

SphereDiffusion: Spherical Geometry-Aware Distortion Resilient Diffusion Model

Tao Wu^{1*}, Xuewei Li^{1*}, Zhongang Qi^{2†}, Di Hu³, Xintao Wang², Ying Shan², Xi Li^{1,4†}

¹College of Computer Science and Technology, Zhejiang University

²ARC Lab, Tencent PCG

³Gaoling School of Artificial Intelligence, Renmin University of China

⁴Zhejiang – Singapore Innovation and AI Joint Research Lab, Hangzhou

taowucs@zju.edu.cn, xueweili@zju.edu.cn, zhongangqi@tencent.com, dihu@ruc.edu.cn, xintaowang@tencent.com, yingsshan@tencent.com, xilizju@zju.edu.cn

Abstract

Controllable spherical panoramic image generation holds substantial applicative potential across a variety of domains. However, it remains a challenging task due to the inherent spherical distortion and geometry characteristics, resulting in low-quality content generation. In this paper, we introduce a novel framework of SphereDiffusion to address these unique challenges, for better generating high-quality and precisely controllable spherical panoramic images. For the spherical distortion characteristic, we embed the semantics of the distorted object with text encoding, then explicitly construct the relationship with text-object correspondence to better use the pre-trained knowledge of the planar images. Meanwhile, we employ a deformable technique to mitigate the semantic deviation in latent space caused by spherical distortion. For the spherical geometry characteristic, in virtue of spherical rotation invariance, we improve the data diversity and optimization objectives in the training process, enabling the model to better learn the spherical geometry characteristic. Furthermore, we enhance the denoising process of the diffusion model, enabling it to effectively use the learned geometric characteristic to ensure the boundary continuity of the generated images. With these specific techniques, experiments on Structured3D dataset show that SphereDiffusion significantly improves the quality of controllable spherical image generation and relatively reduces around 35% FID on average.

Introduction

Spherical panoramic images, also known as 360° panoramic images or omnidirectional panoramic images, are used in various domains such as autonomous driving (de La Garanderie, Abarghouei, and Breckon 2018; Ma et al. 2021), virtual reality (Xu, Zhang, and Gao 2021; Ai et al. 2022), etc. Numerous studies (Yan et al. 2022; Hara, Mukuta, and Harada 2021; Akimoto, Matsuo, and Aoki 2022) have been proposed for the synthesis of spherical panoramic images, with a primary focus on reconstructing scenes from narrow field of view (NFOV) images. However, these generation methods often produce images of inferior quality and lack controllability, which are crucial in real applications.

*These authors contributed equally.

†Corresponding author.

Copyright © 2024, Association for the Advancement of Artificial Intelligence (www.aaai.org). All rights reserved.

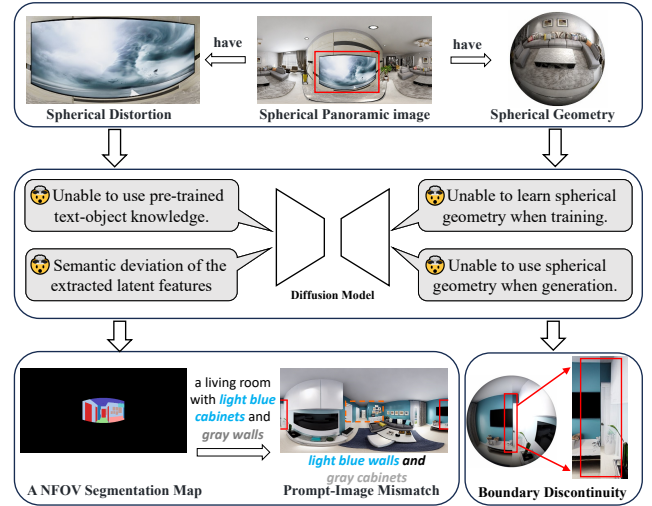


Figure 1: The characteristics of spherical panoramic images and the impact of these characteristics on existing controllable generation methods.

In contrast, extensive works (Zhang, Rao, and Agrawala 2023; Mou et al. 2023) have greatly succeeded in controllable high-quality planar image generation. Most of the existing works resort to fine-tuning the pre-trained large-scale diffusion models to adapt to different application scenarios. However, such a paradigm falls short of expectations for generating the spherical panoramic images, since simply fine-tuning cannot capture the unique characteristics of the spherical panoramic images.

Two characteristics of spherical panoramic images are essentially different from the planar images: **spherical distortion** and **spherical geometry**. As shown in Figure 1, on the one hand, spherical distortion mainly refers to the deformation of objects. Existing controllable generation models are primarily designed and pre-trained based on planar images. Thus, the text-object correspondence knowledge stored in these pre-trained weights cannot be effectively utilized for spherical panoramic images due to the significant deformation of distorted objects. At the same time, spherical distortion makes it difficult to extract effective features of spherical panoramic images, resulting in semantic deviation when

depicting image content. Accordingly, text prompts often fail to correctly guide the generation process, leading to the mismatch between text guidance and generated visual content. On the other hand, spherical geometry means that the visual content of a spherical panoramic image is a projection of the 3D world on a sphere. Such sphere structure shares 3D geometric attributes such as spherical rotation invariance and non-boundary property. Current controllable planar generative models lack geometry-aware training design, resulting in the difficulty of effectively incorporating the spherical geometry characteristic. As a consequence, improving the quality of the generated content using spherical geometry becomes a challenge. Practically, preventing the model from generating spherical images from a global perspective leads to issues like boundary discontinuity. Considering the above issues, one question naturally arises: How can we enable the model to learn and utilize the characteristics of spherical images, then enhance the quality of controllable spherical panoramic image generation?

In this work, we propose the SphereDiffusion framework, which targets to generate high-quality and precisely controllable spherical panoramic images from single NFOV segmentation maps and text prompts. To solve the above issues, we impose the two characteristics of spherical panoramic images into the model design, as well as the training and inference process. Concretely, for spherical distortion characteristic, we introduce Distortion-Resilient Semantic Encoding (DRSE) to enhance the utilization of pre-trained knowledge. It embeds the text semantics into distorted objects, aligning text-object correspondence knowledge of pre-trained planar image generation models and the objects in the spherical panoramic image. Meanwhile, we also introduce a Deformable Distortion-aware Block (DDaB) constructed based on deformable convolution to relieve semantic deviation. The deformability of DDaB helps the model extract effective features from distorted objects with different deviations in spherical panoramic images.

