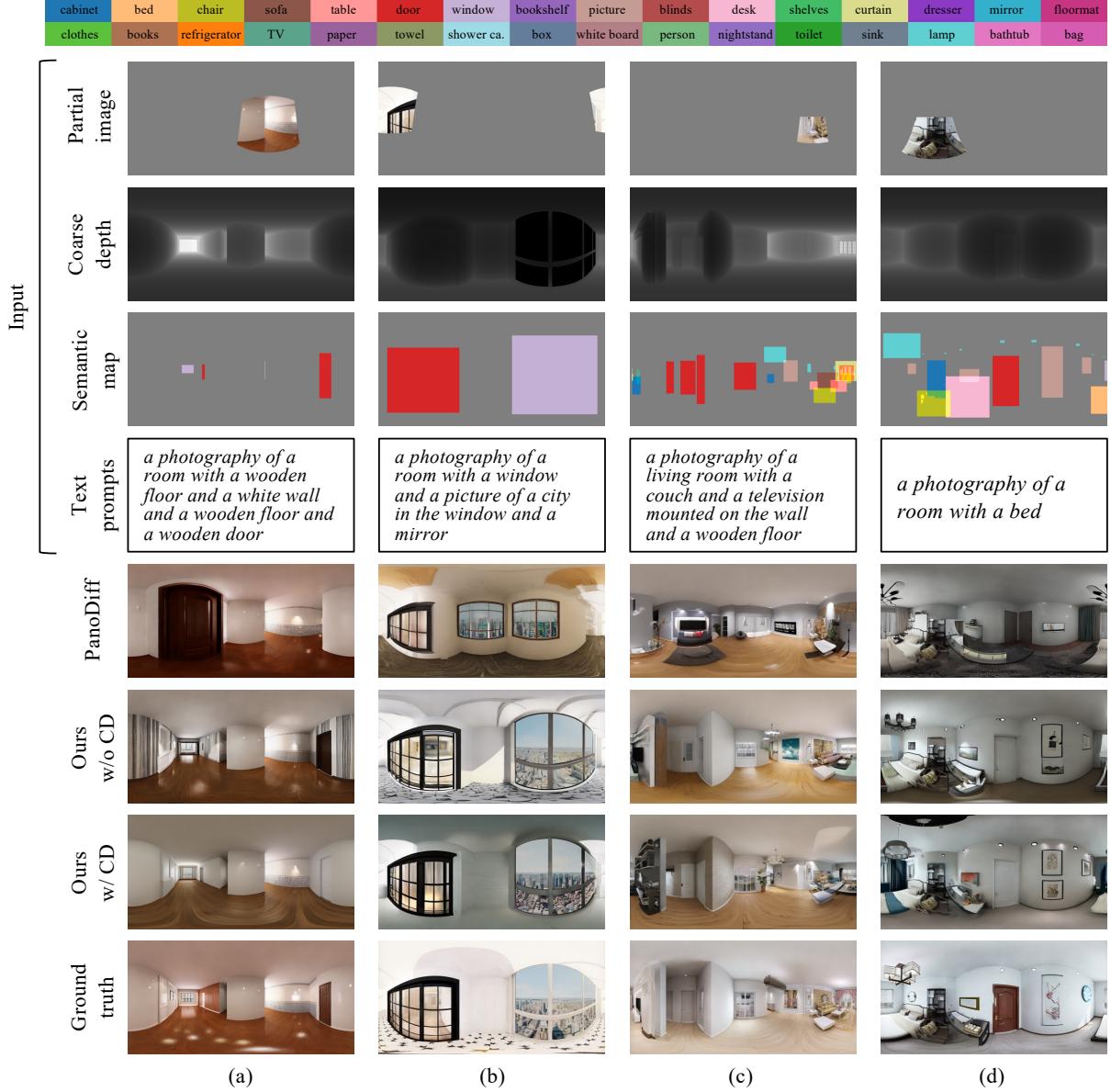


Figure 13: The results of generating a 3D scene for the test set of the Stuructured 3D dataset.

## B.2 Generation Results from Subset of Conditions

To verify the contribution and robustness of each condition of the proposed method, experiments were conducted to generate 360-degree RGB-D from a subset of partial images, layouts, and text prompts. Generation was performed using the proposed method with the CD for the test set of the structured 3D dataset. Because depth estimation in the proposed method requires layout information, LeRes (ERP) [70], a monocular depth estimation of ERP images, was used in the absence of layout conditions. table 4 shows the values of each evaluation metric for the generated results. In terms of FID, it can be seen that the proposed method does not significantly degrade performance when text conditions are included in the generation conditions. This is largely owing to the performance of the text-to-image model used as the base model to ensure the plausibility of the generated image. However, PSNR (whole) decreases in the absence of partial image and layout conditions, indicating that the contribution of these conditions to the composition of the overall structure is high. In addition, CS naturally decreases without the text condition. However, even without the text condition, CS is larger than that in the unconditional generation case, indicating that semantic reproduction is possible to some extent, even from partial images and layout information. For depth generation, the accuracy is significantly

