Single Image Rolling Shutter Removal with Diffusion Models

Zhanglei Yang[1,2*], Haipeng Li[1,2*], Mingbo Hong[2], Chen-Lin Zhang[3], Jiajun Li[4], Shuaicheng Liu[1,2,†]

[1]University of Electronic Science and Technology of China
[2]Megvii Technology
[3] Moonshot AI
[4] Noumena AI
yangzl11974@gmail.com, {lihaipeng@std.,liushuaicheng@}uestc.edu.cn
mingbohong97@gmail.com, zclnjucs@gmail.com, taringlee@noumena.com.cn

Abstract

We present RS-Diffusion, the first Diffusion Models-based method for single-frame Rolling Shutter (RS) correction. RS artifacts compromise visual quality of frames due to the row-wise exposure of CMOS sensors. Most previous methods have focused on multi-frame approaches, using temporal information from consecutive frames for the motion rectification. However, few approaches address the more challenging but important single frame RS correction. In this work, we present an "image-to-motion" framework via diffusion techniques, with a designed patch-attention module. In addition, we present the RS-Real dataset, comprised of captured RS frames alongside their corresponding Global Shutter (GS) ground-truth pairs. The GS frames are corrected from the RS ones, guided by the corresponding Inertial Measurement Unit (IMU) gyroscope data acquired during capture. Experiments show that RS-Diffusion surpasses previous single-frame RS methods, demonstrates the potential of diffusion-based approaches, and provides a valuable dataset for further research.

Code — https://github.com/lhaippp/RS-Diffusion
Datasets — https://huggingface.co/Lhaippp/RS-Diffusion

Introduction

Rolling shutter (RS) is a common effect encountered when capturing images with CMOS sensors. It results from varying exposure times for different rows in each frame, causing artifacts like distorted straight lines and skewed image content, as shown in Fig. 1 (RS Image). These distortions are not only visually unpleasing but also detrimental to downstream tasks (Hedborg et al. 2012; Saurer, Pollefeys, and Lee 2016; Albl, Kukelova, and Pajdla 2015; Saurer et al. 2013). Prior methods for RS removal can be categorized as multi-frame and single-frame based. The former relies on temporal motion in consecutive frames for motion compensation, while the latter solely depends on a single frame for restoration. Single-frame RS correction is particularly challenging yet crucial in data-scarce situations. Many existing methods, including non-learning-based approaches (Rengarajan, Rajagopalan, and Aravind 2016; Rengarajan, Balaji, and Rajagopalan 2017; Purkait, Zach, and Leonardis 2017; Zhuang et al. 2019; Kandula, Kumar, and Rajagopalan 2020), rely on salient structures, such as straight lines. However, they may falter in cases where such salient structures are absent.

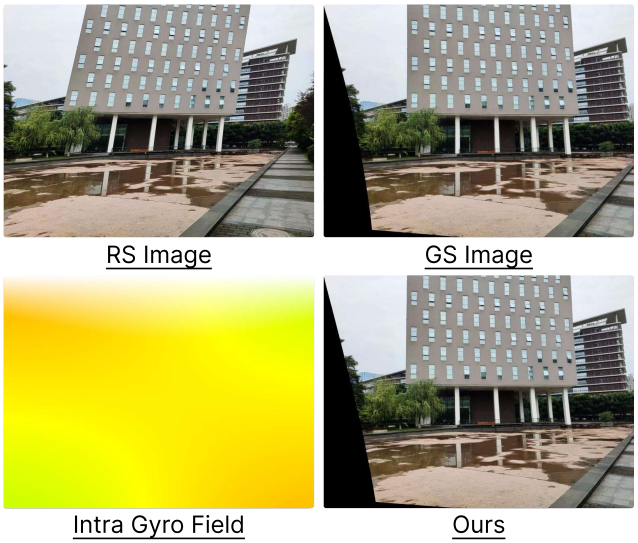

Figure 1: Illustration of the proposed dataset and our results. The first row features a rolling-shutter (RS) image captured in realistic scenes, along with the corresponding ground-truth global-shutter (GS) image. The ground-truth flow used for correcting the RS image to the GS image is displayed on the left side of the second row. In the bottom right, we showcase our corrected RS image.

The majority of deep RS correction methods are multi-frame based (Liang, Chang, and Chen 2008; Ringaby and Forssén 2012; Zhuang, Cheong, and Hee Lee 2017; Vasu, Rajagopalan et al. 2018), with only a few considering a single frame as input. Among single-frame methods, Rengarajan *et al.* (Rengarajan, Balaji, and Rajagopalan 2017) directly learn the mapping between global shutter (GS) and RS frames, facing the challenge of large solution space for row-wise motion estimation. Zhuang *et al.* (Zhuang et al. 2019) addressed this by estimating an additional depth map, although depth estimation from RS frames is inherently ill-

*These authors contributed equally.
†Corresponding author

posed. Recently, Yan *et al.* (Yan et al. 2023) introduced a deep homography mixture model, achieving the current best performance by embedding motion in a subspace and learning coefficients to combine pre-learned motion flow bases.

Recent advancements in generative models (Brown et al. 2020; Goodfellow et al. 2020; Ouyang et al. 2022), including Diffusion Models (DMs) (Ho, Jain, and Abbeel 2020; Song et al. 2020), have significantly advanced the field of Artificial Intelligence. Notably, DMs excel in performing various motion-related tasks, notably in human motion generation (Tevet et al. 2022), in estimating depth/optical flow (Saxena et al. 2023), and in homography data rendering (Li et al. 2024). Meanwhile, DMs have shown strength in solving ill-posed problems, such as image restoration (Gao et al. 2023; Yinhuai, Jiwen, and Jian 2022), where inferring accurate solutions from ambiguous data is difficult. Therefore, drawing on these insights and inspirations, in this work, we introduce a novel single-frame RS correction method based on DMs, namely **RS-Diffusion**. DMs are adept at generating data from Gaussian distributions through multiple sampling steps and can be conditioned on additional information, such as images. Leveraging the transformative capabilities, we are capable of correcting RS images to GS images. Our framework, built on CFG (Ho and Salimans 2022) and DDIM (Song, Meng, and Ermon 2020), uses a downsampled RS image I_{RS} as a conditioning element, concatenated with Gaussian noise. This input is processed by DMs in very few steps to produce a motion field G, which is then applied to remap I_{RS} to eliminate RS artifacts. Additionally, we introduce a patch-attention module based on prior RS motion patterns to enhance results. We visualize our corrected result on the right side of the second row in Fig. 1.

