RecDiffusion: Rectangling for Image Stitching with Diffusion Models

Tianhao Zhou[1*] Haipeng Li[1*] Ziyi Wang[1] Ao Luo[3,2] Chen-Lin Zhang[4] Jiajun Li[4]
Bing Zeng[1] Shuaicheng Liu[1†]
[1] University of Electronic Science and Technology of China
[2] Megvii Technology [3] Southwest Jiaotong University [4] 4Paradigm Inc
{thzhou,lihaipeng,ziyiwang@std.,eezeng,liushuaicheng@}uestc.edu.cn,
{aoluo}@swjtu.edu.cn, {zclnjucs,taringlee}@gmail.com

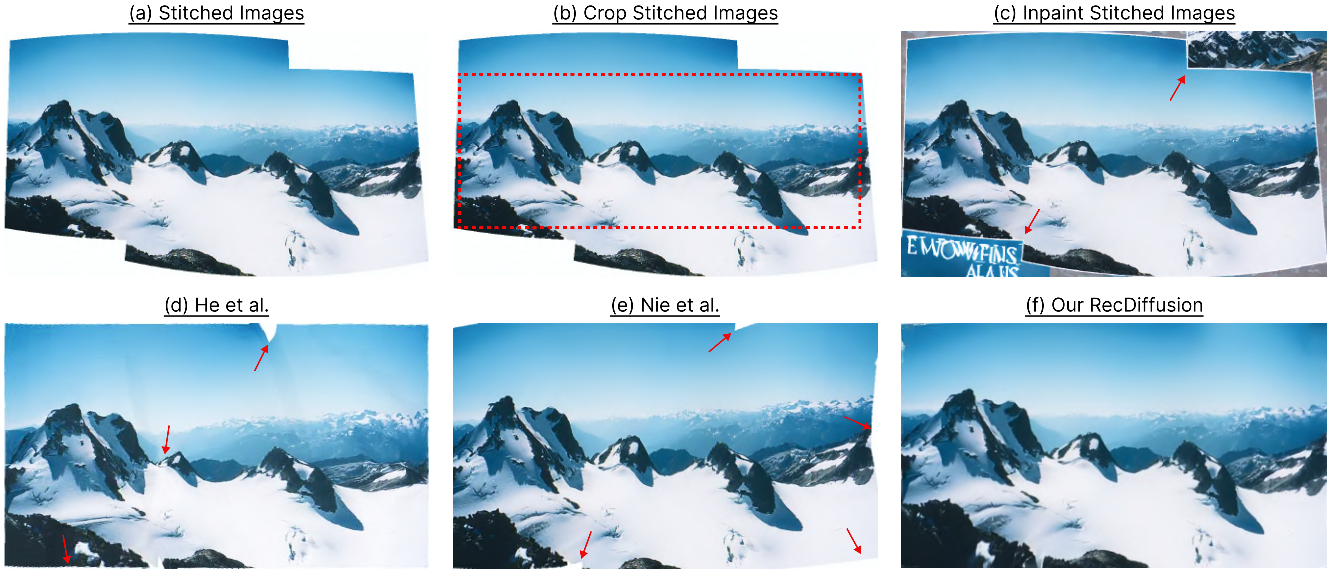


Figure 1. Visual comparisons of our proposed RecDiffusion and previous rectangling approaches including cropping-based, He *et al.* [9], inpainting using Stable Diffusion [37], and Nie *et al.* [32]. We can see that simple cropping reduces the field-of-view, the inpainting-based method introduces unsatisfactory extra contents, He *et al.* [9] presents distortion and edge artifacts, and Nie *et al.* [32] unable to maintain a satisfactory rectangular boundary. In contrast, our method properly complements the boundaries and avoids artifacts

Abstract

*Image stitching from different captures often results in non-rectangular boundaries, which is often considered unappealing. To solve non-rectangular boundaries, current solutions involve cropping, which discards image content, inpainting, which can introduce unrelated content, or warping, which can distort non-linear features and introduce artifacts. To overcome these issues, we introduce a novel diffusion-based learning framework, **RecDiffusion**, for image stitching rectangling. This framework combines Motion Diffusion Models (MDM) to generate motion fields, effectively transitioning from the stitched image's irregular borders to a geometrically corrected intermediary. Followed by Content Diffusion Models (CDM) for image detail refinement. Notably, our sampling process utilizes a weighted map to identify regions needing correction during each iteration of CDM. Our RecDiffusion ensures geometric accuracy and overall visual appeal, surpassing all previous methods in both quantitative and qualitative measures when evaluated on public benchmarks. Code is released at https://github.com/lhaippp/RecDiffusion.*

[*] Equal contribution.
[†] Corresponding author.
[‡] the work is done when Chen-Lin Zhang was in 4paradigm, inc. Chen-Lin Zhang is now at Moonshot AI, Ltd.

1. Introduction

Image stitching is a technique in which multiple overlapping images of the scene are stitched together to generate an image with a wide field of view (FOV) and high resolution [44]. These methods often adopt homography [26] for global or mesh warps [51] for local alignment of overlapping regions. However, the image boundary produced by stitching algorithms is no longer rectangular due to different capture perspectives, which is unpleasant to view and has been tolerated for quite a long time, as shown in Fig. 1(a).

To handle such an issue, the most straightforward way is to crop the image by the largest incised rectangle, as shown in Fig. 1(b). However, the image contents near the boundary have to be discarded, which is also unpleasant due to the loss of image pixels. Another method could be leveraging the recent state-of-the-art (SOTA) generative models to achieve the inpainting, i.e., Stable Diffusion [37], to complete stitched ones, as illustrated in Fig. 1(c). However, it introduces extra content which does not belong to the original images. He *et al.* [9] proposed the concept of image rectangling, where seam carving technique [1] is adopted for inserting an abundant of seams for an initial rectangular shape, and then a mesh is optimized that warps the image for the final rectangling result. In this way, salient image structures can be preserved, while less important regions are either stretched or squeezed to realize the rectangular shape. However, warping-based methods [8, 9, 18] typically preserve only linear structures, such as Manhattan World. Non-linear structures [55] are usually distorted. An example is shown in Fig. 1(d).

Recently, Nie *et al.* [32] proposed a deep learning pipeline that directly minimizes a mesh to warp the image, demonstrating significant improvements over the traditional one [9]. However, the warping-based method could introduce artifacts and noise due to the lack of accuracy of warping motion fields and the issues inherent in the warping operation [6], yielding distortions artifacts (inconsistent boundaries and discontinuous lines) as shown in Fig. 1(e).

In this work, we aim to reformulate the task of image rectangling using diffusion models (DMs) [40, 42, 50]. Our reasons are twofold: 1) Diffusion models have recently achieved notable performances and demonstrated significant potential in various fields, including, but not limited to, image synthesis [11, 27], restoration [28, 29, 47], and enhancement [14]. Specifically, DMs have proven to be effective in various motion-related tasks, such as human motion generation [45], homography synthesis/estimation [19], and depth/optical flow estimation [39]; 2) We believe that predicting motion from a single image is an ill-posed problem that can be adequately addressed by DMs. These models have notably improved the outcomes of classical ill-posed problems, like image restoration [46, 47]. Therefore, based on intuition and previous successes, we propose the first diffusion-based learning algorithm as a baseline to tackle the mentioned challenges. Instead of merely seeking a pair of initial and target meshes for warping, we produce the final rectangular results through motion warping operations and image content refinement.

