Rolling Shutter Image as an IMU: Estimating Camera Motion from a Single Rolling Shutter Image

Prashanth Shyamala
prashanth.shyamala@mail.utoronto.ca

Abstract—Most robotics and VR/AR applications use rolling shutter cameras, which suffer from artifacts under fast motion, causing existing camera pose estimation methods to fail. In this work, we propose a technique that leverages the rolling shutter artifact as a temporal cue for motion estimation rather than treating it as an unwanted artifact. Our approach works by predicting a dense per-row motion flow field and a monocular depth map from a single image. We then recover the per-row instantaneous camera velocity by solving a linear least-squares problem under the small-motion assumption. In essence, our method produces an IMU-like measurement that robustly captures fast and aggressive camera movements. Our approach is able to recover the angular and translational velocities in real-world scenes and outperforms the previous motion-from-single-image method, "Image as an IMU."

Index Terms—Rolling Shutter, Motion Estimation

1 Introduction

Structure from motion and visual odometry methods can accurately estimate camera poses from a sequence of images in most real-world scenes. These methods typically assume that the camera remains stationary during a single exposure, using each captured frame to solve for relative camera motion across time. However, this assumption ignores the fact that a camera does indeed move during exposure, and such intra-frame motion introduces blur and rolling shutter artifacts. Most consumer-grade cameras employ CMOS sensors with rolling shutters. Conventional approaches either discard motion-blurred or rolling-shutter–distorted frames or incorporate inertial measurement units (IMUs) to improve robustness. While effective, IMUs introduce additional challenges such as sensor synchronization and long-term drift.

Most visual odometry techniques fail when inter-frame correspondences are sparse or unreliable due to distortions. A common workaround is to first correct motion blur or rolling-shutter distortion and then apply standard SLAM or odometry methods. However, this additional correction step increases computational cost and is undesirable for real-time applications.

The closest work to estimate motion from a single image - Image as an IMU [1] - predicts camera motion by using motion blur as a cue. When the exposure time is low, motion blur is reduced, but rolling shutter distortions become more pronounced. In this work, we extend IAAI (Image as an IMU) and develop a method for estimating relative motion from a single rolling-shutter image.

Our key insight is that the magnitude and direction of rolling shutter distortion provide direct cues about the underlying camera motion. Building on this observation, we propose a method that estimates the relative motion the camera undergoes during exposure using only the distortions present in a single frame. Our pipeline operates in two stages: first, we predict the rolling-shutter–to–global-shutter flow field and a monocular depth map; then, we recover the scanline-wise relative camera pose by solving a linear least-squares system. Notably, our method avoids explicit rolling shutter correction across multiple frames and, given the sensor readout time, produces instantaneous rotational and translational velocity estimates that can be interpreted as IMU-like measurements.

2 Related Work

Minimal rolling-shutter PnP solvers. Early work on rolling-shutter absolute pose estimation focused on extending minimal PnP solvers to handle row-dependent camera motion. [2] propose a polynomial-time minimal solver that jointly estimates the camera pose along with its translational and angular velocities during readout using a double-linearized RS projection model. Their formulation enables non-iterative estimation from just six 2D–3D correspondences and achieves significantly higher inlier counts than P3P within RANSAC on both synthetic and real data. However, the method assumes the camera orientation is close to identity and commonly relies on a P3P-based initialization, limiting applicability when the motion is large or when robust initialization is unavailable.

Inertial-regularized RS estimation. Subsequent methods increasingly incorporate inertial sensing to constrain the underdetermined RS trajectory. [3] integrates a continuous-time constant-twist rolling-shutter model into a direct visual–inertial odometry framework, jointly optimizing keyframe poses, velocities, IMU biases, and inverse depths in a photometric bundle adjustment that incorporates both RS projection and IMU preintegration. Their results show that explicitly modeling time-varying pose during readout improves trajectory accuracy compared to global-shutter assumptions, though the method requires multi-frame sequences and is not designed for single-image motion inference. In parallel, [4] proposes a learning-based RS correction pipeline that predicts per-pixel depth and row-wise camera poses conditioned on IMU signals; the corrected images substantially improve SLAM performance (e.g., DSO) on RS data. While this demonstrates the value of combining inertial cues with learned priors, the focus is on image rectification rather than directly recovering a physically meaningful motion model from a single frame.

