High-Quality 360-Degree Panoramic Image Generation based on Masks and Stable Diffusion Model

Fuwei Dong
Beijing Technology and Business University
Beijing, China
2230702065@st.btbu.edu.cn

Ming Li
Anran Meng
Beijing Technology and Business University
Beijing, China
2230702018@st.btbu.edu.cn

Siying Zhu (Corresponding Author)
Beijing Vocational College of Labour and Social Security
Beijing, China
2018010308@bvclss.edu.cn

Xiaoming Chen (Corresponding Author)
Beijing Technology and Business University
Beijing, China
xiaoming.chen@btbu.edu.cn

***Abstract*—360-degree panoramic images play a crucial role in multiple technological domains, particularly in virtual reality and augmented reality. However, current generation methodologies encounter significant limitations and are incapable of directly producing 360-degree panoramic images through simple graphical elements such as masks. To address this gap, this study introduces a mask-based 360-degree panoramic image generation method. Utilizing a diffusion model that requires no additional training or fine-tuning, the proposed approach generates high-quality RGB images during the inference phase. By incorporating masks and text prompts, the method efficiently produces complete, high-quality 360-degree panoramic images, substantially improving generation efficiency. Experimental evaluations, employing qualitative metrics such as Absolute Category Rating and Perceptual Consistency Rating, systematically validated the effectiveness of the proposed method across various scenarios. The results demonstrate that the method not only generates high-quality 360-degree panoramic images but also ensures semantic consistency and aligns closely with human visual perception.**

***Index Terms*—360-degree panoramic image, Stable diffusion, Mask image generation.**

I. Introduction

Panoramic images, characterized as wide-angle visual representations synthesized by seamlessly stitching multiple photographs captured from diverse angles to encompass a comprehensive 360-degree horizontal and 180-degree vertical field of view, have emerged as a transformative technology across diverse domains, including virtual reality (VR), urban planning, and computer vision. Their profound impact stems from their unparalleled capacity to deliver immersive visual experiences and rich contextual information, surpassing the capabilities of traditional imaging techniques. These images can not only enhance user engagement through realistic and comprehensive visual effects but also play a crucial role in in tasks such as scene understanding, object recognition, and environmental modeling. Despite their growing significance, conventional panoramic image generation methods encounter significant challenges in handling complex real-world scenarios, which restrict their scalability and practical applicability.

Existing techniques for 360-degree panoramic image generation generally fall into two categories: multi-view image stitching methods and deep learning-based approaches. Multi-view image stitching involves capturing images from multiple viewpoints and producing seamless 360-degree panoramic outputs through geometric correction and alignment. Although effective in controlled settings, these methods rely on precise camera calibration and require computationally intensive geometric adjustments. On the other hand, deep learning-based approaches directly generate 360-degree panoramic images from a single input using neural networks, eliminating the need for multi-view hardware setups and enhancing operational efficiency. However, the performance of these methods is heavily dependent on the quality and diversity of the training dataset, which limits their generalization ability in unseen or complex scenarios.

To overcome the limitations of existing methods, this study introduces a novel mask-based approach for 360-degree panoramic image generation. This method utilizes masks derived from object shapes, enabling the generation model to use these masks as constraints during the image generation process. Unlike traditional methods that rely on multi-view data or extensive model training, the proposed method operates during the inference stage of a stable diffusion model without requiring additional training. By utilizing masks, the proposed method identifies and prioritizes key regions in the input scene, enabling the generation of high-quality 360-degree panoramic images that align with the masked areas, as shown in Figure 1. Experimental results demonstrate the effectiveness of the proposed method in enhancing image quality and maintaining semantic consistency with the input masks. Moreover, the

generated images align closely with human visual perception, significantly enhancing the immersive experience.

Our work and main contributions are sum- marized as follows:

- We introduce a novel 360-degree panoramic image generation approach by leveraging masks derived from object shapes.
- The proposed method is designed to be free of additional training, significantly reducing computational complexity.
- Experiments demonstrats that the generated images show superior quality compaired to traditional methods across multiple metrics.

