CamFreeDiff: Camera-free Image to Panorama Generation with Diffusion Model

Xiaoding Yuan
Johns Hopkins University

Shitao Tang
Simon Fraser University

Kejie Li
ByteDance

Peng Wang
ByteDance

Figure 1. We propose CamFreeDiff that can generate text-guided 360-degree panoramic image with plausible layout and canonical viewpoint from a camera-free input. CamFreeD iff eliminates the restriction on the camera parameters including viewpoints of inputs, showing strong robustness to non-canonical inputs and generates high-quality 360° panoramic image.

Abstract

This paper introduces Camera-free Diffusion (CamFreeDiff) model for 360° image outpainting from a single camera-free image and text description. This method distinguishes itself from existing strategies, such as MVDiffusion, by eliminating the requirement for predefined camera poses. CamFreeDiff seamlessly incorporates a mechanism for predicting homography within the multi-view diffusion framework. The key component of our approach is to formulate camera estimation by directly predicting the homography transformation from the input view to the predefined canonical view. In contrast to the direct two-stage approach of image transformation and outpainting, CamFreeDiff utilizes predicted homography to establish point-level correspondences between the input view and the target panoramic view. This enables consistency through correspondence-aware attention, which is learned in a fully differentiable manner. Qualitative and quantitative experimental results demonstrate the strong robustness and performance of CamFreeDiff for 360° image outpainting in the challenging context of camera-free inputs.

1. Introduction

Unlike traditional movies, 360° movies create an immersive experience that allows viewers to feel as if they are part of the movies's environment rather than merely observing it from a fixed perspective. This immersive aspect has been significantly enhanced with the advent and proliferation of AR/VR devices like VisionPro. Creating 360° movies typically requires specialized equipment, like 360° cameras, making it a highly professional endeavour. An alternative,

yet underexplored, approach is outpainting existing movies into 360° formats.

Image-based panorama outpainting [2, 2, 4, 14, 26, 28, 32, 33, 35] is a necessary step towards video-based 360° movie creation. The recent advance of text-to-image diffusion models [9, 16–20] makes it possible to extrapolate an image into 360° view. PanoDiffusion [32] proposes a panorama outpainting method by fine-tuning a pretrained latent diffusion model [19]. However, with a limited amount of avaiable training data, their fine-tuning disrupts the prior knowledge of the pre-trained model and diminishes its generalization capabilities. MVDiffusion [28], by contrast, leverages the frozen pre-trained latent diffusion model with strong generative capabilities to produce multi-view consistent panoramic images. This method ensures geometric consistency between views through correspondence-aware attention, but still requires known camera intrinsic and pose for the input image, limiting its application to panoramas generated from images with defined camera parameters.

Extending panorama generation to camera-free inputs poses significant challenges. Our method estimates the input camera parameters by estimating the homography transformation from the input image to a predefined canonical view. The homography establishes a correspondence between input views and each targeting panoramic view, allowing for the enforcement of multi-view consistency via correspondence-aware attention [28]. Furthermore, the homographic matrix estimator is seamlessly integrated with the multi-view diffusion model in a fully differentiable manner. By doing so, CamFreeDiff can effectively mitigate the errors introduced to and accumulated in the multi-view generation process by homography estimation. CamFreeDiff is fine-tuned from the pre-trained stable diffusion inpainting model on the high-quality Matterport3D[3] while keeping the U-Net blocks frozen. We demonstrated the strong robustness of CamFreeDiff to random camera parameters. Our model also shows excellent generalization ability when tested on out-of-domain data from Structured3D[36] without training. To summarize, our main contributions include:

- We propose CamFreeDiff, a 360° panoramic image generation model capable of handling camera-free input images. To the best of our knowledge, our model is the first image generation model designed to handle unknown input view and camera parameters.
- We formulate the camera parameter estimation by predicting the homography transformation from the input image to a predefined canonical view. We found that approximating the homography transformation by a three degrees of freedom(3-DoF) parametrization and prediction instead of the full 8-DoF prediction is sufficient in terms of panoramic image outpainting.
- We propose a novel strategy that seamlessly integrates camera prediction into the multi-view image diffusion model to enforce multi-view consistency and produce images with high visual quality.

Figure 2. In the context of panorama generation, we define the camera parameter for a canonical perspective view as $fov = 90°$, x-axis rotation $\phi = 0$ and z-axis rotation $\psi = 0$. The y-axis rotation can be any θ from $0°$ to $360°$.

2. Related Work

Multi-view image generation. The development of large-scale image generation models [16, 19, 20] has paved the way for multi-view image generation [11–13, 22, 23, 27, 28, 31]. A notable advancement in this area is the Zero-1-to-3 model [11], which can deduce novel views of an object given a single image and specified camera pose. Building on this, SyncDreamer [12] introduces a method for generating images at fixed viewpoints through a combination of volumetric representation and epipolar attention, though it occasionally produces artifacts. In an effort to enhance consistency and image fidelity, models such as MVDream [23], ImageDream [31], and Wonder3D [13] have employed comprehensive self-attention mechanisms across all views. Despite their success, these models necessitate extensive fine-tuning of a latent diffusion model's parameters, which can compromise pre-trained priors. While our method also adopts multi-view generation paradigms, our method is able to outpaint panorama for arbitrary images by freezing the pre-trained stable diffusion model.