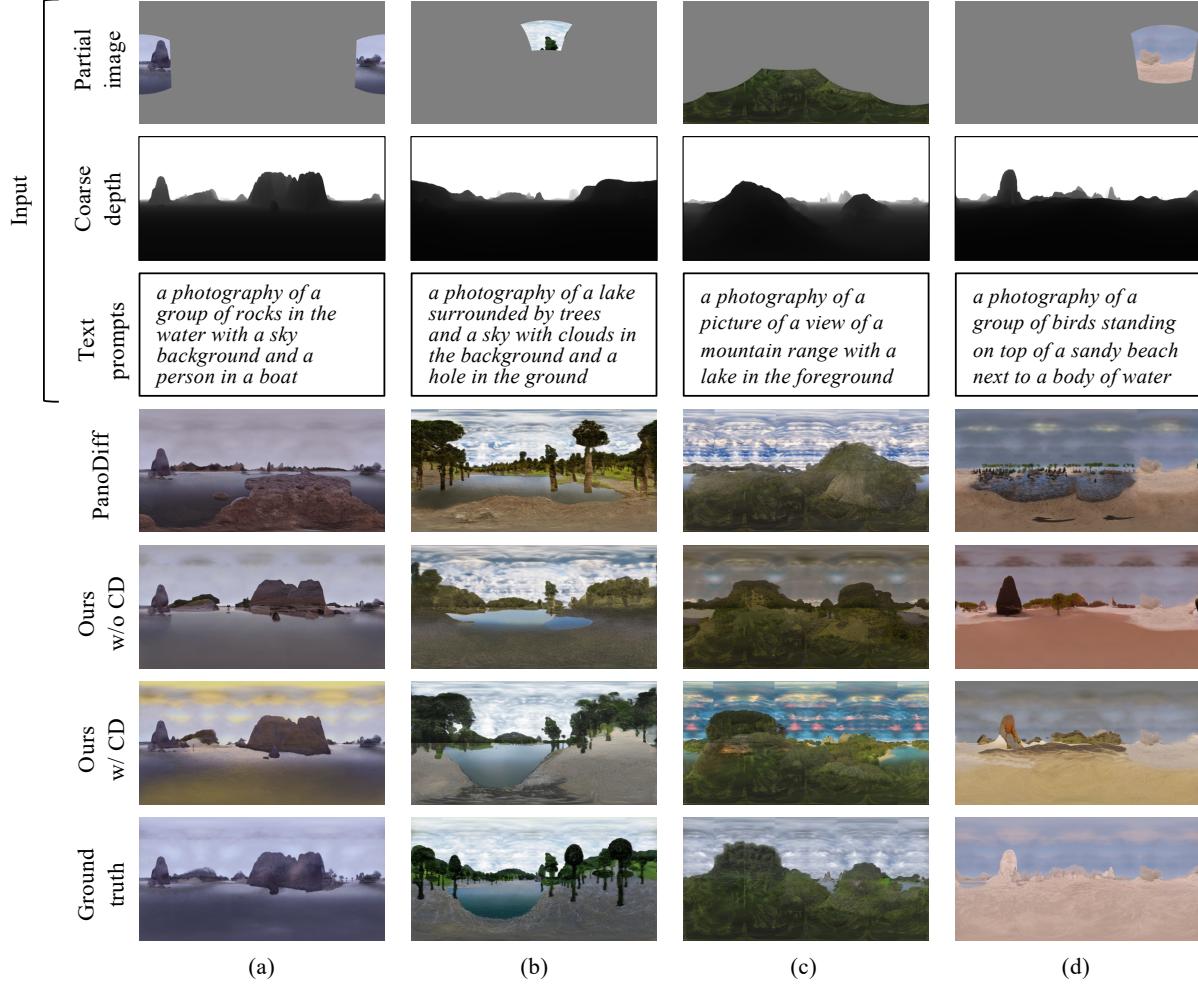


Figure 14: The results of generating a 3D scene for the test set of the SceneDreamer dataset.

degraded because it is impossible to use depth estimation with a coarse depth in the absence of layout conditions. When generated from partial images and text, its performance was comparable to PanoDiff.

