

360-Degree Panorama Generation from Few Unregistered N FoV Images

Jionghao Wang*
Shanghai Jiao Tong University
shanemankiw@sjtu.edu.cn

Ziyu Chen*
Shanghai Jiao Tong University
1252060456@sjtu.edu.cn

Jun Ling
Shanghai Jiao Tong University
lingjun@sjtu.edu.cn

Rong Xie
Shanghai Jiao Tong University
xierong@sjtu.edu.cn

Li Song†
Shanghai Jiao Tong University
song_li@sjtu.edu.cn

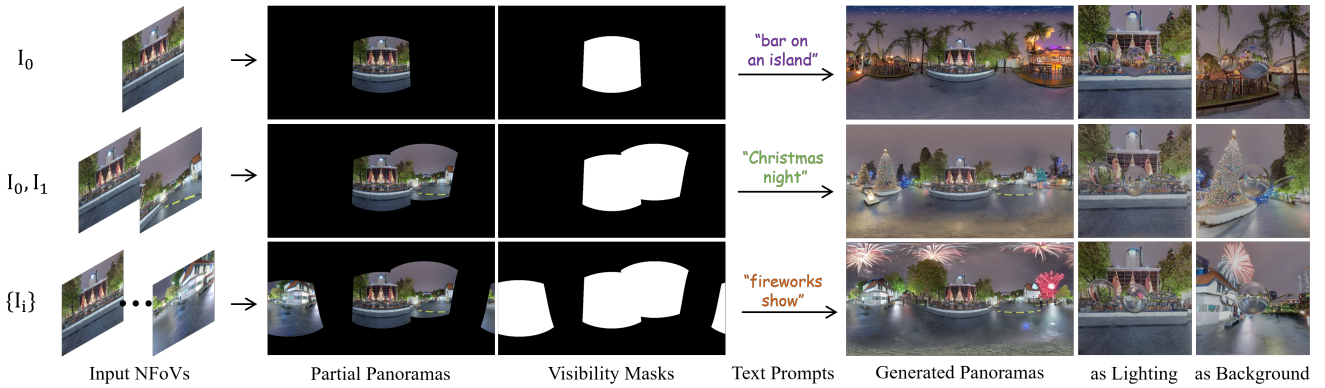


Figure 1: Illustration of our PanoDiff, a novel approach that is capable of synthesizing fine-grained and diverse 360-degree panorama from a(few) unregistered N FoV (Narrow Field of View) image(s) and text prompts.

ABSTRACT

360° panoramas are extensively utilized as environmental light sources in computer graphics. However, capturing a 360° × 180° panorama poses challenges due to the necessity of specialized and costly equipment, and additional human resources. Prior studies develop various learning-based generative methods to synthesize panoramas from a single Narrow Field-of-View (N FoV) image, but they are limited in alterable input patterns, generation quality, and controllability. To address these issues, we propose a novel pipeline called *PanoDiff*, which efficiently generates complete 360° panoramas using one or more unregistered N FoV images captured from arbitrary angles. Our approach has two primary components to

overcome the limitations. Firstly, a two-stage angle prediction module to handle various numbers of N FoV inputs. Secondly, a novel latent diffusion-based panorama generation model uses incomplete panorama and text prompts as control signals and utilizes several geometric augmentation schemes to ensure geometric properties in generated panoramas. Experiments show that PanoDiff achieves state-of-the-art panoramic generation quality and high controllability, making it suitable for applications such as content editing.

CCS CONCEPTS

• **Computing methodologies** → **Computer vision**; *Computer graphics*.

KEYWORDS

360-degree panorama, generative models, multimodal models, latent diffusion, image pose estimation

ACM Reference Format:

Jionghao Wang, Ziyu Chen, Jun Ling, Rong Xie, and Li Song. 2023. 360-Degree Panorama Generation from Few Unregistered N FoV Images. In *Proceedings of the 31st ACM International Conference on Multimedia (MM '23)*, October 29–November 3, 2023, Ottawa, ON, Canada. ACM, New York, NY, USA, 9 pages. <https://doi.org/10.1145/3581783.3612508>

1 INTRODUCTION

Panoramic images, which capture an extensive field of view encompassing a full 360° horizontal by 180° vertical FoV scene, have

*Indicates equal contribution.

†Corresponding author. Affiliated with both School of Electronic Information and Electrical Engineering, Shanghai Jiao Tong University & MoE Key Lab of Artificial Intelligence, AI Institute, Shanghai Jiao Tong University.

Jionghao Wang, Ziyu Chen, Jun Ling and Rong Xie are with School of Electronic Information and Electrical Engineering, Shanghai Jiao Tong University.

Permission to make digital or hard copies of all or part of this work for personal or classroom use is granted without fee provided that copies are not made or distributed for profit or commercial advantage and that copies bear this notice and the full citation on the first page. Copyrights for components of this work owned by others than the author(s) must be honored. Abstracting with credit is permitted. To copy otherwise, or republish, to post on servers or to redistribute to lists, requires prior specific permission and/or a fee. Request permissions from permissions@acm.org.
MM '23, October 29–November 3, 2023, Ottawa, ON, Canada

© 2023 Copyright held by the owner/author(s). Publication rights licensed to ACM.
ACM ISBN 979-8-4007-0108-5/23/10...\$15.00
<https://doi.org/10.1145/3581783.3612508>

become increasingly significant across various applications, such as environment lighting, VR/AR, and autonomous driving system. However, obtaining high-quality panoramic images can be both time-consuming and costly, as they typically necessitate the use of specialized panoramic cameras or stitching software to combine images from multiple perspectives. Our method addresses two main limitations regarding previous generating methods, namely *input pattern* and *generation quality & controllability*.

For *Input pattern*. Most previous works [1, 5, 37] only support a single FoV region in the center of an incomplete panorama as input. However, relying solely on a single input region restricts flexibility when controlling specific contents of the generated scene. Such flexibility is particularly important for applications that require precise control over the visual elements and ensure accurate representations of the intended scene.

In terms of *generation quality & controllability*. Generating a complete 360-degree panorama from N FoV images could be viewed as a large-hole inpainting problem [35]. Previous methods [1, 11, 37] typically all rely on GAN (Generative Adversarial Networks) [10] based methods. Besides, previous methods approach this problem as an image-conditioned generation task, leaving their pipeline with little control and flexibility over the generation results. A more recent work [5] proposes to use additional text guidance for a GAN-based image inpainting method [41]. However, diffusion-based image generation methods [14] have shown impressive results on various generative tasks, and models trained on large datasets [29, 33] have shown better performance and robustness against GAN-based models [7, 15]. Moreover, GANs have limited mode coverage and are difficult to scale for modeling complex multimodal distributions. Unlike GANs, likelihood-based models such as diffusion models are capable of learning the complex distribution of natural images, resulting in the generation of high-quality images, as mentioned in [7, 29].