General rolling-shutter trajectory models. More general RS camera models aim to flexibly parameterize the pose trajectory across image rows. [5] introduces a formulation in which each scanline is assigned its own pose, regularized by smoothness constraints on derivatives of the pose parameters along the readout direction. Their framework optimizes these per-line poses from 2D–3D correspondences under suitable priors, allowing representation of complex non-uniform motions but requiring many correspondences and iterative non-linear optimization—making it unsuitable as a minimal solver.

Relation to our work. Unlike these approaches, the present work targets the single-image setting and seeks to bridge the gap between minimal, closed-form RS solvers such as [2] and highly parameterized dynamic RS models. Our goal is to estimate a physically interpretable row-wise motion trajectory directly from a single rolling-shutter image, without relying on multiple frames or inertial sensing.

3 Background

This work builds upon Image as an IMU [1] in which the image formation process is modeled as the average of a discrete sequence of virtual images across the exposure time -

$$I^B \approx g\left(\frac{1}{N}\sum_{i=1}^{N} I_i^{\nu}\right) \tag{1}$$

such that $\{I_1^v, ..., I_N^v\}$ are the N "virtual images" during the exposure time.

The method jointly predicts a per-pixel motion field from the initial virtual image to the final virtual image in the sequence and dense depth map using a neural network. The motion field can be modeled using the equations from [6], relating image flow (F_x, F_y) to camera translation (t_x, t_y, t_z) and rotation $(\theta_x, \theta_y, \theta_z)$:

$$F_x = \frac{t_z p_x - t_x f}{d} - \theta_y f + \theta_z p_y + \frac{\theta_x p_x p_y}{f} - \frac{\theta_y p_x^2}{f}, \tag{2}$$

$$F_y = \frac{t_z p_y - t_y f}{d} + \theta_x f - \theta_z p_x - \frac{\theta_y p_x p_y}{f} + \frac{\theta_x p_y^2}{f}. \tag{3}$$

where $d \in D$ (depth map) and f is the known focal length.

These can be expressed in matrix form as $Ax = b$, where:

$$A = \begin{bmatrix} -\frac{f}{d} & 0 & \frac{p_x}{d} & \frac{p_x p_y}{f} & -\frac{p_x^2+f^2}{f} & p_y \\ 0 & -\frac{f}{d} & \frac{p_y}{d} & \frac{p_y^2+f^2}{f} & -\frac{p_x p_y}{f} & -p_x \end{bmatrix}, \tag{4}$$

$$x = \begin{bmatrix} t_x \\ t_y \\ t_z \\ \theta_x \\ \theta_y \\ \theta_z \end{bmatrix}, \quad b = \begin{bmatrix} F_x \\ F_y \end{bmatrix}. \tag{5}$$

4 Proposed Method

Our method estimates scanline-wise camera poses by solving the linear system defined in equations 4 and 5. Once these poses are obtained, we consider two strategies for estimating the camera velocity:

1) **Dynamic per-scanline velocity:** For each scanline i, we compute its relative pose with respect to scanline 1. The velocity is then obtained by differencing consecutive poses and dividing by the corresponding time interval between scanline capture times.
2) **Constant-velocity assumption:** Assuming the camera moves with uniform velocity during the frame, we fit a line to the sequence of pose differences. The slope of this line provides the estimated constant velocity.

4.1 Rolling Shutter Camera Model

Following [7] and [5], we model the rolling shutter imaging process as follows.

In a rolling shutter camera, each image row $r \in \{1 \dots n\}$ is captured at a different time, given by

$$t(y) = t_0 + r\,\Delta t, \tag{6}$$

with t_0 the exposure start time and Δt the line readout interval.

Let $X \in \mathbb{R}^3$ be a scene point in world coordinates. During exposure, the camera pose evolves continuously as rotation $R(t)$ and translation $p(t)$. The projection of X onto row r uses the camera pose at time $t(y)$:

$$x(y) \sim K\big(R(t(r))^\top (X - p(t(r)))\big), \tag{7}$$

where K is the intrinsic calibration matrix and $x(r)$ is the homogeneous image coordinate.

