Generative Panoramic Image Stitching

Mathieu Tuli*
LG Electronics
tuli.mathieu@gmail.com

Kaveh Kamali*
LG Electronics
kaveh.kamali@gmail.com

David B. Lindell
LG Electronics
University of Toronto
lindell@cs.toronto.edu

*joint first authors

Figure 1. We introduce a generative method for panoramic image stitching from multiple casually captured reference images that exhibit strong parallax, lighting variation, and style differences. Our approach fine-tunes an inpainting diffusion model to match the content and layout of the reference images. After fine-tuning, we outpaint one reference image (e.g., the leftmost reference image shown here) to create a seamless panorama that incorporates information from the other views. Unlike prior methods such as RealFill [60], which generates content that is inconsistent with the reference images, our method more accurately preserves scene structure and spatial composition (red boxes).

Abstract

We introduce the task of generative panoramic image stitching, which aims to synthesize seamless panoramas that are faithful to the content of multiple reference images containing parallax effects and strong variations in lighting, camera capture settings, or style. In this challenging setting, traditional image stitching pipelines fail, producing outputs with ghosting and other artifacts. While recent generative models are capable of outpainting content consistent with multiple reference images, they fail when tasked with synthesizing large, coherent regions of a panorama. To address these limitations, we propose a method that fine-tunes a diffusion-based inpainting model to preserve a scene's content and layout based on multiple reference images. Once fine-tuned, the model outpaints a full panorama from a single reference image, producing a seamless and visually coherent result that faithfully integrates content from all reference images. Our approach significantly outperforms baselines for this task in terms of image quality and the consistency of image structure and scene layout when evaluated on captured datasets.

1. Introduction

Creating a coherent visual representation from multiple input images is a long-standing problem in computer vision [57, 58], and many techniques have been proposed to combine multiple images from different perspectives to synthesize panoramas [8], multi-perspective images [1, 47, 54], or photo montages [2, 45]. More recently, image generation models make it possible to render or outpaint new image content based on one or more input images [53, 60]. Inspired

by methods for panorama synthesis and recent image generation techniques, we propose to address the task of *generative panoramic image stitching*—i.e., we seek to generate seamless panoramas that are faithful to the content of multiple reference images captured from different viewpoints with strong variations in lighting or style (Figure 1). We focus on the regime where correspondences in the input images exist, but conventional image stitching fails due to parallax [43, 57, 67, 74]. Panoramas in this work are defined as wide-angle representation of a scene, not 360° Panorama or Spherical panoramas [8, 58, 66].

A standard approach for panoramic image stitching involves detecting feature correspondences and estimating geometric transformations between input images [8]. Then, the input images are warped based on the estimated transformation and blended together into a panorama [10, 48]. Conventionally, these techniques use a homography to relate input images, which assumes that there is no parallax (i.e., no translation between captured viewpoints) [57]. Violating this assumption results in artifacts, such as ghosting [17], as shown in Figure 2 (top). Hence, a significant amount of effort has been devoted to improving robustness to viewpoint changes, e.g., by optimizing local warping operations [14, 20, 29, 32–35, 73, 74], by using graph cuts to minimize seams between blended images [17, 21, 74], or optimizing neural networks [42, 44], but completely avoiding artifacts is challenging when images are captured from significantly different positions. Further, standard techniques for image stitching assume that camera acquisition settings and illumination conditions are roughly constant across input images; while image blending can help to mitigate small variations in camera gain, exposure, white balance, or scene illumination [8, 10, 48] it fails to handle strong variations in the lighting or style of input images (Figure 2, bottom).

A separate line of work seeks to create panoramic images via image synthesis. For example, using generative models, recent approaches synthesize panoramas from a text prompt [6, 18, 30, 72] or inpaint masked regions of an input panorama [66]; however, these methods do not handle the stitching of reference images with overlapping fields of view and significant parallax effects. The recent work of Tang et al. [60] uses a pre-trained generative model for reference-guided inpainting, which is close to our task. Specifically, they fine-tune an image diffusion model to inpaint a set of casually captured reference images from different viewpoints and lighting conditions. After fine-tuning, the model can be used to outpaint an existing image in a way that is consistent with the content of the reference images and robust to parallax or lighting variations. However, we find that using this approach for panoramic image stitching fails, as outpainting large missing regions results in artifacts and scene layouts that are not faithful to the reference images (see Figure 1).

Here, we address limitations of conventional methods

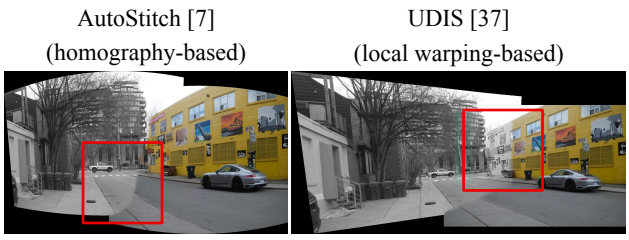

Figure 2. Conventional panoramic image stitching methods [8, 44] fail to account for strong parallax or variations in lighting or style.

for panoramic image stitching as well as more recent, reference-driven outpainting techniques [60]. Given a set of casually captured reference images, we first compute a coarse alignment of the images via conventional feature matching and homography estimation [8], resulting in a set of warped images and their approximate locations on an initial panorama. To correct artifacts in this initial panorama—such as those caused by parallax or lighting inconsistencies—we fine-tune [53] a large, pre-trained inpainting diffusion model [4] to solve a position-aware inpainting task. Specifically, we fine-tune the model to inpaint and outpaint each warped input image while conditioning on positional encodings that reflect the image's location within the panorama. Once fine-tuned, the model is used to iteratively outpaint the panorama from a single reference image, resulting in a seamless composite that integrates content from all references as shown in Figure 1.

In summary, we make the following contributions.

- We propose the task of *generative panoramic image stitching*, which seeks to generate panoramas that are faithful to a set of reference images containing significant parallax effects and variations in illumination or style.
- We address this task with a method that estimates the coarse layout of the reference images within a panorama and then fine-tunes a diffusion model to generate a seamless output panorama via position-aware outpainting.
- We evaluate our approach on a dataset of captured images and show state-of-the-art results for this task compared to baselines based on reference-driven image outpainting and image stitching.

2. Related Work

Our work also connects to other methods for learning-based image stitching, multi-perspective rendering, 3D reconstruction, and reference-driven outpainting.