To overcome these limitations, a pipeline capable of accepting a flexible number of N FoV image(s) as input and generating high-fidelity panoramas is crucial. Nevertheless, this endeavor presents two main challenges: 1) estimating relative camera poses and accurately warping the input N FoV image(s) on the panorama, and 2) using a latent-diffusion-model-based method to generate the entire panorama from input partial panorama of various shapes.

This paper introduces *PanoDiff*, a novel pipeline that efficiently generates complete 360° panoramas using one or more unregistered N FoV images captured from arbitrary angles. The proposed method overcomes the limitations of existing methods, it enables the generation of high-quality panoramas from incomplete 360-degree panoramas that warped from any number of N FoV inputs. This is achieved by addressing two key challenges. First, we propose a robust two-stage angle prediction pipeline that classifies image pairs based on their overlap before regressing specific angle values. Second, we train a hypernetwork that controls a pre-trained large latent-diffusion model, and utilizes geometric augmentation schemes during both training and inference sampling phases to ensure the geometric properties of the generated panoramas.

We summarize our contributions as follows. **Firstly**, the first flexible framework for generating panoramic images from single or multiple N FoV inputs. A two-stage pose estimation module is specially designed for relative pose estimation. **Secondly**, as far

as we know, PanoDiff is the first latent-diffusion-based panorama outpainting model that can not only handle incomplete panoramas of various shapes as input but also supports text prompts. **Finally**, PanoDiff outperforms existing methods in terms of quantitative and qualitative results on different input types, as demonstrated through abundant experiments.

2 RELATED WORKS

2.1 Rotation Estimation

Camera pose estimation has been a long-standing task in computer vision [19]. Most prior works focus on identifying various visual cues from pairs of images. For instance, the classic SIFT [22] extracts scale and direction invariant features from input images and matches them to find correspondences. Some approaches utilize neural networks for more robust feature extraction or graph neural networks for enhanced matching [6, 28, 32]. However, these methods do not account for extreme rotations where the input images have little to no overlap, known as wide baseline scenarios. Some earlier works focus on mining specially handcrafted cues [24] to address this problem, but they are not generalizable for scenes without a specific clue. The work that inspired us the most is [2], which estimates pair-wise rotation angles in both overlap and wide baseline scenarios. However, the estimation network occasionally misidentifies an overlapping pair of images as non-overlapping, which severely impacts the generation quality later on.

2.2 Diffusion Models

Diffusion models have shown impressive ability in generative tasks [7, 14]. Generally, diffusion models represent the image generative process as a denoising process, where a noise map sampled from white Gaussian noise is iteratively denoised using a learned prior distribution $p_\theta(x_{t-1}|x_t)$. The objective function for training diffusion models is simplified as

$$L_{\text{simple}} = E_{t,x_0,\epsilon} [|\epsilon - \epsilon_\theta(x_t, t)|^2] \quad (1)$$

which has been shown to be approximately equivalent to optimizing the prior probability distribution's variational lower bound [14].

To further improve the efficiency and effectiveness of the diffusion model, recent work has proposed to conduct the denoising process in latent space instead of pixel space [29]. In addition, they integrate multi-head cross-attention mechanism [3, 36] in denoising U-Net [25, 30] blocks. This approach enables features of different modalities to guide the generation process [16], e.g. CLIP-ViT [27]. Working on latent space along with some other sampling schedule [34] could also potentially make the sampling more efficient. Controlling the pre-trained model parameters by training hypernetworks [12, 40] can improve its performance on a more specific task while preserving its original generative capability.

2.3 Diffusion-based Inpainting

The task of inpainting was predominantly been done by GAN-based models [21, 35, 42]. Recently, diffusion-based methods have been proposed and have already achieved promising results [23, 38]. However, [23, 38] operate the diffusion and sampling process solely on image space rather than latent space, which limits their generation flexibility. Recently, [31] offered an image inpainting

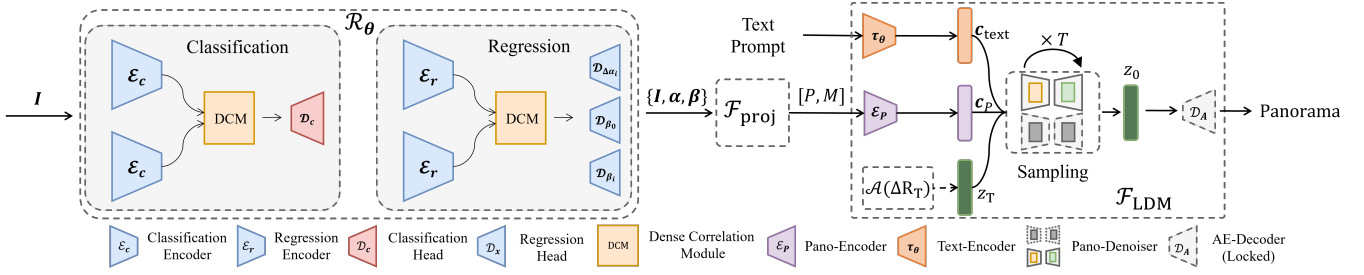


Figure 2: An overview of generating panorama from a few NFOV images. We first calculate their relative rotations based on a two-stage angle prediction network \mathcal{R}_θ (Sec. 3), then project them using backward equirectangular projection $\mathcal{F}_{\text{proj}}$ (Sec. 4) to obtain partial panorama P and visibility mask M . Finally, we feed $[P, M]$ along with text prompts to our control-based latent diffusion model \mathcal{F}_{LDM} (Sec. 4) and sample iteratively in our rotating schedule (Sec. 4.4.2) to get the final generated panorama.

model finetuned from Stable Diffusion [29] and demonstrates strong text-controlled inpainting ability.

2.4 Panorama Generation

With recent advancements in deep generative neural networks, several works adopt different forms of GAN-based methods to generate full panoramas [1, 5, 20, 37], such as VQGAN [8] and CoModGAN [41]. However, these methods are limited to single-view input, in contrast to our alterable input patterns. One work has utilized a few NFOV images as input [11] with accurate camera positions already given, thus cannot deal with unregistered camera inputs.

Recently, some researchers have utilized diffusion models to generate panoramas from text prompts alone [4]. However, this work does not support partial FoV as input, and thus users have limited control over the outcome.