Panorama generation. Previous methods [1, 4, 24, 25, 28, 32, 34] for generating panoramas typically rely on generative adversarial networks [7], auto-regressive [29] or diffusion models. Text2Light [4] emerged as a pioneering approach, utilizing a cascade auto-regressive model for panorama creation, whereas LDM3D [25] and PanoDiffusion [32] have adopted latent diffusion techniques for generating or outpainting panoramas. Despite their innovations, these models often encounter a reduction in their ability to generalize when they are fine-tuned on limited datasets. On the other hand, the process of generating panoramic images inherently benefits from the direct correspondence between different views. Utilizing this, MVDiffusion [28] presents a novel approach by incorporating correspondence-aware attention to maintain multi-view consistency, without modifying the underlying pre-trained model.

However, challenges persist with current outpainting

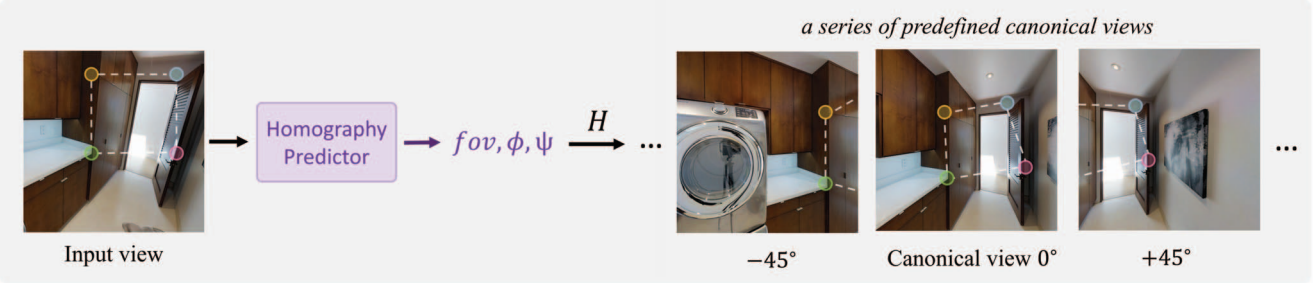

Figure 3. We formulate the camera parameter estimation as estimating the homography matrix H from the input view to a series of predefined canonical views. Coloured points in the figure visualize how the pixels corresponding to the same physical point change across different views. We define the canonical view as the perspective view with an absolute rotation angle of 0°. We use a 3-DoF parameterization of the homography matrix H instead of predicting the standard 8-DoF to predict the transformation from input to canonical views(details in 4.1).

techniques, notably their struggle to generate coherent panoramas due to their inability to infer camera parameters. PanoDiff[30] takes a step forward to estimate the longitude and latitude angles of a view, however still assumes camera roll angle and intrinsic. To address these shortcomings, we introduce a novel, fully differentiable pipeline designed to extrapolate from a single camera-free image to a complete 360° panorama. This method aims to bridge the gap between existing panoramic generation techniques and the need for accurate, consistent, and high-quality panoramic imagery, leveraging the strengths of direct view correspondence while overcoming the limitations of camera parameter estimation.

3. Preliminary

MVDiffusion is a panoramic image diffusion model for coherent multi-view generation. It splits the 360° scene into eight perspective views $\mathbf{I} = \{I_0, I_1, ..., I_7\}$ at fixed viewpoints, each with a 90° horizontal and vertical field of view and overlaps 45° horizontally with neighboring views. For image-based panorama outpainting, MVDiffusion employs a latent diffusion inpainting model and aligns the first view I_0 (canonical view) with the input image. Correspondence-aware attention (CAA) blocks are designed in MVDiffusion to enforce geometry consistency across views. Consider a pair of corresponding points located at p_s and p_t in the source view I_s and target view I_t respectively, the information is aggregated from a $K \times K$ neighborhood $\mathcal{N}(p_s)$ centered at p_s in the source view feature map through a cross-attention mechanism:

$$\text{Attention}(Q, K, V) = \text{softmax}(\frac{QK^T}{\sqrt{d_k}})V \tag{1}$$

$$M = \sum_{p_s^* \in \mathcal{N}(p_s)} \text{Attention}(W_Q f(p_t), W_K f(p_s^*), W_V f(p_s^*)) \tag{2}$$

Where $f(p_t)$ refers to the image feature at location p_t from the target view I_t and $f(p_s^*)$ denotes the image feature at location p_s^* inside the neighbourhood $\mathcal{N}(p_s)$ from the source view I_s. W_Q, W_K, W_V are the query, key and value projections for cross-attention module in CAA blocks. MVDiffusion assumes the camera parameters of the input are the same as the output canonical view, thus not able to generate from camera-free input images.

4. Method

We aim to generate text-guided 360° image from a camera-free input, where we assume unknown camera parameters. In order to address this challenging problem, we propose our Camera-free Diffusion(CamFreeDiff) model. CamFreeDiff predicts the camera parameters of the input image by estimating the homography transformation from the input to a predefined canonical perspective view. The predicted homography matrix provides point-wise correspondences between the input view and multiple targeting views conditionally. CamFreeDiff leverages the correspondence-aware attention(CAA) module to ensure the input texture remains unchanged while enforcing the geometry transformation from the input to the final panorama generation, conditioned on the predicted homography matrix.

4.1. Homography Matrix Estimation

To generate a panorama based on the input image with an unknown camera, we first estimate the homography matrix $H^{3\times3}$ transforming from the unknown input view to a predefined canonical view. Figure 2 Shows how we define a series of canonical views, and 3 provides a visual illustration of our method.