For any time t, the point's camera-coordinate representation is

$$X_c(t) = R(t)^\top \big(X - p(t)\big), \tag{8}$$

which produces the pinhole projection

$$x(t) \sim K\, X_c(t). \tag{9}$$

Substituting $t = t(r)$ yields the rolling shutter projection model:

$$x(y) \sim K\, R(t(r))^\top \big(X - p(t(r))\big). \tag{10}$$

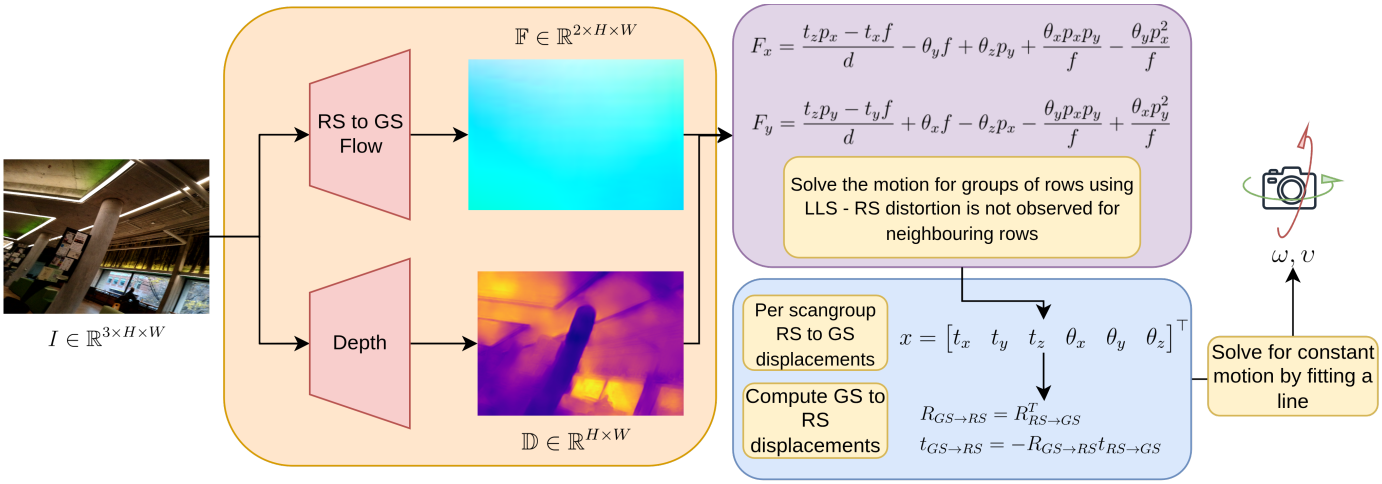

Fig. 1: Method overview. Given a single rolling shutter image, we pass it through the network to predict the flow field and metric depth. These are then formulated in a linear system where the optimal pose displacements are solved for using linear least squares on scangroups. The velocities are calculated based on the slope of the line fit to the pairwise differences of the pose displacements.

4.2 Motion Field and Depth Estimation

To get the required motion field F, we require a dense pixel-wise mapping from rolling shutter (RS) coordinates to their global shutter (GS) counterparts. This motion field is obtained using the method of [8], which employs a diffusion model to denoise a dense RS$\to$GS motion field - $\mathbb{F} \in \mathbb{R}^{2\times H\times W}$. Each flow vector $F = [F_x, F_y]^\top \in \mathbb{F}$ is designed to be the pixel displacement from a given pixel location in the original RS image to the GS image:

$$F = p_{t(0)} - p_{t(y)} \tag{11}$$

Depth estimation must be consistent with the rolling shutter image formation process. Therefore, the depth map, $\mathbb{D} \in \mathbb{R}^{H\times W}$, must be produced from a capture that undergoes the same transformation as in equation 10. We use the depth decoder of [1], which produces plausible depth values aligned with the RS distortions.

4.3 Scanline-wise Pose Estimation

Rolling shutter distortions can be approximated as negligible over a small group of consecutive scanlines [5], [9]. Thus, we process *scan groups* of variable size (including the scanline-wise case with group size 1), balancing computational efficiency and noise robustness. The impact of group size is analyzed in section 5.2.