For spherical geometry characteristic, we aim to let the model adequately learn and use it, then improve both the training process and generation process. On the one hand, we propose Spherical Geometry-aware (SGA) Training, enabling the model to better learn the spherical geometry characteristic. It contains two modules: Spherical Reprojection and Spherical SimSiam Contrastive Learning. Spherical Reprojection applies spherical rotation invariance to the training data, enabling the model to better learn spherical geometry through data diversity. Spherical SimSiam Contrastive Learning ensures spherical rotation invariance in the latent space, increasing the spherical robustness of models at the optimization objective. On the other hand, we introduce SGA Generation, which allows the model to better use the spherical geometry characteristic to improve the generation process. By incorporating spherical rotation invariance into the generation process, we iteratively rotate the intermediate results from the previous denoising step to connect the content located at the two ends of the intermediate results. In this way, the boundary connectivity of the generated image is improved. Our contributions are summarized as follows:

- We propose a novel framework for controllable spherical

panoramic image generation, which takes both spherical geometry and image distortion into consideration.

- We propose DRSE and DDaB to deal with spherical distortion, enabling the model to better use the pre-trained knowledge and reduce the semantic deviation in latent space caused by spherical distortion.
- We introduce SGA Training to make models learn spherical geometry from both data diversity and optimization objectives. We also propose SGA Generation to improve the denoising process of the diffusion model.

Experimental results on the Structured3D dataset (Zheng et al. 2020) demonstrate that our method can significantly improve the quality of controllable spherical image generation and relatively reduces around 35% FID on average compared to previous methods.

Related Work

Conditional Diffusion Probabilistic Model

Diffusion model (Sohl-Dickstein et al. 2015; Dhariwal and Nichol 2021) is a generative probability model, which has attracted many researchers' attention because of its high-quality generative results. Diffusion models can successfully perform conditional image generation when trained with guidance such as semantic layout or class labels (Zheng et al. 2022; Ramesh et al. 2021; Saharia et al. 2022b; Ho and Salimans 2022; Zheng et al. 2023; Xue et al. 2023). A notable example of conditional diffusion models is recent text-to-image diffusion models, which have showcased unprecedented synthetic capabilities (Nichol et al. 2021; Saharia et al. 2022a; Sheynin et al. 2022). Recently, many methods have been observed to enhance user controllability. Existing methodologies can be broadly bifurcated into two primary strategies: (i) Approaches that integrate explicit control by incorporating additional guiding signals into the model (Avrahami et al. 2022; Rombach et al. 2022; Brooks, Holynski, and Efros 2023). However, these studies require costly training on meticulously curated datasets. (ii) Many methods have been proposed to implicitly control the content generated by manipulating the generation process of a pre-trained model (Mokady et al. 2023; Kong et al. 2023) or by conducting lightweight model fine-tuning (Ruiz et al. 2023; Kawar et al. 2023; Zhang, Rao, and Agrawala 2023). Most of these methods only require minimal training overhead and produce high-quality generated content, making them the mainstream approach for controllable generation.

Spherical Panoramic Image Generation

Current spherical panoramic image generation techniques can be divided into two categories: GAN-based generative models and diffusion-based generative models. Kimura et al. (Kimura and Rekimoto 2018) presented a peripheral image generation technique based on pix2pix (Isola et al. 2017). However, the FOV used to generate image was constrained. Sumantri et al. (Sumantri and Park 2020) advanced a spherical image generation technique based on pix2pixHD (Wang et al. 2018), which required a collection of images taken from various directions as input. Hara et al. (Hara,

Mukuta, and Harada 2021) present a novel method to generate spherical images from a single NFOV image by controlling the degree of freedom of generated regions using scene symmetry. Along with the development of the diffusion model, some panoramic image generation methods based on the diffusion model have emerged. Bar-Tal et al. (Bar-Tal et al. 2023) define a new generation process to generate panoramas, which is composed of several reference diffusion generation processes bound together with a set of shared parameters or constraints and without further training or fine-tuning. Zhang et al. (Zhang et al. 2023) propose a combinatorial diffusion model that can take advantage of a trained diffusion model based on factor graph representations to generate spherical panoramic images. Li et al. (Li and Bansal 2023) use recursive overpainting on generated images to create consistent spherical panoramic views by conditioning text descriptions.

Preliminaries

Latent Diffusion Models

Diffusion models are probabilistic models designed to learn a data distribution $p(x)$ by gradually denoising a normally distributed variable and can be interpreted as a sequence of denoising autoencoders $\epsilon_\theta(x_t, t)$. They are trained to predict denoised versions of their inputs x_t , where x_t is a noisy variant of the input x . Latent diffusion models (LDMs) (Rombach et al. 2022) employ a two-stage approach to train diffusion models directly in high-resolution pixel space with acceptable computational cost. First, a learnable autoencoder (consisting of an encoder $\mathbf{E}$ and a decoder $\mathbf{D}$) is trained to compress the image into a smaller latent space representation. Then, a diffusion model of representations $z = \mathbf{E}(x)$ is trained instead of a diffusion model of images x . Moreover, in the forward process, LDM incrementally adds noise to z to get z_t and performs denoising to predict z in the reverse process. New images can be generated by sampling a representation $\tilde{z}$ from the diffusion model and subsequently decoding it into an image using the learned decoder $\tilde{x} = \mathbf{D}(\tilde{z})$. During training, the loss is defined as follows:

$$L_{LDM} = \mathbb{E}_{z_0, \epsilon \sim \mathcal{N}(0,1), t} \left[\|\epsilon - \epsilon_\theta(z_t, t)\|_2^2 \right]. \quad (1)$$