4.1.1. Parameterization of Homography Matrix.

Compared to the most straightforward way of parameterizing a homography into elements of a 3x3 matrix, DeTone Et al. [5] demonstrated that a 4-point parameterization of

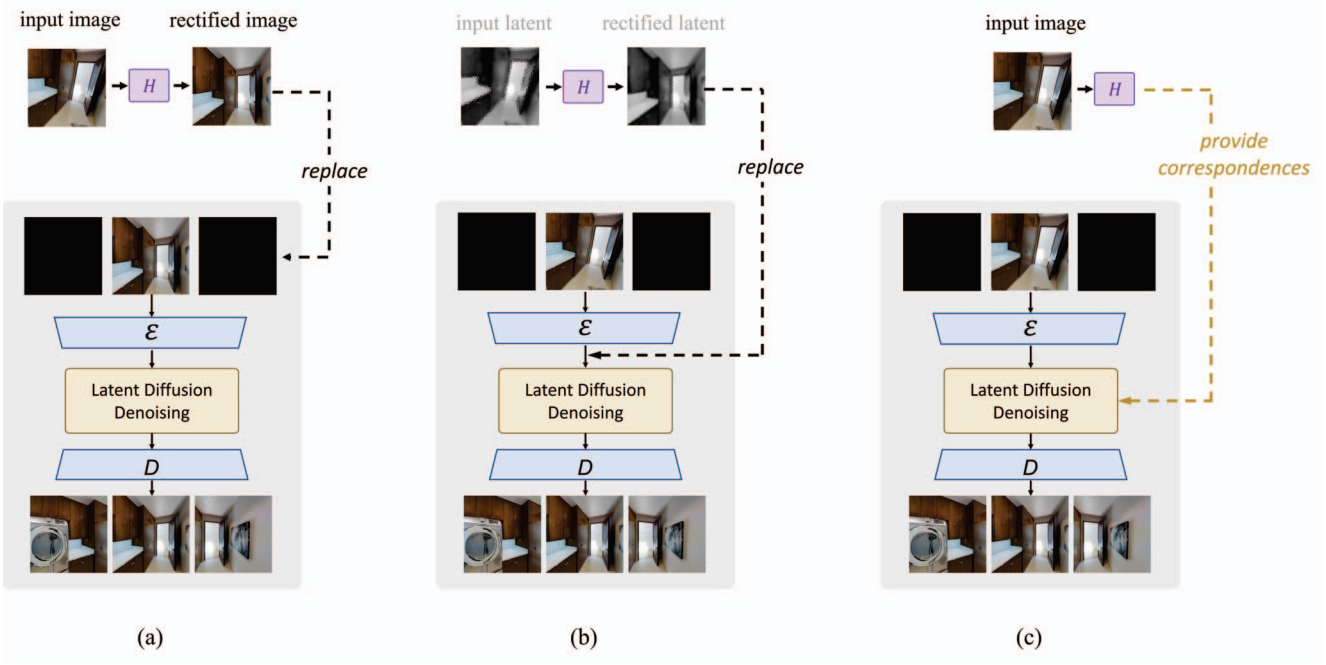

Figure 4. We explored different strategies to generate panorama from a camera-free input. After estimating the homography matrix H from input image to a canonical view, alternatives are: (a) Rectify the original input image by unwarping and use the rectified image for standard panorama outpainting. (b) After encoding the input image to the latent space, unwarp the input latent instead of the original image for later denoising generation. (c) Directly provide point-level correspondences from estimated homography H to the multi-view generation model to enforce generation consistency across target views.

the homography matrix can largely reduce the optimization difficulty for a deep neural network. This is because the 4 points are independent of each other while the 3x3 matrix mixes both the rotational and the translational terms.

We further simplified the parameters for panoramic image generation. When representing homography as $H = K_2 R K_1^{-1}$, where K_2 and K_1 are the camera intrinsics for the canonical view and the input image accordingly, some parameters are constant by default for most common scenarios. Specifically, we define the intrinsic matrix of a canonical view K_2 is

$$\begin{bmatrix} f_x & \gamma & c_x \\ 0 & f_y & c_y \\ 0 & 0 & 1 \end{bmatrix} = \begin{bmatrix} 256 & 0 & 255.5 \\ 0 & 256 & 255.5 \\ 0 & 0 & 1 \end{bmatrix} \tag{3}$$

We assume the intrinsic of the input image satisfies the pinhole camera and thus for K_1, following our canonical view intrinsic, the input axis skew coefficient γ also defaults to 0 and the principal point offsets (c_x, c_y) default to the centre of the image $(w/2,\ h/2)$.

Accordingly, for our task to generate 360° panoramic images the approximated homography matrix from the input view to the canonical view has three degrees of freedom: 1) camera field of view (f), 2) camera rotation along the x-axis (ϕ), 3) camera rotation along the z-axis (ψ). Particularly, based on the condition of a single input image, predicting the absolute rotation around the y-axis (θ) is considered meaningless since the input view can be aligned to any standard canonical view with $0 \leq \theta \leq 360°$ in a 360° panorama.

Therefore, we formulate the model that predicts homography transformation from an input image $I \in \mathbb{R}^{H \times W \times 3}$ to a predefined canonical view as:

$$\mathcal{M}(I) \rightarrow (f,\ \phi,\ \psi) \tag{4}$$

The input image camera intrinsic K_1 can be determined by predicted f. Along with known target perspective camera intrinsic and predicted rotation R from $(\phi,\ \psi,\ \theta = 0)$, the homography transformation H can be recovered from the predictions $(f, \phi,\ \psi)$.