Learning-based image stitching. While conventional image stitching pipelines rely on feature-based homography estimation [57], recent approaches instead learn to regress a homography directly using a neural network [16, 28, 41] or learned feature representations [75], which can improve performance for dynamic scenes images with limited texture. Nie et al.[42, 44] propose a learning-based pipeline

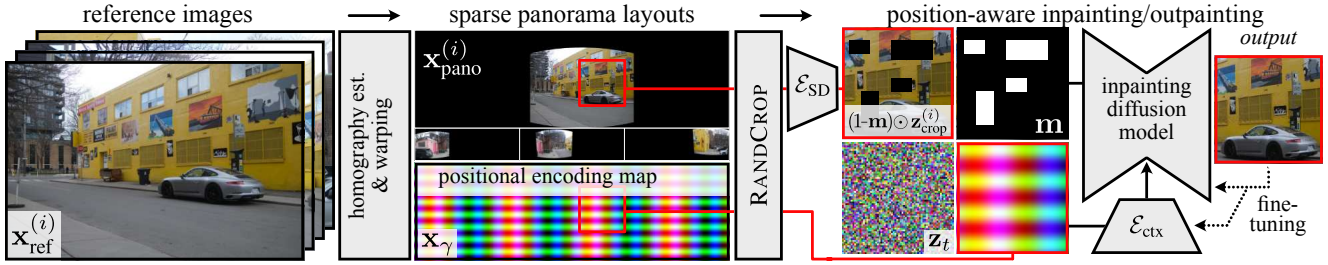

Figure 3. Method overview. Given a set of reference images $\{\mathbf{x}_{\text{ref}}^{(i)}\}_{i=1}^{N}$, we generate sparse panorama layouts $\{\mathbf{x}_{\text{pano}}^{(i)}\}_{i=1}^{N}$ by detecting features [38], estimating homographies, and warping each reference image to its location in a sparse panorama containing only that image. We then fine-tune a pre-trained inpainting diffusion model for a position-aware inpainting/outpainting task. During training, random crops are taken from the sparse panoramas and a positional encoding map $\mathbf{x}_\gamma$. Each panorama crop is processed using an encoder $\mathcal{E}_{\text{SD}}$, and we we multiply the resulting latent image $\mathbf{z}_{\text{crop}}^{(i)}$ with a random binary mask $(1-\mathbf{m})$. We process the crop of $\mathbf{x}_\gamma$ with an encoder $\mathcal{E}_{\text{ctx}}$ and use the result to condition the diffusion model. The other inputs — the masked version of $\mathbf{z}_{\text{crop}}^{(i)}$, the mask $\mathbf{m}$, and the noisy latent image $\mathbf{z}_t$ — are concatenated together and passed as input to the model. After fine-tuning, we generate seamless panoramas by outpainting one of the initial sparse panoramas.

that predicts a homography or warping mesh [43] between two input images, followed by a transformer- or thin-plate-spline–based refinement to reduce stitching artifacts. More recently, diffusion-based approaches [67, 68] have leveraged pre-trained image generation models for image stitching. However, these methods are limited to processing two input images at a time and must rely on sequential registration for multi-image panoramas—a process that can accumulate misalignments and introduce artifacts. Here, we address a different setting than previous work: given multiple reference images, potentially with strong parallax or lighting variations, we jointly fine-tune a diffusion model on all inputs to generate a panorama that remains globally consistent and faithful to the structure of the input images.

Multi-perspective rendering and 3D reconstruction. It is also possible to synthesize panoramas using image-based rendering [1, 5, 36, 45, 50]. Given a sufficiently densely captured set of input images, one can directly capture or estimate the desired set of light rays used to assemble an output panorama or multi-perspective image [5, 7, 31, 51]. Alternatively, one can reconstruct a 3D representation of the scene and render novel views from any desired viewpoint [9, 15, 40, 50, 65]. Still, these techniques cannot be easily applied to our proposed task, where only a few images are provided as input, camera poses are unknown, and the images have inconsistencies, e.g., due to variations in camera capture settings, color palette, or lighting. Finally, while a number of recent techniques propose to generate panoramas from a single image or text prompt [11, 62, 69], our work addresses a different problem—we seek to generate panoramas that are faithful to *multiple* input images of a scene captured from varying viewpoints.

Reference-driven image editing. Rather than directly stitching the input images, our approach synthesizes a panorama by *outpainting* one of the input views using content from the others. This design is inspired by prior work on reference-driven inpainting, where masked regions of an image are filled using information from a reference view of the same scene [77]. Recent methods further improve performance on this task by leveraging pre-trained image diffusion models [37, 70], possibly in conjunction with image correspondences [12]. Most related to our approach, Tang et al. [60] fine-tune a diffusion model for reference-guided outpainting; however, their method does not incorporate scene layout information and therefore fails to generalize to panorama synthesis (see Fig. 1).

3. Generative Panoramic Image Stitching

Here, we describe our method for generative panoramic image stitching based on (1) initial panorama layout estimation via homography estimation and warping, (2) fine-tuning a diffusion model for position-aware panorama inpainting and outpainting, and (3) generating a seamless panorama via iterative outpainting. An overview is shown in Figure 3.

3.1. Layout Estimation & Positional Encoding

Given a set of N input reference images $\{\mathbf{x}_{\text{ref}}^{(i)}\}_{i=1}^{N}$, where $\mathbf{x}_{\text{ref}}^{(i)} \in \mathbb{R}_{+}^{H_{\text{ref}} \times W_{\text{ref}} \times 3}$, we aim to generate a panorama $\mathbf{x}_{\text{pano}} \in \mathbb{R}^{H_{\text{pano}} \times W_{\text{pano}} \times 3}$ via latent diffusion that seamlessly stitches together scene content from the reference views and outpaints uncaptured scene regions.

The first step in this procedure involves producing an initial panorama layout via homography estimation and warping. We adapt the procedure of Brown et al. [8] to detect feature correspondences between the input images, estimate homographies, and warp each image into cylindrical coordinates. The result of this procedure is a set of sparse panoramas $\{\mathbf{x}_{\text{pano}}^{(i)}\}_{i=1}^{N}$, $\mathbf{x}_{\text{pano}}^{(i)} \in \mathbb{R}_{+}^{H_{\text{pano}} \times W_{\text{pano}} \times 3}$, which each contains a single warped reference image (see Figure 3).

We also associate the panorama with a positional encoding map $\mathbf{x}_\gamma$ [40] computed using a function $\gamma(p) = [\cos(\pi f_1 p), \sin(\pi f_1 p), \ldots, \cos(\pi f_F p), \sin(\pi f_F p)]^T$, where $\{f_i\}_{i=1}^{F}$ are the encoding frequencies, and the

function $\gamma(\cdot)$ is applied to each vertical and horizontal pixel coordinate p. Encoding each pixel coordinate results in $\mathbf{x}_\gamma \in \mathbb{R}^{H_{\text{pano}} \times W_{\text{pano}} \times 4F}$, where F is the number of positional encoding frequencies. See Supp. Section S1.1.2 for details.

3.2. Position-aware Inpainting and Outpainting

We use the set of panoramas $\{\mathbf{x}^{(i)}_{\text{pano}}\}_{i=1}^{N}$ and the positional encoding map $\mathbf{x}_\gamma$ to fine-tune an inpainting latent diffusion model [52] for position-aware inpainting and outpainting.

Architecture. Our approach adapts a pre-trained inpainting diffusion model $\Psi(\mathbf{z}_t, t, \mathcal{C})$ with noisy latent $\mathbf{z}_t$, diffusion timestep t, and conditioning signals $\mathcal{C}$ (we use Stable Diffusion 2.1 [4]). The model is conditioned on the input

$$\mathcal{C} = \{\mathbf{m}, (1-\mathbf{m}) \odot \mathbf{z}^{(i)}_{\text{crop}}, \mathbf{c}_{\text{ctx}}\}, \tag{1}$$

where $\mathbf{m}$ is a randomly generated binary mask to be inpainted or outpainted, $\odot$ indicates Hadamard product, and $\mathbf{z}^{(i)}_{\text{crop}}$ is a randomly cropped region of $\mathbf{x}^{(i)}_{\text{pano}}$ that we encode into the latent space using the Stable Diffusion encoder $\mathcal{E}_{\text{SD}}$, or $\mathbf{z}^{(i)}_{\text{crop}} = \mathcal{E}_{\text{SD}}(\textsc{RandCrop}(\mathbf{x}^{(i)}_{\text{pano}}))$. The context embedding tensor $\mathbf{c}_{\text{ctx}}$ is produced as $\mathbf{c}_{\text{ctx}} = \mathcal{E}_{\text{ctx}}(\textsc{RandCrop}(\mathbf{x}_\gamma))$, where we apply the same random crop to the positional encoding map $\mathbf{x}_\gamma$ as for $\mathbf{x}^{(i)}_{\text{pano}}$. We process the cropped version of $\mathbf{x}_\gamma$ using $\mathcal{E}_{\text{ctx}}$, a small three-layer convolutional encoder with a linear layer (see Supp. Section S1.1.3).