### B.3 Results in the Wild

We evaluated the results of the 3D scene generation based on user-generated conditions outside the dataset used for fine-tuning. In this experiment, the end-to-end method was used to estimate the depth in indoor scenes, whereas the depth integration method was applied to outdoor scenes because the SceneDreamer dataset is limited to natural scenery, such as mountainous areas and seashores, using monocular depth estimation models trained on an external dataset. Because CD is effective for fine-tuning with additional text annotations, we used a simpler method without CD in the in-the-wild experiments described in this section. The terrain map  $T \in \mathbb{R}^{H_{\text{ter}} \times W_{\text{ter}}}$  was created as a mixed Gaussian distribution in the following equation:

$$T_p = \sum_{k=1}^K \pi_k \exp \left( -\frac{1}{2} (p - \mu_k)^\top \Sigma_k^{-1} (p - \mu_k) \right), \quad (20)$$

where  $p \in \{1, 2, \dots, H_{\text{ter}}\} \times \{1, 2, \dots, W_{\text{ter}}\}$  is the location on the 2-D map,  $K$  is the number of mixtures, and  $\pi_k \in \mathbb{R}$ ,  $\mu_k \in \mathbb{R}^2$ , and  $\Sigma_k \in \mathbb{R}^{2 \times 2}$  are the parameters of the weights, mean, and covariance matrix of the element distribution, respectively.

Additional examples of 3D scenes generated using the proposed method conditioned on text, partial images, and layouts are presented in figs. 15 to 19. In these figures, the aspect ratios of the ERP images were converted to 2:1 for display

Table 4: Evaluation results for generation from subset of conditions.

Conditions			RGB				Depth	
Partial image	Layout	Text	PSNR↑ (whole)	PSNR↑ (partial)	FID↓	CS↑	RMSE ↓	AbsRel ↓
✓	✓	✓	12.42	33.29	18.84	30.71	5.05	0.0076
✓	✓		12.04	34.46	43.86	28.19	8.96	0.0100
✓	✓	✓	11.45	-	21.71	30.67	8.78	0.0056
✓	✓	✓	11.48	33.64	21.83	30.93	24.56	0.0172
✓			11.40	35.00	55.08	27.00	23.94	0.0158
	✓		11.12	-	59.70	27.59	5.02	0.0086
		✓	10.67	-	25.90	30.85	24.53	0.0171
			10.43	-	87.69	24.40	24.00	0.0180

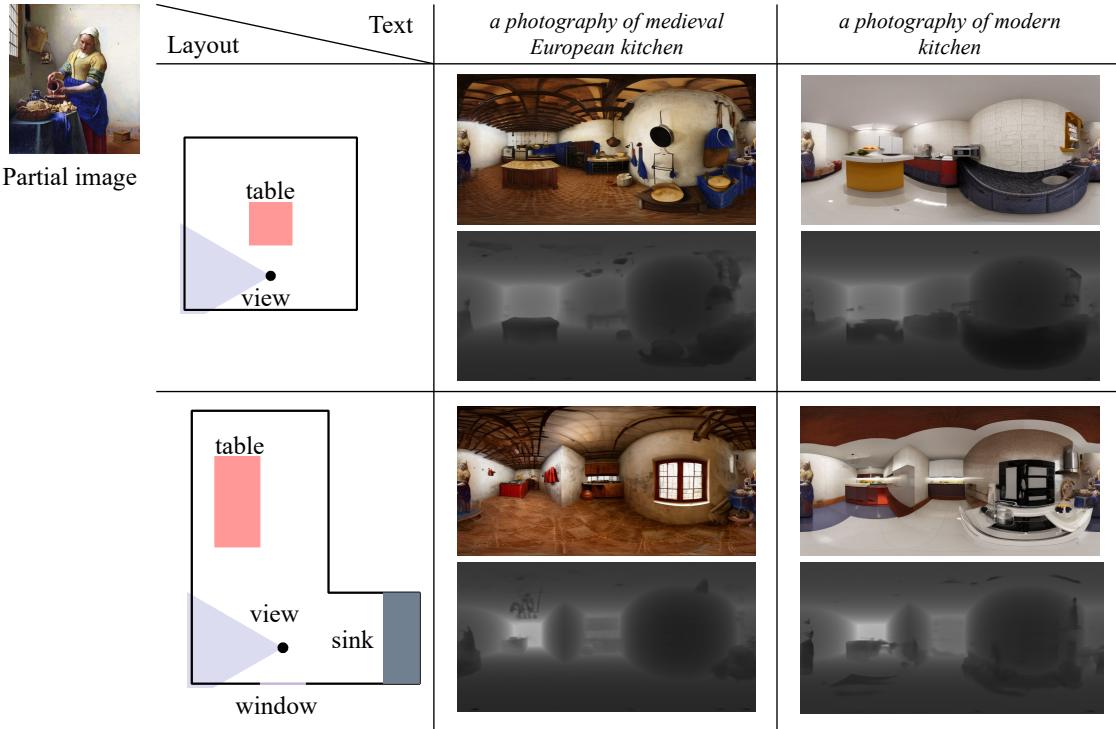


Figure 15: From a given partial image, layout, and text prompt, our method generates the 360-degree RGB space and depth. We used a painting titled "The Milkmaid" by Johannes Vermeer as a partial image. Various 3D scenes can be generated for the same partial image using different layouts and text prompts.

