Anchor Points for Monocular Depth Estimation Under Severe Rolling Shutter Ego-motion

Jacob Carter, Vaishnav Ramesh, Md Jahidul Islam, and Sanjeev J. Koppal[†]

***Abstract*—Rolling shutter (RS) cameras are widely deployed in robotic systems due to their low cost and power efficiency. However, robot ego-motion induces significant RS distortions that degrade geometric consistency and severely challenge visual perception. In this work, we address monocular depth estimation under large rolling shutter distortions caused by ego-motion. We introduce the concept of an anchor point, which establishes a consistent temporal and geometric reference between rolling shutter and global shutter observations. This formulation enables principled adaptation and fine-tuning of deep foundation models for depth estimation under RS distortions. To systematically evaluate the problem, we present a new rolling shutter dataset featuring large motion-induced geometric distortions representative of mobile robotic tasks. Our experimental analysis shows that anchor point selection is a primary factor governing depth accuracy under RS effects. We further characterize anchor configurations for RS-GS data capture and identify stable regimes that yield accurate depth estimates. Using a representative two-stage depth estimation pipeline, we conduct extensive ablation studies demonstrating that anchor-based modeling significantly improves performance, outperforming recent RS correction baselines under strong ego-motion. Finally, we validate our approach on a legged robot platform, demonstrating robust monocular depth estimation in the presence of substantial rolling shutter distortions in under 400 ms.**

I. Introduction

Rolling shutter (RS) effects arise from row-by-row image readout, causing each scanline to be captured at a different time rather than simultaneously as in global shutter (GS) imaging. In dynamic scenes, this temporal offset induces complex geometric distortions that depend on both camera ego-motion and scene dynamics. RS cameras remain ubiquitous in modern robotic systems because they are compact, power-efficient, and cost-effective compared to global shutter alternatives [2].

Prior work addresses RS correction using geometric and learning-based methods. Rolling-shutter-aware pose estimation has also been explored in robotics to mitigate downstream effects. Despite these advances, existing methods struggle with single-frame images under large distortions, often assuming mild motion, requiring extra sensors, or relying on temporal context. Moreover, learning-based methods fix a temporal reference implicitly, yet the impact of this choice on geometric consistency and depth prediction remains underexplored. Due to this lack of exploration, reliable monocular depth estimation from large RS distortions remains a challenge.

A central component of our approach is the **anchor point**, a temporal reference that specifies which GS frame a RS image is aligned to. Existing RS correction methods implicitly select a temporal reference frame, yet the consequences of this choice have not been systematically analyzed [1], [3]. We formalize this reference as an anchor point, derive conditions governing its optimal selection, and demonstrate experimentally that anchor-point choice is a first-order factor affecting both rolling shutter correction and downstream monocular depth estimation. Adjusting this reference alone can yield substantial performance gains without retraining (see Fig. 1). For comprehensive evaluation, we introduce a new collocated RS–GS dataset captured with a beam-splitter rig that provides explicit control over anchor points. Using this dataset and a two-stage visual foundation pipeline for RS correction and depth estimation, we demonstrate improved robustness to large distortions and aggressive ego-motion across synthetic and real-world benchmarks. Our contributions are as follows:

1) We introduce the concept of "*anchor point*" as a temporal reference for rolling shutter to global shutter (RS-GS) alignment, and provide a systematic analysis of its role in rolling shutter correction and monocular depth estimation under severe geometric distortions (Sect. III).
2) We present a new collocated RS-GS dataset with explicit anchor-point control, designed to capture realistic robotic motion with large ego-motion effects, and practical imaging artifacts (Fig. 3-5).
3) We conduct extensive ablation studies across both synthetic and real-world data to quantify the influence of anchor point selection on reconstruction accuracy and depth consistency (Fig. 6-7, Tables 2-3).
4) We develop a two-stage foundation-model-based pipeline and demonstrate its effectiveness for rolling shutter correction and monocular depth estimation under extreme distortions and aggressive ego-motion (Fig. 8, Table 1).
5) We validate the proposed system at 400 ms end-to-end latency on a legged robot with a collocated sensor setup (Fig. 10).