While the context embedding tensor $\mathbf{c}_{\text{ctx}}$ is used by the pre-trained model for text conditioning, our approach repurposes it to encode the positional information, and we provide the tensor as input to the cross-attention layers of the network. The other conditioning signals (i.e., $\mathbf{m}$ and $(1-\mathbf{m}) \odot \mathbf{z}^{(i)}_{\text{crop}}$) are concatenated with the noisy latent image $\mathbf{z}_t$ and passed as input to the diffusion model. More details of the architecture are provided in Supp. Section S1.1.

Optimization. We fine-tune the network jointly across all reference input images to learn the spatial layout of the panorama. Specifically, we minimize the loss function

$$\mathcal{L} = \mathbb{E}_{\boldsymbol{\epsilon}, \mathbf{z}^{(i)}_{\text{crop}}, i, t, \mathbf{m}} \left[\left\| \mathbf{m}_{\text{valid}} \odot \left(\Psi(\mathbf{z}^{(i)}_{\text{crop},t}, t, \mathcal{C}) - \boldsymbol{\epsilon} \right) \right\|_2^2 \right], \tag{2}$$

where $\boldsymbol{\epsilon}$ is noise added in the diffusion process, and $\mathbf{m}_{\text{valid}}$ is a binary mask that restricts the loss to regions of $\mathbf{z}^{(i)}_{\text{crop}}$ that correspond to non-empty areas in the cropped sparse panorama $\mathbf{x}^{(i)}_{\text{pano}}$. Hence, we fine-tune the model to predict the noise added to $\mathbf{z}^{(i)}_{\text{crop}}$ at timestep t (see Supp. Section S1.1.1 for details on this reverse diffusion process).

We use low-rank adaptation (LoRA) [25] to optimize the model's self-attention layers and preserve the capabilities of the Stable Diffusion model's pre-trained weights. The cross-attention layers undergo full-parameter fine-tuning to better adapt to the positional encoding information provided by $\mathbf{c}_{\text{ctx}}$. Last, we initialize and optimize all parameters of the context encoder $\mathcal{E}_{\text{ctx}}$.

Figure 4. Panorama generation. We use the fine-tuned model to iteratively outpaint each tile (green grid) of the panorama, in order of distance from the center (white arrows), to create a seamless result.

3.3. Panorama Generation

After fine-tuning, we generate a seamless panorama $\mathbf{x}_{\text{pano}}$ by outpainting one of the initial sparse panoramas $\mathbf{x}^{(i)}_{\text{pano}}$. The main challenge in this step is that the resolution of the panorama is much larger than the nominal resolution for which the inpainting diffusion model is trained—so we cannot generate the entire panorama in a single inference pass. Instead, similar to prior work [6, 18, 26, 30, 39], we sequentially denoise tiles of the panorama to generate the final output as depicted in Figure 4. We apply the sequential denoising procedure to the sparse panorama containing a centered warped reference image — this is an arbitrary image to which we register the other reference images during the initial layout estimation process (Section 3.1).

Specifically, we generate an evenly spaced grid of overlapping image tiles or boxes $\{\mathbf{b}^{(i)}\}_{i=1}^{B}$ across the panorama, where $\mathbf{b}^{(i)} = \{x^{(i)}, y^{(i)}, H, W\}$ gives the pixel coordinates of the corner of the tile and the height and width of the tile. In practice, we use 20% overlap between tiles, and we set $H = W = 512$. After positioning the tiles, if some tiles extend beyond the extent of the panorama, the overlap is reduced until all tiles fit within the panorama in both the vertical and horizontal dimensions. For each tile in the grid, we run the full reverse diffusion process using the DDPM sampler [22] to inpaint/outpaint the missing regions of the tile. Inpainting/outpainting masks are feathered and composited with the current state of the generated panorama $\mathbf{x}_{\text{pano}}$. Different than training, where we randomly sample the mask values $\mathbf{m}$, during inference we set the $\mathbf{m}$ values to indicate which regions of each input tile have not yet been generated. The tiles are denoised in order of increasing distance from

their centroids to that of the warped reference image. We provide pseudocode for this procedure in Supp. Section S1.2.

3.4. Implementation Details

Masking and augmentation. Inpainting masks are synthesized with randomly generated patterns following Tang et al. [60]. We also introduce an augmentation scheme which perturbs the location of the warped images in the sparse panoramas with a random similarity transformation. We find that this helps to avoid seams from appearing in the final output panoramas at the boundaries of the warped images.

Correspondence-based seed selection. We employ a correspondence-based seed selection process [60] to identify generated panoramas whose layout matches the result of feature-based image registration [8]. Specifically, we generate ten panoramas with different random seeds and take our output to be the panorama with the most feature matches (computed with LoFTR [56]) compared to the reference. See Supp. Section S1.2 for additional implementation details.

Training and inference. We apply LoRA to the Stable Diffusion model's self-attention layers and fully fine-tune the cross-attention layers and $\mathcal{E}_{\text{ctx}}$, using AdamW with learning rates of 1×10^{-4} (LoRA), 3×10^{-4} (cross-attention), and 8×10^{-4} ($\mathcal{E}_{\text{ctx}}$). Training runs for 4,000 iterations with batch size 32 and takes 4.5 hours on 2×A100 GPUs. At inference, a 1000×3000 panorama typically takes 1 minute to generate on a single RTX 2080 Ti. We use classifier-free guidance [23] with $\mathbf{c}_{\text{ctx}} = 0$ and a guidance scale of 1.5. Additional training details are provided in Supp. Section S1.3.

4. Experiments

Dataset. We evaluate our method on two datasets: a newly collected dataset and a subset of the UDIS dataset [42]. Our new dataset comprises eight scenes, each captured under two settings: a set of *tripod-captured* images obtained by rotating a camera on a tripod, and a set of *casually captured* images acquired from varying viewpoints using a handheld camera (Fujifilm X100 VI).

The tripod-captured dataset, with minimal parallax, aligns with assumptions of standard stitching methods and is used to compute a reference panorama for evaluating image quality. In the casually captured dataset, the distance between viewpoints varies by up to one to two meters, and we also introduce other challenging variations, such as capturing images of the same scene with varying illumination conditions, camera white balance, or image color palette. The casually captured dataset also tests robustness to parallax, illumination, and style variations. A detailed description and number of of captured images for each scene is provided in Supp. Section S1.4. The dataset will be released publicly.

To facilitate comparison across output panoramas, we include one tripod-captured image within the set of casually captured images. We configure our method and all baselines so that this shared image is placed at the center of the output panorama, ensuring a consistent layout across output panoramas from both sets of images.