On the other hand, high-quality datasets play a crucial role. To meet the demand, we adhere to two essential criteria (Han et al. 2022): the *label criterion*, which requires precise alignment of the RS correction motion as ground-truth labels between RS-GS pairs, and the *realism criterion*, which ensures that both the image content and RS motion remain realistic. However, existing RS datasets often fall short of meeting these criteria. The synthesized dataset can provide accurate labels but violates realism criterion. In contrast, the captured images can suffice the realism, but the label is not accurate.

To resolve these issues, we introduce Intra Gyro Field (IGF) pipeline, leveraging Inertial Measurement Unit (IMU) gyroscope sensors to record camera rotations during capture. Specifically, we firstly achieve accurate frame-gyro data synchronization. Next, these rotations are translated into a series of homography matrices, further converted into motion fields as RS correction ground-truth (GT) labels. With the help of IGF, we create the **RS-Real** dataset, which satisfies both the label and the realism criteria in RS correction. The dataset contains 40,000 train and 1,000 test samples. A case is shown in Fig. 1, which shows a captured RS image, the IGF between the RS-GS pair and its GS image.

In summary, our diffusion model-based framework advances the state-of-the-art in single-image RS correction, while it could run inference in real-time speed, i.e., up to 28.1 ms per frame on one NVIDIA 2080Ti. The RS-Real dataset, containing high-quality training pairs, addresses the scarcity of qualified datasets in the RS task. Our contributions include:

- We propose the first diffusion-based framework for single image rolling shutter removal, namely **RS-Diffusion**.
- We introduce a pipeline for correcting RS images with recorded IMU gyro readings, which delivers accurate rectified RS images as ground-truth labels for training and testing, yielding a realistic RS dataset, **RS-Real**.
- Experiments show that our approach achieves state-of-the-art performance on public benchmarks, exhibiting both generalizability and applicability.

Related Works

Rolling Shutter Correction Methods

Multi-frame rectification methods use spatial-temporal information, with investigations into per-pixel motion vectors (Liang, Chang, and Chen 2008), image feature tracking (Ringaby and Forssén 2012), and affine models (Baker et al. 2010), now enhanced by 3D data (Zhuang, Cheong, and Hee Lee 2017). Deep learning advances include pixel-wise velocity estimation (Liu et al. 2020) and optical flow for deep feature warping (Fan, Dai, and He 2021). Classical single-image methods mostly utilize salient lines for RS correction (Rengarajan, Rajagopalan, and Aravind 2016). Deep learning for single RS image correction is less common but includes neural networks trained on paired images (Rengarajan, Balaji, and Rajagopalan 2017), employing motion models and depth map estimations (Zhuang et al. 2019). A novel deep homography mixture model shows state-of-the-art results (Yan et al. 2023). We introduce diffusion models (DMs) for single RS rectification, demonstrating their effectiveness and generalizability in motion extraction.

Diffusion Models

Diffusion models (DMs), based on stochastic diffusion processes (Sohl-Dickstein et al. 2015), efficiently transform data distributions through Gaussian transitions and iterative denoising using the data distribution gradient (Song and Ermon 2019). DDPM (Ho, Jain, and Abbeel 2020) employs discrete steps in this process, and Song *et al.* (Song et al. 2020) further refines the methodology via stochastic differential equations (SDE). DDIM (Song, Meng, and Ermon 2020) accelerates reverse sampling through subsequence sampling and ordinary differential equations (ODEs). Conditioned data generation progresses with classifier-based (Dhariwal and Nichol 2021; Liu et al. 2023) and classifier-free (CFG) techniques (Ho and Salimans 2022). DMs also contribute to motion-centric tasks, from video generation (Ni et al. 2023), video frame interpolation (Danier, Zhang, and Bull 2024), object tracking (Luo et al. 2024) to Molecule generation (Huang et al. 2023). Our pioneering work extends DMs to the domain of rolling shutter removal.

Gyroscope-based Motion Methods

Gyroscopes are pivotal for diverse applications such as video stabilization (Karpenko et al. 2011; Bell, Troccoli, and Pulli

Figure 2: Rolling shutter image $\mathbf{I}_{RS}$, is introduced by high-frequency shake with a row-wise exposure CMOS camera. The gyroscope can accurately record these motions, which are then transformed into a motion field, $\mathbf{G} \in \mathbf{R}^{2\times H\times W}$. This field is referred to as the Intra Gyro Field (IGF). With $\mathbf{G}$, we are able to correct $\mathbf{I}_{RS}$, resulting in a Global Shutter Image, $\mathbf{I}_{GS}$.

2014), optical image stabilization (OIS) (Liu et al. 2021), estimating homography/optical flow (Li, Luo, and Liu 2021; Li et al. 2023), and simultaneous localization and mapping (SLAM) (Huang and Liu 2018), with growing using in RS correction if gyroscope data is accessible during the capturing (Mo, Islam, and Sattar 2022). In this work, instead of directly using online captured gyroscope data to promote the performances of different applications. We use the gyroscope data to create a dataset. Our work utilizes gyroscope synchronization at the Android HAL for precise calibration (Jia and Evans 2013; Bloesch et al. 2014), enabling the creation of a **RS-Real** dataset for rolling shutter research, incorporating authentic RS pairings for enhanced training and validation. Note that, our method doesn't need gyro readings during the model inference.

Method

Overview

We present a pipeline named Intra Gyro Field (IGF) designed to create a realistic and precisely annotated dataset through combined engineering and algorithmic efforts, as discussed in Section **Intra Gyro Field**. Furthermore, we propose a novel and general framework based on diffusion models, incorporating a specialized patch attention (PA) module, which is aimed at correcting rolling shutter (RS) images to global shutter (GS) quality, as described in Section **Diffusion Models**.