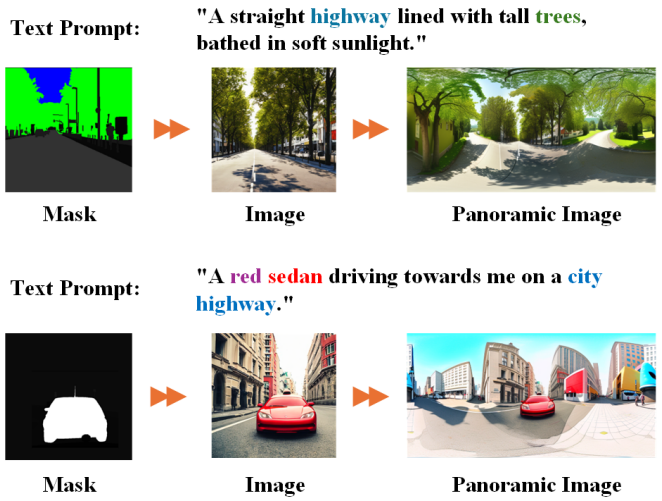

Fig. 1: Example of generating 360-degree panoramic image using masks.

II. Related Work

A. 360-Degree Panoramic Image Generation

A 360-degree panoramic image is an advanced visualization technique that fully renders visual information through spherical projection, enabling viewers to explore a scene from any angle. Recent advancements have explored 360-degree panoramic image generation using diffusion models. MVDiffusion [1] employed eight perspective views as input to generate closed-loop 360-degree panoramic images, though the outputs often resembled wide-angle panoramas, with artifacts such as “sky” and “ground” distortions visible in 360-degree viewers. StitchDiffusion [2] attempted to address continuity issues by globally cropping and aligning the left and right edges of images, but magnified views in 360-degree viewers still revealed visible seams. Additionally, Feng et al. [3] introduced a cyclic padding scheme, fine-tuning the DreamBooth model using standard diffusion processes. During inference, they applied cyclic fusion to improve continuity across seams. Despite these advancements, no existing methods have explored the generation of 360-degree panoramas directly from masks.

B. RGB Image Generation

In recent years, diffusion models [4]–[7] have demonstrated exceptional performance in terms of generation quality and diversity, gradually becoming a research focus in both academia and industry. For instance, techniques like MaskSketch [8] transform hand-drawn sketches into highly realistic images, preserving textures and intricate details. However, this method relies on high-quality sketches with well-defined contours and textures. Further advancements by Wu et al. [9] leveraged the cross-attention mechanism between text and images, enabling diffusion models to generate high-quality images and their corresponding masks from textual descriptions. Xie et al. [10] demonstrated the use of text prompts and manually drawn bounding boxes to guide the generation of realistic images in specified regions using Latent Diffusion Models. However, there is currently no research focused on mask-based image generation methods that do not require training.

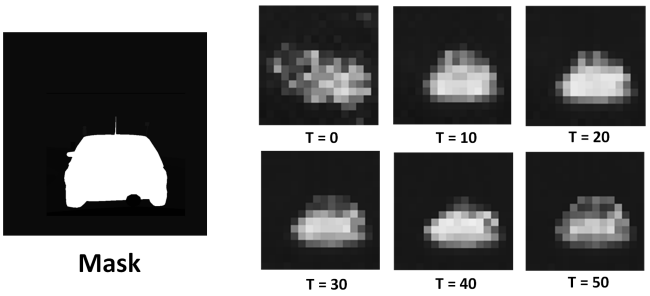

Fig. 2: During the first 50 diffusion steps, the 16×16 latent attention map progressively aligns with the mask, achieving spatial and morphological convergence.

III. Methodology

A. Mask-based High-Quality Image Generation.

Masks are commonly employed to represent the shape and location of objects in images. Masks, however, do not capture texture information. Inspired by the work of Xie et al. [10], we have observed that self-attention maps within the Stable Diffusion model play a pivotal role in image generation, influencing not only spatial positioning but also the creation of content. During the inference process, the model generates corresponding self-attention maps based on the provided textual input, thereby optimizing the latent variables at each timestep and generating corresponding objects at positions with high attention weights. Leveraging this property, we utilize masks as constraints to guide the generation of high-quality images that maintain both spatial and semantic consistency.