Controllable Image Synthesis Diffusion Models

Controllable image synthesis diffusion models allow the creation of diverse images based on text instructions or guidance from a reference image. ControlNet, a trainable adaptor, is specifically designed to function in tandem with Stable Diffusion, which is a representative work of this field. A simple network $\mathcal{F}_{hint}$ is first used to downsample the input control image c to the same size as the input vector z in the latent space of Stable Diffusion, yielding C_{latent} . Subsequently, Controlnet uses its control branch $\mathcal{F}_c$ to process C_{latent} , resulting in multi-scale features $F_c = [F_c^1, F_c^2, \dots, F_c^n]$. These features are then added to the features of the same resolution at the corresponding positions in the middle block and the decoder block of the U-Net structure in Stable Diffusion. This process effectively controls the generation of Stable Diffusion. During training, the main training constraint

is defined as follows:

$$L_C = \mathbb{E}_{z_0, t, c_t, c, \epsilon \sim \mathcal{N}(0,1)} \left[\|\epsilon - \epsilon_\theta(z_t, t, c_t, c)\|_2^2 \right]. \quad (2)$$

In this paper, we adopt ControlNet as our baseline.

Method

In this section, we present the fundamental idea and detailed design of SphereDiffusion. First, we provide an overview of controllable spherical panoramic image generation. Second, we describe our solution to spherical distortion. Finally, we introduce our strategy to allow the model to learn and utilize the characteristic of spherical geometry better.

Overview

SphereDiffusion generates high-quality controllable spherical panoramic images x which simultaneously conform to a corresponding text prompt C_{text} and an NFOV segmentation map C_{mask} . The foundational ControlNet serves as the baseline for this process. In order to improve the quality of controllable spherical panoramic image generation, SphereDiffusion needs to deal with two main characteristics, spherical distortion and spherical geometry.

Spherical distortion causes a certain category of objects in different positions in the spherical panoramic image to show significant and different shape changes compared with the planar image. This poses challenges to the model to effectively utilize the text-object correspondence knowledge stored in pre-trained weights and extract effective features of distorted objects. First, to better use the pre-trained knowledge of planar images, we propose our Distortion-Resilient Semantic Encoding (DRSE), to align the input condition to the pre-trained text-object correspondence knowledge. In addition, to deal with the different shape changes of objects at different locations of a spherical panoramic image, we propose our Deformable Distortion-aware Block (DDaB).

Spherical geometry has several unique properties, such as spherical rotation invariance and non-boundary property. To enable the model to learn and use spherical geometry, we introduce SGA Training and SGA Generation during the training and generation processes, respectively. SGA Training enhances data diversity and optimization objectives during the training process by employing Spherical Reprojection and Spherical SimSiam Contrastive Learning, respectively. This approach enables the model to learn the spherical geometry characteristic better. Furthermore, SGA Generation uses learned geometric characteristics to enhance the boundary connectivity of spherical panoramic images to make the generated content continuous.

Spherical Distortion Properties Solution

As mentioned above, the impact of spherical distortion has two main aspects. First, to utilize the text-object correspondence knowledge stored in pre-trained weights, we replace the original RGB segmentation map with a segmentation map rich in semantic information. Moreover, to reduce the semantic deviation in latent space, models should be specially designed to extract effective features of different locations of a spherical panoramic image differently and adapt

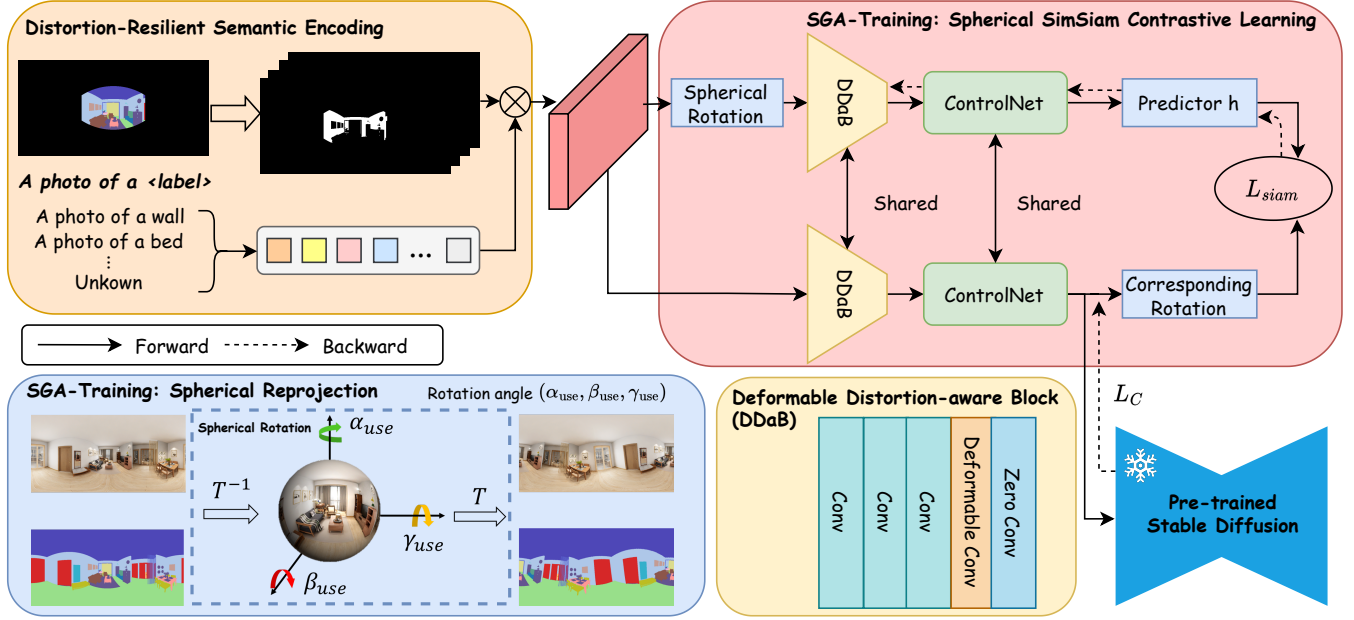


Figure 2: Overall review of SphereDiffusion. (Upper left) Distortion-Resilient Semantic Encoding introduces category information into the representation of segmentation maps to alleviate the issue of text-image mismatch. (Upper right) Spherical SimSiam Contrastive Learning is a part of SGA Training, which constructs contrastive learning in the latent space, equipping SphereDiffusion with spherical geometry at the objective function level. (Lower left) Spherical Reprojection is a part of SGA Training at the data level, and Spherical Rotation serves as the foundation for SGA Training. (Lower middle) DDaB with deformable convolution enhances the model’s perceptual ability of spherical distortion.

tively. Inspired by Trans4PASS (Zhang et al. 2022), we improve the F_{hint} by deformable technique through our Deformable Distortion-aware Block.