Let a scan group y consist of a set of H_y consecutive rolling shutter rows. For this scan group, we obtain the corresponding RS$\to$GS motion field.

$$\mathbb{F}(y) \in \mathbb{R}^{2\times H_y\times W}$$

and the RS-aware depth map

$$\mathbb{D}(y) \in \mathbb{R}^{H_y\times W}.$$

Using $\mathbb{F}(y)$ and $\mathbb{D}(y)$, we construct the linear system in equations 4 and 5. Solving this system yields the relative pose of scan group y with respect to scan group 1, represented as the translation vector $t_{RS\to GS}(y)$ and the Euler angles defining the rotation.

The estimated Euler angles are converted to a rotation matrix by multiplying the rotation matrices in a chained sequence.

$$R_{RS\to GS}(y).$$

The inverse motion (from GS to RS) is obtained by

$$R_{GS\to RS}(y) = R_{RS\to GS}(y)^\top, \tag{12}$$

$$t_{GS\to RS}(y) = -\,R_{GS\to RS}(y)\, t_{RS\to GS}(y). \tag{13}$$

These expressions define the scan group–specific pose.

Given n scan groups, we obtain n such poses. Pairwise pose differences yield per-group dynamic velocities when divided by the capture time difference between the centers of the scan groups. Alternatively, assuming constant velocity, we smooth the sequence of pose differences by taking a moving average, fitting a line, and taking its slope as the frame-level velocity estimate.

5 Experimental Results

5.1 Evaluation

Dataset and Setup. We evaluate our method on three real-world scenes captured by us with an OAK-D camera equipped with a 12MP IMX378 (PY011) rolling shutter sensor. The camera also has an integrated 9-axis IMU - BNO085 - which is used for ground truth measurements. The three sequences include an indoor room, an outdoor scene with natural lighting, and an indoor lounge (New College lounge) featuring significant depth variation. To assess robustness, the camera was shaken vigorously during capture, inducing strong rolling shutter distortions.

Ground-Truth and Sensors. For rotational motion, we compare our estimated angular velocities directly against the gyroscope readings from the onboard IMU. For translation, we compare our predicted translational velocities with