Specifically, the input to the network is the stitched image $I_{\mathbf{S}}$ with irregular boundary, and the output is the corresponding image with rectangular boundary $I_{\mathbf{R}'}$. In particular, we first fed $I_{\mathbf{S}}$ into the proposed Motion Diffusion Models (**MDM**) to produce a motion field. Then we utilize the field to warp $I_{\mathbf{S}}$ to produce a geometrically correct result $I_{\hat{\mathbf{R}}}$, which represents that majority of the content is corrected, but still leaving some details to be optimized, such as the white edges near the boundary, discontinuous lines, and noise. To handle it, we pass $I_{\hat{\mathbf{R}}}$ into another proposed Content Diffusion Model (**CDM**). To be noticed, the sampling procedure is achieved by fusing $I_{\hat{\mathbf{R}}}$ with the output of CDM. As inspired by Rank-Nullity Theorem [47], we compute a weighted map $M_{\hat{\mathbf{R}}}$ to identify the confident regions in $I_{\hat{\mathbf{R}}}$, as a result, for every sampling step of CDM, we keep content of $I_{\hat{\mathbf{R}}}$ according to $M_{\hat{\mathbf{R}}}$, and we extract content from CDM's output via $1 - M_{\hat{\mathbf{R}}}$, then they are combined together to be fed into another sampling iteration. With such a strategy, we could generate geometrically accurate and visually pleasing results that outperform all previous methods in quantitative and qualitative comparisons.

Our contributions can be summarized as follows:

- We propose the first diffusion-based framework for image stitching rectangling, namely, **RecDiffusion**.
- We propose a Motion Diffusion Model (**MDM**) to generate rectangling motion fields, then a Content Diffusion Model (**CDM**) to refine image details.
- Extensive experiments show that our approach achieves state-of-the-art performance on public benchmarks when compared to previous both traditional and deep methods.

2. Related Work

2.1. Image Rectangling

Image irregular boundary is often produced by applying spatial transformations [7], such as image rotation [8], panorama construction [3], video stabilization [54] and image stitching [2]. The most straightforward approach is to crop the empty region for regular boundary. However, some image contents will also be sacrificed. He *et al* [9] introduced the image rectangling task, where a mesh is optimized that can warp the image to realize rectangular boundary while retain image contents. Nie *et al* [32] proposed the first deep learning based rectangling solution, demonstrating superior performances than directly inpaint/synthesize missing regions at borders [43]. Some approaches target at rectangling images under a specific situation, e.g., rotation correction [8, 35] and wild-angle rectification [21]. Video based methods can further rely on temporal information for compensations [30, 48]. In this work, we introduce DM for image stitching rectangling.

2.2. Image Stitching

Image stitching technique aims to create a larger field of view by combining multiple images of a same scene but captured under different perspectives [44]. Most of image stitching methods are traditional ones, which concentrate on several different but important aspects, such as effective image feature utilization [12, 20], dealing with large parallax [17, 23, 52], minimizing distortions [22, 51], preserving shapes of non-overlapping regions [4] and maintaining salient image structures [24, 53]. Deep-based methods can improve the performances under challenging conditions, e.g., low or weak textures [31, 33, 34]. Although many previous works have achieved high quality of stitched images, the shape of the stitching boundary is largely overlooked. In this work, we do not stitch images, but rectanlge the irregular boundary after stitching.

2.3. Diffusion Models

This work is related to Diffusion Models [50], which are generative models, gaining significant popularity recently. DMs work by destroying training data through the successive addition of Gaussian noise, and then learning to recover the data by reversing this noising process [42]. DMs can generate data by simply passing randomly sampled noise through the learned denoising process [11, 25, 41, 47], which iteratively reverse a diffusion process that maps from randomly sampled Gaussian noise to the latent distributions, avoiding issues of instability and model-collapse that often present in previous generative models. Many DMs-based approaches have been proposed with respect to different applications, such as homography synthesis/estimation [19], optical flow estimation [39], human motion synthesis [45], image restoration/enhancement [5, 14, 47], 3D model synthesis [36] and image inpainting [27, 37, 38, 49]. In this work, we introduce DMs for the task of rectangling stitched images.

3. Method

3.1. Overview

The schema of the processing of stitched images is elucidated in Fig. 2. Upon the acquisition of stitched images, we process them by leveraging two diffusion models. In its primary stage, Motion Diffusion Models (MDM) generates motion fields that transform stitched images with irregular edges and white margins into seamlessly rectangular formats devoid of these margins, as delineated in Sec. 3.3. The "image-to-motion" paradigm is adopted during this phase, noted for its efficacy in the delineation of low-level features [39]. However, MDM can introduce noise and morphological errors from imperfect motion fields and the complexity of remapping operations [6], evident in the "Warped Stitched Image" in Fig. 2. To ameliorate these artifacts, a secondary phase is invoked, leveraging Content Diffusion Models (CDM), which specifically target the refinement of the images post-MDM application, especially within regions that present issues. This enhancement is done through a novel strategy employing weighted sampling, predicated on the Rank-Nullity Theorem (RNT) principles [47].

3.2. Diffusion Models

The foundational principles of diffusion models, as explicated by Sohl-Dickstein *et al.* [40] and subsequently refined by Ho *et al.* [11], utilize a Markovian transition process over a total of T steps to instigate the sequential infusion of Gaussian noise into an originating data distribution $\mathbf{x}_0 \sim q(\mathbf{x})$. This method generates an array of incrementally noisier images $\{\mathbf{x}_1, \ldots, \mathbf{x}_T\}$, collectively termed forward diffusion, succinctly expressed as follows:

$$q(\mathbf{x}_{1:T}|\mathbf{x}_0) = \prod_{t=1}^{T} q(\mathbf{x}_t|\mathbf{x}_{t-1}). \tag{1}$$

The induction of noise at each interval adheres to a designated Gaussian distribution delineated by a variance schedule $\{\beta_t \in (0,1)\}_{t=1}^{T}$:

$$q(\mathbf{x}_t|\mathbf{x}_{t-1}) = \mathcal{N}(\mathbf{x}_t; \sqrt{1-\beta_t}\mathbf{x}_{t-1}, \beta_t\mathbf{I}). \tag{2}$$

Employing the reparameterization technique outlined by Kingma *et al.* [16], it becomes feasible to sample from any intermediary distribution $\mathbf{x}_t$ for an arbitrary $t \in [1, T]$:

$$q(\mathbf{x}_t|\mathbf{x}_0) = \mathcal{N}(\mathbf{x}_t; \sqrt{\bar{\alpha}_t}\mathbf{x}_0, (1-\bar{\alpha}_t)\mathbf{I}), \tag{3}$$

where $\alpha_t = 1 - \beta_t$ and $\bar{\alpha}_t = \prod_{i=1}^{t} \alpha_i$. Thereafter, we introduce an optimized denoising model θ to inverse the process of diffusion and thereby generate images conforming to a target data distribution, commencing from isotropic Gaussian noise $\mathbf{x}_T \sim \mathcal{N}(\mathbf{0}, \mathbf{I})$:

$$p_\theta(\mathbf{x}_{0:T}) = p(\mathbf{x}_T) \prod_{t=1}^{T} p_\theta(\mathbf{x}_{t-1}|\mathbf{x}_t), \tag{4}$$

$$p_\theta(\mathbf{x}_{t-1}|\mathbf{x}_t) = \mathcal{N}(\mathbf{x}_{t-1}; \mu_\theta(\mathbf{x}_t, t), \sigma_t^2\mathbf{I}). \tag{5}$$

By executing this inverted transition, the system is endowed with the ability to transform a Gaussian distribution back to the initial data distribution.