Distortion-Resilient Semantic Encoding Distortion-Resilient Semantic Encoding starts from the perspective of input data, upgrading the connection between color information in the segmentation map and the generated objects to the connection between class semantic information in the segmentation map and the generated objects. This allows the model to better utilize the text-object correspondence knowledge stored in pre-trained weights. Specifically, as shown in the upper left of Figure 2, given one NFOV segmentation map C_{mask} , we first set the segmentation maps for the remaining positions in a newly introduced category, "Unknown". This results in our final input segmentation map, C'_{mask} . Then, we downsample the segmentation map to the same resolution as the input vector $z \in \mathbb{R}^{C \times H \times W}$ in the latent space of Stable Diffusion. According to the categories of the labels, we divide the entire image into K two-dimensional binary masks $M = \{m_i \mid m_i \in [0, 1]^{H \times W}\}_{i=1}^K$, where K represents the total number of categories, including the newly added 'Unknown' category. Subsequently, we construct the label texts using the prompt template 'a photo of a {label}' for all categories $\mathcal{L} = \{l_1, l_2, \dots, l_K\}$. These label texts are then encoded using the text encoder of CLIP (Radford et al. 2021), resulting in label embeddings $\mathcal{E}_{label} \in \mathbb{R}^{C_{\mathcal{E}} \times K}$. Finally, we multiply the binary masks M with the label

embeddings $\mathcal{E}_{label}$, resulting in a per-pixel embedding $\mathcal{E}_{pixel} \in \mathbb{R}^{C_{\mathcal{E}} \times H \times W}$. We use $\mathcal{E}_{pixel}$ as the guiding input for the final model (F_{CLIP} is the text encoder of the CLIP model, $\otimes$ is the matrix cross product):

$$C'_{mask} \rightarrow M = \{m_i \mid m_i \in [0, 1]^{H \times W}\}_{i=1}^K, \quad (3)$$

$$\mathcal{E}_{label} = F_{CLIP}(\mathcal{L}), \quad (4)$$

$$\mathcal{E}_{pixel} = \mathcal{E}_{label} \otimes M. \quad (5)$$

Deformable Distortion-aware Block The Deformable Distortion-aware Block starts from the perspective of the model structure, introducing a deformable convolution into the model. This allows the model to better adapt and extract effective features from spherical images. After obtaining $\mathcal{E}_{pixel}$, to align with the original design of ControlNet, we use a learnable block to transform $\mathcal{E}_{pixel}$ into an embedding with the same dimensions as z . To deal with the different shape changes of objects at different locations of a spherical panoramic image, we introduce deformable convolution within this block. In detail, for each image, the offsets $\Delta_{(i,j)}$ of the i^{th} row j^{th} column pixel are defined as:

$$\Delta_{(i,j)} = \begin{bmatrix} \min(\max(-k_D \cdot H, g(f)_{(i,j)}), k_D \cdot H) \\ \min(\max(-k_D \cdot W, g(f)_{(i,j)}), k_D \cdot W) \end{bmatrix}, \quad (6)$$

where $g(\cdot)$ is the offset prediction function. The hyperparameter k_D puts an upper bound on the learnable offsets Δ .

Spherical Geometry-aware Generation

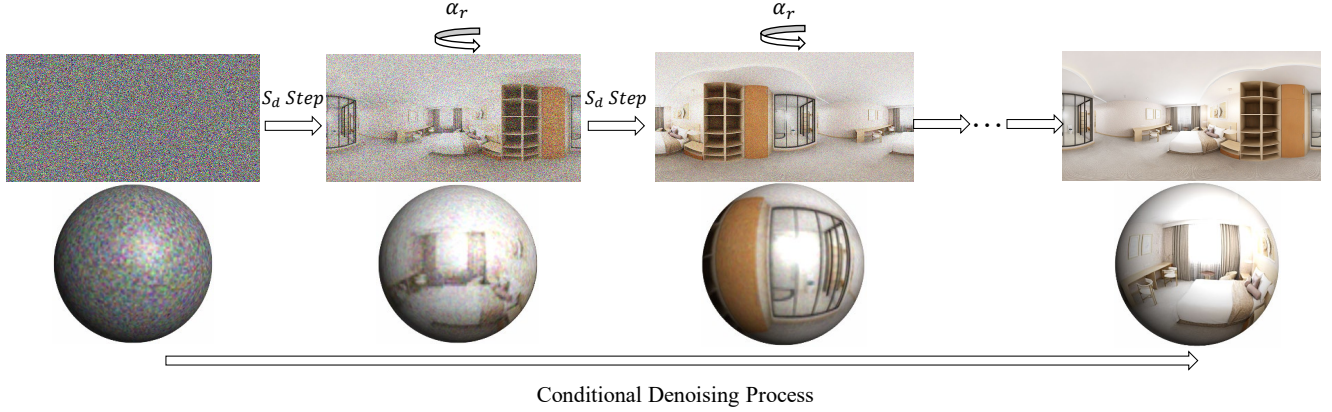


Figure 3: The processing of Spherical Geometry-aware Generation. During the generation process, we uniformly select K steps to rotate an angle α^r to enhance the boundary connectivity of the generated image.

For implementation, as shown in the lower middle of Figure 2, in block $\mathcal{F}_{hint}$, which consists of four convolutional layers and one zero convolution layer, we replace the fourth convolutional layer with a deformable convolution layer.