3 RELATIVE POSE ESTIMATION

3.1 Problem Formulation

Unlike previous panorama generation pipelines, our method takes either single or multiple NFOV images as input. In the case of multiple NFOV (Narrow Field-of-View) images captured at the same location but in different poses, our objective is to estimate the relative poses between them. In the context of omnidirectional images, where no camera translation is involved, we formulate the relative camera poses as rotation angles. Specifically, we view the 360° panorama as a spherical *map* and express the rotation angles as the 1) *lookat* direction, which includes the longitude α , latitude β , and 2) roll angle γ . For a single image, our pipeline estimates only β and places it in lateral center, i.e., $\alpha = \gamma = 0$. Formally, for a set of N NFOV images $\mathbf{I} = I_0, I_1, \dots, I_{N-1}$, we aim to learn a model \mathcal{R}_θ that predicts their relative angles:

$$\mathcal{R}_\theta(\mathbf{I}) \rightarrow [\alpha, \beta, \gamma] \quad (2)$$

where $\alpha = \{\alpha_i\}_{i=0, \dots, N-1}$, $\beta = \{\beta_i\}_{i=0, \dots, N-1}$ and $\gamma = \{\gamma_i\}_{i=0, \dots, N-1}$, respectively. We decompose this task into estimating pair-wise camera relative rotation, which is scalable and can be applied to pair input and sometimes more than two images. We design a pairwise angle prediction network that estimates the relative angles between images. For a set of N NFOV images $(I_0, I_1, \dots, I_{N-1})$, we select one

image I_0 as an anchor and estimate the relative poses $[\Delta\alpha, \Delta\beta, \Delta\gamma]$ of the remaining images with respect to the anchor. Following the assumptions made in [2], we consider that cameras are typically upright, allowing us to estimate the absolute pitch angle (vertical/latitude) instead of the relative angle. We also assume that the camera roll angles are static and remain unchanged at 0, i.e., $\forall i \in 0, \dots, N-1, \gamma_i = 0$. Consequently, our network \mathcal{R}_θ can be expressed as:

$$\mathcal{R}_\theta(I_0, I_i) \rightarrow [\Delta\alpha_{i \rightarrow 0}, \beta_0, \beta_i] \quad (3)$$

where $\Delta\alpha_{i \rightarrow 0}$ is the relative longitude angle between the images, and β_0, β_i are the absolute latitude angle for I_0 and I_i , respectively.

3.2 Two-stage Angle Prediction

In the context of FoV overlapping, we categorize image-pair relationships into two distinct types: *overlap* and *wide baseline*. In *overlap* scenarios (green pair in Fig. 3), two NFOV images share a portion of their FoV, resulting in an overlap on the panorama. Conversely, in *wide baseline* situations (red pair in Fig. 3), the NFOV image pair does not have overlapping FoV, leading to two separate regions on the panorama. The ultimate goal of our method is to create a plausible panorama for further applications. However, a single regression network [2] is not robust enough to differentiate between these two cases. Consequently, images without any overlap might sometimes be predicted to have an overlap, creating severe artifacts in the input for our controlled LDM, which would result in unsatisfactory results. Furthermore, the precision of a single regression network is insufficient.

To address these problems, we propose a two-stage angle prediction pipeline that initially performs classification to determine whether the two images have an overlap, and subsequently regresses the precise angles $[\Delta\alpha_{i \rightarrow 0}, \beta_0, \beta_i]$. This approach significantly reduces the issue of wide baseline images being mistakenly predicted to overlap. The overview of our rotation prediction module can be seen in Fig. 3. We adopt the same backbone structure as [2], which consists of a feature encoder and a 4D dense correlation module. Within the two-stage pipeline, a total of three backbones are utilized, each with its own prediction head. To estimate the relative camera rotation between a pair of images, we first pass the image pair to our classifier to determine if the images are

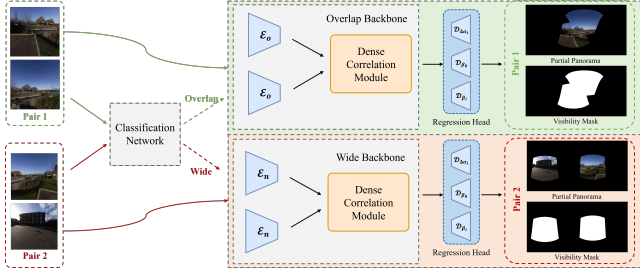


Figure 3: Illustration of our two-stage angle prediction network. $\mathcal{E}_o, \mathcal{E}_n$ represent feature encoders, and $\mathcal{D}_{\Delta\alpha_i}, \mathcal{D}_{\beta_0}, \mathcal{D}_{\beta_i}$ serve as prediction heads for overlap and wide-baseline angle regression, respectively. The pipeline first estimates whether the input image pair has an overlap, then directs the image pair to the corresponding regression network.

overlapping or wide apart. Subsequently, we feed the image pair to its corresponding regressor based on the classification result.

In terms of training, our classifier is trained on our full set of training image pairs. A specialized penalty loss function is employed for misclassified cases, which effectively mitigates failure instances. As for our regressors, we separate our training image pairs into two splits based on their overlapping status and train the overlap branch and wide branch separately on their respective data splits. That is, for the overlap regressor, we train it using primarily overlap data pairs, and vice versa for the wide baseline regressor.

4 PANORAMA GENERATION

4.1 Formulation

With Sec. 3, we can formulate a set of image-angle pairs, i.e., $\{I_0, 0, \beta_0\}$ and $\{I_i, \Delta\alpha_{i \rightarrow 0}, \beta_i\}_{i=1, \dots, N-1}$. In this part, our goal is to take the images and their relative angles to produce a complete $360^\circ \times 180^\circ$ panorama. We divide this problem into two parts: 1) project the image set $I = \{I_i\}$ onto the omnidirectional map based on their estimated angles $\alpha = \{\alpha_i\}_{i=0, \dots, N-1}$ and $\beta = \{\beta_i\}_{i=0, \dots, N-1}$; and 2) generate the full panorama using the incomplete input as a control signal. Specifically, we formulate this problem as:

$$\begin{aligned} \mathcal{F}_{\text{proj}}(I, \alpha, \beta) &\rightarrow P, M, \\ \mathcal{F}_{\text{LDM}}(P, M, \text{text}) &\rightarrow \text{Pano}, \end{aligned} \quad (4)$$

where $\mathcal{F}_{\text{proj}}$ and \mathcal{F}_{LDM} represent the projection and diffusion generation operations, respectively. The projection operation $\mathcal{F}_{\text{proj}}$ takes the NFoV image set and their angles as input and produces a partial panorama image P and a visibility mask M , indicating which parts of the FoV are missing. Subsequently, the LDM (Latent Diffusion Model) operation \mathcal{F}_{LDM} accepts the incomplete P , visibility mask M , and a text prompt as input, iteratively generating the full panorama Pano.

4.2 Recap: Controlling Stable Diffusion

We propose a generative process based on training a hypernetwork over a pre-trained Stable Diffusion (SD) model[29], a large text-to-image generative model utilizing latent diffusion. Latent diffusion models iteratively perform denoising sampling in latent space. At

time step t , given input latent z_t and condition y , the denoiser network is formulated as $\epsilon_\theta(z_t, t, y)$. The denoiser in SD employs a U-Net bottleneck architecture.