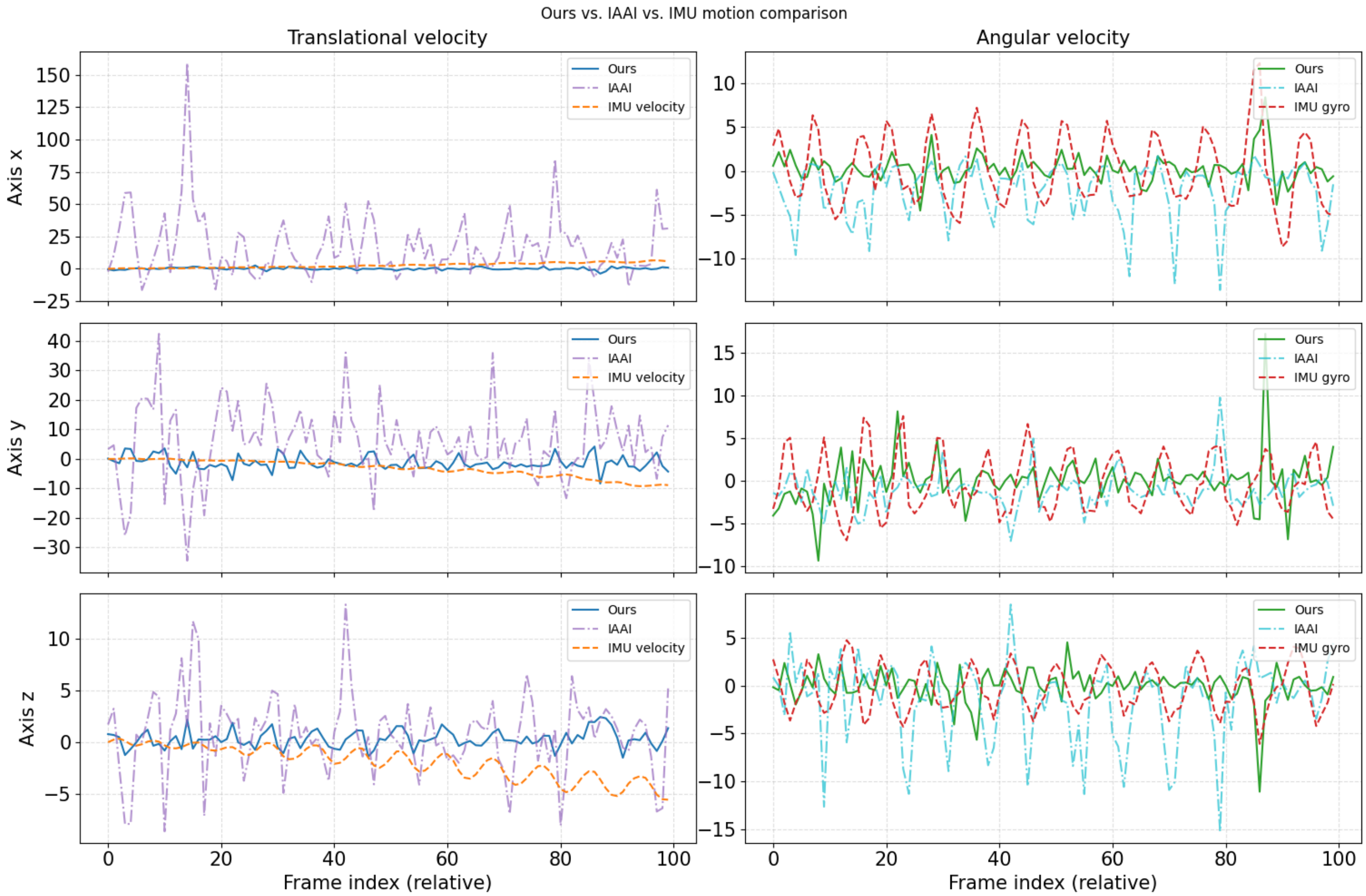

Fig. 2: Predicted velocities for the outdoor scene using our method and IAAI. The plots show angular and translational velocities along the three axes. Our method closely follows the ground-truth motion—especially in translation—with only minor deviations caused by depth prediction errors and slight camera–IMU timing synchronization error. In contrast, IAAI produces highly erratic translational estimates and inconsistent or near-zero angular predictions.

TABLE 1: RMSE comparison of angular and translational velocities (lower is better). For angular motion, our method achieves lower error than IAAI in most axes across the evaluated scenes. For translational motion, our method consistently outperforms IAAI in all captured scenes.

Method	Outdoor	Indoor	New College	Avg
Angular Velocities RMSE (rad/s)				
	$\omega_x/\omega_y/\omega_z$	$\omega_x/\omega_y/\omega_z$	$\omega_x/\omega_y/\omega_z$	$\omega_x/\omega_y/\omega_z$
IAAI	3.95/**3.15**/3.70	**4.22**/**3.71**/4.76	**4.03**/3.52/4.23	4.07/**3.46**/4.23
Ours	**3.72**/3.95/**2.75**	4.53/5.39/**3.74**	4.15/**2.70**/**3.39**	**4.13**/4.01/**3.29**
Translational Velocities RMSE (m/s)				
	$v_x/v_y/v_z$	$v_x/v_y/v_z$	$v_x/v_y/v_z$	$v_x/v_y/v_z$
IAAI	23.03/127.73/4.60	56.64/151.12/4.22	42.42/188.19/21.37	40.70/155.68/10.73
Ours	**3.39/3.68/2.90**	**5.27/7.81/1.69**	**3.52/9.35/18.37**	**4.06/6.95/7.65**

accelerometer-derived velocities. Because the accelerometer measures instantaneous acceleration, integration introduces bias and drift, making absolute comparison unreliable. Thus, we focus on matching peaks, troughs, and temporal structure rather than absolute scale.

Scanline vs. Frame-Level Evaluation. The IMU sampling rate is considerably lower than the effective row sampling rate of a rolling shutter sensor, so true scanline-level ground truth is unavailable. Moreover, per-row velocity estimates are noisy. For these reasons, we evaluate the aggregated frame-level (global) velocity estimates. Table 1 summarizes RMSE values for angular and translational motion across all scenes.