Baselines. We compare our approach to the following methods (see Supp. Section S1.5 for details).

- *AutoStitch* [8]: A conventional image stitching pipeline using feature matching, homography estimation, warping, and blending. Its bundle adjustment procedure provides robust multi-image alignment and informs our own panorama layout estimation.
- *Stable Diffusion 2 Inpainting* [4]: The backbone of our method, used here without positional encoding or fine-tuning; we follow the same iterative outpainting process as our approach.
- *RealFill* [60]: A reference-guided inpainting model fine-tuned for outpainting. We use their inpainting-based fine-tuning strategy and then perform inference using our iterative outpainting procedure. We use RealFill as it is closest to our task and outperforms other related generative inpainting methods [70, 77].
- *Agisoft+NeRF*: A geometry-based baseline using Agisoft Metashape [3] (commercial photogrammetry software) for camera pose estimation and NeRF [40] for novel view synthesis. We render novel views corresponding to rotating the camera from an initial viewpoint and stitch them together with AutoStitch.
- *RDIStitcher* [67]: A reference-driven diffusion model applied iteratively to stitch overlapping image pairs. To adapt this method to the multi-image setting, we warp the input images onto a panorama canvas, manually crop overlapping regions for image pairs, and iteratively paste the stitched results into the panorama.

Metrics. We evaluate our method using standard image quality metrics, learning-based metrics that assess high-level image structure, and feature-matching-based metrics that assess how well our approach preserves the scene layout. Specifically, we use standard image quality metrics: peak signal-to-noise ratio, structure similarity [64], and learned perceptual image patch similarity [76]. To evaluate high-level image structure, we use DreamSim [19], which assesses similarity in semantic content and layout. We also compute the cosine similarity between the DINO [13] and CLIP [49] full-image embeddings. Additionally, we use image feature matches from LoFTR [56] to assess how well the layout of the output panorama matches a reference. We report both the L2 distance between the pixel coordinates of matching features and the number of matched features divided by the total number of features in the reference image (see Supp. Section S1.6 for more details).

Qualitative results. We show qualitative results on the tripod-captured dataset in Figure 5 and on the casually-

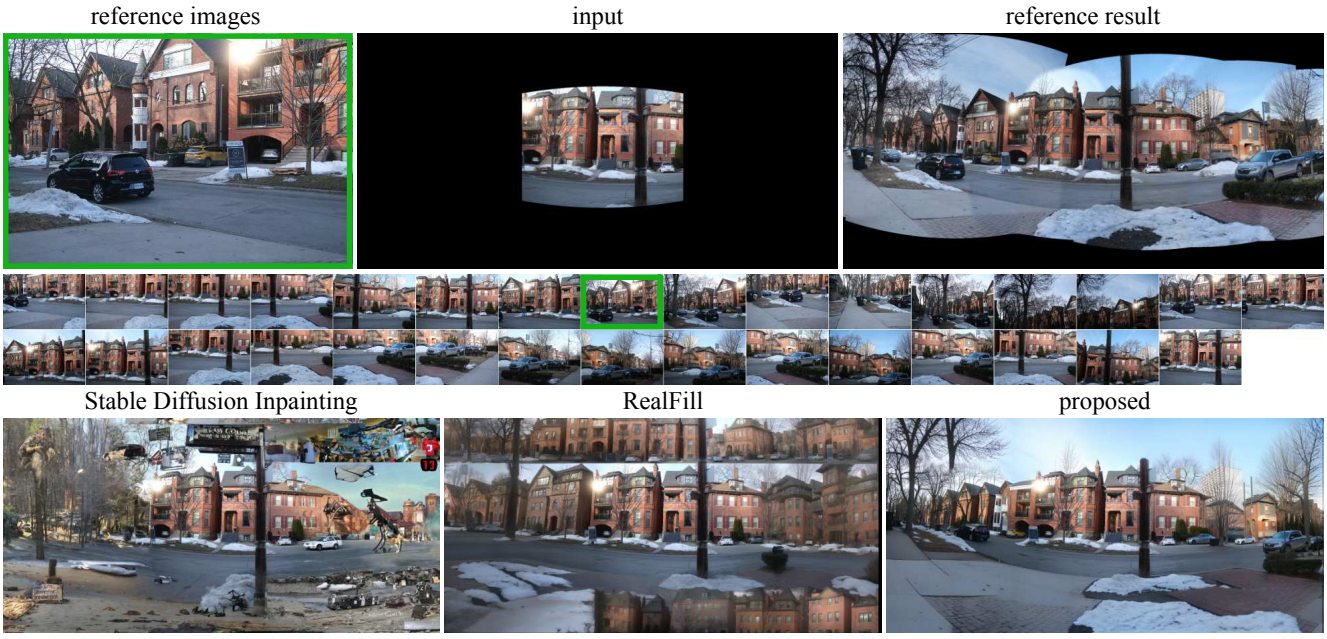

Figure 5. Qualitative results on the tripod-captured dataset. We find that our approach produces panoramas that are more consistent with the layout and content of the reference panorama than baseline approaches based on inpainting/outpainting.

Method	PSNR (dB) ↑	SSIM ↑	LPIPS ↓	DreamSim ↓	DINO ↑	CLIP ↑	LoFTR (L2 Distance) ↓	LoFTR (Matching) ↑
SD2	9.97 (0.74)	0.267 (0.099)	0.650 (0.040)	0.295 (0.050)	0.916 (0.031)	0.859 (0.074)	85.45 (56.00)	0.012 (0.003)
RealFill	11.71 (1.61)	0.366 (0.143)	0.559 (0.069)	0.198 (0.040)	0.952 (0.020)	0.918 (0.048)	43.01 (35.07)	0.030 (0.010)
proposed	12.11 (2.05)	0.388 (0.136)	0.453 (0.077)	0.107 (0.031)	0.974 (0.019)	0.941 (0.033)	15.11 (7.60)	0.181 (0.080)

Table 1. Quantitative assessment of generative panoramic image stitching on the tripod-captured image dataset. We outpaint a single warped reference image (see Figure 5), and compare the generated result to a reference panorama produced using AutoStitch [8]. Our approach generates panoramas that are most faithful to the reference images. Standard deviations are reported in parentheses.

captured image datasets in Figures 1 and 6. For the tripod-captured dataset we observe that the Stable Diffusion inpainting model [4] produces image content that is locally plausible, but fails to adhere to the layout and content of the actual scene. RealFill [60] improves on this result, but tends to repeat scene content from the reference images without respecting the actual scene layout. Our approach provides a much closer match to the layout of the reference panorama while also resolving seams and avoiding ghosting artifacts.

For the casually captured results in Figure 6, we compare to AutoStitch [8], which fails to convincingly blend between the different image regions, resulting in ghosting and other artifacts. We see similar artifacts for RealFill as in the tripod-captured dataset, and we find that our approach produces seamless results that are more consistent with the layout and content of the scene. Additional results for all scenes and Agisoft+NeRF and RDIStitcher baselines are included in Supp. Section S2. We find that the Agisoft+NeRF baseline exhibits artifacts due to the sparse number of viewpoints, and our extension of the RDIStitcher baseline to multi-image stitching results in seams due to the requirement to stitch image pairs sequentially compared to our global optimization.