Spherical Geometry-aware Diffusion Model

We made improvements to both the training and inference process. SGA Training fuses spherical geometry in data diversity and optimization objective. We adopt random rotations in 3D space to enhance data diversity. In terms of optimization objective, we propose Spherical SimSiam Contrastive Learning to make the extracted features equipped with spherical rotation invariance. we introduce SGA Generation, which allows the model to better use the spherical geometry characteristic to improve the generation process.

Spherical Geometry-aware (SGA) Training Traditional training strategies treat the input as a planar image. It results in the model overfitting to images of a single projection way, thereby inhibiting the model’s ability to learn the spherical geometry characteristic. Therefore, we introduce SGA training with the aim of enhancing the ability of the model’s control branch to learn the spherical rotation invariance of spherical images. First, we introduce the spherical rotation. As shown in the lower left of Figure 2, let T denote the forward transformation of the Equirectangular Projection (ERP), which entails the conversion of spherical coordinates to planar coordinates. T^{-1} signify the inverse one. Given an input panoramic image processed through ERP, we initially convert the image I to spherical coordinates by applying the inverse ERP transformation. Subsequently, benefiting from (Li et al. 2023), we employ a three-dimensional rotation matrix within the spherical coordinate system to execute a three-dimensional rotation. For a generic rotation in three-dimensional space, the angles of yaw, pitch, and roll are represented by α_{use} , β_{use} , and γ_{use} , respectively. The associated rotation matrix is denoted by $R(\alpha_{use}, \beta_{use}, \gamma_{use})$. By multiplying R with the data in the spherical coordinate system, we acquire the rotated data within the same coordinate

system. Ultimately, we apply the ERP forward transformation to convert the rotated spherical coordinate system image into a panoramic image, thereby obtaining a specific rotated image of the real input of the model. The corresponding point in the input image of a pixel in the rotated image may not possess integer coordinates; thus, we choose the nearest pixel as its corresponding pixel. In summary, the rotation process of a spherical image I can be defined as follows: $O_{3D}(I, \alpha_{use}, \beta_{use}, \gamma_{use}) = T(R(\alpha_{use}, \beta_{use}, \gamma_{use}) \cdot T^{-1}(I))$. During training, two methods help the model learn the geometric property, including spherical rotation invariance.

Spherical Reprojection: Given a data pair (x, C_{mask}, C_{text}) , we can rotate both x and C'_{mask} by a random rotation angle chosen randomly within the maximum rotation angle $(\alpha_d, \beta_d, \gamma_d)$ to obtain x^r and c'_{mask} , thereby generating more data $(x^r, C'_{mask}, C_{text})$. This method allows the model to learn geometric properties directly.

Spherical SimSiam Contrastive Learning: Benefiting from SimSiam (Chen and He 2021), we use C_{latent} representing the input of the control branch $\mathcal{F}_c$ in ControlNet. As shown in the upper right part of Figure 2, we randomly rotate C_{latent} using a random rotation $(\alpha_{use}, \beta_{use}, \gamma_{use})$ chosen randomly within the maximum rotation angle $(\alpha_c, \beta_c, \gamma_c)$ to obtain a new view $C'_{latent} = O_{3D}(C_{latent}, \alpha_{use}, \beta_{use}, \gamma_{use})$. The encoders $\mathcal{F}_c$ of the two branches share the same weights. A prediction MLP head h transforms the output of one view and matches it to the other view. We maximize the cosine similarity between the two branches as follows:

$$\mathcal{D}(p_1, z_2) = -\frac{p_1}{\|p_1\|_2} \cdot \frac{z_2}{\|z_2\|_2}, \quad (7)$$

where $\|\cdot\|_2$ is ℓ_2 -norm, $p_1 = h(\mathcal{F}_c(C_{latent}))$, $z_2 = O_{3D}(\mathcal{F}_c(C'_{latent}), \alpha_{use}, \beta_{use}, \gamma_{use})$. Then we define a symmetrized loss as follows:

$$L_{siam} = \frac{1}{2}\mathcal{D}(p_1, \text{stop}(z_2)) + \frac{1}{2}\mathcal{D}(p_2, \text{stop}(z_1)) \quad (8)$$

where stop represents the stop-gradient operation and prevents a degenerate solution due to model collapse. We set

our total loss as (λ is a hyperparameter):

$$L_{all} = L_c + \lambda \cdot L_{siam}. \quad (9)$$

Spherical Geometry-aware (SGA) Generation During the generation process with diffusion models, the output is not produced in a single step; rather, it involves multiple iterations. Therefore, inspired by the non-boundary property of the spherical panoramic image, we have also improved the generation process. As shown in Figure 3, assuming that we need N steps $\{t^1, t^2, \dots, t^N\}$ to complete the generation of the diffusion model, we will uniformly select K steps throughout the process $S = \{s^1, s^2, \dots, s^K\}$, $S_d = \frac{N}{K+1}$ steps between each step. When the current iteration step is $t \in S$, we simultaneously rotate the latent space vector and the guided segmentation map at an angle of $\alpha_r = \frac{360^\circ}{K}$. Such an approach can enhance the boundary connectivity of the spherical image during the generation process.

Experiments

Datasets, Protocols, and Evaluation Metrics

We evaluated our model on the Structured3D dataset (Zheng et al. 2020), which provides 196k spherical panoramic images of 21,835 rooms in 3,500 scenes. We use scene_00000 to scene_03249 for training, and scene_03250 to scene_03499 for testing. Our experiments are conducted with a server with eight NVIDIA A100 GPUs, and training epochs are 20. The base model is Stable Diffusion 1.5, and text prompts are annotated with BLIP (Li et al. 2022). Following the settings in (Hara, Mukuta, and Harada 2021), during the training process, we extract an NFOV image from a spherical image with a field of view ranging from 30° to 120° and an aspect ratio of 2 : 1. Subsequently, the viewpoint direction was arbitrarily established on the sphere and projected onto the equirectangular image. We set $(\alpha_c, \beta_c, \gamma_c) = (360^\circ, 3^\circ, 3^\circ)$ and $(\alpha_d, \beta_d, \gamma_d) = (360^\circ, 10^\circ, 10^\circ)$. $\lambda / N / K$ are set to 0.1 / 50 / 4, respectively. We choose widely used metrics to evaluate image generation quality, including Fréchet Inception Distance (FID) (Heusel et al. 2017), spatial Fréchet Inception Distance (sFID) (Nash et al. 2021), and Inception Score (IS) (Salimans et al. 2016).