[40] proposed to use combine trainable copies from pre-trained SD model parameters, and zero-convolution blocks, whose parameters are initialized as all zero. Concretely, suppose there is a neural network block $\mathcal{F}(\cdot; \Theta)$ from the pre-trained SD denoiser U-Net, which takes an input feature map $x \in \mathbb{R}^{h \times w \times c}$ and outputs a feature y , i.e. $y = \mathcal{F}(x; \Theta)$. To add more conditions as control signals, two zero-convolution layers $\mathcal{Z}_1, \mathcal{Z}_2$ and their parameters Θ_{z1}, Θ_{z2} , and a trainable copy of the original neural block and its parameters as \mathcal{F}_c and Θ_c are introduced. Now given a new condition signal c , the new controlled output feature y_c can be written as:

$$y_c = \mathcal{F}(x; \Theta) + \mathcal{Z}_2(\mathcal{F}_c(x + \mathcal{Z}_1(c; \Theta_{z1}); \Theta_c); \Theta_{z2}) \quad (5)$$

It is worth noting that the original operation $\mathcal{F}(\cdot; \Theta)$ is locked and does not produce any gradient for backpropagation. In this setting, the new set of trainable parameters are $\Theta_c, \Theta_{z1}, \Theta_{z2}$, where the initial Θ_c is copied from Θ , and both Θ_{z1} and Θ_{z2} are initialized as zeros. This allows the controlled input at the first step to be identical to the original output, i.e., $y_c = y$. This property is favorable as it maintains the generation ability of the pre-trained model and makes the updating process of $\Theta_c, \Theta_{z1}, \Theta_{z2}$ more stable. An intuitive illustration of this scheme can be seen in Fig. 4.

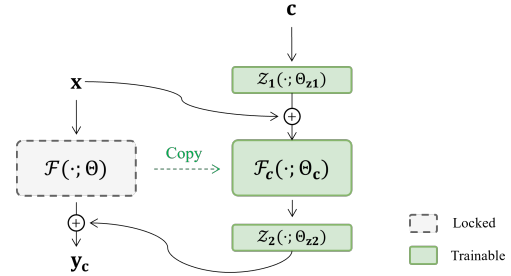


Figure 4: An intuitive explanation of integrating a new control signal into an existing model block.

4.3 Controlling LDM with Partial FoV

Given a noisy latent space feature z_t at time step t , we want to learn a denoiser that could predict the noise ϵ_t as in:

$$\epsilon_t = \epsilon_\theta(z_t, t, c_{\text{text}}, c_p) \quad (6)$$

Here, c_{text} and $c_p = [M, P]$ represent the text conditions and input panorama conditions, respectively. We introduce two components for our partial-FoV-controlled LDM, namely Pano-Denoiser ϵ_θ and Pano-Encoder \mathcal{E}_p .

Our Pano-Denoiser is designed as discussed in Sec. 4.2, where we employ control units to introduce control signals to the pre-trained Stable Diffusion model. In practice, we follow the approach of [40] and add control units to the four encoder blocks and one middle block of the denoising U-Net model used in Stable Diffusion, while utilizing zero-convolutions for the other four decoder blocks.

In addition to the Pano-Denoiser, we incorporate a shallow Pano-Encoder \mathcal{E}_p to construct our controlling condition. In order to feed

image-space signals, such as P and M , into the latent space as control signals, we use \mathcal{E}_p to encode them into latent features that are compatible with the original SD U-Net and consequently our control unit. The latent features are then passed to our learnable encoders and zero convolutions connected to the middle and decoder parts of the original SD U-Net.

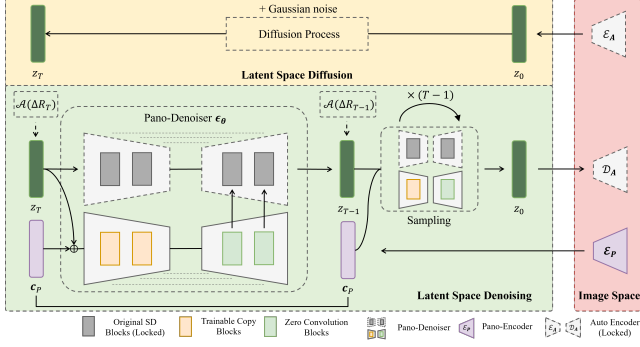


Figure 5: An overview of our partial-FoV-controlled latent diffusion model. The upper part and lower part are the diffusion process and sampling process of the model, respectively. Note that the text conditions and classifier-free guidance process are omitted from the figure for simplicity.

As depicted in Fig. 5, during the diffusion process (upper part of the figure), the encoder \mathcal{E}_A of the AutoEncoder transforms the image into feature space, and then Gaussian noises are added iteratively to the latent feature, as described in [14]. For the denoising process, our Pano-Denoiser utilizes the partial panorama P and visibility mask M as control signals and denoises the input noisy latent z_T iteratively in our rotating schedule (Sec. 4.4.2) for T steps until z_0 is generated. Finally, z_0 is decoded by the decoder \mathcal{E}_D of the AutoEncoder to produce the final panorama.

4.4 Denoising in 360-Degree

In this subsection, we focus on denoising 360-degree panoramas while preserving their unique geometric characteristics. During the training stage, we introduce a rotation equivariance loss to enforce rotation consistency in the latent space. During the inference stage, we employ a customized rotating schedule to enhance the robustness and maintain the geometric integrity of the generated panoramas. Additionally, we implement a circular padding technique during inference to mitigate edge effects and prevent geometric discontinuity.

4.4.1 Rotation Equivariance Loss. Panoramas are captured to depict a completely spherical environment, which means they exhibit rotation equivariance. This property implies that if we apply a transformation based on an $SO(3)$ rotation (limited to 2 degrees of freedom corresponding to longitude and latitude rotations) to a panorama image, it should still be able to represent the same scene.

As our method functions within the latent space, the constraint on our denoiser can be expressed as:

$$\epsilon_\theta(\mathcal{A}(R)z_t, t, c_{\text{text}}, \mathcal{A}(R)c_p) = \mathcal{A}(R)\epsilon_\theta(z_t, t, c_{\text{text}}, c_p), \quad (7)$$

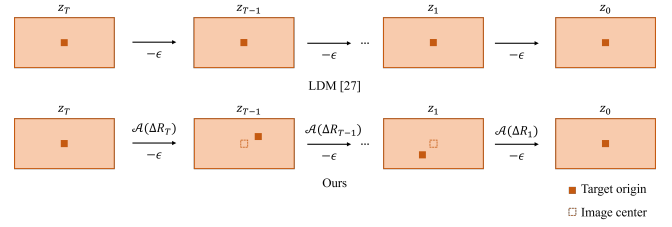


Figure 6: An illustration of our rolling schedule. The target origin represents the position of the original pixel corresponding to longitude and latitude coordinate ($\theta = 0, \phi = 0$) on the panorama.