The proposed framework demonstrates promise for near-real-time RSC by mobile robots under challenging motion conditions. Its validation on a legged robot, subjected to rapid, irregular ego-motion, highlights its viability in real-world applications.

All authors are affiliated with the Dept. of Electrical and Computer Engineering, University of Florida, Gainesville, FL 32611, USA. E-mail: carterjacob@ufl.edu (corresponding author).

[†] Dr. Sanjeev J. Koppal holds concurrent appointments as an Associate Professor of ECE at the University of Florida and as an Amazon Scholar at Amazon Robotics. This project was performed at the University of Florida and is not associated with Amazon Robotics.

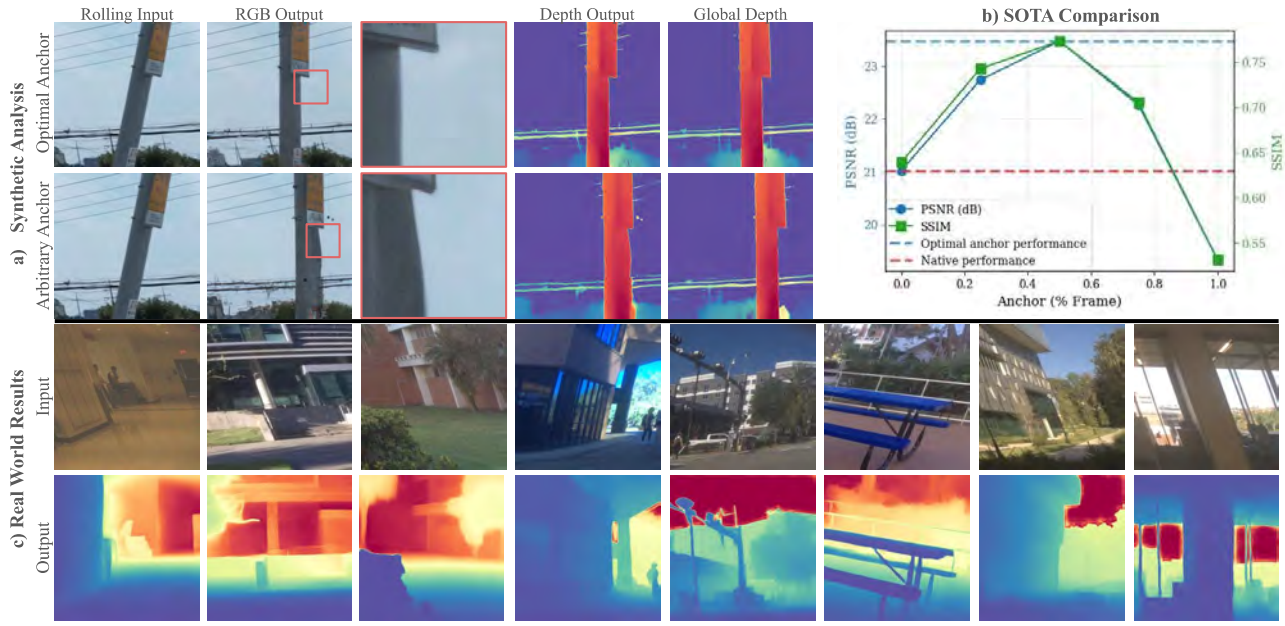

Fig. 1: **Anchor point formulation for rolling shutter correction**: **a)** We introduce the concept of an *anchor point*, which defines the temporal reference used to align rolling shutter observations during correction. **b)** Quantitative analysis on SOTA rolling shutter correction method [1] shows that performance varies systematically with anchor-point selection, providing a significant metric boost around the proposed optimal anchor when compared to native implementation. This modification is **one line of code**, requiring no retraining of the model. **c)** We introduce Visual Foundation Models to rolling shutter correction, providing geometrically consistent monocular depth from large rolling shutter distortions.