Intra Gyro Field

Gyroscope records the relative camera 3D rotation in a relative time and it is possible to convert the gyroscope readings into a 2D motion field to achieve alignment between two consecutive frames (Li et al. 2023). In this chapter, we demonstrate how to leverage the gyroscope to correct RS effect for a single image and thus propose an Intra Gyro Field (IGF). Specifically, given the 3-axis angular velocity readings (row $\mathbf{v_r}$, pitch $\mathbf{v_p}$ and yaw $\mathbf{v_y}$) and relative time intervals Δt between gyro readings. The 3-axis angular angles $(\angle\mathbf{r}, \angle\mathbf{p}, \angle\mathbf{y})$ can be computed as:

$$\angle\mathbf{r} = \mathbf{v_r} \cdot \Delta t, \quad \angle\mathbf{p} = \mathbf{v_p} \cdot \Delta t, \quad \angle\mathbf{y} = \mathbf{v_y} \cdot \Delta t. \tag{1}$$

Then we utilize the Rodrigues Formula to convert rotation angles into the rotation matrix $\mathbf{R}(\Delta t) \in SO(3)$. In order to bridge the relationship between 3D camera pose and 2D image motion, we choose to use the homography matrix $\mathbf{H}$. The homography matrix represents the relationship between two different perspectives of the same scene, more specifically, the conditions for its validity are satisfied when the camera motion is purely rotational or when the contents lie in the same plane (Hartley and Zisserman 2003). Theoretically, given 3D rotation matrix $\mathbf{R}(\Delta t)$ and translation vector $\mathbf{t}(\Delta t)$), we can represent $\mathbf{H}$ as:

$$\mathbf{H} = \mathbf{K}\left(\mathbf{R}(\Delta t) + \mathbf{t}(\Delta t)\frac{\mathbf{n}^T}{d}\right)\mathbf{K}^{-1}, \tag{2}$$

where $\mathbf{n}^T$ represents the normal vector of a certain plane, d is the distance between the camera center and the plane, while $\mathbf{K}$ is the camera intrinsic matrix.

In this work, we do not incorporate translations and only use gyroscope data to model the homography matrix, the reasons are threefold. Firstly, the rolling shutter effect caused by camera shake occurs primarily due to rotational movements (Karpenko et al. 2011). Secondly, even though translations can be gathered from accelerometer data, they are significantly less accurate than rotational measurements (Forssén and Ringaby 2010; Joshi et al. 2010). Lastly, according to Eq. 2, we can find that translation is correlated to depth, but accurately estimating depth maps can be another non-trivial problem. As a result, we can formulate a rotational-only homography:

$$\mathbf{H} = \mathbf{K}\left(\mathbf{R}(\Delta t)\right)\mathbf{K}^{-1}. \tag{3}$$

The gyroscope records N readings during the capture of one image, it thus can produce N camera rotation matrices $(\mathbf{R}_1(\Delta t), \mathbf{R}_2(\Delta t), \cdots, \mathbf{R}_N(\Delta t))$. If we split the image from top to bottom into N patches $(\mathbf{p}_1, \mathbf{p}_2, \cdots, \mathbf{p}_N)$, these rotation matrices correspond to the motion between the first row and bottom row of patches as illustrated in Fig. 2. Moreover, the inter-patch motion is approximately to be smooth (Liu et al. 2021; Dai, Li, and Kneip 2016), so we can apply the spherical linear interpolation (SLERP) to interpolate the motion to avoid the discontinuities across row patches. Subsequently, we follow Eq. 3 to produce N homography matrices $(\mathbf{H}_1, \mathbf{H}_2, \cdots, \mathbf{H}_N)$ to model the 2D relationship between patches. To this end, we follow previous method (Li, Luo, and Liu 2021) to convert the array of homography matrix into a motion field, it is achieved by transforming grid points by their corresponding homography and subtracting them. Lastly, this motion field is IGF,

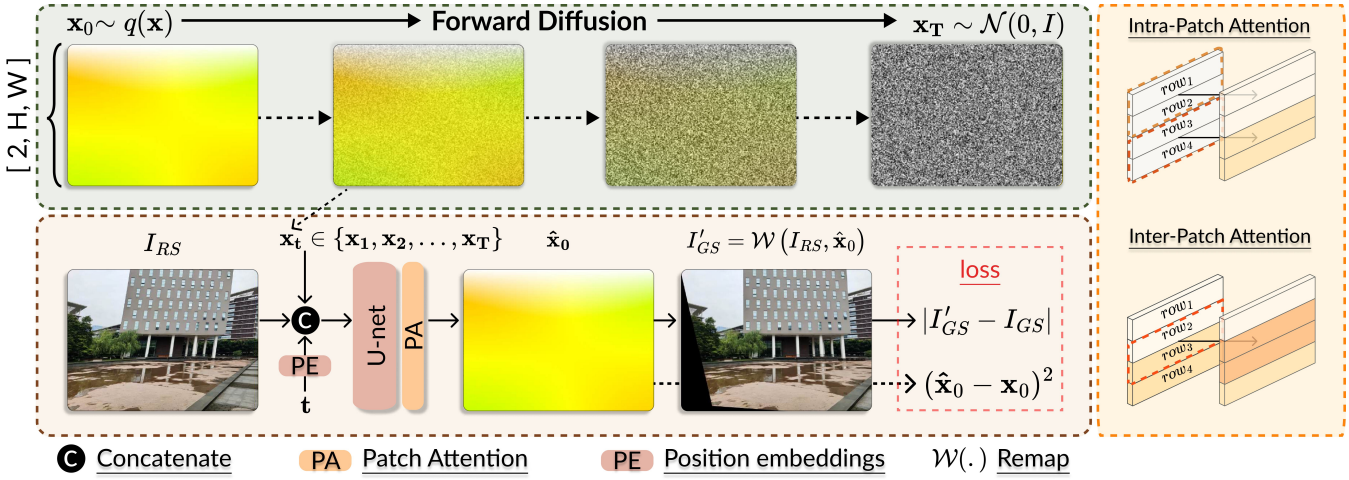

Figure 3: Illustration of the framework: During training, $\mathbf{x}_0$ undergoes forward diffusion to become $\mathbf{x}_t$. The network θ processes the concatenated input. The Patch-Attention module, which includes both Intra-Patch and Inter-Patch attention mechanisms, enhances the relationships between patches. The resulting output, $\hat{\mathbf{x}}_0$, can be used to correct $\mathbf{I}_{RS}$. The loss function comprises the MSELoss, calculated between $\hat{\mathbf{x}}_0$ and $\mathbf{x}_0$, and the photometric loss, computed between the GT GS image and $\mathbf{I'}_{GS}$.