To bolster the model's control over the generative procedure and improve the fidelity of the resultant imagery, we introduce additional conditioning variables $\mathbf{y}$ into our architectural framework, following methods advocated by Ho *et*

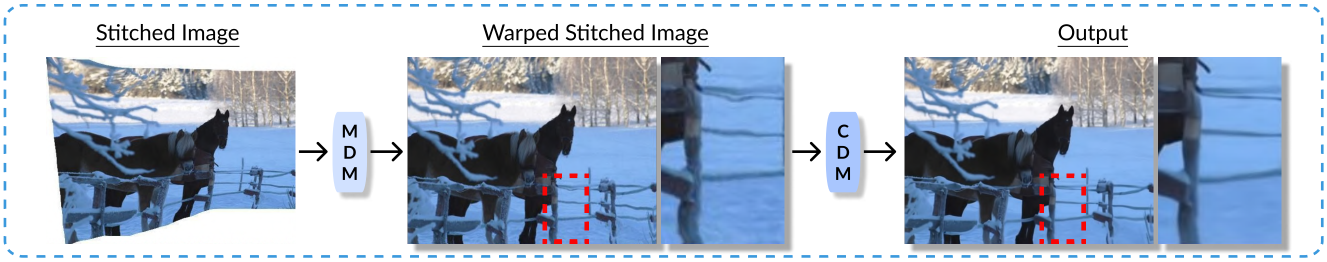
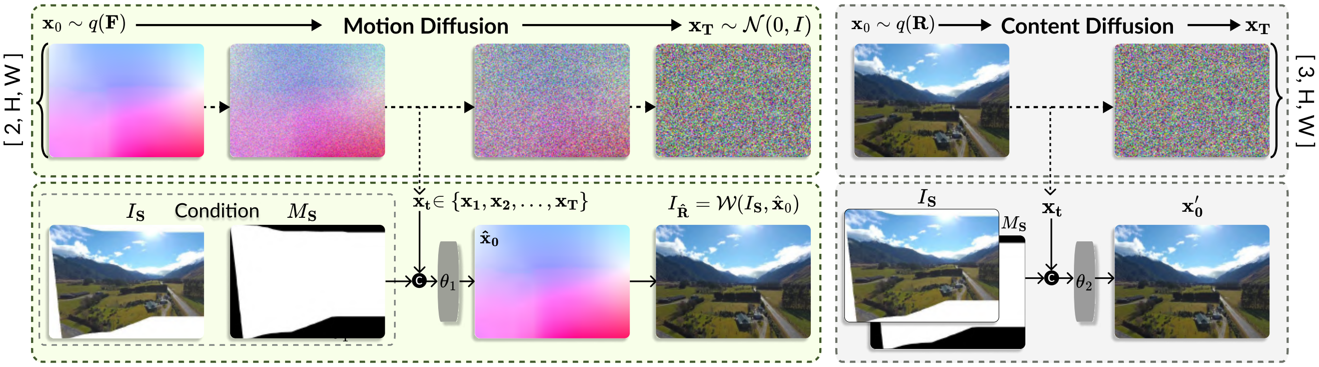


Figure 2. Workflow of RecDiffusion. Initially, Motion Diffusion Models (MDM) are employed to convert irregularly-bordered stitched images into a seamless rectangular form via generated motion fields, which occasionally introduce artifacts like distortion (highlighed by the red box). Content Diffusion Models (CDM) subsequently refine these images.



$\mathbf{x}_0 \sim q(\mathbf{F})$
Motion Diffusion
$\mathbf{x_T} \sim \mathcal{N}(0, I)$
[2, H, W]
$\mathbf{x}_0 \sim q(\mathbf{R})$
Content Diffusion
$\mathbf{x_T}$
[3, H, W]
$I_\mathbf{S}$
Condition
$M_\mathbf{S}$
$\mathbf{x_t} \in \{\mathbf{x_1}, \mathbf{x_2}, \dots, \mathbf{x_T}\}$
θ_1
$\hat{\mathbf{x}}_0$
$I_{\hat{\mathbf{R}}} = \mathcal{W}(I_\mathbf{S}, \hat{\mathbf{x}}_0)$
$I_\mathbf{S}$
$M_\mathbf{S}$
$\mathbf{x_t}$
θ_2
$\mathbf{x}'_0$

Figure 3. Overview of training procedures. The left block illustrates the training of MDM, which generates motion fields $\hat{\mathbf{x}}_\mathbf{0}$ from stitched images $I_\mathbf{S}$ and their masks $M_\mathbf{S}$, transforming $I_\mathbf{S}$ into rectangling images $I_{\hat{\mathbf{R}}}$. The right block shows the training of CDM under the same conditions ($I_\mathbf{S}$, $M_\mathbf{S}$) to directly generate a rectangling result $\mathbf{x}'_\mathbf{0}$. Both methods aim to reconstruct high-definition rectangling images from stitched inputs, respectively realizing it via motion and content-based manners.

al. [10]. The conditioning mechanism operates by merging these variables with the intermediary noisy data, yielding enhanced results:

$$p_\theta(\mathbf{x}_{t-1}|\mathbf{x}_t, \mathbf{y}) = \mathcal{N}(\mathbf{x}_{t-1}; \mu_\theta(\mathbf{x}_t, t, \mathbf{y}), \sigma_t^2\mathbf{I}). \quad (6)$$

3.3. Rectangling Diffusion Models

In our approach, the stitched images denoted as $\mathbf{S}$ can be regarded as a degraded counterpart of rectangling images $\mathbf{R}$, where the composite degradation is attributed to both motion warping and content degradation. Consequently, the proposed framework is designed to learn the transformation from $\mathbf{S}$ back to $\mathbf{R}$, which it achieves by training a motion diffusion model (MDM) and a content diffusion model (CDM) for their respective degradation processes.

Training Process: We initiate the training of MDM by constructing "image-to-motion" diffusion models tasked with generating the rectangling motion fields $\mathbf{F}$ that reverse stitched images, $\mathbf{S}$, to rectangling images, $\mathbf{R}$. As depicted in the left block of Fig. 3, based on the conditional framework defined in Eq. 6, starting from a random sampling of data points, $\mathbf{x}_0 \sim q(\mathbf{F})$, we iteratively introduce noise following Eq. 3. Inputs to the network θ_1 include the associated stitched images $I_\mathbf{S}$, their corresponding masks $M_\mathbf{S}$ delineating validated image content, along with noised motion fields $\mathbf{x_t}$. The output of this network, generated motion fields $\hat{\mathbf{x}}_\mathbf{0}$, are then utilized to rectangle $I_\mathbf{S}$ via a warping function $\mathcal{W}(.)$, thereby yielding the rectangled image $I_{\hat{\mathbf{R}}}$:

$$I_{\hat{\mathbf{R}}} = \mathcal{W}(I_\mathbf{S}, \hat{\mathbf{x}}_\mathbf{0}). \quad (7)$$