II. Related Work

Rolling shutter correction. Early single-frame approaches estimate camera motion from visual cues such as line geometry or motion blur [4]–[6], while multi-frame methods rely on tracking or interpolation across image sequences to recover parametric motion models [7]–[9]. Recent learning-based methods address RS correction using paired RS-GS data [1], [3]. Diffusion-based approaches have shown promise for single-frame RS correction [1], but assume moderate distortions or rely on additional sensing such as IMU measurements. In contrast, our work focuses on large RS distortions without external sensor input, emphasizing the role of temporal reference selection in downstream depth estimation.

Rolling shutter in robotics. Classical visual odometry (VO) and SLAM pipelines often assume global shutter imagery, leading to degraded pose and structure estimation under RS effects [10], [11]. Continuous-time trajectory representations and rolling-shutter-aware bundle adjustment have been proposed to mitigate these effects [12], [13], while minimal solvers for RS relative pose enable robust RANSAC-based geometry estimation [14], [15]. In visual-inertial pipelines, models that explicitly account for RS readout improve pose and map accuracy under aggressive motion [16]–[18]. This work complements these approaches by focusing on single-frame correction in large-distortion regimes, enabling downstream monocular depth estimation without inertial measurements.

Rolling shutter datasets. Progress in RS correction is closely tied to the availability of representative datasets. Existing datasets such as BS-RSCD [19], Fastec-RS and Carla-RS [3], and Real-RS [1] primarily target mild-to-moderate distortions or rely on synthetic motion profiles. In this work, we introduce a new collocated RS-GS dataset designed to capture large RS distortions arising from realistic robotic motion, with explicit control over temporal alignment through anchor point selection (see Fig. 5). This enables systematic analysis of RS correction and depth estimation under extreme distortions.

Rolling shutter correction and related imaging tasks. RS correction is related to motion deblurring and computational photography techniques that exploit coded exposure or specialized sensing [20]–[22]. Joint RS correction and motion deblurring has been explored [23], [24]. In this work, we assume sharp images without motion blur, consistent with prior RS correction studies [1], [3], [5], and focus on correcting geometric distortions. Unlike approaches that exploit RS artifacts for sensing or coding [25]–[28], our goal is to recover conventional geometry for downstream perception tasks.

Monocular depth estimation under distortions. Monocular depth estimation has seen significant advances through large-scale training and foundation models, including MiDaS [29], DPT [30], AdaBins [31], Depth-Anything [32], UDepth [33], and diffusion methods such as Marigold [34]. While these models generalize well, they assume GS imagery and degrade due to RS distortions. Our work addresses monocular depth estimation by first correcting severe RS artifacts and analyzing how temporal alignment influences geometric consistency.

III. Anchor Point for Rolling Shutter Correction

Rolling-shutter (RS) cameras expose rows sequentially, producing geometric distortions that depend on motion and row index. Any RSC algorithm must therefore choose a reference time specifying which global shutter (GS) frame the algorithm is attempting to reconstruct. Despite its importance, prior work often selects this arbitrarily (e.g., the first row or midpoint). Below we provide a theoretical definition and analysis of this reference, which we call the *anchor point*. Our analysis explains the importance of anchor point selection and derives why the midpoint $\frac{N}{2}$ is a strong default choice under mild motion assumptions.

A. Our discrete rolling shutter model

Following [35], the observed RS image can be modeled as an integration over exposure time T of the scene radiance E weighted by the shutter function S:

$$I(x,y) = \int_0^T E(x,y,t)\, S(x,y,t)\, dt. \tag{1}$$

We reinterpret each RS image as a discrete sampling from a hypothetical GS video recorded at the same rate as the row readout, so that row y of RS image is sampled from frame y:

$$I(x,y) = G(x,y,y). \tag{2}$$

Thus, correcting RS distortion amounts to choosing a GS frame anchor point t_{anch} and warping all rolling shutter rows toward that frame.