4.1.2. Homography Estimator

We build an MLP classifier with three hidden layers upon a general image encoder. The U-Net encoder pre-trained by stable diffusion model for image generation also serves as our homography estimation image encoder, but with weights frozen for efficiency. Only the MLP classifier is optimized to learn a homography estimator. Feature dimensions for each hidden layer in MLP are set to 5120, 2560 and 1280. $SiLU$[6] is used as activation functions in MLP block. Cross-entropy loss is applied as learning objectives to fov, ϕ and ψ, respectively.

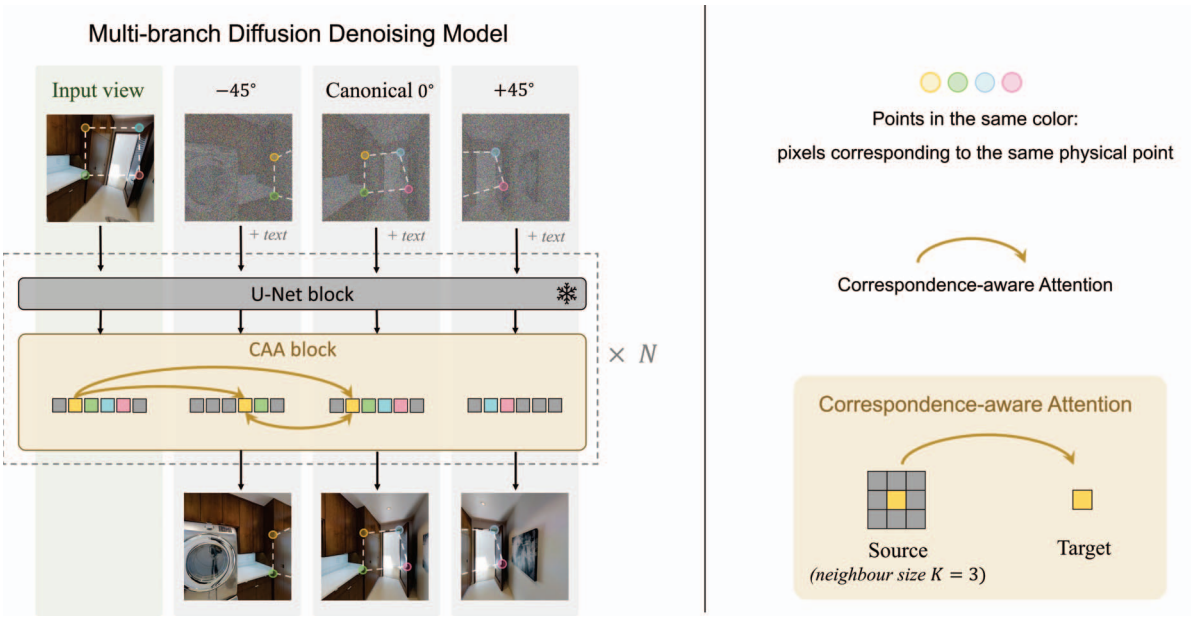

Figure 5. Our panorama generation pipeline based on multi-view diffusion denoising model. With the predicted homography matrix from the input view to a predefined canonical view, point-wise information can be aggregated from the input view to all target canonical views through correspondence-aware attention(CAA). Note that this figure only shows the cross-attention between one group of corresponding points for clear visualization. However, the same process is applied to all groups of corresponding points.

4.2. Camera-free 360-Degree Image Outpainting

Homography matrix from the camera-free input image to a predefined canonical image provides pixel-level correspondences between them. Consider the homography transformation from view I_a to view I_b as $H_{a,b}$. The projection from a point at location $p_a = (u_a,\ v_a)$ in view I_a and its corresponding point at location $p_b = (u_b,\ v_b)$ in view I_b is formulated as:

$$\begin{bmatrix} u_b \\ v_b \\ 1 \end{bmatrix} = H_{a,b} \begin{bmatrix} u_a \\ v_a \\ 1 \end{bmatrix} \tag{5}$$

Based on the estimated correspondences, we design three variants to generate 360° panoramic images, as shown in Figure 4.

Variant 1 (unwarp image): The most straightforward way of utilization the predicted homography transformation is to unwarp and rectify the input image. As depicted in Fig. Figure 4(a), the input image is first transformed to a canonical view using the predicted homography. Subsequently, the latent diffusion inpainting model generates eight panoramic views (eight diffusion branches with the same shared weight) conditioned on the rectified input image, following the design of MVDiffusion [28] image-based panorama outpainting model. For the diffusion branch of canonical view 0°, input comprises a concatenation of noisy latent, the latent of the rectified input, and a binary mask that identifies the area requiring inpainting. In the binary mask, *zero* indicates the visible region and *one* otherwise. The inputs for the remaining seven diffusion branches consist of the noisy latent, the latent of a pure white image, and a uniformly one-valued mask indicating that all locations are invisible. It is important to note that the latent diffusion inpainting model is designed to preserve existing image content where the mask value is set to *zero* while generating new content in areas where the mask value is *one*.

Variant 2 (unwarp latent): Unlike Variant 1 in which the input image is first rectified by unwarping and then encoded into a latent, in Variant 2, we explore another strategy that first encodes the original input into the latent space, followed by unwarping the latent feature map into the canonical view 0°. Aside from this distinction, all other aspects related to multi-view image diffusion remain consistent with those outlined for Variant 1. See details of Variant 2 in Figure 4(b).