The training loss incorporates two components: the mean square error ℓ_{mse} quantifying the divergence between input and output motion fields defined by:

$$\ell_{mse} = ||\hat{\mathbf{x}}_\mathbf{0} - \mathbf{x}_0||^2, \quad (8)$$

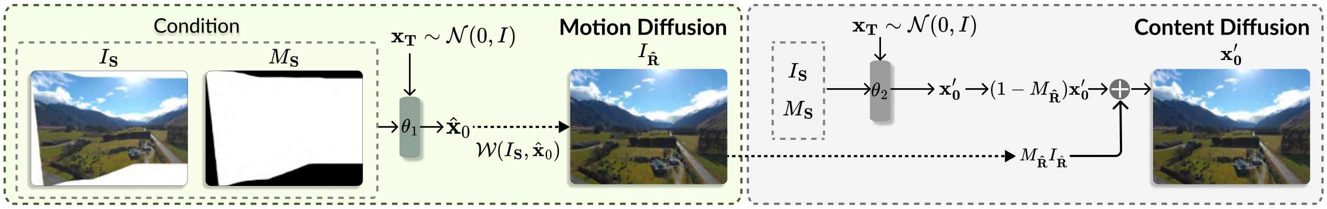
and a Photometric loss assessing the disparity between resulting rectangled imagery and ground truth, given as:

$$\ell_{pl} = \left| I_{\hat{\mathbf{R}}} - I_\mathbf{R} \right|. \quad (9)$$

The composite loss function is therefore presented as a weighted sum:

$$\ell_{mdm} = \ell_{mse} + \frac{|\ell_{mse}|}{|\ell_{pl}|} \cdot \ell_{pl}, \quad (10)$$

where the norm of $\frac{|\ell_{mse}|}{|\ell_{pl}|}$ is used to balance the contribution of each loss component to the overall training objective.



Condition
$I_{\mathbf{S}}$
$M_{\mathbf{S}}$
$\mathbf{x_T} \sim \mathcal{N}(0, I)$
Motion Diffusion
$I_{\hat{\mathbf{R}}}$
θ_1
$\hat{\mathbf{x}}_0$
$\mathcal{W}(I_{\mathbf{S}}, \hat{\mathbf{x}}_0)$
$\mathbf{x_T} \sim \mathcal{N}(0, I)$
Content Diffusion
$\mathbf{x}'_0$
$I_{\mathbf{S}}$
$M_{\mathbf{S}}$
θ_2
$\mathbf{x}'_0$
$(1 - M_{\hat{\mathbf{R}}})\mathbf{x}'_0$
$M_{\hat{\mathbf{R}}} I_{\hat{\mathbf{R}}}$

Figure 4. Illustration of the sampling procedure. Initially, stitching images $I_{\mathbf{S}}$ and masks $M_{\mathbf{S}}$ are processed by MDM, which generates motion fields $\hat{\mathbf{x}}_0$ iteratively and warps $I_{\mathbf{S}}$ to form preliminary rectangling images $I_{\hat{\mathbf{R}}}$ with corresponding confidence masks $M_{\hat{\mathbf{R}}}$. Secondly, for each sampling step, CDM polishes these images by keeping confidence regions $M_{\hat{\mathbf{R}}}$ of $I_{\hat{\mathbf{R}}}$ and updating non-confidence regions $(1 - M_{\hat{\mathbf{R}}})$ via the output of CDM $\mathbf{x}'_0$. As a result, we are capable of iteratively reconstructing ideal rectangling images.

For the training of the Content Diffusion Model (CDM), we employ a parallel strategy, as demonstrated in the right block of Fig. 3. Here, the CDM governs an "image-to-image" diffusion process involving the model, θ_2, which aims to refine the MDM rectangling images, $I_{\hat{\mathbf{R}}}$. Different from MDM, in CDM, we direct the generation process towards the original rectangling images through sampling, $\mathbf{x}_0 \sim q(\mathbf{R})$, while retaining the same conditional inputs, specifically the stitched images $I_{\mathbf{S}}$ and masks $M_{\mathbf{S}}$. Consequently, the model produces an enhanced version of the rectangling images, denoted as $\mathbf{x}'_0$. The associated training loss is the MSELoss measuring the distance between these enhanced images and the ground truth rectangling images:

$$\ell_{cdm} = ||\mathbf{x}'_0 - \mathbf{x}_0||^2. \tag{11}$$

In harnessing power of diffusion models to capture and correct motion-based and content-based degradations, the integrated training process enables the reconstruction of high-fidelity rectangling images from input stitched counterparts.

Sampling Process: Fig. 4 delineates the procedure we follow for the sampling process. After adequately training both the Motion Diffusion Model (MDM) and the Content Diffusion Model (CDM), we progress through two principal steps to transform stitching images towards refined rectangling images. Initially, pairs of stitching images $I_{\mathbf{S}}$ along with their corresponding masks $M_{\mathbf{S}}$ serve as the input conditions. From here, the correction motion fields $\hat{\mathbf{x}}_0$ are iteratively estimated from Gaussian noise over a series of steps. These fields then warp $I_{\mathbf{S}}$ to produce preliminary rectangling results, $I_{\hat{\mathbf{R}}}$. Considering that this process may introduce noise and artifacts due to the accuracy of the generated motion fields and the properties of the remapping operation, our strategy includes computing a confidence mask, $M_{\hat{\mathbf{R}}}$, to categorize regions according to their reliability in terms of confidence levels.

More specifically, $M_{\hat{\mathbf{R}}}$ is computed via 3 different masks: 1) input stitched image masks $M_{\mathbf{S}}$, 2) intensity map of $\hat{\mathbf{x}}_0$, M_0, which is obtained by the normalized displacement of grids, 3) white edge mask of $I_{\hat{\mathbf{R}}}$ as M_1. Then we can formulate $M_{\hat{\mathbf{R}}}$ as:

$$M_{\hat{\mathbf{R}}} = 1 - \max\left\{M_1, \frac{\omega_0 \cdot 1 + M_0 + M_{\mathbf{S}}}{\omega_0 + 2}\right\}, \tag{12}$$

where ω_0 is a hyper-parameter to be tuned.

Secondly, we utilize CDM to refine the noise and artifacts present within $I_{\hat{\mathbf{R}}}$. To facilitate this, we deploy a weighted sampling technique inspired by the Rank-Nullity Theorem (RNT) [47]. We commence by establishing two primary constraints: a consistency constraint (Eq. 13), which asserts that $\hat{\mathbf{r}}$ (representative of vectorized $I_{\hat{\mathbf{R}}}$) should match the vectorized desired rectangling images $\mathbf{r}'$, after the degradation $\mathbf{A}$. Moreover, we implement a realism constraint (Eq. 14), proposing that the generated results for $\mathbf{r}'$ conform with the expected distribution:

$$Consistency : \hat{\mathbf{r}} = \mathbf{A}\mathbf{r}', \tag{13}$$

$$Realism : \mathbf{r}' \sim q\left(\mathbf{R}\right). \tag{14}$$

We then formulate an equation based on Rank-Nullity considerations (Eq. 15) that serves to merge the constraints of consistency and realism:

$$\mathbf{r}' = \mathbf{A}^{\dagger}\mathbf{A}\mathbf{r}' + \left(\mathbf{I} - \mathbf{A}^{\dagger}\mathbf{A}\right)\mathbf{r}', \tag{15}$$

where $\mathbf{A}^{\dagger}$ represents the Pseudo-inverse of $\mathbf{A}$. This equation expresses $\mathbf{r}'$ as a combination of its projection into the range-space of a matrix $\mathbf{A}$ and its projection into the corresponding null-space.