B. Theoretical claims

Claim 1. *The optimal anchor point lies in a discrete set*

Proof Sketch of Claim 1. Each row is captured at integer time index y, so choosing

$$t_{anch} = k \in 1, ..., N \tag{3}$$

forces row k to be perfectly aligned with its GS counterpart ($D_k(k) = 0$, where D can be *any* non-negative distortion metric). Non-integers do not provide this same property of perfect alignment. Under standard temporal smoothness assumptions, the expected distortion of other rows depends only on $|y-k|$, yielding

$$\mathbb{E}[D_y(t)] = h(|y-t|), \qquad h(0) = 0,\ h(u) > 0\ \forall u > 0. \tag{4}$$

Thus the set of viable anchors is the set of row indices. □

Corollary 1.1. *Because the feasible set contains only N elements, the optimal solution can be obtained by a grid search over candidate anchors.*

Claim 2 (General Optimal Anchor)**.** *For any rolling-shutter image, there exists an optimal global-shutter anchor point t^*_{anch} that minimizes the total expected distortion.*

Proof Sketch of Claim 2. Let X_n denote the horizontal position of a tracked structure in row n, and $X(t)$ be the GS reference position at anchor t. By defining a general distortion metric $D(t) = \sum_{n=1}^{N} |X_n - X(t)|$, we can write

$$X_n = vn + \alpha_n, \qquad X(t) = vt, \tag{5}$$

where v is nominal linear motion and α_n encodes accelerations, biases, or asymmetries. Then

$$D(t) = \sum_{n=1}^{N} |(vn - vt) + \alpha_n|. \tag{6}$$

Minimizing $D(t)$ over $t \in \{1, \ldots, N\}$ defines the general optimal anchor optimization pipeline as:

$$t^*_{\text{anch}} = \arg\min_{t \in \{1,\ldots,N\}} \sum_{n=1}^{N} |(vn - vt) + \alpha_n|. \tag{7}$$

□

Claim 3. *Under mild symmetry assumptions on scene motion (e.g., zero-mean deviations about constant velocity), the distortion-minimizing anchor coincides with the temporal midpoint of the exposure, $t_{anch} = \frac{N}{2}$*

Proof Sketch of Claim 3. For the derivation above, if the deviations satisfy $\mathbb{E}[\alpha_n] = 0$ for all n, then

$$\mathbb{E}[D(t)] = \sum_{n=1}^{N} |vn - vt|, \tag{8}$$

which is minimized by the median of $\{1, \ldots, N\}$. For evenly spaced rows, this is the temporal midpoint $t^*_{\text{anch}} = \frac{N}{2}$. □

Corollary 3.1. *The midpoint $\frac{N}{2}$ arises as a special-case solution under zero-mean, symmetric motion statistics.*

These claims show that anchor-point selection is a fundamental factor in RS correction. Our method serves as a strong instantiation demonstrating these effects, while the dataset provides a controlled benchmark for systematic evaluation.

Assumptions and Limitations Claim 3 should be interpreted as a special-case result rather than a universal optimum. Under strongly asymmetric motion profiles, the scene-dependent optimum described in Claim 2 may deviate from the temporal midpoint. However, our experiments on the REAL dataset and legged robot platform indicate that the midpoint anchor remains a robust practical choice, likely because deviations average out across exposure intervals and image content.

C. Anchor points as a method-agnostic improvement

Anchor-point selection is not specific to our pipeline, but is implicitly present in any RS correction method that maps a RS image to a single GS reference. We demonstrate with RS-Diffusion [1], whose output is anchored to the first row.

Without retraining, we re-centered the predicted flow so that zero displacement occurs at the image midpoint rather than the top row, changing the anchor from $t_{\text{anch}} = 0$ to $t_{\text{anch}} = \frac{N}{2}$. This simple post-processing step improves PSNR and SSIM (Fig. 1b), demonstrating that anchor-point selection is a first-order factor in RS correction. The broader significance is that anchor selection constitutes an overlooked degree of freedom shared by many RS correction methods.