Variant 3 (new view): Variant 3 introduces a novel approach that diverges from the straightforward strategies that involve unwarping on the image space or the latent space employed in variants 1 and 2. This design is structured into one additional condition branch and eight standard generation branches. In Variants 1 and 2, the branch of canonical view 0° is conditioned on the rectified image or latent. In Variant 3, however, the canonical view branch functions as a standard generation branch without any image conditioning. Instead, the additional condition branch depends on the original input image. Inputs for all other branches align with

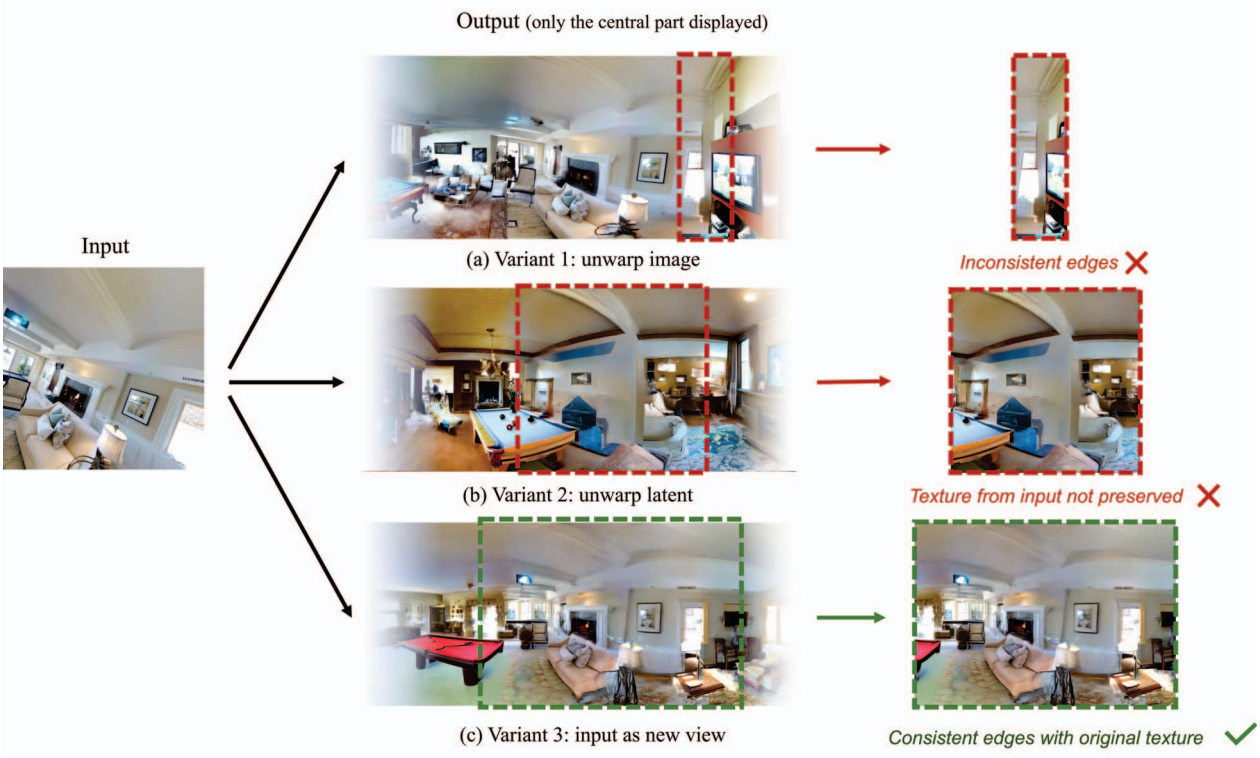

Figure 6. Qualitative comparison between different approaches for integrating camera prediction and multi-view image generation. For hard samples, as shown above, The approaches involves unwarping image space or latent space (Variant 1 and 2) resulting in different problems. In contrast, the seamless integration by Variant 3 produces a consistent output layout and preserves the original texture from the input.

those specified in Variant 1. Particularly, Variant 3 ensures consistency across views by implementing CAA not only between the standard generation branches but also between the condition branch and the generation branches of canonical views. This approach effectively reduces the inaccuracies in homography estimation and results in superior generation quality, including smooth edges and better preservation of the original input textures. We also illustrate the Variant 3 in Figure 4(c).

5. Experiments

We provide experiment details and results in this section, including experiment setup (5.1), baselines (5.2), and qualitative and experiment results (5.3).

5.1. Experiment setup

Dataset. We fine-tune our CamFreeView model on the real-world Matterport3D[3] dataset, which contains 90 building-scale indoor scenes with 10,912 high-resolution panoramic images. 9820 and 1092 images are split for training and evaluation, respectively, following MVDiffusion [28]. Each room in the dataset provides six distinct non-overlapping perspective images taken from identical camera positions, while each has a 90° horizontal and vertical field of view. To reach our goal of learning 360° image-to-panorama outpainting from camera-free input, we apply a random warp on each input perspective image with a field of view ranging from 60° to 110° and camera rotations from $\pm 15^\circ$ to create camera-free inputs. After random warping, The input image is resized in 512×512 resolution. Our model outputs 8 perspective views, each with 512×512 resolution and has a field of view of 90° with a horizontal 45° overlap. They are combined and transferred into a single horizontal 360° panoramic image.

