Spherical Manifold Guided Diffusion Model for Panoramic Image Generation

Supplementary Material

A. Experiments on Structured3D dataset

To further validate the robustness of our model, we conducted extensive evaluations on the Structured3D dataset. Specifically, we utilized a total of $21,133$ images with a resolution of 1024×512, splitting them into $19,019$ images for training and $2,114$ images for testing. The adoption of a test split highlights the generalizability of our approach and demonstrates its robustness against overfitting. For comparison, we focused on the second-best performing method, Panfusion, as a baseline. The quantitative results for the test images are presented in Table S1. As evidenced by the table, our method consistently outperforms the baseline, achieving the state-of-the-art performance across all evaluated metrics.

B. In-depth analysis on SMConv

To further highlight the distinction between the proposed SMConv and the spherical convolution introduced in SphereNet, we visualize the kernels generated by both methods in Fig. S1. Both kernels are derived by projecting a grid with the same unit length onto the tangent plane at the kernel's center point. As demonstrated in the figure, the length-preserving property of the exponential mapping leads to notable differences in the kernel patterns, particularly at lower resolutions. Specifically, the kernel of our SMConv exhibits a significantly larger receptive field compared to that of SphereNet. At the highest resolution of 64×32, while the differences between the kernels of SMConv and SphereNet are relatively subtle, the kernel patterns are not identical. It is worth emphasizing that even this slight discrepancy can play a significant role in influencing the overall performance. This observation underscores the importance of precise kernel design in spherical convolutions, as minor variations can lead to measurable impacts on the model's effectiveness.

After inspecting the differences in kernel patterns across various resolutions, we conduct a detailed analysis of the SMUNet architecture to evaluate the performance improvements achieved by our SMGD model in comparison to LDM+SphereNet. Specifically, in SMUNet, the resolution settings are configured as follows: 64×32 for the SPB, 64×32 and 32×16 for the circular encoder/decoder, 16×8 and 8×4 for the SME and SMD, and 8×4 for the SMRB. Given this configuration, there are 2 SMConv layers at resolution 64×32, 11 SMConv layers at resolution 16×8 and 14 SMConv layers at resolution 8×4. It can be concluded that the majority of SMConv layers operate at lower resolutions, which contributes significantly to the performance gain of the SMGD model compared to LDM+SphereNet. This performance gain is primarily attributed to the significant disparities between the two models at lower resolutions. Specifically, the distinctive properties of SMConv, including the adoption of exponential mapping and optimized kernel patterns, as illustrated in Fig. S1, play a pivotal role in enhancing feature extraction capabilities and the overall performance of the model.

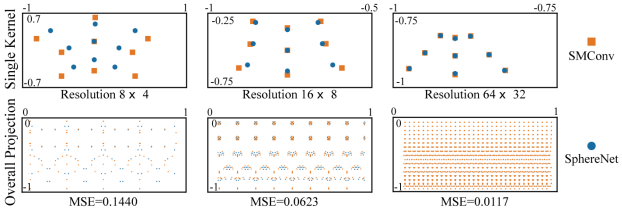


Figure S1. The projection difference in practice, given all possible latent resolutions and 3×3 kernels used in practice. The first row is for a single kernel. The second row shows the overall projection points. We normalize the range to $[-1, 1]$ to calculate the MSE between SMConv and SphereNet projected positions.

Furthermore, we provide additional qualitative comparison results between the proposed SMGD model and LDM+SphereNet to further demonstrate the superior performance of our SMGD model. As shown in Fig. S2, images generated by LDM+SphereNet exhibit blurriness, unnatural global structures, and artifacts in fine details, highlighting its limitations. In contrast, images produced by our SMGD model achieve significantly higher quality and demonstrate better alignment with the provided text prompts, underscoring the effectiveness of our approach.

C. More Qualitative Results for SMGD

We provide additional results to further demonstrate the effectiveness of our approach in generating panoramic images

Table S1. Comparisons on the test images from Structured3D dataset. We only compared with the second-best PanFusion method, both re-trained on the Structured3D dataset. We represent the best numbers by **red color**.

Methods	FID ↓	FID_{avg} ↓	FID_{cent} ↓	FID_{bord} ↓	FID_{rand} ↓	FID_{equ} ↓	FID_{pole} ↓	CS ↑
SMGD (Ours)	**18.65**	**17.89**	**17.62**	**18.27**	**17.77**	**15.64**	**46.94**	**0.1884**
PanFusion	31.05	30.39	29.50	31.44	30.23	37.28	49.15	0.1882

Figure S2. More qualitative comparison results between our SMGD model and LDM+SphereNet. The first row presents results generated by the proposed SMGD model, while the second row displays results from LDM+SphereNet. Each column corresponds to images synthesized using the text prompt provided at the bottom.