purposes. These conditions were created freely by the authors. It can be seen that the generated scene contains the given partial image and conforms to the instructions of the text prompt according to the given layout. In addition to the coarse depth created by the room shape or terrain alone, the geometry of objects such as chairs, tables, trees, and buildings can be seen. fig. 15 shows how various scenes can be generated in a controlled manner by changing the combination of layout and text for the same partial image. figs. 16 to 19 shows that our method can generate a variety of 3D scenes from photos on the web, photos taken in the real world, and fanciful paintings, taking into account the layout and text requirements we give. These results show that the proposed method can generate 360-degree RGB-D images with appearance, geometry, and overall context controlled according to the input information, even outside the dataset used for fine-tuning.

## C Discussion

### C.1 Advantages of Using 360-Degree Images

The proposed method uses a trained text-to-image model to generate a 2D image, from which the depth is generated. The proposed method is unique because it uses a 360-degree image as the 2D image for generation. Using 360-degree images is advantageous over perspective projection images in terms of scene consistency and reduced computational costs. fig. 20 shows examples of the generated scene from a partial image by the incremental multi-view inpainting and MVDiffusion [56]. Incremental multi-view inpainting is a method of repeating SD inpainting by projecting an input image from a different viewpoint. In the example shown in this figure, the road disappears, indicating that the scene is inconsistent. This is due to the fact that inpainting is performed on each perspective projection image; therefore, the overall consistency cannot be guaranteed. In addition, inpainting must be applied repeatedly, which is computationally expensive and difficult to parallelize. MVDiffusion, on the other hand, takes cross-attention among multiple views and generates multiple views that are simultaneously consistent using SD. This method is computationally expensive because it requires running SD for each view and paying cross-attention to the combinations of multiple views. The order of computational complexity is  $O(N^2)$ , where  $N$  is the number of viewpoints. Because the proposed method generates a single 360-degree image, it is easy to achieve scene consistency at a low computational cost. However, the resolution of the generated image using ERP is lower than that of multiview images, and a higher resolution is a future challenge.

### C.2 Limitation

Although the performance of the proposed method was promising, it had several limitations.

fig. 21 shows examples of problems in RGB generation. First, if the objects specified in the layout are in overlapping positions from a viewpoint, they cannot be separated and drawn in the correct number and position. This is because the 2D layout information is converted to ERP for input, which requires additional ingenuity, such as generating a 3D scene jointly from multiple viewpoints. Second, when using conditions outside the dataset, the specified conditions may not be reflected, depending on the interaction between each condition. For example, there is the phenomenon that certain text prompts do not produce certain objects. Third, it is not possible to specify the regions where objects do not exist. Except for the regions where objects are specified, object generation is controlled by other conditions such as partial image, depth, and text.

fig. 22 shows examples of problems in 6 DoF 3D scene generation. It is difficult to synthesize plausible views when generating 3D scenes from 360-degree RGB-D images with large missing regions that exceed image completion capabilities.

We hope that these limitations will be addressed in future studies.