In addition to the primary dataset, we conducted further generalization testing on the Structured3D[36] dataset, a photo-realistic compilation of 3,500 indoor scenes encompassing 21,835 rooms, each rendered with panoramic images. Perspective images for each room are also generated from random camera positions and poses. This step aims to rigorously assess the model's generalization abilities to out-of-domain data. We applied the same random warp approach in Structured3D as in Matterport3D.

We use BLIP-2 [10] captioning model to generate per-view text descriptions for both datasets mentioned above.

Training Details. Our CamFreeDiff model is fine-tuned from the stable diffusion image inpainting model. We retain the weights of VAE image encoder/decoder and the latent

Figure 7. Qualitative results of CamFreeDiff with Variant 3(new view). It prevents the latent diffusion model from modifying the texture from original input image and generates consistent edges and reasonable scene layout for panorama outpainting.

Model	FID ↓	IS ↑	CS↑	PSNR↑
PanoDiffusion	48.7	3.1	-	-
MVDiffusion	42.4	5.4	21.9	-
CamFreeDiff				
+ unwarp image	35.2	5.5	23.6	18.7
+ unwarp latent	34.3	5.6	22.4	15.6
+ new view	**27.0**	**5.6**	**24.4**	**19.3**

Table 1. We report the comparisons on panorama generation performance on Matterport3D dataset. For baseline PanoDiffusion and MVDiffusion, we use our homography predictor to rectify the camera-free input image first, and then evaluate their panorama generations only. Therefore, PSNR is not reported for baselines.

Model	Train	Cam free	FID↓	IS↑	CS↑	PSNR↑
PanoDiffusion	✓	✗	35.3	3.2	-	-
CamFreeDiff	✗	✓	**31.1**	**6.2**	25.5	19.8

Table 2. Our CamFreeDiff tests directly on the Structured3D dataset without applying fine-tuning or domain transfer techniques to demonstrate the generalization ability. In contrast, PanoDiffusion is trained on Structured3D without text guidance. We provide the rectified input image predicted by our Homography Estimator to PanoDiffusion, therefore no CLIP score (CS) or PSNR evaluation provided for PanoDiffusion.

denoising U-Net blocks frozen as pre-trained. Only the MLP block for homography prediction and the CAA blocks for multi-view consistency are trained with a learning rate 2×10^{-4} for 30 epochs.

Evaluation Metrix. Our text-guided 360° image outpainting employs a series of standard image generation metrics to evaluate visual quality: Frechet Inception Distance (FID)[8], which quantifies the distance between real and generated images; Inception Score (IS)[21] offering insight into the diversity and quality of generated images; and CLIP score[15] to measure the alignment between text description and corresponding images. In addition, we use the Peak signal-to-noise ratio (PSNR) on the corresponding region between the generated and target canonical view 0° to evaluate our view estimation error. We also use Mean Absolute Error to assess the accuracy of homography estimation only.

5.2. Baselines

We consider the following baselines: *MVDiffusion* and *PanoDiffusion*. *MVDiffusion* is a multi-view text-to-image diffusion model to generate view-consistent 360° scenes. In contrast, our approach enables panoramic image generation from camera-free input with unknown camera parameters. *PanoDiffusion* is designed for RGB-D panorama outpainting from input images with different types of masks. A super-resolution model further enhances the outpainting results with higher resolution.

To the best of our knowledge, no existing panorama generation model was designed or has learned to handle input

Model	FID↓	IS↑	CS↑	PSNR↑
HomographyNet	29.2	5.6	24.2	19.2
SD encoder(frozen) + MLP	**27.0**	5.6	**24.4**	**19.3**

Table 3. Comparison of the architecture design for homography estimator. Using a three-layer MLP after the existing frozen Stable Diffusion encoder as the homography estimator performs better than using an independent HomographyNet.

camera parameters including different camera poses (rotation) and focal lengths (field of view). Therefore, by default, we provide all baselines with rectified input images using the predicted camera parameters by our homography estimator for a fair comparison of the image generation quality. Importantly, we also compare and analyze the three variants of approaches that combine camera prediction with diffusion-based image generation.

5.3. Results

5.3.1. Qualitative Results

We report qualitative results for different variants of outpainting strategies as introduced in 4.2. Qualitative comparison for a hard sample is shown in Figure 6. We observe that the homography prediction error can be propagated into later panorama generation for Variant 1(unwarping image), resulting in inconsistent scene layout. For Variant 2(unwarping latent), inaccurate homography prediction could lead to changed textures for the input view. Compared to the above variants, introducing a new conditional branch for the input view and leveraging correspondence-aware attention(Variant 3) enables the most consistent and robust multi-view generation. We provide more qualitative results for CamFreeDiff in Appendix Figure 7.

5.3.2. Quantitative Results

We show the numerical results for panorama generation between baselines and different variants of our CamFreeDiff model in Table 1. Among them, CamFreeDiff with Variant 3, treating the input as a new view, achieved the best results in terms of visual quality for panorama generation (FID, IS, CS) and reconstruction quality (PSNR).

To demonstrate the generalization ability of our model, we also tested our model on the Structured3D[36] dataset. Note that our model was never trained on or applied with domain transfer techniques from Structured3D. Results shown in Table 2 indicate the strong generalization ability of CamFreeDiff to out-of-domain data, even surpassing PanoDiffusion, which is trained directly on Structured3D but without learning from camera-free input.