Rotational Motion Results. The IAAI baseline performs competitively for angular velocity and occasionally matches or slightly surpasses our method in individual axes. This is partly due to near-periodic motion during the recordings, where many angular velocities cluster near zero. Nevertheless, qualitative comparisons show that our predictions align more faithfully with the magnitude and direction of the ground-truth IMU signal. A slight temporal offset remains, stemming from image-based apparent motion being captured later and residual sensor-sync differences.

Translational Motion Results. Translation estimation is

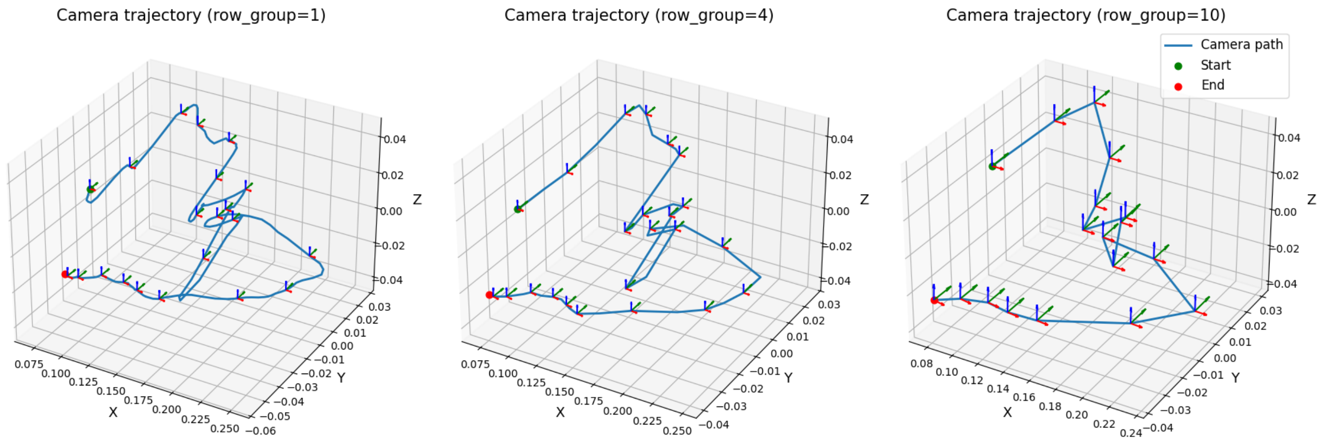

Fig. 3: Intra-frame camera poses estimated using different scan-group sizes. The camera path does not vary as we decrease the group size, which shows that the rolling shutter distortions are not significant for consecutive rows. A greater group size leads to smoother trajectories.

substantially more challenging, and here our method clearly outperforms IAAI. Our approach recovers the major peaks and troughs and maintains realistic magnitudes consistent with IMU-derived velocities, despite the inherent drift of integrated accelerometer data. In contrast, IAAI frequently produces erratic and physically implausible translational outputs—especially along the x and y axes. Even along z, where IAAI appears more stable, the predictions fail to track the actual temporal dynamics. These findings show that while inertial-aided rolling-shutter techniques handle rotation reasonably well, they struggle significantly with translation, where our method provides consistent improvements.

5.2 Varying Group size

Figure 3 shows the recovered intra-frame camera trajectory for several choices of scan-group size for an image. We observe that the overall trajectory shape remains consistent. This holds across all configurations and indicates that rolling shutter distortions are relatively small within short ranges of consecutive rows. This supports the assumption that scan groups of modest size can be treated as approximately distortion-free when estimating local camera motion.

As the group size increases, the number of pose estimates per frame naturally decreases, resulting in a sparser trajectory. However, these estimates exhibit reduced noise and improved stability, since each larger group aggregates information over more rows and therefore averages out local pixel-level uncertainty. In contrast, group size 1 provides the finest temporal granularity but is also the most sensitive to noise in the motion field and depth estimates.

Overall, the trend illustrates a clear trade-off between temporal resolution and measurement robustness. Smaller scan groups capture rapid intra-frame changes but may suffer from jitter, while larger groups provide smoother, more reliable poses at the cost of reduced sampling density.