Back to our method, we desire to produce favorable rectangling images $I_{\mathbf{R}'}$ (where $\mathbf{r}'$ represents the vectorizing format) via RNT. To achieve it, we consider the confident regions $M_{\hat{\mathbf{R}}}$ of $I_{\hat{\mathbf{R}}}$ to be the range space, and the rest regions as the null space. Therefore, the degradation matrix $\mathbf{A}$ is replaced with confidence masks, producing a novel relationship between $\hat{\mathbf{r}}$ and $\mathbf{r}'$ for each sample step as Eq. 16, in which the multiplication between $\mathbf{M}$ and $\mathbf{r}$ is element-wise:

$$\mathbf{r}' = \mathbf{M}\sqrt{\bar{\alpha}_t}\hat{\mathbf{r}} + \left(\mathbf{I} - \mathbf{M}\right)\mathbf{r}', \tag{16}$$

in essence, the diagonal matrix $\mathbf{M}$ that arises from vectorizing $M_{\hat{\mathbf{R}}}$ helps integrate the confidence levels associated with different regions of the image. With this revised relationship, the refinement process iteratively adjusts $\hat{\mathbf{r}}$ towards the ultimate target, $\mathbf{r}'$.

This iterative process is graphically depicted in the right panel of Fig. 4. Specifically, for each iteration, the algorithm preserves pixels of $\hat{\mathbf{r}}$ with high confidence, as indicated by the mask $\mathbf{M}$, which is equal to $M_{\hat{\mathbf{R}}} I_{\hat{\mathbf{R}}}$, and is then diffused to timestep t by multiplying $\sqrt{\bar{\alpha}_t}$. Conversely, for the remaining pixels identified by $(\mathbf{I} - \mathbf{M})$, the output of the CDM is used to substitute values with the goal of reducing noise and enhancing realism. It can be realized as $(1 - M_{\hat{\mathbf{R}}})\mathbf{x}'_0$. Such a sampling approach allows us to progressively reconstruct $\mathbf{r}'$, thus achieving step-by-step refinement of $\mathbf{x}'_0$, accomplishing the rectangling images $I_{\mathbf{R}'}$.

4. Experiment

We provide configurations of the experiment in Sec. 4.1. The comparisons with other methods are shown in Sec. 4.2 and Sec. 4.3. In addition, we test the generalizability of the models in Sec. 4.4 and compare with inpainting methods in Sec. 4.5. Lastly we conduct ablation studies in Sec. 4.6. Furthermore, we provide a brief introduction to datasets and dynamic visualizations in **Supplementary Materials** to better demonstrate the results.

4.1. Implementation Details

The proposed framework consists of MDM and CDM. The design of them follows DDIM [41] and classifier-free method (CFG) [10]. Both of them are trained using Adam optimizer [15] with parameters $\beta_1 = 0.9$, $\beta_2 = 0.99$. The generated pseudo motion fields used to train MDM is from the previous state-of-the-art method, i.e., Nie *et al.* [32]. For the configuration of MDM, the condition scaling of CFG is 6, learning rate is 2.0×10^{-4}, batch size is 64, sampling step is 2, the number of training steps is $320,000$. For CDM, the batch size is 32, learning rate is 1.0×10^{-5}, sampling step is 200, and the number of training steps is $450,000$. The time taken to train on 8 NVIDIA A100s are 3 and 4 days for MDM and CDM, respectively. More details will be demonstrated in Supplementary Materials.

4.2. Quantitative Comparison

We adopt the evaluation settings from previous studies, utilizing the Fréchet inception distance (FID), Structural Similarity Index (SSIM), and Peak Signal-to-Noise Ratio (PSNR) to assess these methods. Our evaluation on the **DIR-D** dataset, presented in Table 1, compares our approach with both the traditional rectangling method [9] and deep learning-based technique [32]. Specifically, we calculate FID on trainset as 519 test cases are not enough to compute a meaningful score. Previous methods tend to treat rectangling as a regression problem, addressing it with specialized architectures and task-specific loss functions, such as local-to-global strategies, feature warps, perception loss, or grid constraints. In contrast, our generative framework does not rely on specialized components nor regression frameworks, relying exclusively on diffusion models and achieving superior performance across all metrics, establishing a new state-of-the-art. It offers a novel potential technological path to solving the problem.

Method	FID ↓	SSIM ↑	PSNR ↑
Reference	12.25	0.3245	11.30
He *et al.* [9]	-	0.3775	14.70
Nie *et al.* [32]	4.14	0.7141	21.28
Ours	**3.63**	**0.7733**	**22.21**

Table 1. Quantitative comparisons of PSNR, SSIM, and FID between our method and other rectangling methods on the DIR-D [32] test set. “Reference” denotes that the metrics are computed by using input stitched images as rectangling results. The best results are highlighted in **bold**.