Figure 4: A glance at the **RS-Real** dataset reveals that it contains data pairs featuring various RS motion patterns and different intensities, all of which are captured in diverse scenes.

denoted as $\mathbf{G} \in \mathbf{R}^{2\times H\times W}$, which is capable of correcting the RS effect within an image:

$$\mathbf{I}_{GS} = \mathcal{R}emap(\mathbf{I}_{RS}, \mathbf{G}). \tag{4}$$

Dataset Collecting. Our proposed IGF effectively addresses the challenges found in previous datasets, such as unrealistic data distributions and spatio-temporal synchronization issues. By recording the RS effect with a gyroscope and subsequently recovering from it, our approach circumvents synchronization challenges by using only one camera at a time. As a result, we can flexibly capture a diverse range of realistic scenes, spanning various indoor and outdoor environments, and subjecting the camera to multiple motion patterns and speeds. This method enables us to gather a comprehensive dataset of realistic RS images, denoted as $\mathbf{X}_{RS} = \{\mathbf{I}^0_{RS}, \mathbf{I}^1_{RS}, \ldots, \mathbf{I}^k_{RS}\}$, along with their corresponding IGFs, $\mathbf{X}_{IGF} = \{\mathbf{G}^0, \mathbf{G}^1, \ldots, \mathbf{G}^k\}$. After applying RS removal using Eq. 4, we obtain the set of GS images, $\mathbf{X}_{GS} = \{\mathbf{I}^0_{GS}, \mathbf{I}^1_{GS}, \ldots, \mathbf{I}^k_{GS}\}$. We refer to this dataset as **RS-Real**. Examples from the dataset are illustrated in Fig. 4.

Diffusion Models

To correct distortions resulting from the rolling shutter effect in general cases, we utilize diffusion models adept at managing the distributional transformations of images captured under both RS and GS conditions.

Specifically, we leverage the diffusion models as proposed by Sohl-Dickstein *et al.* (Sohl-Dickstein et al. 2015) and Ho *et al.* (Ho, Jain, and Abbeel 2020), which define a Markov chain process over T steps to incrementally introduce Gaussian noise into the original data distribution $\mathbf{x}_0 \sim q(\mathbf{x})$. This stepwise infusion of noise generates a sequence of increasingly distorted states $\mathbf{x}_1, \ldots, \mathbf{x}_T$, defined as forward diffusion, and is mathematically represented as:

$$q\left(\mathbf{x}_{1:T} \mid \mathbf{x}_0\right) = \prod_{t=1}^{T} q\left(\mathbf{x}_t \mid \mathbf{x}_{t-1}\right). \tag{5}$$

Following this, a denoising model, denoted as θ, is then trained to reverse the diffusion process $p_\theta\left(\mathbf{x}_{t-1} \mid \mathbf{x}_t\right)$ to construct desired data samples from isotropic Gaussian noise $\mathbf{x}_T \sim \mathcal{N}(\mathbf{0}, \mathbf{I})$:

$$p_\theta\left(\mathbf{x}_{0:T}\right) = p\left(\mathbf{x}_T\right) \prod_{t=1}^{T} p_\theta\left(\mathbf{x}_{t-1} \mid \mathbf{x}_t\right). \tag{6}$$

This inversion transforms the signal from the noise-dominant Gaussian distribution back into the target data distribution. To refine this reconstruction process and improve the quality of the generated samples, we can integrate conditional variables $\mathbf{y}$ into the model as described by CFG (Ho and Salimans 2022). These conditions are merged with the

noisy data transitions in the model:

$$p_\theta\left(\mathbf{x}_{t-1} \mid \mathbf{x}_t, \mathbf{y}\right)=\mathcal{N}\left(\mathbf{x}_{t-1} ; \boldsymbol{\mu}_\theta\left(\mathbf{x}_t, t, \mathbf{y}\right), \sigma_t^2 \mathbf{I}\right). \quad (7)$$

By incorporating supplementary conditions, we are able to exert enhanced control over the sample generation phase, which results in an improvement in the fidelity of the final reconstructions.

Rolling Shutter Removal Module

Our framework is built upon CFG (Ho and Salimans 2022) and DDIM (Song, Meng, and Ermon 2020). We illustrate the pipeline in the Fig. 3, we designate IGFs as the initial data distribution, $\mathbf{x}_0$, and the RS images $\mathbf{I}_{RS}$ as conditions $\mathbf{y}$. This configuration endows the model with an "*image-to-motion*" capability. During the training process, $\mathbf{x}_0$ is noised via forward diffusion to $\mathbf{x}_t$, then we concatenate the $\mathbf{I}_{RS}$, $\mathbf{x}_t$ and time embedding to fed into the U-net network θ. It is imperative to mention that this concatenation is introduced at various layers within the network. Finally, the output of θ is processed through the Patch-Attention module, resulting in the generated outputs $\hat{\mathbf{x}}_0$.

Patch-Attention Module. As described in previous sections, we observe that the motion patterns in RS images are consistently correlated between rows. In practice, rows are often regrouped into patches to mitigate the effects of RS. Capitalizing on this a priori, we propose an attention block to foster the inter-relationships between these patches. Our attention block operates in two stages: Firstly, the intra-patch attention phase involves evenly splitting the input into non-overlapping patches. As exemplified on the right side of Fig. 3, the feature is divided into patches (4 rows into 2 patches in this example). Within each patch, self-attention is employed to process features internally. Subsequently, the inter-patch attention stage is applied. In this phase, self-attention mechanisms are utilized to facilitate the exchange of information between consecutive rows across different patches. This approach ensures that the relationship between adjacent patches is effectively enhanced.

Loss Function. After computing $\hat{\mathbf{x}}_0$, we first apply the mean squared error (MSE Loss) as follows:

$$\ell_{mse}=(\hat{\mathbf{x}}_0-\mathbf{x}_0)^2. \quad (8)$$

In addition to the standard MSE Loss, which facilitates learning the distribution transformation from RS images to their respective IGF, we further propose to constrain the network with an extra conditional loss. Specifically, we warp the original RS images $\mathbf{I}_{RS}$ via the computed IGF $\hat{\mathbf{x}}_0$ to produce corrected RS images $\mathbf{I}'_{GS}$:

$$\mathbf{I}'_{GS}=\mathcal{W}\left(\mathbf{I}_{RS}, \hat{\mathbf{x}}_0\right), \quad (9)$$

where $\mathcal{W}(.)$ represents the remapping operation. Then we calculate the photometric loss between $\mathbf{I}'_{GS}$ and $\mathbf{I}_{GS}$:

$$\ell_{pl}=|\mathbf{I}'_{GS}-\mathbf{I}_{GS}|. \quad (10)$$

Consequently, the overall loss can be computed as a dynamically weighted sum. In other words, it continually adjusts ℓ_{pl} to be equal to ℓ_{mse}, formulated as:

$$\ell_{overall}=\ell_{mse}+\frac{|\ell_{mse}|}{|\ell_{pl}|}\cdot\ell_{pl}. \quad (11)$$

Experiments

We begin by outlining existing and our proposed datasets in Section **Dataset**. Our method's performance is compared against others on varied benchmarks in Section **Quantitative Comparison**, with qualitative evaluations presented in Section **Qualitative Comparison**. An in-depth analysis of our framework's intricate designs is described in Section **Ablation Studies**. Finally, we list implementation details and demonstrate the utility of our framework in a video stabilization application (Liu et al. 2013) within the **supplementary materials**.