Neighbour Size	FID↓	IS↑	CS↑	PSNR↑
$K = 1$	66.4	5.5	22.3	19.2
$K = 3$	51.1	5.8	23.9	**19.5**
$K = 5$	48.0	**6.2**	**24.1**	19.4
$K = 7$	**47.3**	6.1	**24.1**	19.4

Table 4. Ablation study on the neighbour size K in Correspondence-aware attention(CAA) blocks for information aggregation between two views. Larger neighbour size K aggregates more information from the source view to the target view, but is less efficient for cross-attention operation.

5.4. Ablation Study

5.4.1. Homography Estimator

We ablate on the architecture choise of the homography estimator. We compared our design of building an MLP block on a frozen stable diffusion image encoder with the HomographyNet, which is designed to predict homography matrix between views. Our designs achieve better generation results as shown in Table 3. We also compared classification and regression losses in the Appendix.

5.4.2. Neighborhood Size in CAA blocks

Given correspondences between views, Correspondence-aware attention (CAA) aggregates information from source point p_s neighborhood to target point p_t, which is the key to yielding consistency between multiple views. The neighborhood in CAA refers to the $K \times K$ neighboring points centered at p_s. We ablate the neighborhood size K for $K = 1, 3, 5, 7$. From results shown in Table 4, we find larger neighborhood size generally leads to better multi-view generation quality, but the improvement is limited. Note that larger K results in more computational and time complexity for CAA operation.

6. Conclusion

We introduce CamFreeDiff, a novel diffusion-based method for generating panoramic image from a camera-free input. We formulate the camera parameter estimation of the input as an estimation of the homography transformation from the input view to a predefined canonical view. Our method builds upon the MVDiffusion model for consistent multi-view image generation and seamlessly incorporates correspondences between the input and target canonical views for coherent and high-quality panorama generation. We also demonstrate the strong robustness of CamFreeDiff to camera-free inputs and generalization ability to out-of-domain data.

Acknowledgement. We thank Professor Alan L. Yuille for the support of this work and the authors. This paper is supported by Office of Naval Research (ONR) with N000142412696.

References

[1] Naofumi Akimoto, Seito Kasai, Masaki Hayashi, and Yoshimitsu Aoki. 360-degree image completion by two-stage conditional gans. In *2019 IEEE International Conference on Image Processing (ICIP)*, pages 4704–4708. IEEE, 2019. 2

[2] Naofumi Akimoto, Yuhi Matsuo, and Yoshimitsu Aoki. Diverse plausible 360-degree image outpainting for efficient 3dcg background creation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 11441–11450, 2022. 2

[3] Angel Chang, Angela Dai, Thomas Funkhouser, Maciej Halber, Matthias Niessner, Manolis Savva, Shuran Song, Andy Zeng, and Yinda Zhang. Matterport3d: Learning from rgb-d data in indoor environments. *arXiv preprint arXiv:1709.06158*, 2017. 2, 6

[4] Zhaoxi Chen, Guangcong Wang, and Ziwei Liu. Text2light: Zero-shot text-driven hdr panorama generation. *ACM Transactions on Graphics (TOG)*, 41(6):1–16, 2022. 2

[5] Daniel DeTone, Tomasz Malisiewicz, and Andrew Rabinovich. Deep image homography estimation. *arXiv preprint arXiv:1606.03798*, 2016. 3

[6] Stefan Elfwing, Eiji Uchibe, and Kenji Doya. Sigmoid-weighted linear units for neural network function approximation in reinforcement learning. *Neural networks*, 107:3–11, 2018. 4

[7] Ian Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. Generative adversarial networks. *Communications of the ACM*, 63(11):139–144, 2020. 2

[8] Martin Heusel, Hubert Ramsauer, Thomas Unterthiner, Bernhard Nessler, and Sepp Hochreiter. Gans trained by a two time-scale update rule converge to a local nash equilibrium. *Advances in neural information processing systems*, 30, 2017. 7

[9] Jonathan Ho, Ajay Jain, and Pieter Abbeel. Denoising diffusion probabilistic models. *Advances in neural information processing systems*, 33:6840–6851, 2020. 2

[10] Junnan Li, Dongxu Li, Silvio Savarese, and Steven Hoi. Blip-2: Bootstrapping language-image pre-training with frozen image encoders and large language models. *arXiv preprint arXiv:2301.12597*, 2023. 6

[11] Ruoshi Liu, Rundi Wu, Basile Van Hoorick, Pavel Tokmakov, Sergey Zakharov, and Carl Vondrick. Zero-1-to-3: Zero-shot one image to 3d object. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 9298–9309, 2023. 2

[12] Yuan Liu, Cheng Lin, Zijiao Zeng, Xiaoxiao Long, Lingjie Liu, Taku Komura, and Wenping Wang. Syncdreamer: Generating multiview-consistent images from a single-view image. *arXiv preprint arXiv:2309.03453*, 2023. 2

[13] Xiaoxiao Long, Yuan-Chen Guo, Cheng Lin, Yuan Liu, Zhiyang Dou, Lingjie Liu, Yuexin Ma, Song-Hai Zhang, Marc Habermann, Christian Theobalt, et al. Wonder3d: Single image to 3d using cross-domain diffusion. *arXiv preprint arXiv:2310.15008*, 2023. 2

[14] Changgyoon Oh, Wonjune Cho, Yujeong Chae, Daehee Park, Lin Wang, and Kuk-Jin Yoon. Bips: Bi-modal indoor panorama synthesis via residual depth-aided adversarial learning. In *European Conference on Computer Vision*, pages 352–371. Springer, 2022. 2

[15] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. Learning transferable visual models from natural language supervision. In *International conference on machine learning*, pages 8748–8763. PMLR, 2021. 7

[16] Aditya Ramesh, Prafulla Dhariwal, Alex Nichol, Casey Chu, and Mark Chen. Hierarchical text-conditional image generation with clip latents. *arXiv preprint arXiv:2204.06125*, 1(2):3, 2022. 2

[17] Aditya Ramesh, Mikhail Pavlov, Gabriel Goh, Scott Gray, Chelsea Voss, Alec Radford, Mark Chen, and Ilya Sutskever. Zero-shot text-to-image generation. In *International Conference on Machine Learning*, pages 8821–8831. PMLR, 2021.