Performance Comparison

In this section, we compare the image generation quality with the latest work, and Table 1 shows the performance comparison between our method and other approaches. To ensure a fair comparison, we use the official implementation of ControlNet with the same hyperparameters and training iterations on the Structured3D dataset. We select four FOV sizes, 30, 60, 90, and 120 degrees for comparison. As can be seen, our method outperforms other methods in all metrics. Among them, in the test of four different FOV sizes, the most widely used FID score, our method improved significantly by 14.312 on average compared to ControlNet.

Furthermore, the visualization of the generated images also intuitively shows that our generation quality is more consistent with the textual descriptions and semantic segmentation maps compared to ControlNet. As shown in Figure 4, when we want to generate a bedroom with white walls

Method	FOV	FID↓	sFID↓	IS↑
ControlNet	30°	44.801	174.841	3.006
Ours		29.156	121.607	3.323
ControlNet	60°	41.917	158.873	2.957
Ours		26.262	111.318	3.325
ControlNet	90°	39.450	142.747	2.954
Ours		25.042	105.165	3.234
ControlNet	120°	35.690	123.075	2.853
Ours		24.147	92.039	3.246

Table 1: Comparison with the existing methods on Structure3D dataset. We use the same hyperparameter settings and number of training epochs for a fair comparison.

DRSE	DDaB	SR	SSCL	SGAG	FID↓
✗	✗	✗	✗	✗	39.450
✓	✗	✗	✗	✗	38.805
✓	✓	✗	✗	✗	35.076
✓	✓	✓	✗	✗	32.468
✓	✓	✓	✓	✗	25.763
✓	✓	✓	✓	✓	25.043

Table 2: Effect of each module of SphereDiffusion. ‘DRSE’ / ‘DDaB’ / ‘SR’ / ‘SSCL’ / ‘SGAG’ represent our Distortion-Resilient Semantic Encoding / Deformable Distortion-aware Block / Spherical Reprojection / Spherical SimSiam Contrastive Learning / Spherical Geometry-aware Generation.

and a pink bed, we can see that ControlNet erroneously generates a room with pink walls, and the area originally labeled ‘curtain’ in the semantic segmentation does not correctly generate the specified object. In contrast, since our method achieves a better object understanding of spherical panoramic images, our method accurately generates white walls and a pink bed, with the corresponding ‘curtain’ area correctly generating the specified object, resulting in a more reasonable overall output. When we try to generate a kitchen with gray walls and white cabinets, ControlNet mistakenly generates gray cabinets. In contrast, our method correctly generates gray cabinets and gray walls. Furthermore, the boundary connectivity of our generated images is significantly better than that generated by ControlNet.

Ablation Study

Effect of Four Modules in Training Process

As shown in Table 2, we validate Distortion-Resilient Semantic Encoding, Deformable Distortion-aware Block, Spherical SimSiam Contrastive Learning, and Spherical Reprojection, respectively. We selected a FOV size of 90° for the ablation experiment. The baseline FID score is 39.450. The FID score improves to 38.805, only including Distortion-Resilient Semantic Encoding, indicating the positive impact of incorporating semantic representation in the model. Adding our Deformable Distortion-aware Block to the model with Distortion-Resilient Semantic Encoding fur-

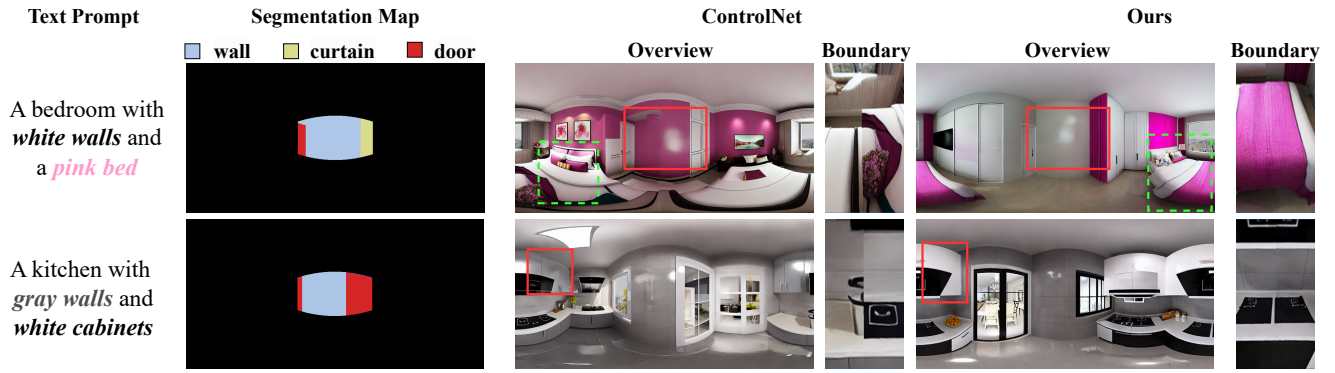


Figure 4: Visualization comparison of comparing SphereDiffusion with ControlNet. The images generated by our SphereDiffusion are more closely aligned with the guidance provided by the segmentation maps and text prompts (highlighted by red line boxes and green dotted boxes). ‘Overview’ is generated image, and ‘Boundary’ displays the boundary of the generated image.



Figure 5: Visualization of generated image results with or without the Spherical Geometry-aware Generation. We use the same SphereDiffusion model, employing consistent text prompts, segmentation maps, and random seeds for generation. The first row shows images generated without incorporating SGA Generation, while the second row presents images generated with the inclusion of SGA Generation. ‘Rotated Image’ is obtained by rotating the generated ‘Overall Image’ by $\alpha = 180^\circ$.

ther enhances the performance, achieving an improvement in the FID score of 3.7. The combination of Distortion-Resilient Semantic Encoding, Deformable Distortion-aware Block, and Spherical Reprojection results in an improvement, with the FID score dropping to 32.468. This demonstrates the effectiveness of incorporating Spherical Reprojection in the model. When we add all our components, the FID score further improves to 25.763, highlighting the importance of Spherical SimSiam Contrastive Learning in significantly enhancing the model’s performance.