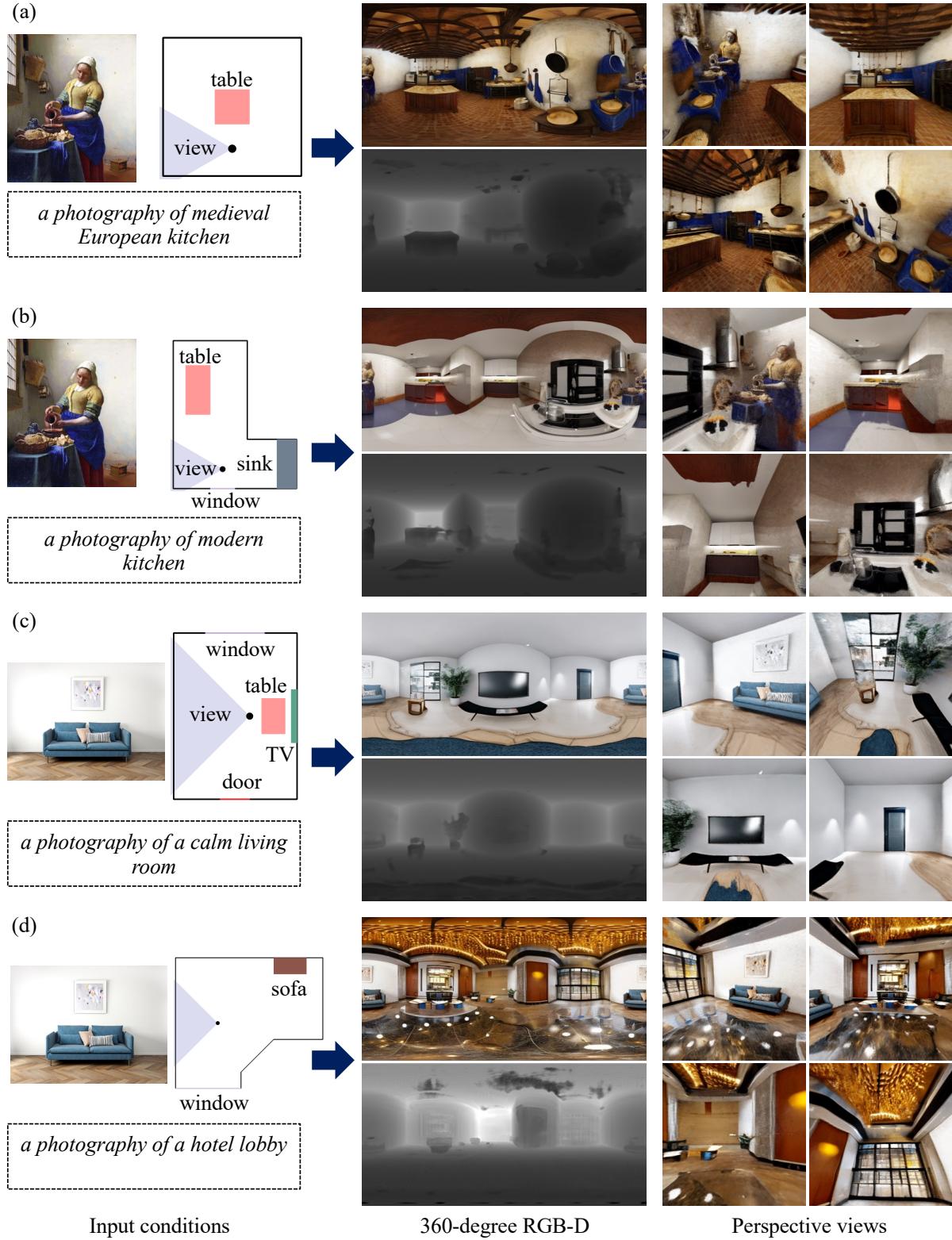


Figure 16: The various generated indoor 3D scenes represented by 360-degree RGB-D images and free perspective images rendered using NeRF owing to conditions outside the used dataset. (a) (b) We used a painting titled "The Milkmaid" by Johannes Vermeer as a partial image. (c) (d) A photo of sofas downloaded from the web (<https://www.photo-ac.com/>) was provided as a partial image.