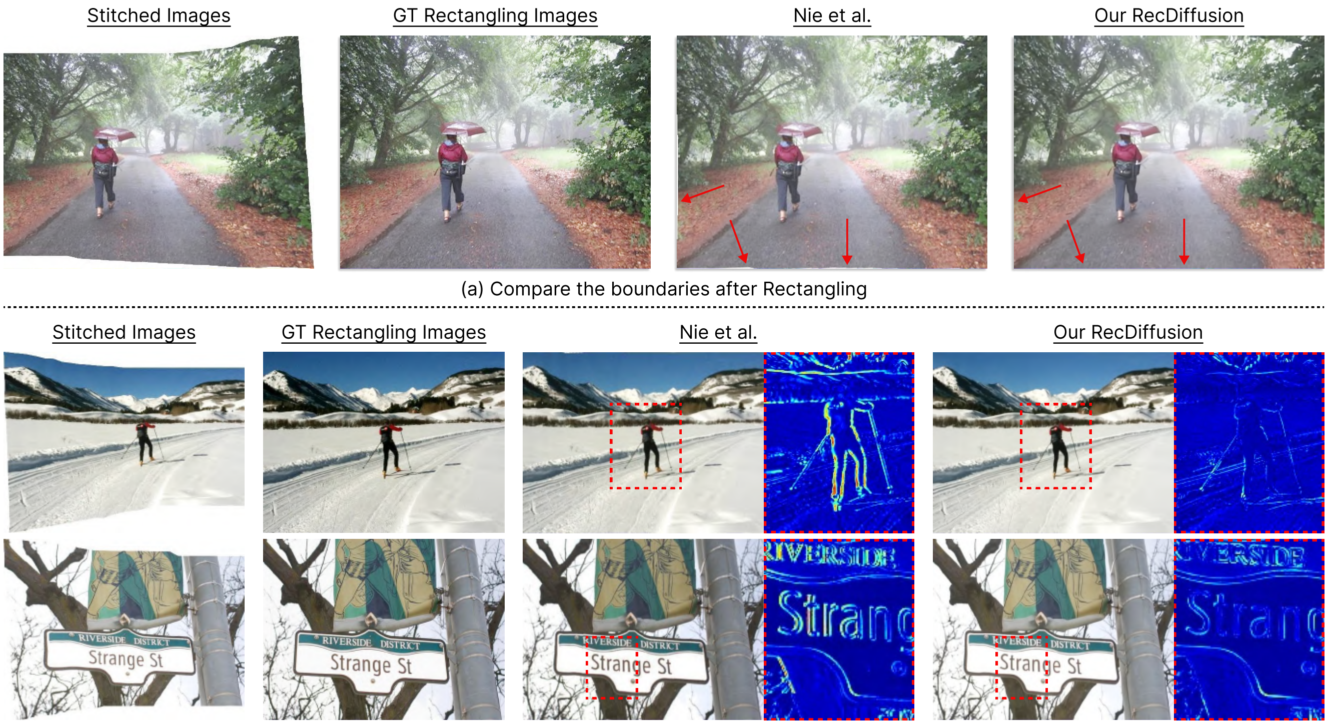
4.3. Qualitative Comparison

Our method is evaluated against the previous state-of-the-art method on **DIR-D** [32]. The visual comparisons are respectively illustrated in Fig. 5. For comparisons in Fig. 5 (a), we mainly compare whether the corrected stitched images are seamless rectangular ones or not, because as far as we know one of the most key aspects of the rectangling task is the complete elimination of irregular boundaries of the stitched images. However, Nie *et al.* [32] leveraging the warping meshes to achieve rectangling, naturally faces the risk of irregular boundary artifacts due to the accuracy of correcting motion and the inherent problems with warp operations. We use red arrows to indicate those white edging regions in the figure. On the contrary, our diffusion models-based framework locates the issue at the schematic side and is capable of generating desired rectangling images.

On the other hand, despite the incomplete white edges, artifacts could occur within the images. For example, line discontinuities and local distortions can occur, due to the lack of accuracy and smoothness of the warping motion fields. We demonstrate related images in the Fig. 5 (b). More specifically, to vividly demonstrate the similarities between produced results and GT images, we adopt the alignment heatmap [13], where darker regions correspond to better similarity. We encircle some of the ROIs in the graphs, which are subject contents. From the results, we can observe that our produced results are closer to the GT rectangling images, thus suffering from less artifacts. More dynamic results in GIF format can be found in **Supplementary Materials**.

4.4. Generalizability Experiments

Experiments involve zero-shot inference on **APAP-conssite** dataset [51] using He *et al.* [9], while Nie *et al.* [32] and



(a) Compare the boundaries after Rectangling

(b) Compare the local similarity with GT Rectangling images

Figure 5. Comparative Evaluation of Nie *et al.* [32] on the DIR-D Dataset. The input stitched images and the GT rectangling references are displayed in the first two columns. The third column shows the rectangling results by Nie *et al.*, while our proposed diffusion models-based outcomes are exhibited in the last column. In figure (a), red arrows accentuate white edge artifacts present in the outcomes of the previous state-of-the-art. Figure (b) scrutinizes the presence of internal artifacts such as line discontinuities and local distortions, highlighted within Regions of Interest (ROIs) circled on alignment heatmaps where darker shades signal higher fidelity to the ground truth. Our results demonstrate enhanced similarity to the ground truth, indicating a significant reduction in artifacts compared to the previous method.

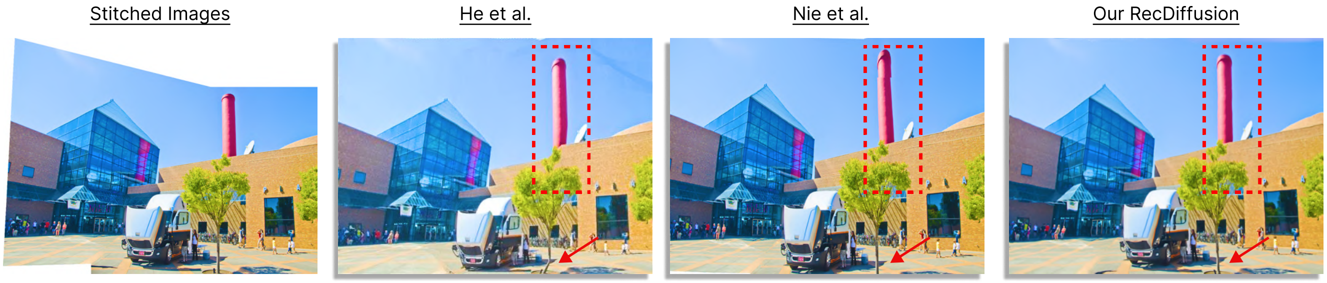


Figure 6. We test the zero-shot capacity of different methods on **APAP-conssite** [51], including He *et al.* [9], Nie *et al.* [32] and our RecDiffusion, trained on **DIR-D** [32]. Roofs and branches twisted (red boxes and arrows in He *et al.* result), chimney breakage and flower bed moved out of figure (red boxes and arrows in Nie *et al.* result) exist in their outputs. Our results performs the best among them.

our RecDiffusion are pre-trained on the **DIR-D** dataset [32]. Outcomes are illustrated in Fig. 6. From the results, we observe that other methods produce artifacts as highlighted by red box and arrow, for example, chimneys and branches are twisted in the result of He *et al.*. The output of Nie *et al.* also contains line discontinuity (red box) and flower bed is removed from the bottom of the figure (red arrow). On the contrary, our framework's robust backbone ensures its generalizability across different datasets.

4.5. Comparison with Inpainting Methods

Image rectangling aims to eliminate irregular boundaries while maintaining as much data consistency and achieving good qualitative results as possible, therefore, previous methods [9, 32] choose to warp stitching images. While inpainting methods [37, 38] are powerful at generating visu-

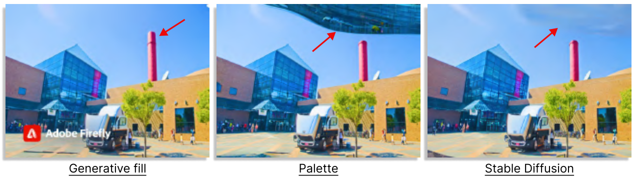


Figure 7. Inpainted stitching images by Adobe commercial software - Generative fill, Palette [38] and Stable Diffusion 2.1 [37].

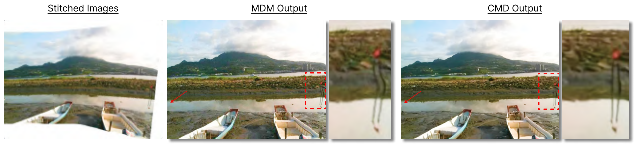


Figure 8. Illustration of the output of MDM and CDM. We can find that the local distortion is well handled by CDM.

Method	FID ↓	SSIM ↑	PSNR ↑
Reference	12.25	0.3245	11.30
Palette [38]	-	0.3315	14.49
Stable Diffusion 2.1 [37]	15.58	0.3276	14.23
Ours	3.63	0.7733	22.21

Table 2. Quantitative comparisons of PSNR, SSIM, and FID with inpainting methods on the DIR-D dataset.

ally pleasing outcomes, they tend to introduce extra content into the stitched ones, as demonstrated in Fig. 7, thus affecting the data consistency negatively. As shown in Table 2, inpainted stitching images (row 2 and 3) result in the much lower PSNR/SSIM metrics than by RecDiffusion. Moreover, their FID scores (computed on the trainset) are higher than the FID scores comparing the stitched input images to the ground truth rectangling images, indicating a significant discrepancy in image quality.

4.6. Ablation Study

We evaluate our framework designs through experiments on test set of **DIR-D** dataset [32], starting with comparisons under the details of Motion Diffusion Models (MDM). Specifically, we conduct experiments on different resolutions and the effectiveness of stitched image masks M_{S} as conditions. Then we explore the design of Content Diffusion Models (MDM), we evaluate different combinations, including solely leveraging CDM, streamlining MDM with CDM, and the effectiveness of weight sampling masks.