Effect of Spherical Geometry-aware Generation

We evaluate our SGA Generation through Table 2 and Figure 3. As shown in Table 2, without retraining the model and only incorporating SGA Generation during the testing process, the FID score improves by almost 0.7. Visualizations are shown in Figure 3. Without SGA Generation, the generated images exhibit discontinuity at the boundary, with a clear demarcation line. However, once using SGA Generation, the generated images exhibit better connectivity. This demonstrates that SGA Generation can use the spherical ge-

ometric characteristic to enhance the boundary connectivity of generated spherical panoramic images.

Conclusion

Generating spherical panoramic images is a challenging task, as it requires considering spherical distortion and geometric characteristics. We propose SphereDiffusion, a framework that accounts for these characteristics, generating high-quality, controllable spherical panoramic images from single NFOV segmentation maps and text prompts. For spherical distortion characteristic, we introduce Distortion-Resilient Semantic Encoding and Deformable Distortion-aware Block. For spherical geometry characteristic, we leverage the spherical rotation invariance of spherical panoramic images and propose SGA Training, which includes Spherical Reprojection and Spherical SimSiam Contrastive Learning. Additionally, we introduce SGA Generation to improve the generation process. Through experiments, we verified that our method can significantly improve the quality of the generated images.

Acknowledgements

This work is supported in part by National Natural Science Foundation of China under Grant U20A20222, National Science Foundation for Distinguished Young Scholars under Grant 62225605, Zhejiang Key Research and Development Program under Grant 2023C03196, Research Fund of ARC Lab, Tencent PCG, and sponsored by CCF-AFSG Research Fund as well as The Ng Teng Fong Charitable Foundation in the form of ZJU-SUTD IDEA Grant, 188170-11102.

References

- Ai, H.; Cao, Z.; Zhu, J.; Bai, H.; Chen, Y.; and Wang, L. 2022. Deep Learning for Omnidirectional Vision: A Survey and New Perspectives. *arXiv preprint arXiv:2205.10468*.
- Akimoto, N.; Matsuo, Y.; and Aoki, Y. 2022. Diverse plausible 360-degree image outpainting for efficient 3DCG background creation. In *Proc. CVPR*, 11441–11450.
- Avrahami, O.; Hayes, T.; Gafni, O.; Gupta, S.; Taigman, Y.; Parikh, D.; Lischinski, D.; Fried, O.; and Yin, X. 2022. Spa-Text: Spatio-Textual Representation for Controllable Image Generation. *arXiv preprint arXiv:2211.14305*.
- Bar-Tal, O.; Yariv, L.; Lipman, Y.; and Dekel, T. 2023. MultiDiffusion: Fusing Diffusion Paths for Controlled Image Generation. *arXiv preprint arXiv:2302.08113*.
- Brooks, T.; Holynski, A.; and Efros, A. A. 2023. Instruct-pix2pix: Learning to follow image editing instructions. In *Proc. CVPR*, 18392–18402.
- Chen, X.; and He, K. 2021. Exploring simple siamese representation learning. In *Proc. CVPR*, 15750–15758.
- de La Garanderie, G. P.; Abarghouei, A. A.; and Breckon, T. P. 2018. Eliminating the blind spot: Adapting 3d object detection and monocular depth estimation to 360 panoramic imagery. In *Proc. ECCV*, 789–807.
- Dhariwal, P.; and Nichol, A. 2021. Diffusion models beat gans on image synthesis. In *Proc. NeurIPS*, volume 34, 8780–8794.
- Hara, T.; Mukuta, Y.; and Harada, T. 2021. Spherical image generation from a single image by considering scene symmetry. In *Proc. AAAI*, volume 35, 1513–1521.
- Heusel, M.; Ramsauer, H.; Unterthiner, T.; Nessler, B.; and Hochreiter, S. 2017. Gans trained by a two time-scale update rule converge to a local nash equilibrium. In *Proc. NeurIPS*, volume 30, 6629–6640.
- Ho, J.; and Salimans, T. 2022. Classifier-free diffusion guidance. *arXiv preprint arXiv:2207.12598*.
- Isola, P.; Zhu, J.-Y.; Zhou, T.; and Efros, A. A. 2017. Image-to-image translation with conditional adversarial networks. In *Proc. CVPR*, 1125–1134.
- Kawar, B.; Zada, S.; Lang, O.; Tov, O.; Chang, H.; Dekel, T.; Mosseri, I.; and Irani, M. 2023. Imagic: Text-based real image editing with diffusion models. In *Proc. CVPR*, 6007–6017.
- Kimura, N.; and Rekimoto, J. 2018. ExtVision: augmentation of visual experiences with generation of context images for a peripheral vision using deep neural network. In *Proc. CHI*, 1–10.
- Kong, C.; Jeon, D.; Kwon, O.; and Kwak, N. 2023. Leveraging off-the-shelf diffusion model for multi-attribute fashion image manipulation. In *Proc. WACV*.
- Li, J.; and Bansal, M. 2023. PanoGen: Text-Conditioned Panoramic Environment Generation for Vision-and-Language Navigation. *arXiv preprint arXiv:2305.19195*.
- Li, J.; Li, D.; Xiong, C.; and Hoi, S. 2022. BLIP: Bootstrapping Language-Image Pre-training for Unified Vision-Language Understanding and Generation. In *Proc. ICML*.
- Li, X.; Wu, T.; Qi, Z.; Wang, G.; Shan, Y.; and Li, X. 2023. SGAT4PASS: Spherical Geometry-Aware Transformer for Panoramic Semantic Segmentation. In *Proc. of IJCAI*, 1125–1133.
- Ma, C.; Zhang, J.; Yang, K.; Roitberg, A.; and Stiefelhagen, R. 2021. Densepass: Dense panoramic semantic segmentation via unsupervised domain adaptation with attention-augmented context exchange. In *Proc. ITSC*, 2766–2772. IEEE.
- Mokady, R.; Hertz, A.; Aberman, K.; Pritch, Y.; and Cohen-Or, D. 2023. Null-text inversion for editing real images using guided diffusion models. In *Proc. CVPR*, 6038–6047.
- Mou, C.; Wang, X.; Xie, L.; Zhang, J.; Qi, Z.; Shan, Y.; and Qie, X. 2023. T2i-adapter: Learning adapters to dig out more controllable ability for text-to-image diffusion models. *arXiv preprint arXiv:2302.08453*.
- Nash, C.; Menick, J.; Dieleman, S.; and Battaglia, P. W. 2021. Generating images with sparse representations. *arXiv preprint arXiv:2103.03841*.
- Nichol, A.; Dhariwal, P.; Ramesh, A.; Shyam, P.; Mishkin, P.; McGrew, B.; Sutskever, I.; and Chen, M. 2021. Glide: Towards photorealistic image generation and editing with text-guided diffusion models. *arXiv preprint arXiv:2112.10741*.
- Radford, A.; Kim, J. W.; Hallacy, C.; Ramesh, A.; Goh, G.; Agarwal, S.; Sastry, G.; Askell, A.; Mishkin, P.; Clark, J.; et al. 2021. Learning transferable visual models from natural language supervision. In *Proc. ICML*, 8748–8763. PMLR.
- Ramesh, A.; Pavlov, M.; Goh, G.; Gray, S.; Voss, C.; Radford, A.; Chen, M.; and Sutskever, I. 2021. Zero-shot text-to-image generation. In *Proc. ICML*, 8821–8831. PMLR.
- Rombach, R.; Blattmann, A.; Lorenz, D.; Esser, P.; and Ommer, B. 2022. High-resolution image synthesis with latent diffusion models. In *Proc. CVPR*, 10684–10695.
- Ruiz, N.; Li, Y.; Jampani, V.; Pritch, Y.; Rubinstein, M.; and Aberman, K. 2023. Dreambooth: Fine tuning text-to-image diffusion models for subject-driven generation. In *Proc. CVPR*, 22500–22510.
- Saharia, C.; Chan, W.; Saxena, S.; Li, L.; Whang, J.; Denton, E. L.; Ghasemipour, K.; Gontijo Lopes, R.; Karagol Ayan, B.; Salimans, T.; et al. 2022a. Photorealistic text-to-image diffusion models with deep language understanding. In *Proc. NeurIPS*, volume 35, 36479–36494.
- Saharia, C.; Ho, J.; Chan, W.; Salimans, T.; Fleet, D. J.; and Norouzi, M. 2022b. Image super-resolution via iterative refinement. *IEEE Trans. Pattern Anal. Mach. Intell.*, 45(4): 4713–4726.