Dataset

In our experiments, we validate the proposed RS correction method on 3 datasets, including one long-standing public datasets and two new ones for comprehensive evaluation. To ascertain the reliability of our approach, we synchronize our evaluations with established RS image correction techniques (Rengarajan, Balaji, and Rajagopalan 2017; Kandula, Kumar, and Rajagopalan 2020; Yan et al. 2023), utilizing the **Building** dataset (Xiao et al. 2010; Shao, Svoboda, and Van Gool 2003; Philbin et al. 2007) as a standard benchmark. Additionally, we extend our test scope with a challenging dataset **RS-Homo** from Yan *et al.* (Yan et al. 2023), crafted to simulate adverse conditions like low lighting and scant textures, critical for testing diffusion model robustness. A notable contribution of this paper is the **RS-Real** dataset, detailed in Section , which exploits sensor information during image capture to address common synchronization challenges and offers a new methodology for data gathering in RS scenarios.

Quantitative Comparison

In our study, we evaluate our RS image correction method using key metrics for visual quality and motion accuracy. Visual comparisons are done via the Peak Signal-to-Noise Ratio (PSNR) and the Structural Similarity Index Measure (SSIM) against GT GS images. The motion-based assessment considers the Endpoint Error (EPE) for correction flow accuracy. Pixels on black edges, lacking information, are omitted from our analysis.

Building dataset. Our study compares with traditional and deep learning-based RS correction methods, as shown in Table 1. Approaches using curve detection (Rengarajan, Rajagopalan, and Aravind 2016; Purkait, Zach, and Leonardis 2017) to discern motion patterns struggle in low-structure environments, showing subpar PSNR and SSIM, alongside elevated EPE. The homography mixtures method (Grundmann et al. 2012) fares better with its video sequence foundation and refined feature correspondences. However, learning methods (Rengarajan, Balaji, and Rajagopalan 2017; Kandula, Kumar, and Rajagopalan 2020) have proven robust to strong outliers and complex camera motions. Yan *et al.* model (Yan et al. 2023), merging homography mixtures with learning, showcases significant strides in metric performance due to constrained motion spaces from learned bases. Our proposed technique outperforms existing methods, setting a new state-of-the-art (SOTA) across evaluated metrics and emphasizing its effectiveness.

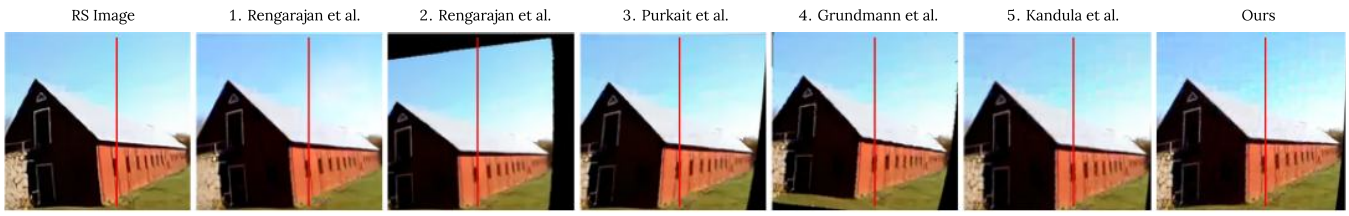

Figure 5: Comparison with existing methods: 1. Rengarajan *et al.* (Rengarajan, Rajagopalan, and Aravind 2016), 2. Rengarajan *et al.* (Rengarajan, Balaji, and Rajagopalan 2017), 3. Purkait *et al.* (Purkait, Zach, and Leonardis 2017), 4. Grundmann *et al.* (Grundmann et al. 2012) and 5. Kandula *et al.* (Kandula, Kumar, and Rajagopalan 2020) on a RS building image. Red vertical lines to highlight correction results.

Method	PSNR(dB)↑	SSIM↑	EPE↓
1. Rengarajan *et al.*	29.82	0.67	11.89
2. Purkait *et al.*	29.22	0.55	8.32
3. Grundmann *et al.*	32.57	0.72	3.34
4. Rengarajan *et al.*	32.25	0.70	3.76
5. Kandula *et al.*	32.85	0.73	2.84
6. Yan *et al.*	33.34	0.75	1.25
RS-Diffusion	**34.92**	**0.79**	**0.90**

Table 1: Comparison of PSNR, SSIM, and EPE between our method and existing ones: 1. Rengarajan *et al.* (Rengarajan, Rajagopalan, and Aravind 2016), 2. Purkait *et al.* (Purkait, Zach, and Leonardis 2017), 3. Grundmann *et al.* (Grundmann et al. 2012), 4. Rengarajan *et al.* (Rengarajan, Balaji, and Rajagopalan 2017), 5. Kandula *et al.* (Kandula, Kumar, and Rajagopalan 2020) and 6. Yan *et al.* (Yan et al. 2023) on **Building** dataset.

Method	PSNR(dB)↑	SSIM↑	EPE↓
Yan *et al.*	26.15	0.77	4.10
RS-Diffusion	**36.60**	**0.94**	**1.02**

Table 2: Comparison of PSNR, SSIM, and EPE between our method and Yan *et al.* (Yan et al. 2023) on **RS-Homo**.

RS-Homo dataset. The results in Table 2 show our diffusion models-based framework substantially improves upon Yan *et al.*'s state-of-the-art single-image model. Demonstrating great proficiency in managing challenging scenarios, our proposed method surpasses specifically designed architectures, enhancing the capability of diffusion models to address rolling shutter (RS) effects effectively.

Method	PSNR(dB)↑	SSIM↑	EPE↓
Yan *et al.*	18.48	0.55	4.18
RS-Diffusion	**22.02**	**0.69**	**2.12**

Table 3: Comparison of PSNR, SSIM, and EPE between our method and Yan *et al.* (Yan et al. 2023) on **RS-Real**.