where $\mathcal{A}(R)$ represents an image space projecting operation, similar to $\mathcal{F}_{\text{proj}}$. This equation implies that when the latent feature z_t and all input image-space conditions undergo a transformation by $\mathcal{A}(R)$, the predicted noise should also transform accordingly. To enforce this behavior, we introduce a rotation perturbation for latent features during the training phase. Consequently, the training objective evolves from Eq. 1 into:

$$\mathcal{L} =$$

$$\mathbb{E}_{z_0, t, c_{\text{text}}, c_p, \epsilon} [\|\mathcal{A}(\Delta R)\epsilon - \epsilon_\theta(\mathcal{A}(\Delta R)z_t, t, c_{\text{text}}, \mathcal{A}(\Delta R)c_p)\|_2^2]$$

4.4.2 Rotating Schedule. During the inference process, we employ a customized schedule as depicted in Fig. 6.

As illustrated in Figure 6, the denoising steps are executed while the latent feature z_i undergoes a scheduled transformation $\mathcal{A}(\Delta R_i)$ in a step-by-step manner. It is important to note that this schedule is consistent with the rotation constraint incorporated during the training phase, as both involve the same type of transformation in the latent space. The rotating schedule enhances the robustness of our method and facilitates the generation of panoramas with improved geometric integrity.

4.4.3 Circular Padding. As discussed in [11], generative methods operating in the latent space may lead to geometric discontinuity due to border effects of convolution operations. To address this issue, during inference time, we implement circular padding to mitigate edge effects. Specifically, the right portion of the latent feature is concatenated to the left side, while the left part of the original latent feature is concatenated to the right side. This process is illustrated in Fig. 7.

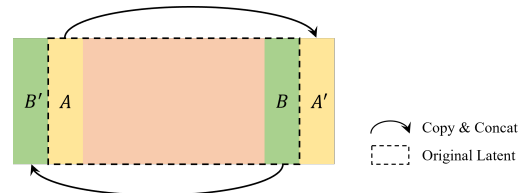


Figure 7: An illustration of our circular padding technique. Thin slices of the feature, denoted as A and B , from both the left and right ends of the latent feature, are copied and concatenated to the right and left sides of the latent feature as A' , B' , respectively.

Table 1: FID↓ results compared with other generation methods for quantitative evaluation.

Methods	SUN360 [39]			Laval [9]		
	Single	Pair (GT rots)	Pair (Pred rots)	Single	Pair (GT rots)	Pair (Pred rots)
SIG-SS [11]	13.06	15.94	16.50	-	-	-
StyleLight [37]	-	-	-	22.53	34.10	33.88
Omni-Comp [1]	14.69	12.38	12.63	18.91	14.62	14.73
Omni-Adjust [1]	11.27	9.93	10.63	9.60	10.47	10.92
ImmerseGAN [5]	9.26	10.46	10.57	12.69	24.23	24.51
PanoDiff (Ours)	8.61	6.94	7.04	7.72	6.96	7.18

After the sampling process, the decoder generates an image of shape $(w + 2w_p) \times h$, where w_p is the extra width from the padded feature. To produce a standard panorama, the extra width is removed from the final image.

5 EXPERIMENTS

5.1 Implementation Details

5.1.1 Data Preparation. We conducted experiments using real-world 360-degree panoramic image datasets SUN360 [39] and Laval Indoor [9]. SUN360 comprises both indoor and outdoor scenes, while Laval Indoor comprises solely indoor scenes. For SUN360, we randomly selected 2000/500 panoramas for training/testing, respectively. For Laval Indoor, we followed the approach in [11] and chose 289 images for testing. Notably, we **do not** train our model using the Laval Indoor dataset as there are already indoor scenes within our selected training data from the SUN360 dataset.

We used two input types in our experiments: a single input, which is a single NFoV image with a 90° FoV placed at the center of the panorama, and paired input, which is a pair of NFoVs with relative rotation, allowing us to validate our method’s capability to generate a panorama from multiple input images. We generated five pairs for each of our training/testing panoramas, resulting in 10,000/2,500 pairs of training/testing inputs on SUN360.

5.1.2 Training. We train our angle prediction network and latent diffusion model separately. The angle prediction network is trained with the strategies mentioned in 3.2. Our Pano-denoiser and Pano-Encoder are trained for 7 epochs on pair inputs. For both single and pair input, we take in NFoV images of shape 256×256 and produce 1024×512 panoramas. During training, we generate input text prompts using BLIP [18].

5.1.3 Metrics. We evaluate our method using two kinds of metrics. **Panorama Generation.** We use Fréchet Inception Distance (FID) [13, 17] as our quantitative metric since FID can report the visual quality of generated panorama images to some extent. Besides, it has also been adopted by prior studies [1, 11, 37].

Rotation Estimation. Following Cai *et al.* [2], we evaluate the geodesic error of the estimated rotation matrix (denoted by $\hat{\mathbf{R}}$) and the ground truth matrix (denoted by \mathbf{R}) using $\arccos(\frac{\text{tr}(\mathbf{R}^T \hat{\mathbf{R}}) - 1}{2})$.

5.1.4 Baselines. We compared our method to three previous SOTA methods. Omni-Dreamer [1] is trained on SUN360 with 47938 images. For the Laval dataset, Omni-Dreamer is finetuned on the model trained on SUN360 and with 1837 training samples. SIG-SS [11] is only trained on SUN360 with 50000 training images. Stylelight [37]

Table 2: Evaluation of relative pose estimation on SUN360 [39] and Laval Indoor [9]. We present the Average geodesic error Avg($^\circ$ ↓) and the percentage of pairs with errors under 10 degrees 10° (%↑).

Pair Type	Method	SUN360 [39]		Laval [9]	
		Avg($^\circ$ ↓)	10° (%↑)	Avg($^\circ$ ↓)	10° (%↑)
Overlap	1-stage	7.19	90.27	4.21	96.74
	2-stage	3.58	95.61	3.66	96.88
Wide Baseline	1-stage	41.52	40.64	32.31	62.34
	2-stage	27.12	49.77	29.46	62.62
All	1-stage	24.29	65.54	18.00	79.86
	2-stage	15.31	72.77	16.32	80.07

is trained on the Laval dataset, the same training set as used in Omni-Dreamer. We inferred these models with their officially released trained models on our test split. Since their training split is unknown to us, images in our test set could be in their training set. We also reached out to the authors of ImmerseGAN [5] and acquired their model’s results on our test set.

5.2 Quantitative Evaluation

We examine the performance of our approach from two perspectives, namely the accuracy of rotation estimation and the panorama generation quality.