- Salimans, T.; Goodfellow, I.; Zaremba, W.; Cheung, V.; Radford, A.; and Chen, X. 2016. Improved techniques for training gans. In *Proc. NeurIPS*, volume 29, 2234–2242.
- Sheynin, S.; Ashual, O.; Polyak, A.; Singer, U.; Gafni, O.; Nachmani, E.; and Taigman, Y. 2022. Knn-diffusion: Image generation via large-scale retrieval. *arXiv preprint arXiv:2204.02849*.
- Sohl-Dickstein, J.; Weiss, E.; Maheswaranathan, N.; and Ganguli, S. 2015. Deep unsupervised learning using nonequilibrium thermodynamics. In *Proc. ICML*, 2256–2265. PMLR.
- Sumantri, J. S.; and Park, I. K. 2020. 360 Panorama synthesis from a sparse set of images with unknown field of view. In *Proc. WACV*, 2386–2395.
- Wang, T.-C.; Liu, M.-Y.; Zhu, J.-Y.; Tao, A.; Kautz, J.; and Catanzaro, B. 2018. High-resolution image synthesis and semantic manipulation with conditional gans. In *Proc. CVPR*, 8798–8807.
- Xu, Y.; Zhang, Z.; and Gao, S. 2021. Spherical DNNs and Their Applications in 360 Images and Videos. *IEEE Trans. Pattern Anal. Mach. Intell.*
- Xue, H.; Huang, Z.; Sun, Q.; Song, L.; and Zhang, W. 2023. Freestyle Layout-to-Image Synthesis. In *Proc. CVPR*, 14256–14266.
- Yan, K.; Ji, L.; Wu, C.; Liang, J.; Zhou, M.; Duan, N.; and Ma, S. 2022. HORIZON: A High-Resolution Panorama Synthesis Framework. *arXiv preprint arXiv:2210.04522*.
- Zhang, J.; Yang, K.; Ma, C.; Reiß, S.; Peng, K.; and Stiefelhagen, R. 2022. Bending Reality: Distortion-aware Transformers for Adapting to Panoramic Semantic Segmentation. In *Proc. CVPR*, 16917–16927.
- Zhang, L.; Rao, A.; and Agrawala, M. 2023. Adding Conditional Control to Text-to-Image Diffusion Models. In *Proc. ICCV*, 3836–3847.
- Zhang, Q.; Song, J.; Huang, X.; Chen, Y.; and Liu, M.-Y. 2023. DiffCollage: Parallel Generation of Large Content with Diffusion Models. *arXiv preprint arXiv:2303.17076*.
- Zheng, G.; Li, S.; Wang, H.; Yao, T.; Chen, Y.; Ding, S.; and Li, X. 2022. Entropy-driven sampling and training scheme for conditional diffusion generation. In *Proc. ECCV*, 754–769. Springer.
- Zheng, G.; Zhou, X.; Li, X.; Qi, Z.; Shan, Y.; and Li, X. 2023. LayoutDiffusion: Controllable Diffusion Model for Layout-to-image Generation. In *Proc. CVPR*, 22490–22499.
- Zheng, J.; Zhang, J.; Li, J.; Tang, R.; Gao, S.; and Zhou, Z. 2020. Structured3D: A Large Photo-realistic Dataset for Structured 3D Modeling. In *Proc. ECCV*.