[18] Scott Reed, Zeynep Akata, Xinchen Yan, Lajanugen Logeswaran, Bernt Schiele, and Honglak Lee. Generative adversarial text to image synthesis. In *International conference on machine learning*, pages 1060–1069. PMLR, 2016.

[19] Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. High-resolution image synthesis with latent diffusion models. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 10684–10695, 2022. 2

[20] Chitwan Saharia, William Chan, Saurabh Saxena, Lala Li, Jay Whang, Emily L Denton, Kamyar Ghasemipour, Raphael Gontijo Lopes, Burcu Karagol Ayan, Tim Salimans, et al. Photorealistic text-to-image diffusion models with deep language understanding. *Advances in Neural Information Processing Systems*, 35:36479–36494, 2022. 2

[21] Tim Salimans, Ian Goodfellow, Wojciech Zaremba, Vicki Cheung, Alec Radford, and Xi Chen. Improved techniques for training gans. *Advances in neural information processing systems*, 29, 2016. 7

[22] Ruoxi Shi, Hansheng Chen, Zhuoyang Zhang, Minghua Liu, Chao Xu, Xinyue Wei, Linghao Chen, Chong Zeng, and Hao Su. Zero123++: a single image to consistent multi-view diffusion base model. *arXiv preprint arXiv:2310.15110*, 2023. 2

[23] Yichun Shi, Peng Wang, Jianglong Ye, Mai Long, Kejie Li, and Xiao Yang. Mvdream: Multi-view diffusion for 3d generation. *arXiv preprint arXiv:2308.16512*, 2023. 2

[24] Gowri Somanath and Daniel Kurz. Hdr environment map estimation for real-time augmented reality. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 11298–11306, 2021. 2

[25] Gabriela Ben Melech Stan, Diana Wofk, Scottie Fox, Alex Redden, Will Saxton, Jean Yu, Estelle Aflalo, Shao-Yen Tseng, Fabio Nonato, Matthias Muller, et al. Ldm3d: Latent diffusion model for 3d. *arXiv preprint arXiv:2305.10853*, 2023. 2

[26] Julius Surya Sumantri and In Kyu Park. 360 panorama synthesis from a sparse set of images with unknown field of view. In *Proceedings of the IEEE/CVF winter conference on applications of computer vision*, pages 2386–2395, 2020. 2

[27] Shitao Tang, Jiacheng Chen, Dilin Wang, Chengzhou Tang, Fuyang Zhang, Yuchen Fan, Vikas Chandra, Yasutaka Furukawa, and Rakesh Ranjan. Mvdiffusion++: A dense high-resolution multi-view diffusion model for single or sparse-view 3d object reconstruction. *arXiv preprint arXiv:2402.12712*, 2024. 2

[28] Shitao Tang, Fuayng Zhang, Jiacheng Chen, Peng Wang, and Furukawa Yasutaka. Mvdiffusion: Enabling holistic multi-view image generation with correspondence-aware diffusion. *arXiv preprint 2307.01097*, 2023. 2, 5, 6

[29] Aaron Van den Oord, Nal Kalchbrenner, Lasse Espeholt, Oriol Vinyals, Alex Graves, et al. Conditional image generation with pixelcnn decoders. *Advances in neural information processing systems*, 29, 2016. 2

[30] Jionghao Wang, Ziyu Chen, Jun Ling, Rong Xie, and Li Song. 360-degree panorama generation from few unregistered nfov images. *arXiv preprint arXiv:2308.14686*, 2023. 3

[31] Peng Wang and Yichun Shi. Imagedream: Image-prompt multi-view diffusion for 3d generation. *arXiv preprint arXiv:2312.02201*, 2023. 2

[32] Tianhao Wu, Chuanxia Zheng, and Tat-Jen Cham. Panodiffusion: 360-degree panorama outpainting via diffusion. In *The Twelfth International Conference on Learning Representations*, 2023. 2

[33] Zhenqiang Ying and Alan Bovik. 180-degree outpainting from a single image. *arXiv preprint arXiv:2001.04568*, 2020. 2

[34] Chuanxia Zheng, Tat-Jen Cham, and Jianfei Cai. Pluralistic image completion. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 1438–1447, 2019. 2

[35] Chuanxia Zheng, Tat-Jen Cham, Jianfei Cai, and Dinh Phung. Bridging global context interactions for high-fidelity image completion. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 11512–11522, 2022. 2

[36] Jia Zheng, Junfei Zhang, Jing Li, Rui Tang, Shenghua Gao, and Zihan Zhou. Structured3d: A large photo-realistic dataset for structured 3d modeling. In *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part IX 16*, pages 519–535. Springer, 2020. 2, 6, 8