Figure S3. Additional results for text-conditioned panoramic image generation with text prompts drawn from our dataset. We illustrate the panoramic images generated by our model in both the ERP format and the converted CMP format.

conditioned on text descriptions. These results include examples derived from text prompts within our dataset, illustrating the alignment between the generated images and text descriptions, as shown in Figs. S3 to S5. Moreover, we evaluate the zero-shot generalization capacity of our model by employing novel text prompts that were not encountered during training, as depicted in Figs. S6 and S7. These results demonstrate the capacity of model to generalize effectively to previously unseen prompts. We present the panoramic images in both the generated equirectangular projection (ERP) format and the converted cubemap projection (CMP) format. This dual representation effectively illustrates the continuity of the left and right image boundaries, as well as the spherical geometry intrinsic to panoramic images.

CVPR
#7468

"a bedroom with a mirror and a bed"

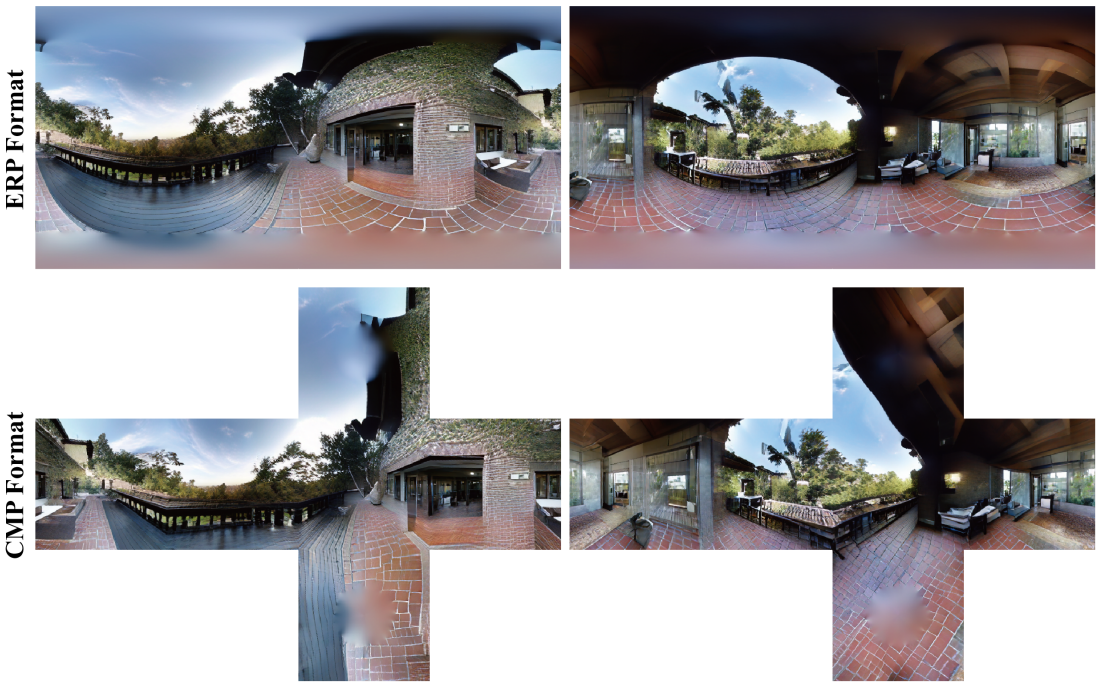
Figure S4. Additional results for text-conditioned panoramic image generation with text prompts drawn from our dataset. We illustrate the panoramic images generated by our model in both the ERP format and the converted CMP format.

"a hallway with paintings on the wall"

Figure S5. Additional results for text-conditioned panoramic image generation with text prompts drawn from our dataset. We illustrate the panoramic images generated by our model in both the ERP format and the converted CMP format.

"A house designed with both luxury and relaxation in mind, featuring a swimming pool that serves as the centerpiece of the backyard."

Figure S6. Results for zero-shot text-conditioned panoramic image generation. We illustrate the panoramic images generated by our model in both the ERP format and the converted CMP format.



"Lush garden viewed from a cozy balcony, verdant plants, tranquil setting, sunlight filtering through leaves, peaceful retreat."

Figure S7. Results for zero-shot text-conditioned panoramic image generation. We illustrate the panoramic images generated by our model in both the ERP format and the converted CMP format.