4.6.1 Motion Diffusion Models

While implementing MDM, conditional stitched image masks M_{S} and resolution are important factors impacting performance as shown in Table 3. Without the mask, the model cannot even outperform the baseline, i.e., the model to generate pseudo-motion fields for the train set. The larger resolution delivers better results as well as expected, but smaller resolution could lead to much faster inference.

Condition Mask	Resolution	SSIM↑	PSNR↑
	256 × 192	0.6125	20.23
✓	256 × 192	0.7337	21.97
✓	512 × 384	**0.7580**	**22.03**

Table 3. Comparison of different resolutions and conditions.

4.6.2 Content Diffusion Models

Results are demonstrated in Table 4. We find that CDM alone is ineffective as it generates images with different illuminations. Besides, we find that the output of MDM could be directly improved via CDM, and the weighted sampling mask (WSM) could further improves the performance as illustrated in Fig. 8, where local distortions are eliminated, and missing content has been restored (mark by red arrow).

MDM	CDM	WSM	SSIM ↑	PSNR ↑
	✓		0.3129	14.70
✓	✓		0.7618	22.03
✓	✓	✓	**0.7733**	**22.21**

Table 4. Comparison of the effectiveness of CDM and WSM.

5. Conclusion

In this work, we present **RecDiffusion**, the first diffusion models-based approach for rectangling stitched images. Compared to previous methods specialized for this task, which include special network structures and loss functions, we demonstrate that a typical diffusion model based on generative motion outperforms these methods. Furthermore, to address the problem of artifacts introduced by motion inaccuracy and the warping operation, we propose a strategy that uses a weighted sampling mask. This strategy combines the advantages of warping methods and generative modeling, effectively improving performance. This approach could potentially be applied to other motion-related tasks. Overall, we have achieved state-of-the-art performance in comparison to previous methods on public benchmarks. Code and model weights are available at https://github.com/lhaippp/RecDiffusion.

Acknowledgements. This work was supported by National Natural Science Foundation of China (NSFC) under grant No.62372091, 61872405, the "111" Project under Grant B17008 and Sichuan Science and Technology Program of China under grant No.2023NSFSC0462.

References

[1] Shai Avidan and Ariel Shamir. Seam carving for content-aware image resizing. *ACM Trans. Graphics (Proc. of SIGGRAPH)*, 26(3):10, 2007. 2

[2] Matthew Brown and David G Lowe. Automatic panoramic image stitching using invariant features. *International Journal of Computer Vision*, 74:59–73, 2007. 2

[3] Matthew Brown, David G Lowe, et al. Recognising panoramas. In *Proc. ICCV*, page 1218, 2003. 2

[4] Che-Han Chang, Yoichi Sato, and Yung-Yu Chuang. Shape-preserving half-projective warps for image stitching. In *Proc. CVPR*, pages 3254–3261, 2014. 3

[5] Sicheng Gao, Xuhui Liu, Bohan Zeng, Sheng Xu, Yanjing Li, Xiaoyan Luo, Jianzhuang Liu, Xiantong Zhen, and Baochang Zhang. Implicit diffusion models for continuous super-resolution. In *Proc. CVPR*, pages 10021–10030, 2023. 3

[6] Yunhui Han, Kunming Luo, Ao Luo, Jiangyu Liu, Haoqiang Fan, Guiming Luo, and Shuaicheng Liu. RealFlow: EM-based realistic optical flow dataset generation from videos. In *Proc. ECCV*, pages 288–305, 2022. 2, 3

[7] Richard Hartley and Andrew Zisserman. *Multiple view geometry in computer vision*. Cambridge university press, 2003. 2

[8] Kaiming He, Huiwen Chang, and Jian Sun. Content-aware rotation. In *Proc. ICCV*, pages 553–560, 2013. 2, 3

[9] Kaiming He, Huiwen Chang, and Jian Sun. Rectangling panoramic images via warping. *ACM Trans. Graphics*, 32 (4):1–10, 2013. 1, 2, 6, 7

[10] Jonathan Ho and Tim Salimans. Classifier-free diffusion guidance. *arXiv preprint arXiv:2207.12598*, 2022. 4, 6

[11] Jonathan Ho, Ajay Jain, and Pieter Abbeel. Denoising diffusion probabilistic models. In *Proc. NeurIPS*, pages 6840–6851, 2020. 2, 3

[12] Qi Jia, ZhengJun Li, Xin Fan, Haotian Zhao, Shiyu Teng, Xinchen Ye, and Longin Jan Latecki. Leveraging line-point consistence to preserve structures for wide parallax image stitching. In *Proc. CVPR*, pages 12186–12195, 2021. 3

[13] Hai Jiang, Haipeng Li, Yuhang Lu, Songchen Han, and Shuaicheng Liu. Semi-supervised deep large-baseline homography estimation with progressive equivalence constraint. In *Proc. AAAI*, pages 1024–1032, 2023. 6

[14] Hai Jiang, Ao Luo, Haoqiang Fan, Songchen Han, and Shuaicheng Liu. Low-light image enhancement with wavelet-based diffusion models. *ACM Trans. Graphics*, 42 (6), 2023. 2, 3

[15] Diederik P. Kingma and Jimmy Ba. Adam: A method for stochastic optimization, 2017. 6

[16] Durk P Kingma, Tim Salimans, and Max Welling. Variational dropout and the local reparameterization trick. In *Proc. NeurIPS*, pages 1–9, 2015. 3

[17] Kyu-Yul Lee and Jae-Young Sim. Warping residual based image stitching for large parallax. In *Proc. CVPR*, pages 8198–8206, 2020. 3

[18] Dongping Li, Kaiming He, Jian Sun, and Kun Zhou. A geodesic-preserving method for image warping. In *Proc. CVPR*, pages 213–221, 2015. 2

[19] Haipeng Li, Hai Jiang, Ao Luo, Ping Tan, Haoqiang Fan, Bing Zeng, and Shuaicheng Liu. Dmhomo: Learning homography with diffusion models. *ACM Trans. Graphics*, 2024. 2, 3

[20] Shiwei Li, Lu Yuan, Jian Sun, and Long Quan. Dual-feature warping-based motion model estimation. In *Proc. ICCV*, pages 4283–4291, 2015. 3

[21] Kang Liao, Lang Nie, Chunyu Lin, Zishuo Zheng, and Yao Zhao. Recrecnet: Rectangling rectified wide-angle images by thin-plate spline model and dof-based curriculum learning. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 10800–10809, 2023. 3

[22] Chung-Ching Lin, Sharathchandra U Pankanti, Karthikeyan Natesan Ramamurthy, and Aleksandr Y Aravkin. Adaptive as-natural-as-possible image stitching. In *Proc. CVPR*, pages 1155–1163, 2015. 3

[23] Kaimo Lin, Nianjuan Jiang, Loong-Fah Cheong, Minh Do, and Jiangbo Lu. Seagull: Seam-guided local alignment for parallax-tolerant image stitching. In *Proc. ECCV*, pages 370–385, 2016. 3

[24] Kaimo Lin, Shuaicheng Liu, Loong-Fah Cheong, and Bing Zeng. Seamless video stitching from hand-held camera inputs. In *Computer Graphics Forum*, pages 479–487, 2016. 3

[25] Gongye Liu, Haoze Sun, Jiayi Li, Fei Yin, and Yujiu Yang. Accelerating diffusion models for inverse problems through shortcut sampling. *arXiv preprint arXiv:2305.16965*, 2023. 3