To achieve this, the total inference timesteps of the model are divided into two distinct stages. In the initial half of the denoising process, we extract the attention map corresponding to the textual description of the object specified in the prompt. Subsequently, the mask associated with the object label is retrieved and aligned with the latent variable’s attention map. A matrix product is then computed between the aligned mask and the attention map. To further impose spatial constraints,

we identify the top-k highest values from the resulting matrix product and compute their average, which serves as the optimization target. This process enables the model to adjust its attention to focus on specific locations during each timestep, as shown in Figure 2. In the latter half of the denoising process, the model incrementally refines the image at these targeted locations, generating high-quality outputs that align with the semantic content of the prompt. The entire process is shown in the Stage 1 of Figure 3.

The spatial constraints imposed by this approach are formalized in the following loss function:

$$L_{fi}^t = 1 - \frac{1}{k} \sum \text{topk}(A_i^t \cdot M_i), \quad L_F = \sum L_{fi}^t \tag{1}$$

$$L_{bi}^t = \frac{1}{k} \sum \text{topk}(A_i^t \cdot M_i), \quad L_B = \sum L_{bi}^t \tag{2}$$

where L_{fi}^t represents the foreground loss of the i-th text at time t, L_{bi}^t represents the background loss of the i-th text at time t, A_i^t denotes the attention map of the i-th index at time t, and M_i is the mask of the i-th object. Subsequently, the foreground loss L_{fi}^t and the background loss L_{bi}^t are summed to form the overall loss L:

$$L = L_F + L_B \tag{3}$$

B. Single Image-based 360-degree Panoramic Image Generation.

To generate high-quality 360-degree panoramic images, we drew inspiration from the method described in [2]. During the inference phase, a single input image is first transformed into a central cubic map, which effectively preserves the spatial structure of the image. This transformation serves as a solid foundation for the subsequent image generation process. The resulting central cubic map is then fed into the model, where, after processing by the control network, it generates the complete 360-degree panoramic image.

In the image generation process, the Variational Autoencoder (VAE) decoder ensures seamless alignment of the left and right boundaries of the image, mitigating common issues with visual seams. To maintain geometric continuity and prevent distortion at the boundaries, we employed a circular blending strategy. This technique applies a weighted averaging approach at the edges, enabling a smooth transition of pixel values between adjacent regions and resulting in a natural, seamless stitching effect. Finally, we generate high-definition 360-degree panoramic images using super-resolution techniques. The entire process is shown in the second stage of Figure 3. The formula for the circular blending strategy is as follows:

$$B = w \cdot L + (1 - w) \cdot R \tag{4}$$

where B represents the blended result, L and R denote the feature blocks extracted from the left and right parts of the image, respectively, w is an adaptive weight used to balance the blending of left and right features. This weight is adaptively adjusted based on the position to ensure a seamless transition in the boundary regions.

Table 1: Average scores of Absolute Category Rating (ACR), Visual Attribute Rating (VAR), Spatial Semantic Differential Scale (SSDS), and Perceived Consistency Rating (PCR) in four scenarios.

Scenes	vehicle	street	indoor	mountainous
ACR [13]	4.05	3.82	3.93	4.02
VAR [14]	3.2	3.92	3.79	3.85
SSDS [15]	4.32	3.89	3.96	4.202
PCR	4.23	3.96	4.02	4.58

IV. Experiments

Given the limited research on generating panoramic images from masks and the lack of ground-truth of generated images, this study primarily adopts subjective experiments. Masks were collected from a variety of scenes across different datasets [11], [12], including vehicle routes, open streets, indoor spaces, and mountainous environments. A total of 32 participants (average age: 27, SD = 2.41) wore HTC VIVE Pro Eye VR headsets to view the generated panoramic images. Participants rated the images from different scenes using Absolute Category Rating (ACR) [13], Visual Attribute Rating (VAR) [14], Spatial Semantic Differential Scale (SSDS) [15], and Perceived Consistency Rating (PCR), with scores ranging from 1 to 5, where 5 represents the highest quality. The specific evaluation metrics used are shown in Table 1.