RS-Real dataset. Our dataset, designed for realism in content, RS-motion, and label accuracy, comprises 40,000 training and 1,000 test pairs across diverse scenes. It includes RS and GS images along with GT flow. We benchmark against Yan *et al.* (Yan et al. 2023) by retraining their model using the default settings on our trainset. The results confirm that our approach consistently outperforms theirs in both photometric and motion measures as demonstrated in Table 3.

Qualitative Comparison

Our method is evaluated against current methods on different benchmarks. We follow the settings used by Yan *et al.* (Yan et al. 2023) to compare with existing methods (Rengarajan, Balaji, and Rajagopalan 2017; Rengarajan, Rajagopalan, and Aravind 2016; Purkait, Zach, and Leonardis 2017; Grundmann et al. 2012; Kandula, Kumar, and Rajagopalan 2020) in Fig. 5, and ours outperforms the others. In addition, due to limited open-source options, we compare with Yan *et al.* (Yan et al. 2023) method on the remaining datasets. We visualize RS and ground-truth GS images in the first two columns, then present Yan *et al.* results paired with alignment heatmaps that use darker shades to indicate greater similarity compared to GT images. We then display our own results alongside their corresponding alignment heatmaps, illustrating our approach's consistency with benchmarks and precision in alignment. This visual comparison highlights the effectiveness of our methodology.

RS-Real and RS-Homo Datasets. In the first three rows of Fig. 6, we present the results on the **RS-Real** dataset, while the last row shows the results on the **RS-Homo** dataset. Our approach consistently outperforms the method of Yan *et al.* (Yan et al. 2023) across varied RS challenges, including right-skewed images, dynamic scenes with moving cars, complex motion patterns from quick device movement, and low-texture scenarios, proving its robustness and superior correction capabilities.

Ablation Studies

We evaluate our framework design through experiments, starting with comparisons under diverse generative objectives, paired with analytical analyses. The effectiveness of the Patch-Attention Block is then scrutinized.

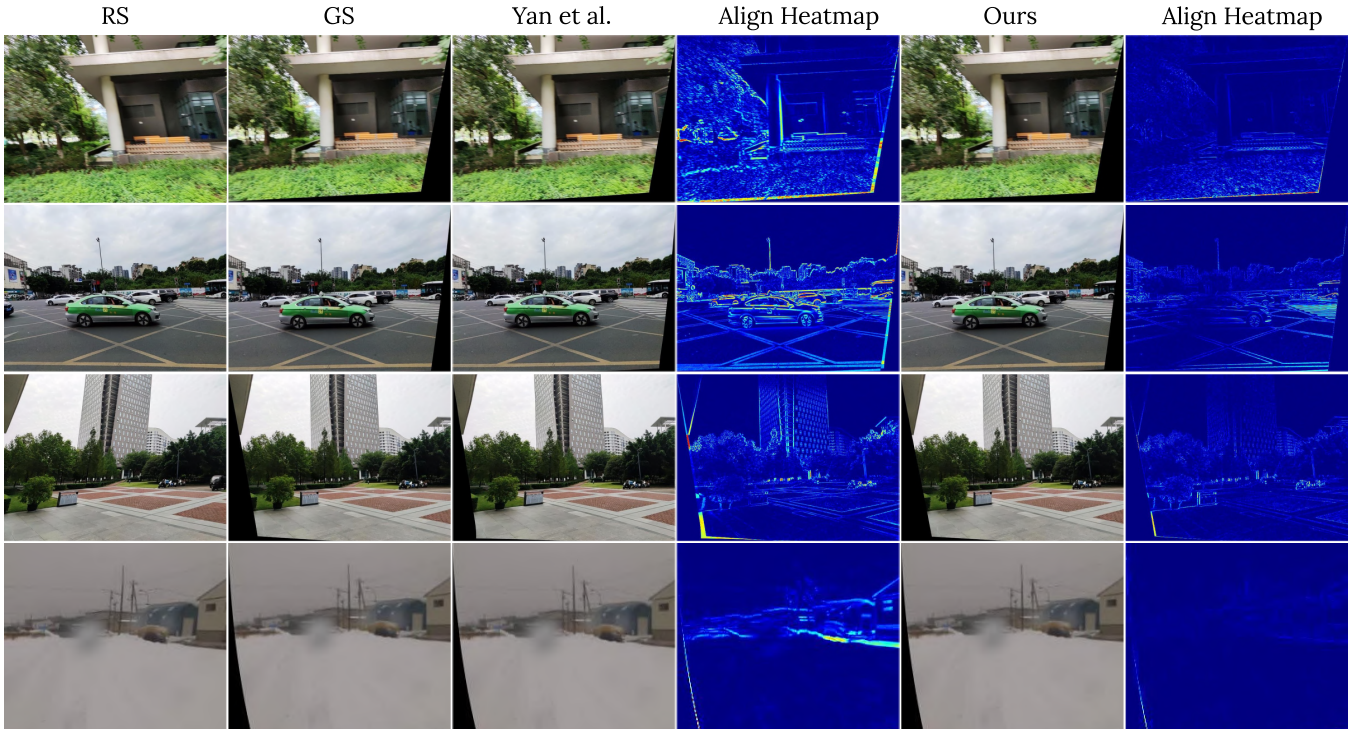

Figure 6: Comparison on **RS-Real** and **RS-Homo** dataset. Column 1 shows the input RS image and Column 2 shows the ground-truth GS image. Column 3 and 5 are results by Yan *et al.* (Yan et al. 2023) and our method with the alignment heatmaps that use darker shades to indicate greater similarity compared to ground-truth GS images.

Method	PSNR(dB)↑	SSIM↑
"image-to-image"	16.73	0.47
"image-to-motion"	**20.78**	**0.63**

Table 4: Comparison between different generating objects. "image-to-image" indicates transforming RS images to GS. "image-to-motion" represents predicting IGFs from RS.

Predicting IGF vs. GS Image We contrast our "image-to-motion" pipeline with traditional "image-to-image" methods that use diffusion models to convert RS images to GS images at a fixed resolution of 256×256, later upscaled to 600×800 for visual metric comparison with GT GS images, as shown in Table 4. Our approach outperforms the conventional framework in the RS removal task for two main reasons: 1) It is not constrained by diffusion model resolutions, since we can upsample IGF to correct RS images at their original size; 2) While "image-to-image" models learn the joint probability distribution between RS and GS images, our method gains additional improvements by learning from the distribution involving IGFs, RS and GS images, thus leveraging more information for enhanced results.