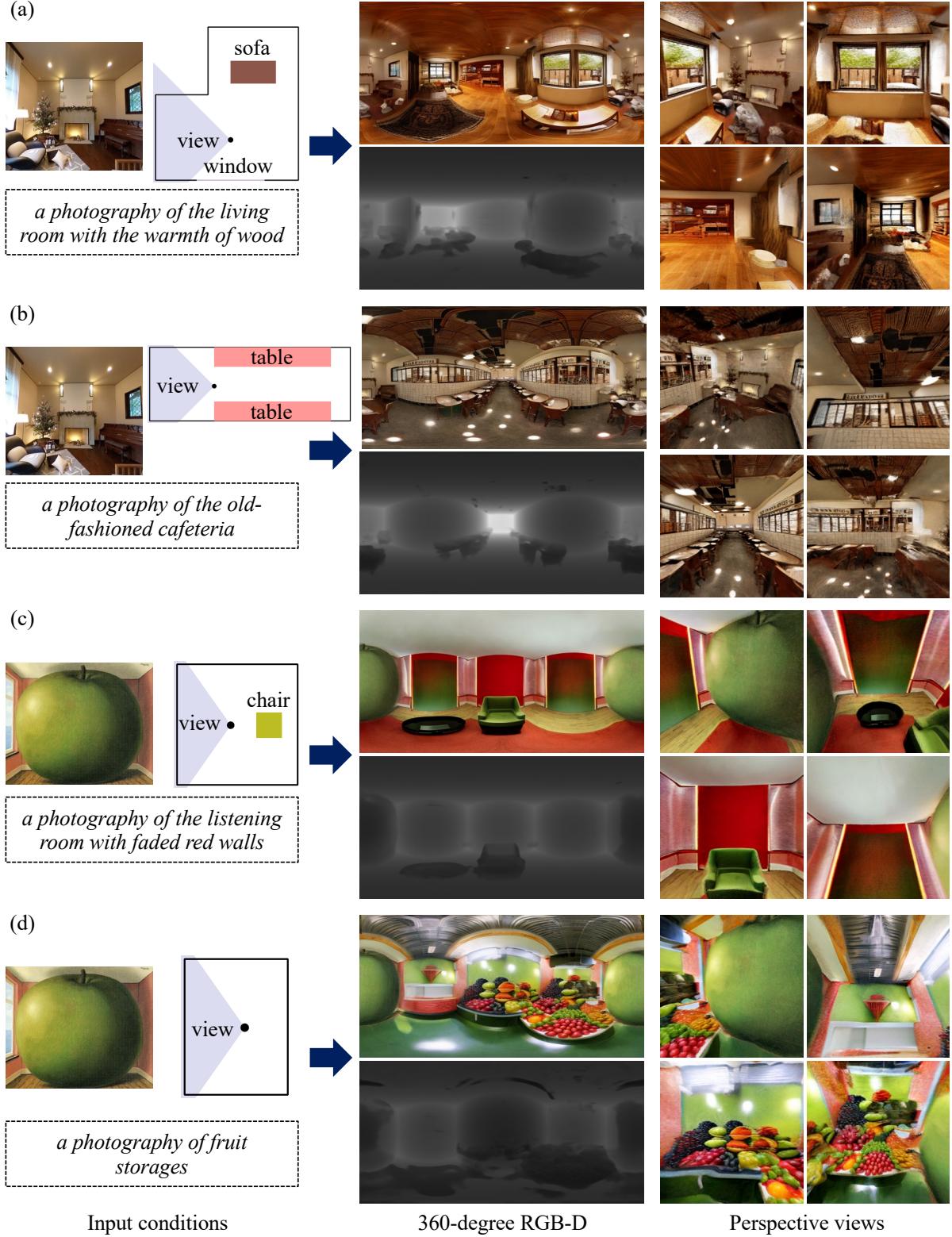


Figure 17: The various generated indoor 3D scenes represented by 360-degree RGB-D images and free perspective images rendered using NeRF owing to conditions outside the used dataset. (a) (b) An image captured by the author using a camera is shown as a partial image. (e) (f) We presented a painting titled "The Listening Room" by René Magritte as a partial image.