As shown in Table 1, the 360-degree panoramic images generated using the mask-based approach received higher scores in terms of overall quality and visual perception. Specifically, the images demonstrated significant advantages across multiple dimensions, including clarity, color reproduction, and detail rendering, thereby enhancing the visual experience for viewers. Regarding spatial-semantic consistency, the mask-based spatial constraint mechanism effectively controlled the alignment between the image content and the mask region, ensuring that the spatial structure of the generated image closely matched the original mask. This was particularly evident in the retention of object boundaries and shapes, such as those of vehicles. In terms of perceptual consistency, the extended control network model led to 360-degree panoramic images with generally smoother transitions, particularly in complex scenes, such as mountainous environments. By fine-tuning the fusion methods for different regions, the model effectively minimized unnatural transitions and abrupt edges, enhancing the visual smoothness of the image and contributing to higher ratings.

V. Conclusion

In this study, we propose a novel mask-based method for 360-degree panoramic image generation. This method utilizes masks to selectively constrain attention distribution, ensuring both spatial and semantic consistency in the generated images.

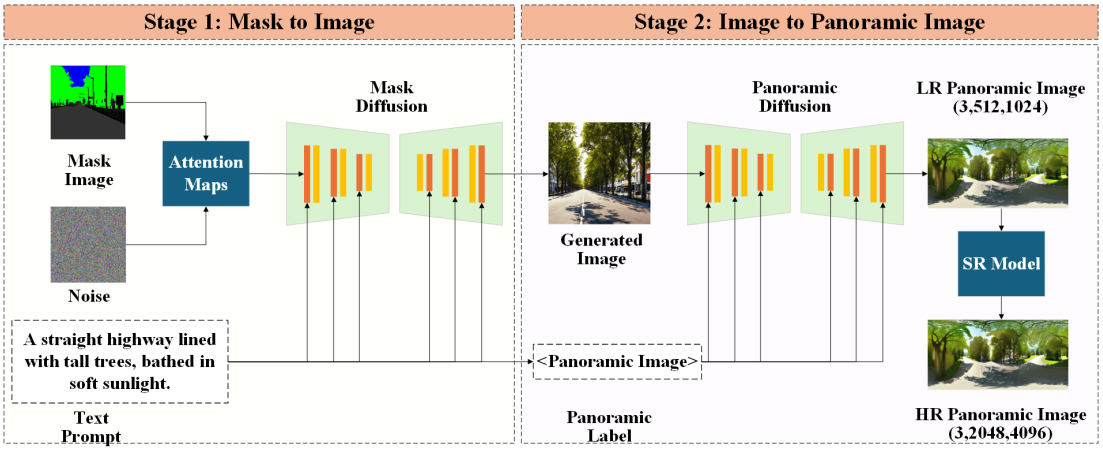

Fig. 3: The overall architecture for generating 360-degree panoramic images from masks. In the first stage, the attention map is constrained by the mask, and high-quality RGB images are generated at specified locations using Stable Diffusion. In the second stage, the RGB image is first converted into a low-quality 360-degree panoramic image, then its quality is enhanced using super-resolution techniques [2] to generate the final high-quality 360-degree panoramic image.

The generated outputs are subsequently transformed into high-resolution 360-degree panoramic images through an extended control network model. Experimental results demonstrate that our method achieves high-quality 360-degree panoramic image generation while preserving semantic consistency with the original input, providing an innovative solution to key challenges in 360-degree panoramic image generation.

Despite its advantages, the proposed method has certain limitations. On the one hand, as the Stable Diffusion model primarily performs denoising on latent variables, the constraints imposed by the mask are applied only at the latent variable level. This requires downsampling the mask to a lower resolution, which may lead to incomplete spatial alignment between the decoded output and the mask. On the other hand, since background elements typically receive lower attention values, the application of mask-based constraints can sometimes result in insufficient spatial consistency in specific scenarios. Future research will focus on addressing these limitations to further enhance the robustness and applicability of the proposed method.