Rolling-Shutter Patch Attention As demonstrated in Table 5, our proposed Patch Attention module proves to be effective, as it not only aggregates features within individual patches but also facilitates interaction between features across different patches. The experimental results align well with our prior understanding of RS motion.

Intra	Inter	PSNR(dB)↑	SSIM↑	EPE↓
		20.78	0.63	2.62
✓		21.10	0.65	2.54
✓	✓	**22.02**	**0.69**	**2.12**

Table 5: The effectiveness of Patch Attention. Intra refers to intra-patch attention and Inter indicates inter-patch attention.

Conclusion

In this work, we have presented RS-Diffusion, the first Diffusion Model based approach for single frame rolling shutter rectification. We have captured a novel dataset, namely RS-Real, to accomplish this task. The RS-Real dataset contains captured RS images, and the corresponding ground-truth GS images can be created according to synchronized gyroscope data that recorded during the RS frame capturing, yielding RS-GS image pairs. Uniquely, this dataset fulfills both accuracy and authenticity requirements for RS research. In addition, we have presented RS-Diffusion for real-time RS correction using just one RS frame. We have achieved state-of-the-art performances when compared to previous single RS correction methods. We publicly share our code and dataset with the community at https://github.com/lhaippp/RS-Diffusion.

Acknowledgements

This work was supported in part by National Natural Science Foundation of China (NSFC) under Grant Nos. 62372091, 62071097 and in part by Sichuan Science and Technology Program under Grant Nos. 2023NSFSC0462, 2023NSFSC0458, 2023NSFSC1972.

References

Albl, C.; Kukelova, Z.; and Pajdla, T. 2015. R6P-rolling shutter absolute camera pose. In *Proc. CVPR*, 2292–2300.

Baker, S.; Bennett, E.; Kang, S. B.; and Szeliski, R. 2010. Removing rolling shutter wobble. In *Proc. CVPR*, 2392–2399.

Bell, S.; Troccoli, A.; and Pulli, K. 2014. A non-linear filter for gyroscope-based video stabilization. In *Proc. ECCV*, 294–308.

Bloesch, M.; Omari, S.; Fankhauser, P.; Sommer, H.; Gehring, C.; Hwangbo, J.; Hoepflinger, M. A.; Hutter, M.; and Siegwart, R. 2014. Fusion of optical flow and inertial measurements for robust egomotion estimation. In *Proc. IROS*, 3102–3107.

Brown, T.; Mann, B.; Ryder, N.; Subbiah, M.; Kaplan, J. D.; Dhariwal, P.; Neelakantan, A.; Shyam, P.; Sastry, G.; Askell, A.; et al. 2020. Language models are few-shot learners. *Proc. NeurIPS*, 33: 1877–1901.

Dai, Y.; Li, H.; and Kneip, L. 2016. Rolling shutter camera relative pose: Generalized epipolar geometry. In *Proc. CVPR*, 4132–4140.

Danier, D.; Zhang, F.; and Bull, D. 2024. Ldmvfi: Video frame interpolation with latent diffusion models. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 38, 1472–1480.

Dhariwal, P.; and Nichol, A. 2021. Diffusion models beat GANs on image synthesis. In *Proc. NeurIPS*, 8780–8794.

Fan, B.; Dai, Y.; and He, M. 2021. SUNet: symmetric undistortion network for rolling shutter correction. In *Proc. CVPR*, 4541–4550.

Forssén, P.-E.; and Ringaby, E. 2010. Rectifying rolling shutter video from hand-held devices. In *Proc. CVPR*, 507–514.

Gao, S.; Liu, X.; Zeng, B.; Xu, S.; Li, Y.; Luo, X.; Liu, J.; Zhen, X.; and Zhang, B. 2023. Implicit diffusion models for continuous super-resolution. In *Proc. CVPR*, 10021–10030.

Goodfellow, I.; Pouget-Abadie, J.; Mirza, M.; Xu, B.; Warde-Farley, D.; Ozair, S.; Courville, A.; and Bengio, Y. 2020. Generative adversarial networks. *Communications of the ACM*, 63(11): 139–144.

Grundmann, M.; Kwatra, V.; Castro, D.; and Essa, I. 2012. Calibration-free rolling shutter removal. In *Proc. ICCP*, 1–8.

Han, Y.; Luo, K.; Luo, A.; Liu, J.; Fan, H.; Luo, G.; and Liu, S. 2022. RealFlow: EM-based realistic optical flow dataset generation from videos. In *Proc. ECCV*, 288–305.

Hartley, R.; and Zisserman, A. 2003. *Multiple view geometry in computer vision*. Cambridge university press.

Hedborg, J.; Forssén, P.-E.; Felsberg, M.; and Ringaby, E. 2012. Rolling shutter bundle adjustment. In *Proc. CVPR*, 1434–1441.

Ho, J.; Jain, A.; and Abbeel, P. 2020. Denoising diffusion probabilistic models. In *Proc. NeurIPS*, 6840–6851.

Ho, J.; and Salimans, T. 2022. Classifier-free diffusion guidance. *arXiv preprint arXiv:2207.12598*.

Huang, L.; Zhang, H.; Xu, T.; and Wong, K.-C. 2023. Mdm: Molecular diffusion model for 3d molecule generation. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 37, 5105–5112.

Huang, W.; and Liu, H. 2018. Online initialization and automatic camera-IMU extrinsic calibration for monocular visual-inertial SLAM. In *Proc. ICRA*, 5182–5189.

Jia, C.; and Evans, B. L. 2013. Online calibration and synchronization of cellphone camera and gyroscope. In *IEEE Global Conference on Signal and Information Processing*, 731–734.

Joshi, N.; Kang, S. B.; Zitnick, C. L.; and Szeliski, R. 2010. Image deblurring using inertial measurement sensors. *ACM Transactions on Graphics (TOG)*, 29(4): 1–9.

Kandula, P.; Kumar, T. L.; and Rajagopalan, A. 2020. Deep end-to-end rolling shutter rectification. *JOSA A*, 37(10): 1574–1582.

Karpenko, A.; Jacobs, D.; Baek, J.; and Levoy, M. 2011. Digital video stabilization and rolling shutter correction using gyroscopes. *Computer and Structures*, 1(2): 13.

Li, H.; Jiang, H.; Luo, A.; Tan, P.; Fan, H.; Zeng, B.; and Liu, S. 2024. Dmhomo: Learning homography with diffusion models. *ACM Transactions on Graphics*, 43(3): 1–16.