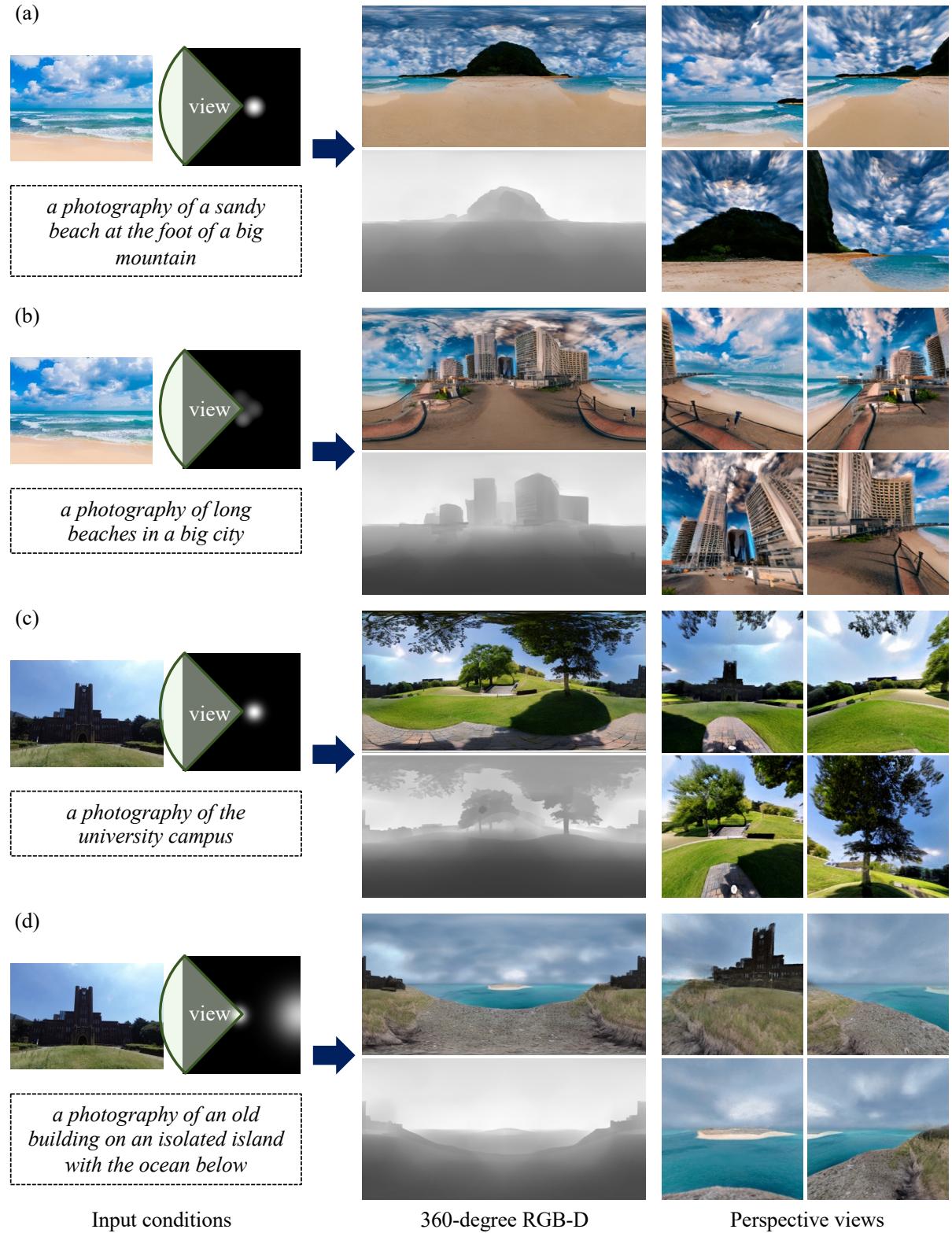


Figure 18: The various generated outdoor 3D scenes represented by 360-degree RGB-D images and free perspective images rendered using NeRF owing to conditions outside the used dataset. (a) (b) A photo of a sandy beach downloaded from the web (<https://www.photo-ac.com/>) was given as a partial image. (c) (d) An image captured by the author using a camera is shown as a partial image.

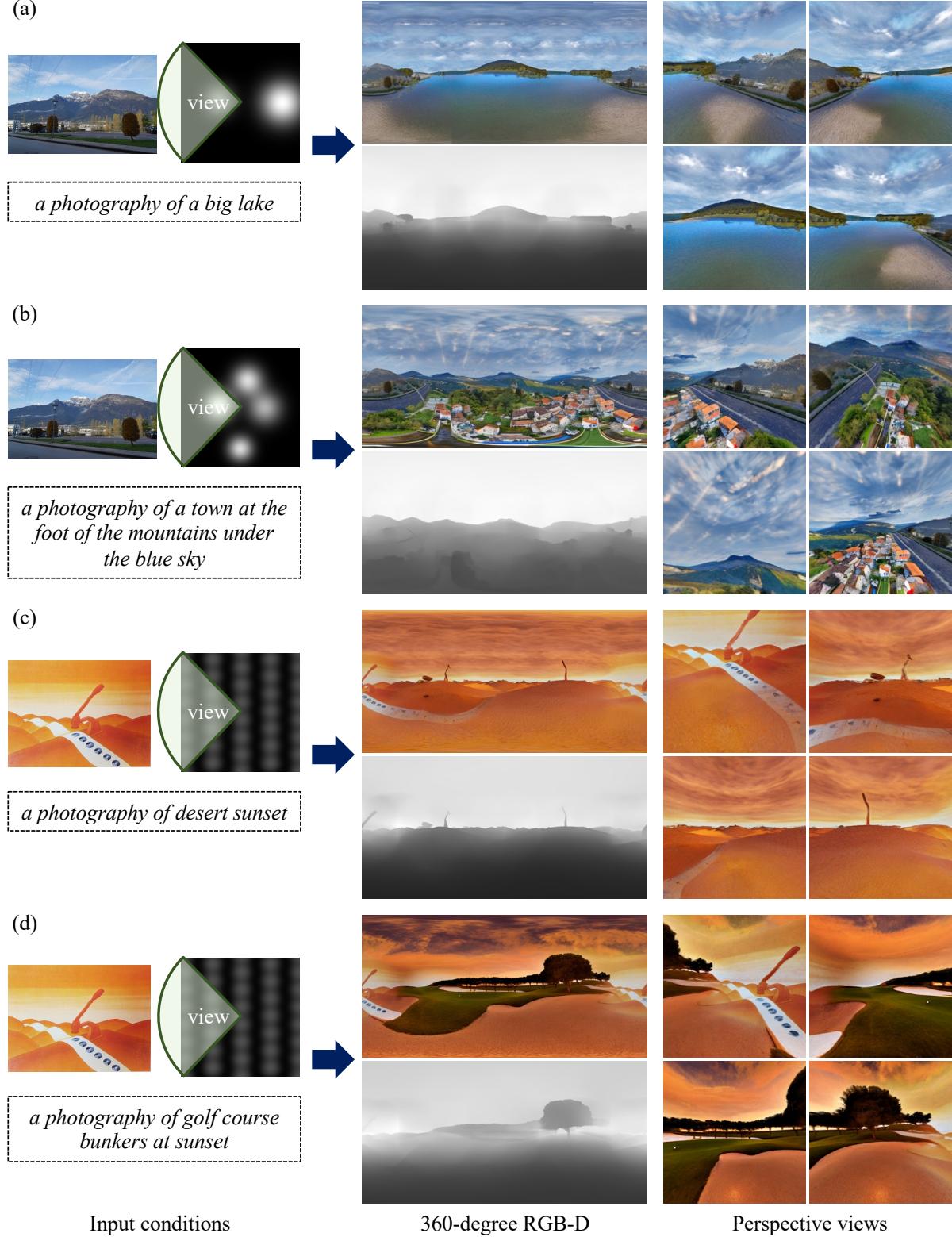
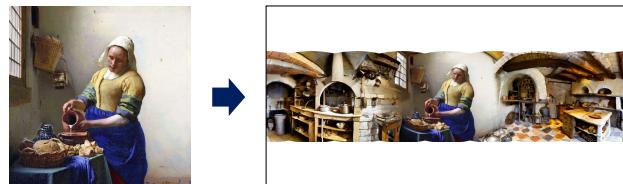