Quantitative results. We report quantitative results on the tripod-captured and casually captured image datasets in Tables 1 and Tables 2, respectively. For the tripod-captured dataset, we construct a reference panorama using the method AutoStitch [8], which is well-suited to these images, as they have minimal parallax or variations in illumination. We find that the proposed approach generates panoramas that are significantly more consistent with the layout of the reference panorama than the baselines. This trend is clear from the qualitative results as well as the metrics that assess similarity in high-level image structure (e.g., DreamSim, CLIP) and in layout based on feature matching (LoFTR). We note that low-level image quality metrics (e.g., PSNR, SSIM) are perhaps less useful for assessing performance on this task because small variations in layout can produce large changes in the pixel values. Nevertheless, these metrics follow the same trend as the high-level metrics, and are included for completeness.

For the casually captured image dataset, we compare the output of our approach and baselines to the same reference panorama as before (i.e., computed with the tripod-captured dataset). For RealFill and our proposed method, we em-

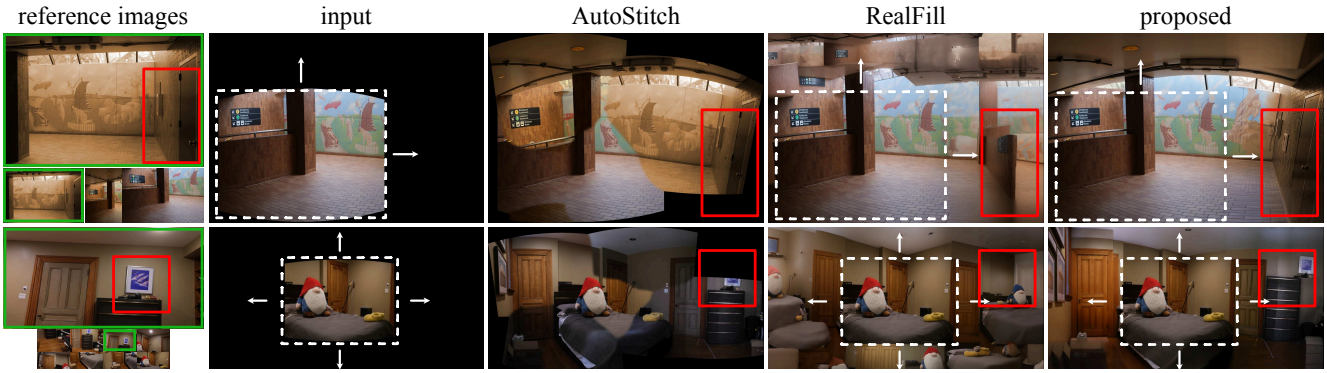

Figure 6. Qualitative results on the casually captured dataset. Even in this challenging scenario, where the input images have strong parallax effects and variations in style, illumination, color palette, or camera capture settings, our approach reconstructs seamless panoramas that preserve the content and layout of the reference.

Method	PSNR ↑	SSIM ↑	LPIPS ↓	DreamSim ↓	DINO ↑	CLIP ↑	LoFTR (L2 Distance) ↓	LoFTR (Matching) ↑
Agisoft+NeRF	10.22 (1.43)	0.282 (0.129)	0.748 (0.058)	0.210 (0.066)	0.903 (0.027)	0.874 (0.035)	75.30 (41.33)	0.104 (0.054)
Omni^2 Inpainting	10.12 (1.24)	0.302 (0.042)	0.808 (0.044)	0.355 (0.040)	0.802 (0.021)	0.809 (0.023)	68.06 (22.95)	0.062 (0.027)
Omni^2 Outpainting	8.87 (0.64)	0.264 (0.013)	0.848 (0.018)	0.360 (0.021)	0.829 (0.013)	0.829 (0.018)	84.49 (12.28)	0.006 (0.000)
RDIStitcher	11.31 (1.49)	0.035 (0.126)	0.697 (0.081)	0.212 (0.059)	0.926 (0.036)	0.819 (0.063)	17.55 (10.03)	0.108 (0.046)
SD2	9.97 (0.74)	0.267 (0.099)	0.650 (0.040)	0.295 (0.050)	0.916 (0.031)	0.859 (0.074)	85.45 (56.00)	0.012 (0.003)
RealFill	11.47 (1.50)	0.360 (0.141)	0.578 (0.058)	0.197 (0.024)	0.943 (0.024)	0.912 (0.036)	30.75 (7.82)	0.026 (0.008)
AutoStitch	10.61 (1.74)	0.335 (0.128)	0.554 (0.060)	0.202 (0.070)	0.949 (0.017)	0.906 (0.042)	16.57 (8.34)	0.168 (0.048)
proposed	11.35 (2.15)	0.374 (0.143)	0.508 (0.076)	0.137 (0.033)	0.971 (0.013)	0.917 (0.035)	17.97 (5.14)	0.130 (0.056)

Table 2. Quantitative assessment of generative panoramic image stitching from casually captured images. We compare the generated results to a reference panorama using AutoStitch [8] on the tripod-captured dataset. Our approach generates panoramas that are close to the reference despite operating on images with parallax and variations in style or lighting.

ploy correspondence-based seed selection. Since the set of input images differs from that of the reference panorama, we notice worse performance in the low-level image quality metrics on this dataset. However, our approach still outperforms baselines for most metrics. We notice similar trends in the high-level image quality metrics to those of the tripod-captured dataset, which suggests that our approach retains the same layout and structure as the reference despite the significantly more challenging setting. While AutoStitch [8] performs slightly better than our method on the feature-matching based metrics, it achieves this at a cost of seams and other artifacts because it imperfectly accounts for parallax and variations in capture settings or illumination. Finally, metrics for Agisoft+NeRF are lowered by artifacts from viewpoint sparsity, and for RDIStitcher due to seams introduced by sequential stitching. Stitching failed in about $\sim 15\%$ of Agisoft+NeRF scenes due to artifacts and inability to match features well.

Ablation study. We conduct an ablation study on four scenes from the casually captured dataset (see Table 3). We evaluate (1) not perturbing the warped image positions (Section 3.4), (2) replacing LoRA with full fine-tuning of the self-attention layers, and (3) using a single random seed instead of correspondence-based seed selection. Without perturbation, the model is less robust to misalignments in the initial layout estimation; full fine-tuning shows no significant advantage over LoRA and is more computationally expensive; correspondence-based seed selection improves the overall image quality and feature similarity. In Supp. Section S2.4 we provide additional ablations to evaluate the effects of guidance scale, seed selection, tiling strategy, and positional encoding frequencies.

No-reference image quality evaluation. We evaluate our method on no-reference image quality metrics using the UDIS [42] dataset. We selected two outdoor scenes from the UDIS test set and reorganized the associated images to form multi-view input groups that could be stitched together (see Supp. Section S1.4). Since the UDIS dataset does not provide a reference panorama, we evaluated the stitched outputs using no-reference image quality metrics, including HyperIQA [55], CLIP-IQA [63], and SIQS-Q [67]. Across both scenes, our method achieved the highest image quality scores, outperforming AutoStitch and RDIStitcher in terms of perceptual consistency and overall panorama quality. We provide qualitative results in Supp. Section S2.