[26] Shuaicheng Liu, Nianjin Ye, Chuan Wang, Jirong Zhang, Lanpeng Jia, Kunming Luo, Jue Wang, and Jian Sun. Content-aware unsupervised deep homography estimation and its extensions. *IEEE Trans. on Pattern Analysis and Machine Intelligence*, 45(3):2849–2863, 2022. 2

[27] Andreas Lugmayr, Martin Danelljan, Andres Romero, Fisher Yu, Radu Timofte, and Luc Van Gool. Repaint: Inpainting using denoising diffusion probabilistic models. In *Proc. CVPR*, pages 11461–11471, 2022. 2, 3

[28] Ziwei Luo, Fredrik K Gustafsson, Zheng Zhao, Jens Sjölund, and Thomas B Schön. Image restoration with mean-reverting stochastic differential equations. *International Conference on Machine Learning*, 2023. 2

[29] Ziwei Luo, Fredrik K Gustafsson, Zheng Zhao, Jens Sjölund, and Thomas B Schön. Refusion: Enabling large-size realistic image restoration with latent-space diffusion models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops*, pages 1680–1691, 2023. 2

[30] Yasuyuki Matsushita, Eyal Ofek, Weina Ge, Xiaoou Tang, and Heung-Yeung Shum. Full-frame video stabilization with motion inpainting. *IEEE Trans. on Pattern Analysis and Machine Intelligence*, 28(7):1150–1163, 2006. 3

[31] Lang Nie, Chunyu Lin, Kang Liao, Shuaicheng Liu, and Yao Zhao. Unsupervised deep image stitching: Reconstructing stitched features to images. *IEEE Trans. on Image Processing*, 30:6184–6197, 2021. 3

[32] Lang Nie, Chunyu Lin, Kang Liao, Shuaicheng Liu, and Yao Zhao. Deep rectangling for image stitching: a learning base-

line. In *Proc. CVPR*, pages 5740–5748, 2022. 1, 2, 6, 7, 8

[33] Lang Nie, Chunyu Lin, Kang Liao, Shuaicheng Liu, and Yao Zhao. Depth-aware multi-grid deep homography estimation with contextual correlation. *TCSVT*, 32(7):4460–4472, 2022. 3

[34] Lang Nie, Chunyu Lin, Kang Liao, Shuaicheng Liu, and Yao Zhao. Parallax-tolerant unsupervised deep image stitching. In *Proc. ICCV*, pages 7399–7408, 2023. 3

[35] Lang Nie, Chunyu Lin, Kang Liao, Shuaicheng Liu, and Yao Zhao. Deep rotation correction without angle prior. *IEEE Trans. on Image Processing*, 32:2879–2888, 2023. 3

[36] Ben Poole, Ajay Jain, Jonathan T. Barron, and Ben Mildenhall. Dreamfusion: Text-to-3d using 2d diffusion. In *The Eleventh International Conference on Learning Representations*, 2023. 3

[37] Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. High-resolution image synthesis with latent diffusion models. In *Proc. CVPR*, pages 10684–10695, 2022. 1, 2, 3, 7, 8

[38] Chitwan Saharia, William Chan, Huiwen Chang, Chris Lee, Jonathan Ho, Tim Salimans, David Fleet, and Mohammad Norouzi. Palette: Image-to-image diffusion models. In *Proc. ACM SIGGRAPH*, pages 1–10, 2022. 3, 7, 8

[39] Saurabh Saxena, Charles Herrmann, Junhwa Hur, Abhishek Kar, Mohammad Norouzi, Deqing Sun, and David J Fleet. The surprising effectiveness of diffusion models for optical flow and monocular depth estimation. *Advances in Neural Information Processing Systems*, 36, 2024. 2, 3

[40] Jascha Sohl-Dickstein, Eric Weiss, Niru Maheswaranathan, and Surya Ganguli. Deep unsupervised learning using nonequilibrium thermodynamics. In *Proc. ICML*, pages 2256–2265, 2015. 2, 3

[41] Jiaming Song, Chenlin Meng, and Stefano Ermon. Denoising diffusion implicit models. In *International Conference on Learning Representations*, 2021. 3, 6

[42] Yang Song and Stefano Ermon. Generative modeling by estimating gradients of the data distribution. In *Proc. NeurIPS*, pages 1–9, 2019. 2, 3

[43] Roman Suvorov, Elizaveta Logacheva, Anton Mashikhin, Anastasia Remizova, Arsenii Ashukha, Aleksei Silvestrov, Naejin Kong, Harshith Goka, Kiwoong Park, and Victor Lempitsky. Resolution-robust large mask inpainting with fourier convolutions. In *WACV*, pages 2149–2159, 2022. 2

[44] Richard Szeliski et al. Image alignment and stitching: A tutorial. *Foundations and Trends® in Computer Graphics and Vision*, 2(1):1–104, 2007. 2, 3

[45] Guy Tevet, Sigal Raab, Brian Gordon, Yonatan Shafir, Daniel Cohen-Or, and Amit H Bermano. Human motion diffusion model. *arXiv preprint arXiv:2209.14916*, 2022. 2, 3

[46] Yinhuai Wang, Jiwen Yu, Runyi Yu, and Jian Zhang. Unlimited-size diffusion restoration. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 1160–1167, 2023. 2

[47] Yinhuai Wang, Jiwen Yu, and Jian Zhang. Zero-shot image restoration using denoising diffusion null-space model. In *The Eleventh International Conference on Learning Representations*, 2023. 2, 3, 5

[48] Jin-Liang Wu, Jun-Jie Shi, and Lei Zhang. Rectangling irregular videos by optimal spatio-temporal warping. *Computational Visual Media*, 8:93–103, 2022. 3

[49] Shaoan Xie, Zhifei Zhang, Zhe Lin, Tobias Hinz, and Kun Zhang. SmartBrush: Text and shape guided object inpainting with diffusion model. In *Proc. CVPR*, pages 22428–22437, 2023. 3

[50] Ling Yang, Zhilong Zhang, Yang Song, Shenda Hong, Runsheng Xu, Yue Zhao, Wentao Zhang, Bin Cui, and Ming-Hsuan Yang. Diffusion models: A comprehensive survey of methods and applications. *ACM Computing Surveys*, 56(4): 1–39, 2023. 2, 3

[51] Julio Zaragoza, Tat-Jun Chin, Michael S Brown, and David Suter. As-projective-as-possible image stitching with moving dlt. In *Proc. CVPR*, pages 2339–2346, 2013. 2, 3, 6, 7

[52] Fan Zhang and Feng Liu. Parallax-tolerant image stitching. In *Proc. CVPR*, pages 3262–3269, 2014. 3

[53] Yun Zhang, Yu-Kun Lai, and Fang-Lue Zhang. Content-preserving image stitching with piecewise rectangular boundary constraints. *IEEE Trans. on Visualization and Computer Graphics*, 27(7):3198–3212, 2020. 3

[54] Zhuofan Zhang, Zhen Liu, Ping Tan, Bing Zeng, and Shuaicheng Liu. Minimum latency deep online video stabilization. In *Proc. ICCV*, pages 23030–23039. 2

[55] Fushun Zhu, Shan Zhao, Peng Wang, Hao Wang, Hua Yan, and Shuaicheng Liu. Semi-supervised wide-angle portraits correction by multi-scale transformer. In *Proc. CVPR*, pages 19689–19698, 2022. 2