Rotation Estimation. We evaluate the performance of relative rotation estimation on SUN360 [39] and Laval Indoor [9] datasets. For clarity, we separate the input pairs into two categories based on whether they belong to the ‘overlap’ or ‘wide baseline’. We compare two relative rotation estimation models: our proposed two-stage model, and a single-stage model which is designed the same as [2]. Table. 2 reports the evaluation results, which demonstrate that the two-stage relative rotation estimation method significantly outperforms the single-stage approach in terms of average relative pose estimation error for both ‘overlap’ and ‘wide baseline’. It is noted that both methods trained only on the SUN360 dataset.

Panorama Generation. The primary results are summarized in Table. 1. As demonstrated in the table, our method attains the most favorable FID metrics with all three input types on both datasets. The terms *GT rots* and *Pred rots* denote the utilization of ground truth relative angles and predicted angles from our network \mathcal{R}_θ , respectively. Despite the imperfect nature of the FID metric [17, 26], the significant margin of our method’s superiority substantiates its overall effectiveness. It is noteworthy that our model is NOT trained on the Laval Indoor dataset, yet it surpasses the performance of previous methods that were specifically trained on this dataset.

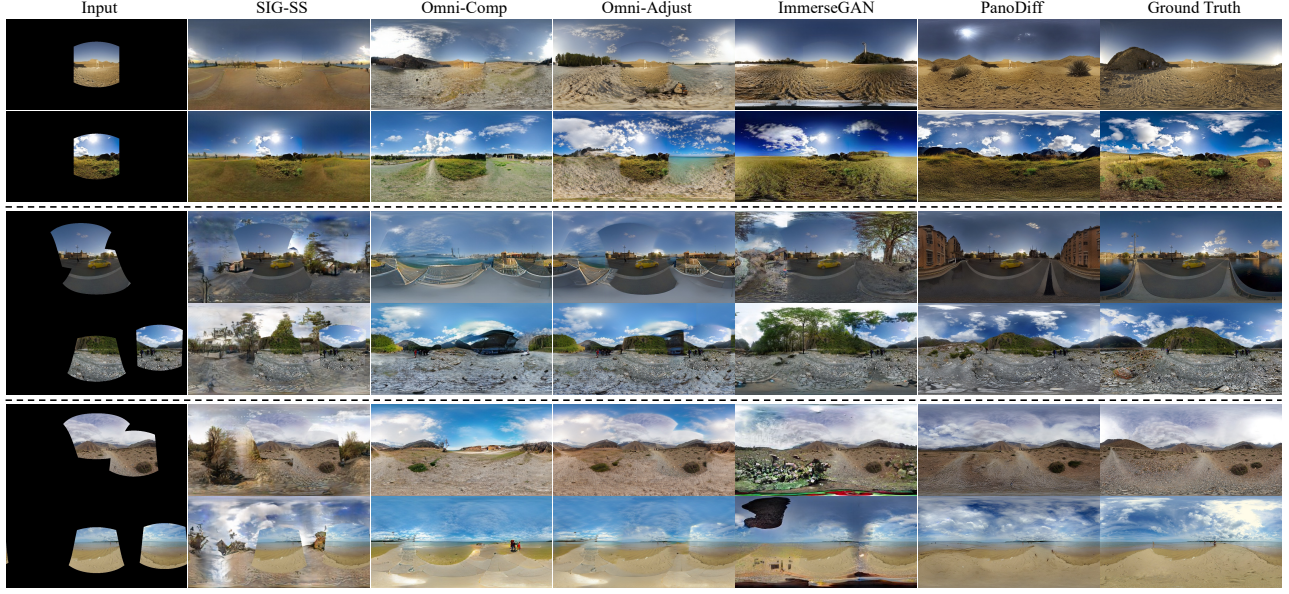


Figure 8: Out-painting results on SUN360 Dataset [39]. *Omni-Comp* and *Omni-Adjust* denote the CompletionNet and AdjustmentNet outputs of Omni-Dreamer [1], respectively.

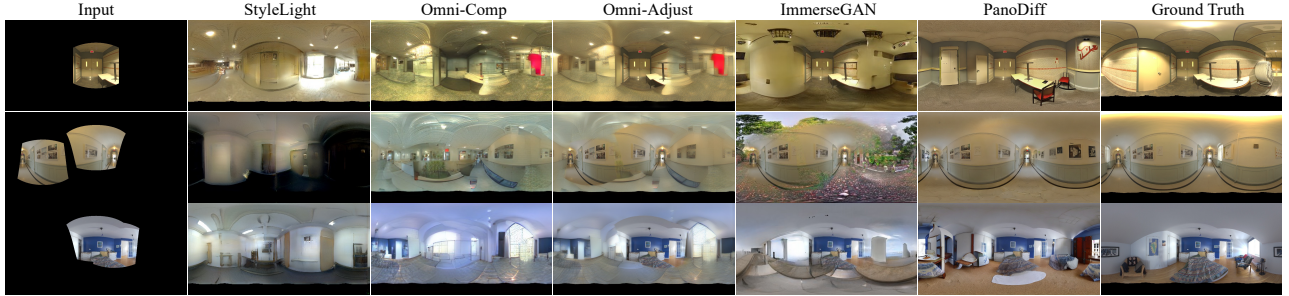


Figure 9: Out-painting results on Laval Indoor Dataset [9]. We evaluate the models of *StyleLight*, *Omni-Comp*, and *Omni-Adjust*, by utilizing their officially released models trained on Laval Indoor Dataset. The results obtained from *ImmerseGAN* were graciously provided by the authors who ran their model for our evaluation.

5.3 Qualitative Comparison

The visual results are shown in Figure 8. To illustrate the generation quality of our method, we compare our method to previous works under three kinds of inputs: single NFOV input (top two rows), the paired NFOV input with *GT rots* (middle two rows), and the paired NFOV input with *Pred rots* (bottom two rows). As can be observed from the results, both SIG-SS and Omni-Adjust exhibit inconsistencies in generating boundaries between the input patches and the out-painting content. While the generated images of Omni-Comp show a marginal improvement in boundary issues, they compromise the quality and realism of the generated content to achieve smooth output boundaries. Omni-Adjust produces relatively more realistic content, but it occasionally out-paints the wrong content. For instance, in the second row, Omni-Adjust wrongly out-paints a ‘beach’ while the input region indicates a grassland scene. ImmerseGAN excels at generating satisfactory results with a single FoV input, but struggles with paired FoV input due to the lack

of specific training for handling multiple FoVs. PanoDiff outperforms previous approaches in content consistency, realism, and texture continuity, maintaining high-quality and realistic content generation.

We proceed to look at the generalizing performance of our approach on Laval Indoor Dataset [9] and present the results in Figure 9 with both single and paired inputs. Note that our model was NOT trained on the Laval Indoor dataset but solely on SUN360. Nonetheless, our method still achieves high-quality panorama generation with consistent visual properties, e.g., lighting, temperature, and geometrical consistency.

5.4 Ablation Study

We validate the efficacy of key components in our approach, i.e., the denoising strategies, and the circular padding strategy.