Additional experiments. We further analyze our method's robustness to layout errors, misalignment, panorama resolution, and correspondance-based seed selection in Supp. Sections S2.6- S2.9. Results show that performance degrades gracefully under moderate homography perturbations and pixel misalignment, retains performance trends between baselines with and without correspondence-based seed selec-

Method	PSNR ↑	SSIM ↑	LPIPS ↓	DreamSim ↓	DINO ↑	CLIP ↑	LoFTR (L2 Distance) ↓	LoFTR (Matching) ↑
proposed (w/o perturb)	10.83 (1.76)	0.394 (0.100)	0.547 (0.058)	0.146 (0.011)	0.971 (0.014)	0.920 (0.042)	20.90 (3.76)	0.099 (0.040)
proposed (w/o LoRA)	11.00 (1.89)	0.420 (0.102)	0.534 (0.054)	0.142 (0.019)	0.972 (0.012)	0.923 (0.032)	19.37 (5.06)	0.117 (0.042)
proposed (random seed)	11.24 (2.07)	0.375 (0.140)	0.514 (0.076)	0.139 (0.034)	0.971 (0.012)	0.922 (0.034)	19.95 (7.48)	0.121 (0.052)
proposed	11.35 (2.15)	0.374 (0.143)	0.508 (0.076)	0.137 (0.033)	0.971 (0.013)	0.917 (0.035)	17.97 (5.14)	0.130 (0.056)

Table 3. Ablation study. We evaluate the effects of omitting (1) the similarity transform used to perturb the location of the warped images in the sparse panoramas (w/o perturb.), (2) LoRA and instead using full fine-tuning on the model's self-attention layers (w/o LoRA), and (3) correspondence-based seed selection (random seed). We compare the generated results to a reference panorama produced using AutoStitch [8] on the tripod-captured dataset.

	Backyard	Bedroom	Cops	College	Livingroom	Street	TTC	Waterfront
Tripod Captured	0.0363	0.0062	0.0008	0.0017	0.0070	0.0008	0.0029	0.0012
Casually Captured	0.0341	0.0076	0.0124	0.0553	0.0712	0.0166	0.0991	0.0054

Table 4. Average Baseline-to-Depth Ratio (↑ higher means more parallax; see Supp. Figs. S4 and S5 for images for each scene).

Scene	Method	HyperIQA (↑)	CLIP-IQA (↑)	SIQS-Q (↑)
Scene1	RDIStitcher	55.99 (52.00)	0.38 (0.29)	7.0 (5.22)
	AutoStitch	66.96 (0.67)	0.26 (0.0)	6.4 (2.31)
	Proposed	67.27 (66.08)	0.52 (0.50)	10.0 (10.0)
Scene2	RDIStitcher	47.93 (45.56)	0.20 (0.18)	6.0 (5.11)
	AutoStitch	54.35 (0.66)	0.42 (0.0)	8.1 (0.54)
	Proposed	56.10 (54.85)	0.60 (0.58)	8.00 (7.20)

Table 5. No-reference image quality evaluation. We evaluate our proposed method using no-reference image quality metrics on the UDIS dataset using two sample scenes. For each scene, we manually selected a set of overlapping images that can be stitched together to form a larger panorama.

tion, and remains stable across increasing panorama sizes, demonstrating strong structural consistency and scalability.

5. Discussion

Our work overcomes several failure cases associated with conventional panoramic image stitching methods and shows the utility of image generation methods for this low-level computer vision task. We see multiple promising directions for future work. While our method is currently fine-tuned on a single scene, future extensions could train a more general model that incorporates layout and content from multiple reference images in a feed-forward fashion. Additionally, we show how our method can handle input images with large variations in viewpoint, lighting, white balance, and color palette. However, strong variations in scene content, such as dynamic scenes with moving objects, can be challenging to handle with our layout estimation scheme, which leverages conventional feature matching and homography estimation.

Handling transient objects. Addressing dynamic scenes with transient objects remains an interesting challenge. In Fig. 7 we show that our method is relatively robust to scene changes, e.g., using images captured in winter and spring. There are a number of inconsistencies/transient objects that the model must resolve to create the stitched output; for example, cars are in different locations, and the snow banks appear in the winter images but not the spring images. Despite these challenges, we achieve a coherent output panorama. Other scenes with strong lighting or white balance changes across the input image set demonstrate this as well (e.g., Figs. 1, 6, and S9). Finally, we include an example of panorama stitching from an online photo-collection with transient objects in Fig. S14.

Figure 7. Generative stitching result (bottom) with changing scene content in the reference images (top). See Supp. Section S1.4 for all reference images for this scene.

Parallax. We include a quantitative assessment of parallax in our dataset to better assess the problem setting and limitations of our approach. Specifically, in Table 4, we report the baseline-to-depth ratio for each scene, averaged across all overlapping image pairs. The metric is computed by estimating poses with Agisoft Metashape and taking the ratio of the baseline and average scene depth (across triangulated features) for each image pair. Our casually captured dataset covers a range of baseline-to-depth ratios, where a large ratio indicates a greater amount of parallax. Ultimately, if the parallax between views becomes so significant that we cannot find feature correspondences and produce a coarse alignment via homography, our method will fail.

Broader impact. In contrast to conventional image-stitching methods, we use an image generation model that can hallucinate scene content. Hence, our method should be used for applications where the qualitative appearance of a panorama is more important than strict pixel-level fidelity.

References

[1] Aseem Agarwala, Maneesh Agrawala, Michael Cohen, David Salesin, and Richard Szeliski. Photographing long scenes with multi-viewpoint panoramas. In *Proc. ACM SIGGRAPH*. 2006. 1, 3

[2] Aseem Agarwala, Mira Dontcheva, Maneesh Agrawala, Steven Drucker, Alex Colburn, Brian Curless, David Salesin, and Michael Cohen. Interactive digital photomontage. In *Proc. ACM SIGGRAPH*. 2004. 1

[3] Agisoft LLC. Agisoft metashape. https://www.agisoft.com/, 2024. 5

[4] Stability AI. Stable-diffusion-2-inpainting, 2022. 2, 4, 5, 6, 1, 10

[5] Robert Anderson, David Gallup, Jonathan T Barron, Janne Kontkanen, Noah Snavely, Carlos Hernández, Sameer Agarwal, and Steven M Seitz. Jump: virtual reality video. *ACM Trans. Graph.*, 35(6):1–13, 2016. 3

[6] Omer Bar-Tal, Lior Yariv, Yaron Lipman, and Tali Dekel. Multidiffusion: Fusing diffusion paths for controlled image generation. In *Proc. ICML*, 2023. 2, 4

[7] James R Bergen and Edward H Adelson. The plenoptic function and the elements of early vision. *Comput. Models Vis. Process.*, 1(8):3, 1991. 3

[8] Matthew Brown and David G Lowe. Automatic panoramic image stitching using invariant features. *IJCV*, 74:59–73, 2007. 1, 2, 3, 5, 6, 7, 8, 4, 10, 12, 13, 14

[9] Chris Buehler, Michael Bosse, Leonard McMillan, Steven Gortler, and Michael Cohen. Unstructured lumigraph rendering. In *Proc. SIGGRAPH*, 2001. 3

[10] Peter J Burt and Edward H Adelson. A multiresolution spline with application to image mosaics. *ACM Trans. Graph.*, 2(4):217–236, 1983. 2

[11] Zhipeng Cai, Matthias Mueller, Reiner Birkl, Diana Wofk, Shao-Yen Tseng, Junda Cheng, Gabriela Ben-Melech Stan, Vasudev Lai, and Michael Paulitsch. L-magic: Language model assisted generation of images with coherence. In *Proc. CVPR*, 2024. 3