Li, H.; Luo, K.; and Liu, S. 2021. GyroFlow: gyroscope-guided unsupervised optical flow learning. In *Proc. ICCV*, 12869–12878.

Li, H.; Luo, K.; Zeng, B.; and Liu, S. 2023. GyroFlow+: Gyroscope-Guided Unsupervised Deep Homography and Optical Flow Learning. *arXiv preprint arXiv:2301.10018*.

Liang, C.-K.; Chang, L.-W.; and Chen, H. H. 2008. Analysis and compensation of rolling shutter effect. *IEEE Trans. on Image Processing*, 17(8): 1323–1330.

Liu, P.; Cui, Z.; Larsson, V.; and Pollefeys, M. 2020. Deep shutter unrolling network. In *Proc. CVPR*, 5941–5949.

Liu, S.; Li, H.; Wang, Z.; Wang, J.; Zhu, S.; and Zeng, B. 2021. DeepOIS: Gyroscope-guided deep optical image stabilizer compensation. *IEEE Transactions on Circuits and Systems for Video Technology*, 32(5): 2856–2867.

Liu, S.; Yuan, L.; Tan, P.; and Sun, J. 2013. Bundled camera paths for video stabilization. *ACM Trans. Graphics*, 32(4): 1–10.

Liu, X.; Park, D. H.; Azadi, S.; Zhang, G.; Chopikyan, A.; Hu, Y.; Shi, H.; Rohrbach, A.; and Darrell, T. 2023. More control for free! image synthesis with semantic diffusion guidance. In *Proc. WACV*, 289–299.

Luo, R.; Song, Z.; Ma, L.; Wei, J.; Yang, W.; and Yang, M. 2024. Diffusiontrack: Diffusion model for multi-object tracking. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 38, 3991–3999.

Mo, J.; Islam, M. J.; and Sattar, J. 2022. IMU-Assisted Learning of Single-View Rolling Shutter Correction. In *Proc. ICRL*, 861–870.

Ni, H.; Shi, C.; Li, K.; Huang, S. X.; and Min, M. R. 2023. Conditional Image-to-Video Generation with Latent Flow Diffusion Models. In *Proc. CVPR*, 18444–18455.

Ouyang, L.; Wu, J.; Jiang, X.; Almeida, D.; Wainwright, C.; Mishkin, P.; Zhang, C.; Agarwal, S.; Slama, K.; Ray, A.; et al. 2022. Training language models to follow instructions with human feedback. *Proc. NeurIPS*, 35: 27730–27744.

Philbin, J.; Chum, O.; Isard, M.; Sivic, J.; and Zisserman, A. 2007. Object retrieval with large vocabularies and fast spatial matching. In *Proc. CVPR*, 1–8.

Purkait, P.; Zach, C.; and Leonardis, A. 2017. Rolling shutter correction in manhattan world. In *Proc. ICCV*, 882–890.

Rengarajan, V.; Balaji, Y.; and Rajagopalan, A. 2017. Unrolling the shutter: CNN to correct motion distortions. In *Proc. CVPR*, 2291–2299.

Rengarajan, V.; Rajagopalan, A. N.; and Aravind, R. 2016. From bows to arrows: Rolling shutter rectification of urban scenes. In *Proc. CVPR*, 2773–2781.

Ringaby, E.; and Forssén, P.-E. 2012. Efficient video rectification and stabilisation for cell-phones. *International Journal of Computer Vision*, 96(3): 335–352.

Saurer, O.; Koser, K.; Bouguet, J.-Y.; and Pollefeys, M. 2013. Rolling shutter stereo. In *Proc. ICCV*, 465–472.

Saurer, O.; Pollefeys, M.; and Lee, G. H. 2016. Sparse to dense 3D reconstruction from rolling shutter images. In *Proc. CVPR*, 3337–3345.

Saxena, S.; Herrmann, C.; Hur, J.; Kar, A.; Norouzi, M.; Sun, D.; and Fleet, D. J. 2023. The Surprising Effectiveness of Diffusion Models for Optical Flow and Monocular Depth Estimation. *arXiv preprint arXiv:2306.01923*.

Shao, H.; Svoboda, T.; and Van Gool, L. 2003. Zubud-Zurich buildings database for image based recognition. *Tech. Rep(Swiss Federal Institute of Technology*, 260: 6–8.

Sohl-Dickstein, J.; Weiss, E.; Maheswaranathan, N.; and Ganguli, S. 2015. Deep unsupervised learning using nonequilibrium thermodynamics. In *Proc. ICML*, 2256–2265.

Song, J.; Meng, C.; and Ermon, S. 2020. Denoising diffusion implicit models. *arXiv preprint arXiv:2010.02502*.

Song, Y.; and Ermon, S. 2019. Generative modeling by estimating gradients of the data distribution. In *Proc. NeurIPS*, 1–9.

Song, Y.; Sohl-Dickstein, J.; Kingma, D. P.; Kumar, A.; Ermon, S.; and Poole, B. 2020. Score-based generative modeling through stochastic differential equations. *arXiv preprint arXiv:2011.13456*.

Tevet, G.; Raab, S.; Gordon, B.; Shafir, Y.; Cohen-Or, D.; and Bermano, A. H. 2022. Human motion diffusion model. *arXiv preprint arXiv:2209.14916*.

Vasu, S.; Rajagopalan, A.; et al. 2018. Occlusion-aware rolling shutter rectification of 3d scenes. In *Proc. CVPR*, 636–645.

Xiao, J.; Hays, J.; Ehinger, K. A.; Oliva, A.; and Torralba, A. 2010. SUN database: Large-scale scene recognition from abbey to zoo. In *Proc. CVPR*, 3485–3492.

Yan, W.; Tan, R. T.; Zeng, B.; and Liu, S. 2023. Deep Homography Mixture for Single Image Rolling Shutter Correction. In *Proc. ICCV*, 9868–9877.

Yinhuai, W.; Jiwen, Y.; and Jian, Z. 2022. Zero-shot image restoration using denoising diffusion null-space model. *arXiv preprint arXiv:2212.00490*.

Zhuang, B.; Cheong, L.-F.; and Hee Lee, G. 2017. Rolling-shutter-aware differential SFM and image rectification. In *Proc. ICCV*, 948–956.

Zhuang, B.; Tran, Q.-H.; Ji, P.; Cheong, L.-F.; and Chandraker, M. 2019. Learning structure-and-motion-aware rolling shutter correction. In *Proc. CVPR*, 4551–4560.