Denoising Strategies in 360-Degree. We conduct quantitative comparisons in Table. 3 regarding the usage of our rotation equivariance loss, and rotating schedule. These denoising strategies assist



Figure 10: Control panorama generation with text prompts. This demonstrates the capability of our method to generate diverse high-quality results with various text prompts, while still able to maintain the geometric characteristics of panoramas.



Figure 11: Circular padding solves the problem of discontinuity between the left and right sides in panorama generation.

our model to understand the geometric characteristics of panoramas in latent space. As could be seen from the table, both our strategies contribute significantly to the quality of our panorama generation.

Table 3: FID↓ results of PanoDiff with different strategies. ‘Equi.’ denotes the usage of our rotation equivariance loss in Sec. 4.4.1, and ‘Schedule’ denotes our rotating denoising schedule in Sec. 4.4.2.

Strategy	Equi.	Schedule	FID ↓
	-	-	7.88
Choices	✓	-	6.73
	✓	✓	6.56

Circular Padding For Continuity. When decoding the latent feature z_0 , the inherited border effect caused by the domain gap in latent space and image space frequently occurs. To alleviate this issue, we implement a circular padding strategy as described in 4.4.3. In Figure 11, we validate the necessity of circular padding with visual result comparisons by rotating 180° horizontally from the output image. As observed, circular padding seamlessly eliminates discontinuity between both ends.

5.5 Applications

Text Editing. We extend the editing capability of our method. As depicted in Figure 10, our method not only inherits the powerful text-to-image generation ability from Stable Diffusion but also preserves the sound geometric properties of panoramas.

Environment Texture. Panoramic images can serve as environment textures in 3DCG software, which can provide background

lighting for 3D assets. Figure 12 shows some examples of our generated panoramas as environmental textures. As can be found, our method produces diverse panoramas that not only serve as plausible rendering backgrounds (first two) but also provide environmental lighting (last two).

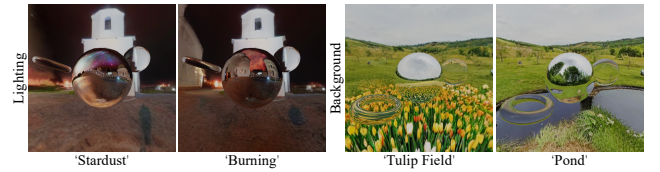


Figure 12: Examples of panoramic images used as environment textures.

Multiple NFoV As Input. Our framework includes a two-stage relative pose estimation module, allowing it to handle multiple NFoV images (e.g., >2) of the same scene as input. As shown in Figure 1, our pipeline can robustly generate high-quality panoramic images using multiple NFoV images. Please refer to the supplementary for more generated 360-degree panoramas.

6 CONCLUSION

In this paper, we present PanoDiff, a novel framework that generates 360° panoramas from one or more NFoV inputs. The pipeline consists of two main modules, namely rotation estimation and panorama generation. Our two-stage rotation estimation network first classifies the input image pairs into overlap and wide baseline scenarios and then performs precise angle prediction. In panorama generation, we use incomplete partial panoramas along with text prompts as signals to generate diverse panoramas. We hope that our work inspires further research in panorama generation for advanced applications, including style control, direct HDRI panorama generation, and related areas.

Acknowledgement. This work was supported by the Fundamental Research Funds for the Central Universities, STCSM under Grant 22DZ2229005, 111 project BP0719010.