VI. Acknowledgement

This work was supported by National Natural Science Foundation of China under Grant 62177001.

References

[1] Z. Deng, X. He, Y. Peng, X. Zhu, and L. Cheng, "Mv-diffusion: Motion-aware video diffusion model," in *Proceedings of the 31st ACM International Conference on Multimedia*, 2023, pp. 7255–7263.

[2] H. Wang, X. Xiang, Y. Fan, and J.-H. Xue, "Customizing 360-degree panoramas through text-to-image diffusion models," in *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, 2024, pp. 4933–4943.

[3] M. Feng, J. Liu, M. Cui, and X. Xie, "Diffusion360: Seamless 360 degree panoramic image generation based on diffusion models," *arXiv preprint arXiv:2311.13141*, 2023.

[4] A. Ramesh, P. Dhariwal, A. Nichol, C. Chu, and M. Chen, "Hierarchical text-conditional image generation with clip latents," *arXiv preprint arXiv:2204.06125*, vol. 1, no. 2, p. 3, 2022.

[5] Y. Gu, X. Wang, J. Z. Wu, Y. Shi, Y. Chen, Z. Fan, W. Xiao, R. Zhao, S. Chang, W. Wu *et al.*, "Mix-of-show: Decentralized low-rank adaptation for multi-concept customization of diffusion models," *Advances in Neural Information Processing Systems*, vol. 36, 2024.

[6] R. Rombach, A. Blattmann, D. Lorenz, P. Esser, and B. Ommer, "High-resolution image synthesis with latent diffusion models," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2022, pp. 10 684–10 695.

[7] J. Sohl-Dickstein, E. Weiss, N. Maheswaranathan, and S. Ganguli, "Deep unsupervised learning using nonequilibrium thermodynamics," in *International conference on machine learning*. PMLR, 2015, pp. 2256–2265.

[8] D. Bashkirova, J. Lezama, K. Sohn, K. Saenko, and I. Essa, "Masksketch: Unpaired structure-guided masked image generation," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2023, pp. 1879–1889.

[9] W. Wu, Y. Zhao, M. Z. Shou, H. Zhou, and C. Shen, "Diffumask: Synthesizing images with pixel-level annotations for semantic segmentation using diffusion models," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2023, pp. 1206–1217.

[10] J. Xie, Y. Li, Y. Huang, H. Liu, W. Zhang, Y. Zheng, and M. Z. Shou, "Boxdiff: Text-to-image synthesis with training-free box-constrained diffusion," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2023, pp. 7452–7461.

[11] M. Cordts, M. Omran, S. Ramos, T. Rehfeld, M. Enzweiler, R. Benenson, U. Franke, S. Roth, and B. Schiele, "The cityscapes dataset for semantic urban scene understanding," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2016, pp. 3213–3223.

[12] T.-Y. Lin, M. Maire, S. Belongie, J. Hays, P. Perona, D. Ramanan, P. Dollár, and C. L. Zitnick, "Microsoft coco: Common objects in context," in *Computer Vision–ECCV 2014: 13th European Conference, Zurich, Switzerland, September 6-12, 2014, Proceedings, Part V 13*. Springer, 2014, pp. 740–755.

[13] T.-J. Liu, K.-H. Liu, H.-H. Liu, and S.-C. Pei, "Comparison of subjective viewing test methods for image quality assessment," in *2015 IEEE International Conference on Image Processing (ICIP)*. IEEE, 2015, pp. 3155–3159.

[14] X. Jin, L. Wu, G. Zhao, X. Li, X. Zhang, S. Ge, D. Zou, B. Zhou, and X. Zhou, "Aesthetic attributes assessment of images," in *Proceedings of the 27th ACM international conference on multimedia*, 2019, pp. 311–319.

[15] J. Stoklasa, T. Talášek, and J. Stoklasová, "Semantic differential for the twenty-first century: scale relevance and uncertainty entering the semantic space," *Quality & Quantity*, vol. 53, pp. 435–448, 2019.