Figure 19: The various generated outdoor 3D scenes represented by 360-degree RGB-D images and free perspective images rendered using NeRF owing to conditions outside the used dataset. (a) (b) An image captured by the author using a camera is shown as a partial image. (c) and (d) We provided a painting titled "Day after Day" by Jean-Michel Folon as a partial image.



(a) Incremental multi-view inpainting



(b) MVDiffusion (Cross attention between multi-views)

Figure 20: Examples of the scene generation from a partial image through the generation of perspective projection images. The generated scenes were displayed in ERP format. (a) In incremental multiview inpainting of the perspective image downloaded from the web ([https://unsplash.com/@overture\\_creations/](https://unsplash.com/@overture_creations/)), the road disappears on the other side, indicating that the scene is not consistent. (b) MVDiffusion maintains consistency between multiple views; however, the computational cost is high because cross attention is required for each combination of multiple views.

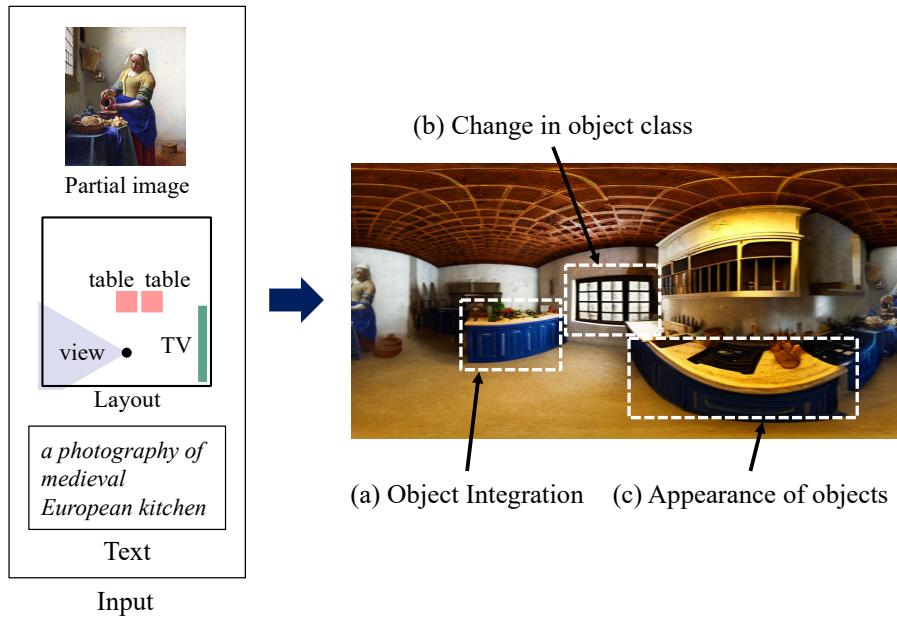


Figure 21: Examples of limitations of 360-degree RGB-D generation from multimodal conditions. (a) When two tables specified in the layout condition overlap in the ERP, they are merged and generated as a single table. (b) Although the layout conditions dictate the placement of a television, it is generated and converted to a window because it does not conform to the context of “a medieval European kitchen,” which is presented in the text prompt. (c) Where nothing is specified in the layout conditions, objects may be generated automatically according to text prompts. It is impossible to specify areas where no objects exist.

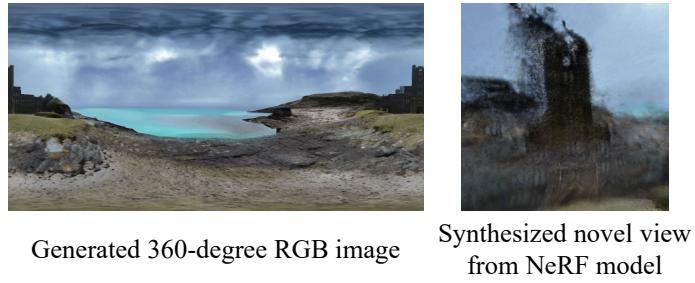


Figure 22: Examples of limitations of synthesized novel views from the NeRF model trained on the generated 360-degree image. It is difficult to synthesize plausible views when generating 3D scenes from 360-degree RGB-D images with large missing regions that exceed image completion capabilities. In this example, the image quality is significantly reduced in the occluded region at the back of the building.