REFERENCES

- [1] Naofumi Akimoto, Yui Matsuo, and Yoshimitsu Aoki. 2022. Diverse plausible 360-degree image outpainting for efficient 3deg background creation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 11441–11450.
- [2] Ruojin Cai, Bharath Hariharan, Noah Snavely, and Hadar Averbuch-Elor. 2021. Extreme rotation estimation using dense correlation volumes. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 14566–14575.
- [3] Chun-Fu Richard Chen, Quanfu Fan, and Rameswar Panda. 2021. Crossvit: Cross-attention multi-scale vision transformer for image classification. In *Proceedings of the IEEE/CVF international conference on computer vision*. 357–366.
- [4] Zhaoxi Chen, Guangcong Wang, and Ziwei Liu. 2022. Text2light: Zero-shot text-driven HDR panorama generation. *ACM Transactions on Graphics (TOG)* 41, 6 (2022), 1–16.
- [5] Mohammad Reza Karimi Dastjerdi, Yannick Hold-Geoffroy, Jonathan Eisenmann, Siavash Khodadadeh, and Jean-François Lalonde. 2022. Guided Co-Modulated GAN for 360° Field of View Extrapolation. In *2022 International Conference on 3D Vision (3DV)*. IEEE, 475–485.
- [6] Daniel DeTone, Tomasz Malisiewicz, and Andrew Rabinovich. 2018. Superpoint: Self-supervised interest point detection and description. In *Proceedings of the IEEE conference on computer vision and pattern recognition workshops*. 224–236.
- [7] Prafulla Dhariwal and Alexander Nichol. 2021. Diffusion models beat gans on image synthesis. *Advances in Neural Information Processing Systems* 34 (2021), 8780–8794.
- [8] Patrick Esser, Robin Rombach, and Bjorn Ommer. 2021. Taming transformers for high-resolution image synthesis. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*. 12873–12883.
- [9] Marc-André Gardner, Kalyan Sunkavalli, Ersin Yumer, Xiaohui Shen, Emiliano Gambaretto, Christian Gagné, and Jean-François Lalonde. 2017. Learning to predict indoor illumination from a single image. *arXiv preprint arXiv:1704.00090* (2017).
- [10] Ian Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherilj Ozair, Aaron Courville, and Yoshua Bengio. 2020. Generative adversarial networks. *Commun. ACM* 63, 11 (2020), 139–144.
- [11] Takayuki Hara, Yusuke Mukuta, and Tatsuya Harada. 2022. Spherical Image Generation From a Few Normal-Field-of-View Images by Considering Scene Symmetry. *IEEE Transactions on Pattern Analysis and Machine Intelligence* (2022).
- [12] Heathen. [n.d.]. Discussion on Stable Diffusion WebUI. <https://github.com/automatic1111/stable-diffusion-webui/discussions/2670>
- [13] Martin Heusel, Hubert Ramsauer, Thomas Unterthiner, Bernhard Nessler, and Sepp Hochreiter. 2017. Gans trained by a two time-scale update rule converge to a local nash equilibrium. *Advances in neural information processing systems* 30 (2017).
- [14] Jonathan Ho, Ajay Jain, and Pieter Abbeel. 2020. Denoising diffusion probabilistic models. *Advances in Neural Information Processing Systems* 33 (2020), 6840–6851.
- [15] Jonathan Ho, Chitwan Saharia, William Chan, David J Fleet, Mohammad Norouzi, and Tim Salimans. 2022. Cascaded Diffusion Models for High Fidelity Image Generation. *J. Mach. Learn. Res.* 23, 47 (2022), 1–33.
- [16] Jonathan Ho and Tim Salimans. 2022. Classifier-free diffusion guidance. *arXiv preprint arXiv:2207.12598* (2022).
- [17] Tuomas Kynkäänniemi, Tero Karras, Miika Aittala, Timo Aila, and Jaakko Lehtinen. 2022. The Role of ImageNet Classes in Fréchet Inception Distance. *arXiv preprint arXiv:2203.06026* (2022).
- [18] Junnan Li, Dongxu Li, Caiming Xiong, and Steven Hoi. 2022. BLIP: Bootstrapping Language-Image Pre-training for Unified Vision-Language Understanding and Generation. In *ICML*.
- [19] Kang Liao, Lang Nie, Shujuan Huang, Chunyu Lin, Jing Zhang, Yao Zhao, Moncef Gabbouj, and Dacheng Tao. 2023. Deep Learning for Camera Calibration and Beyond: A Survey. *arXiv preprint arXiv:2303.10559* (2023).
- [20] Kang Liao, Xiangyu Xu, Chunyu Lin, Wenqi Ren, Yunchao Wei, and Yao Zhao. 2022. Cylind-Painting: Seamless 360 {°} Panoramic Image Outpainting and Beyond with Cylinder-Style Convolutions. *arXiv preprint arXiv:2204.08563* (2022).
- [21] Guilin Liu, Aysegül Dundar, Kevin J Shih, Ting-Chun Wang, Fitsum A Reda, Karan Sapra, Zhiding Yu, Xiaodong Yang, Andrew Tao, and Bryan Catanzaro. 2022. Partial convolution for padding, inpainting, and image synthesis. *IEEE Transactions on Pattern Analysis and Machine Intelligence* (2022).
- [22] David G Lowe. 2004. Distinctive image features from scale-invariant keypoints. *International journal of computer vision* 60 (2004), 91–110.
- [23] Andreas Lugmayr, Martin Danelljan, Andres Romero, Fisher Yu, Radu Timofte, and Luc Van Gool. 2022. Repaint: Inpainting using denoising diffusion probabilistic models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 11461–11471.
- [24] Wei-Chiu Ma, Anqi Joyce Yang, Shenlong Wang, Raquel Urtasun, and Antonio Torralba. 2022. Virtual correspondence: Humans as a cue for extreme-view geometry. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 15924–15934.
- [25] Ozan Oktay, Jo Schlemper, Loïc Le Folgoc, Matthew Lee, Mattias Heinrich, Kazunari Misawa, Kensaku Mori, Steven McDonagh, Nils Y Hammerla, Bernhard Kainz, et al. 2018. Attention u-net: Learning where to look for the pancreas. *arXiv preprint arXiv:1804.03999* (2018).
- [26] Gaurav Parmar, Richard Zhang, and Jun-Yan Zhu. 2022. On Aliased Resizing and Surprising Subtleties in GAN Evaluation. In *CVPR*.
- [27] Alec Radford, Jong Wook Kim, Chris Hallacy, A. Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, Gretchen Krueger, and Ilya Sutskever. 2021. Learning Transferable Visual Models From Natural Language Supervision. In *ICML*.
- [28] Chris Rockwell, Justin Johnson, and David F. Fouhey. 2022. The 8-Point Algorithm as an Inductive Bias for Relative Pose Prediction by ViTs. In *3DV*.
- [29] Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. 2021. High-Resolution Image Synthesis with Latent Diffusion Models. *arXiv:2112.10752 [cs.CV]*
- [30] Olaf Ronneberger, Philipp Fischer, and Thomas Brox. 2015. U-net: Convolutional networks for biomedical image segmentation. In *Medical Image Computing and Computer-Assisted Intervention—MICCAI 2015: 18th International Conference, Munich, Germany, October 5–9, 2015, Proceedings, Part III* 18. Springer, 234–241.
- [31] RunwayML. 2021. Stable Diffusion. <https://github.com/runwayml/stable-diffusion>.
- [32] Paul-Edouard Sarlin, Daniel DeTone, Tomasz Malisiewicz, and Andrew Rabinovich. 2020. SuperGlue: Learning feature matching with graph neural networks. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*. 4938–4947.
- [33] Christoph Schuhmann, Romain Beaumont, Richard Vencu, Cade Gordon, Ross Wightman, Mehdi Cherti, Theo Coombes, Aarush Katta, Clayton Mullis, Mitchell Wortsman, et al. 2022. Laion-5b: An open large-scale dataset for training next generation image-text models. *arXiv preprint arXiv:2210.08402* (2022).
- [34] Jiaming Song, Chenlin Meng, and Stefano Ermon. 2020. Denoising diffusion implicit models. *arXiv preprint arXiv:2010.02502* (2020).
- [35] Roman Suvorov, Elizaveta Logacheva, Anton Mashikhin, Anastasia Remizova, Arsenii Ashukha, Aleksei Silvestrov, Naejin Kong, Harshith Goka, Kiwoong Park, and Victor Lempitsky. 2022. Resolution-robust large mask inpainting with fourier convolutions. In *Proceedings of the IEEE/CVF winter conference on applications of computer vision*. 2149–2159.
- [36] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. *Advances in neural information processing systems* 30 (2017).
- [37] Guangcong Wang, Yinuo Yang, Chen Change Loy, and Ziwei Liu. 2022. Stylelight: Hdr panorama generation for lighting estimation and editing. In *Computer Vision—ECCV 2022: 17th European Conference, Tel Aviv, Israel, October 23–27, 2022, Proceedings, Part XV*. Springer, 477–492.
- [38] Yinhuai Wang, Jiwen Yu, and Jian Zhang. 2022. Zero-Shot Image Restoration Using Denoising Diffusion Null-Space Model. *arXiv preprint arXiv:2212.00490* (2022).
- [39] Jianxiong Xiao, Krista A Ehinger, Aude Oliva, and Antonio Torralba. 2012. Recognizing scene viewpoint using panoramic place representation. In *2012 IEEE Conference on Computer Vision and Pattern Recognition*. IEEE, 2695–2702.
- [40] Lvmin Zhang and Maneesh Agrawala. 2023. Adding Conditional Control to Text-to-Image Diffusion Models. *arXiv:2302.05543 [cs.CV]*
- [41] Shengyu Zhao, Jonathan Cui, Yilun Sheng, Yue Dong, Xiao Liang, Eric I Chang, and Yan Xu. 2021. Large Scale Image Completion via Co-Modulated Generative Adversarial Networks. In *International Conference on Learning Representations (ICLR)*.
- [42] Shengyu Zhao, Jonathan Cui, Yilun Sheng, Yue Dong, Xiao Liang, Eric I Chang, and Yan Xu. 2021. Large scale image completion via co-modulated generative adversarial networks. *arXiv preprint arXiv:2103.10428* (2021).