[12] Chenjie Cao, Yunuo Cai, Qiaole Dong, Yikai Wang, and Yanwei Fu. Leftrefill: Filling right canvas based on left reference through generalized text-to-image diffusion model. In *Proc. CVPR*, 2024. 3

[13] Mathilde Caron, Hugo Touvron, Ishan Misra, Hervé Jégou, Julien Mairal, Piotr Bojanowski, and Armand Joulin. Emerging properties in self-supervised vision transformers. In *Proc. ICCV*, 2021. 5

[14] Che-Han Chang, Yoichi Sato, and Yung-Yu Chuang. Shape-preserving half-projective warps for image stitching. In *Proc. CVPR*, 2014. 2

[15] Paul E Debevec, Camillo J Taylor, and Jitendra Malik. Modeling and rendering architecture from photographs: A hybrid geometry-and image-based approach. In *Proc. SIGGRAPH*, 1996. 3

[16] Daniel DeTone, Tomasz Malisiewicz, and Andrew Rabinovich. Deep image homography estimation. *arXiv preprint arXiv:1606.03798*, 2016. 2

[17] Ashley Eden, Matthew Uyttendaele, and Richard Szeliski. Seamless image stitching of scenes with large motions and exposure differences. In *Proc. CVPR*, 2006. 2

[18] Stanislav Frolov, Brian B Moser, and Andreas Dengel. Spotdiffusion: A fast approach for seamless panorama generation over time. In *Proc. WACV*, 2025. 2, 4

[19] Stephanie Fu, Netanel Tamir, Shobhita Sundaram, Lucy Chai, Richard Zhang, Tali Dekel, and Phillip Isola. Dreamsim: Learning new dimensions of human visual similarity using synthetic data. In *Proc. NeurIPS*, 2023. 5, 9

[20] Junhong Gao, Seon Joo Kim, and Michael S Brown. Constructing image panoramas using dual-homography warping. In *Proc. CVPR*, 2011. 2

[21] Junhong Gao, Yu Li, Tat-Jun Chin, and Michael S Brown. Seam-driven image stitching. In *Eurographics*, 2013. 2

[22] Jonathan Ho, Ajay Jain, and Pieter Abbeel. Denoising diffusion probabilistic models. *Proc. NeurIPS*, 2020. 4, 1

[23] Jonathan Ho and Tim Salimans. Classifier-free diffusion guidance. In *Proc. NeurIPS Workshop on Deep Generative Models and Downstream Applications*, 2021. 5

[24] Alain Hore and Djemel Ziou. Image quality metrics: Psnr vs. ssim. In *2010 20th international conference on pattern recognition*, pages 2366–2369. IEEE, 2010. 9

[25] Edward J Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, Weizhu Chen, et al. Lora: Low-rank adaptation of large language models. In *Proc. ICLR*, 2022. 4

[26] Nikolai Kalischek, Michael Oechsle, Fabian Manhardt, Philipp Henzler, Konrad Schindler, and Federico Tombari. Cubediff: Repurposing diffusion-based image models for panorama generation. In *The Thirteenth International Conference on Learning Representations*, 2025. 4

[27] Alex Krizhevsky, Ilya Sutskever, and Geoffrey E Hinton. Imagenet classification with deep convolutional neural networks. *Advances in neural information processing systems*, 25, 2012. 9

[28] Hoang Le, Feng Liu, Shu Zhang, and Aseem Agarwala. Deep homography estimation for dynamic scenes. In *Proc. CVPR*, 2020. 2

[29] Kyu-Yul Lee and Jae-Young Sim. Warping residual based image stitching for large parallax. In *Proc. CVPR*, 2020. 2

[30] Yuseung Lee, Kunho Kim, Hyunjin Kim, and Minhyuk Sung. Syncdiffusion: Coherent montage via synchronized joint diffusions. *Proc. NeurIPS*, 2023. 2, 4

[31] Marc Levoy and Pat Hanrahan. Light field rendering. In *Proc. SIGGRAPH*, 1996. 3

[32] Jing Li, Zhengming Wang, Shiming Lai, Yongping Zhai, and Maojun Zhang. Parallax-tolerant image stitching based on robust elastic warping. *IEEE Trans. Multimedia*, 20(7):1672–1687, 2017. 2

[33] Tianli Liao and Nan Li. Single-perspective warps in natural image stitching. *IEEE TIP*, 29:724–735, 2019.

[34] Chung-Ching Lin, Sharathchandra U Pankanti, Karthikeyan Natesan Ramamurthy, and Aleksandr Y Aravkin. Adaptive as-natural-as-possible image stitching. In *Proc. CVPR*, 2015.

[35] Wen-Yan Lin, Siying Liu, Yasuyuki Matsushita, Tian-Tsong Ng, and Loong-Fah Cheong. Smoothly varying affine stitching. In *Proc. CVPR*, 2011. 2

[36] Feng Liu, Michael Gleicher, Hailin Jin, and Aseem Agarwala. Content-preserving warps for 3d video stabilization. *ACM Trans. on Graph.*, 28(3):1–9, 2009. 3

[37] Kuan-Hung Liu, Cheng-Kun Yang, Min-Hung Chen, Yu-Lun Liu, and Yen-Yu Lin. Corrfill: Enhancing faithfulness in reference-based inpainting with correspondence guidance in diffusion models. In *Proc. WACV*, 2025. 3

[38] David G Lowe. Distinctive image features from scale-invariant keypoints. *Int. J. Comput. Vis.*, 60:91–110, 2004. 3, 1

[39] Or Madar and Ohad Fried. Tiled diffusion. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 7795–7804, 2025. 4

[40] Ben Mildenhall, Pratul P Srinivasan, Matthew Tancik, Jonathan T Barron, Ravi Ramamoorthi, and Ren Ng. NeRF: Representing scenes as neural radiance fields for view synthesis. *Commun. ACM*, 65(1):99–106, 2021. 3, 5

[41] Ty Nguyen, Steven W Chen, Shreyas S Shivakumar, Camillo Jose Taylor, and Vijay Kumar. Unsupervised deep homography: A fast and robust homography estimation model. *IEEE Robot. Autom. Lett.*, 3(3):2346–2353, 2018. 2

[42] Lang Nie, Chunyu Lin, Kang Liao, Shuaicheng Liu, and Yao Zhao. Unsupervised deep image stitching: Reconstructing stitched features to images. *IEEE Transactions on Image Processing*, 30:6184–6197, 2021. 2, 5, 7, 10

[43] Lang Nie, Chunyu Lin, Kang Liao, Shuaicheng Liu, and Yao Zhao. Deep rectangling for image stitching: A learning baseline. In *Proc. CVPR*, 2022. 2, 3

[44] Lang Nie, Chunyu Lin, Kang Liao, Shuaicheng Liu, and Yao Zhao. Parallax-tolerant unsupervised deep image stitching. In *Proc. ICCV*, 2023. 2

[45] Yoshikuni Nomura, Li Zhang, and Shree K Nayar. Scene collages and flexible camera arrays. In *Proc. Eurographics*, 2007. 1, 3