5.3 Limitations

Figure 3 shows the scan-group-wise camera trajectory. Ideally, the motion should follow a smooth curve; however, we observe short segments of smooth motion interrupted by abrupt direction changes. This pattern appears consistently across most of the rolling shutter images in our captured dataset, and consequently, the estimated velocities are inherently noisy. The primary source of this noise stems from limitations in the underlying models.

The RS→GS diffusion model from [10] was trained exclusively on outdoor pedestrian area scenes containing predominantly rotation-induced rolling shutter distortions. As a result, it fails to provide reliable corrections for translational motion or indoor environments. Similarly, the IAAI depth model performs well in scenes with small depth variation but becomes unreliable when applied to out-of-distribution scenes or environments with large depth ranges. The combination of these two factors contributes significantly to the noise observed in our motion estimates.

An important direction for improvement is the development of an RS-aware depth model trained on a synthetic dataset. Such a dataset can be constructed by applying the transformation in equation 10 to 3D scenes, thereby generating consistent RS-aware depth maps. Since both least-squares estimation and regression are differentiable, the entire pipeline can be trained end-to-end using motion supervision, potentially yielding more stable and accurate results.

6 Conclusion

We have presented a method for estimating instantaneous camera velocity directly from individual rolling shutter images, effectively producing IMU-like measurements without requiring an actual IMU. Our evaluations demonstrate that the method can recover rotational and translational motion with correct magnitude and direction, and in many cases, even outperform the previous state-of-the-art (IAAI) as well as the IMU-derived ground truth. This suggests that rolling shutter distortions, when properly leveraged, contain rich information about camera motion.

Despite limitations stemming from the underlying correction and depth models, our approach offers a strong foundation for exploiting motion-induced rolling shutter effects. With improved RS-aware depth estimation and end-to-end training, we expect the performance to increase further. Moreover, integrating our method with existing visual odometry pipelines has the potential to significantly enhance motion estimation.

ACKNOWLEDGMENTS

The authors would like to thank Kelly Zhu for supervising the project, and Professor David Lindell.

REFERENCES

[1] J. Chen and R. Clark, "Image as an imu: Estimating camera motion from a single motion-blurred image," 2025. [Online]. Available: https://arxiv.org/abs/2503.17358

[2] C. Albl, Z. Kukelova, and T. Pajdla, "R6p — rolling shutter absolute camera pose," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2015, pp. 2292–2300.

[3] D. Schubert, N. Demmel, L. von Stumberg, V. Usenko, and D. Cremers, "Rolling-shutter modelling for direct visual-inertial odometry," *CoRR*, vol. abs/1911.01015, 2019. [Online]. Available: http://arxiv.org/abs/1911.01015

[4] J. Mo, M. J. Islam, and J. Sattar, "Learning rolling shutter correction from real data without camera motion assumption," *CoRR*, vol. abs/2011.03106, 2020. [Online]. Available: https://arxiv.org/abs/2011.03106

[5] L. Magerand and A. Bartoli, "A generic rolling shutter camera model and its application to dynamic pose estimation," 2010. [Online]. Available: https://api.semanticscholar.org/CorpusID:199569139

[6] E. Trucco and A. Verri, *Introductory Techniques for 3-D Computer Vision*. USA: Prentice Hall PTR, 1998.

[7] P.-E. Forssén and E. Ringaby, "Rectifying rolling shutter video from hand-held devices," in *2010 IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, 2010, pp. 507–514.

[8] Z. Yang, H. Li, M. Hong, C.-L. Zhang, J. Li, and S. Liu, "Single image rolling shutter removal with diffusion models," 2025. [Online]. Available: https://arxiv.org/abs/2407.02906

[9] A. Karpenko and D. E. Jacobs, "Ctsr 2011-03 digital video stabilization and rolling shutter correction using gyroscopes," 2011. [Online]. Available: https://api.semanticscholar.org/CorpusID:887233

[10] Z. Yang, H. Li, M. Hong, C. Zhang, J. Li, and S. Liu, "Single image rolling shutter removal with diffusion models," *arXiv preprint arXiv:2407.02906*, 2024.

Fig. 4: Rolling-shutter images from the indoor, outdoor, and New College sequences that highlight the rolling shutter distortion extent in the evaluation sequences. The first column of images are taken without motion.