[46] Maxime Oquab, Timothée Darcet, Théo Moutakanni, Huy Vo, Marc Szafraniec, Vasil Khalidov, Pierre Fernandez, Daniel Haziza, Francisco Massa, Alaaeldin El-Nouby, et al. Dinov2: Learning robust visual features without supervision. *arXiv preprint arXiv:2304.07193*, 2023. 9

[47] Shmuel Peleg, Benny Rousso, Alex Rav-Acha, and Assaf Zomet. Mosaicing on adaptive manifolds. *IEEE TPAMI*, 22(10):1144–1154, 2000. 1

[48] Patrick Pérez, Michel Gangnet, and Andrew Blake. Poisson image editing. In *Proc. ACM SIGGRAPH*. 2003. 2

[49] 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 *Proc. ICML*, 2021. 5, 9

[50] Alex Rav-Acha, Giora Engel, and Shmuel Peleg. Minimal aspect distortion (MAD) mosaicing of long scenes. *IJCV*, 78:187–206, 2008. 3

[51] Christian Richardt, Yael Pritch, Henning Zimmer, and Alexander Sorkine-Hornung. Megastereo: Constructing high-resolution stereo panoramas. In *Proc. CVPR*, 2013. 3

[52] Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. High-resolution image synthesis with latent diffusion models. In *Proc. CVPR*, 2022. 4

[53] Nataniel Ruiz, Yuanzhen Li, Varun Jampani, Yael Pritch, Michael Rubinstein, and Kfir Aberman. Dreambooth: Fine tuning text-to-image diffusion models for subject-driven generation. In *Proc. CVPR*, 2023. 1, 2, 5

[54] Steven M Seitz and Jiwon Kim. Multiperspective imaging. *IEEE Comput. Graph. Appl.*, 23(6):16–19, 2003. 1

[55] Shaolin Su, Qingsen Yan, Yu Zhu, Cheng Zhang, Xin Ge, Jinqiu Sun, and Yanning Zhang. Blindly assess image quality in the wild guided by a self-adaptive hyper network. In *Proc. CVPR*, 2020. 7

[56] Jiaming Sun, Zehong Shen, Yuang Wang, Hujun Bao, and Xiaowei Zhou. Loftr: Detector-free local feature matching with transformers. In *Proc. CVPR*, 2021. 5, 3, 9

[57] Richard Szeliski et al. Image alignment and stitching: A tutorial. *Found. Trends Comput. Graph. Vis.*, 2(1):1–104, 2007. 1, 2

[58] Richard Szeliski and Heung-Yeung Shum. Creating full view panoramic image mosaics and environment maps. In *Proc. ACM SIGGRAPH*, 1997. 1, 2

[59] Matthew Tancik, Ethan Weber, Evonne Ng, Ruilong Li, Brent Yi, Terrance Wang, Alexander Kristoffersen, Jake Austin, Kamyar Salahi, Abhik Ahuja, et al. Nerfstudio: A modular framework for neural radiance field development. In *ACM SIGGRAPH 2023 conference proceedings*, pages 1–12, 2023. 8

[60] Luming Tang, Nataniel Ruiz, Qinghao Chu, Yuanzhen Li, Aleksander Holynski, David E Jacobs, Bharath Hariharan, Yael Pritch, Neal Wadhwa, Kfir Aberman, et al. Realfill: Reference-driven generation for authentic image completion. *ACM Trans. Graph.*, 43(4):1–12, 2024. 1, 2, 3, 5, 6, 10

[61] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. *Advances in neural information processing systems*, 30, 2017. 2

[62] Hai Wang, Xiaoyu Xiang, Yuchen Fan, and Jing-Hao Xue. Customizing 360-degree panoramas through text-to-image diffusion models. In *Proc. WACV*, 2024. 3

[63] Jianyi Wang, Kelvin CK Chan, and Chen Change Loy. Exploring clip for assessing the look and feel of images. In *Proc. AAAI*, 2023. 7

[64] Zhou Wang, Alan C Bovik, Hamid R Sheikh, and Eero P Simoncelli. Image quality assessment: from error visibility to structural similarity. *IEEE Trans. Image Process.*, 13(4):600–612, 2004. 5, 9

[65] Daniel N Wood, Daniel I Azuma, Ken Aldinger, Brian Curless, Tom Duchamp, David H Salesin, and Werner Stuetzle. Surface light fields for 3d photography. In *Proc. SIGGRAPH*, 2000. 3

[66] Tianhao Wu, Chuanxia Zheng, and Tat-Jen Cham. Panodiffusion: 360-degree panorama outpainting via diffusion. In *Proc. ICLR*, 2024. 2

[67] Ziqi Xie, Xiao Lai, Weidong Zhao, Siqi Jiang, Xianhui Liu, and Wenlong Hou. Modification takes courage: Seamless image stitching via reference-driven inpainting. *arXiv preprint arXiv:2411.10309*, 2024. 2, 3, 5, 7, 8

[68] Ziqi Xie, Weidong Zhao, Jian Zhao, and Ning Jia. Reconstructing the image stitching pipeline: Integrating fusion and rectangling into a unified inpainting model. 2024. 3

[69] Zhexiao Xiong, Zhang Chen, Zhong Li, Yi Xu, and Nathan Jacobs. Panodreamer: Consistent text to 360-degree scene generation. In *Proc. CVPR*, 2025. 3

[70] Binxin Yang, Shuyang Gu, Bo Zhang, Ting Zhang, Xuejin Chen, Xiaoyan Sun, Dong Chen, and Fang Wen. Paint by example: Exemplar-based image editing with diffusion models. In *Proc. CVPR*, 2023. 3, 5

[71] Liu Yang, Huiyu Duan, Yucheng Zhu, Xiaohong Liu, Lu Liu, Zitong Xu, Guangji Ma, Xiongkuo Min, Guangtao Zhai, and Patrick Le Callet. Omni2: Unifying omnidirectional image generation and editing in an omni model. In *Proceedings of the 33rd ACM International Conference on Multimedia*, pages 10103–10112, 2025. 13, 14

[72] Weicai Ye, Chenhao Ji, Zheng Chen, Junyao Gao, Xiaoshui Huang, Song-Hai Zhang, Wanli Ouyang, Tong He, Cairong Zhao, and Guofeng Zhang. Diffpano: Scalable and consistent text to panorama generation with spherical epipolar-aware diffusion. In *Proc. NeurIPS*, 2024. 2

[73] Julio Zaragoza, Tat-Jun Chin, Michael S Brown, and David Suter. As-projective-as-possible image stitching with moving DLT. In *Proc. CVPR*, 2013. 2

[74] Fan Zhang and Feng Liu. Parallax-tolerant image stitching. In *Proc. CVPR*, 2014. 2

[75] Jirong Zhang, Chuan Wang, Shuaicheng Liu, Lanpeng Jia, Nianjin Ye, Jue Wang, Ji Zhou, and Jian Sun. Content-aware unsupervised deep homography estimation. In *Proc. ECCV*, 2020. 2

[76] Richard Zhang, Phillip Isola, Alexei A Efros, Eli Shechtman, and Oliver Wang. The unreasonable effectiveness of deep features as a perceptual metric. In *Proc. CVPR*, 2018. 5, 9

[77] Yuqian Zhou, Connelly Barnes, Eli Shechtman, and Sohrab Amirghodsi. Transfill: Reference-guided image inpainting by merging multiple color and spatial transformations. In *Proc. CVPR*, 2021. 3, 5