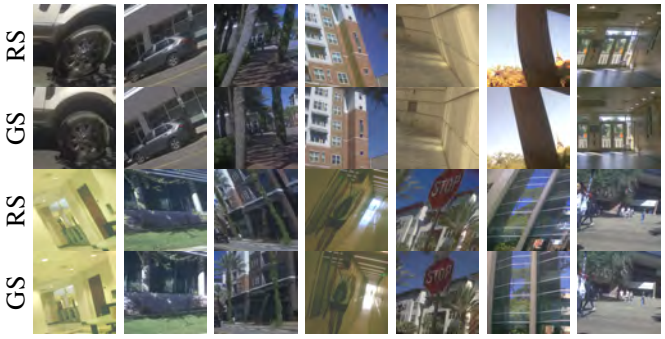

Fig. 2: Highlights of the Real Dataset. Each pair of rows shows the rolling shutter capture and the corresponding global shutter capture. The dataset will be released to the public.

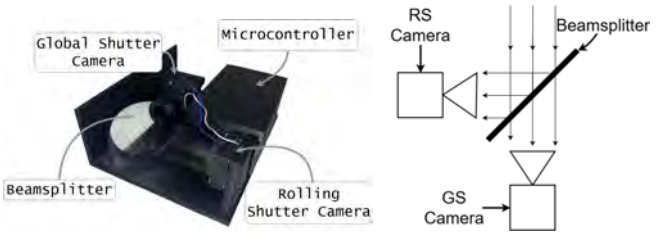

Fig. 3: Collocated RS–GS camera setup used for REAL. Hardware synchronization allows explicit control of anchor points for each capture. This device will be provided as a demo if accepted.

IV. Methodology

A. Dataset Creation and Analysis

To evaluate rolling shutter correction under large distortions, we provide three datasets containing paired RS and GS images.

SYN1: Synthetic Benchmark. *SYN1* contains RS–GS pairs rendered in Blender [36]. Each single-view scene (table fan, ceiling fan, propeller plane) includes two GS images for different anchor points and eight RS variants with increasing distortion, producing $(8+2) \times 200 = 2000$ images per scene.

SYN2: Realistic RS Simulation. *SYN2* introduces realistic motions and textures while maintaining full control over RS effects. RS frames are synthesized from the high-speed X4K1000FPS dataset [37] by row-wise stitching of consecutive GS frames. This method allows the selection of multiple anchor points for a single RS image. The dataset includes **22,040** GS frames and **4,408** RS images, covering a variety of motions and scene types.

REAL: Collocated RS-GS Capture. *REAL* is a real-world dataset captured using a **dual-camera collocated system** (Fig. 3), where rolling-shutter and global-shutter cameras are beam-splitter-aligned and hardware-synchronized. Varying trigger delays provides explicit control of the anchor point, producing multiple RS–GS pairs per scene. . The dataset contains 40k image pairs at 300×300 resolution (20k outdoor, 10k indoor) with additional images for anchor-point ablations. Optical-flow analysis (Fig. 5) shows that **REAL** exhibits the largest motion among RS datasets, making it suitable for evaluating large-distortion RS correction and depth estimation.

B. Integration of Vision Foundation Models

To address RS correction and monocular depth estimation, we leverage pre-existing vision foundation models (VFMs). Our pipeline consists of two sequential stages:

1) **RS Correction Stage:** We adopt a Marigold U-Net [38] trained to map RS images to corresponding to GS frames. The model operates on RGB images and focuses on removing row-wise distortions while preserving structure.
2) **Monocular Depth Estimation Stage:** Corrected RS frames from the first stage are passed to a transformer-based Depth-Anything-V2 [32]. This stage produces dense depth maps, benefiting from the reduction in RS-induced geometric distortions.

Since metric depth ground truth is unavailable for the RS-GS datasets, we evaluate depth consistency relative to a frozen Depth-Anything-V2 teacher operating on the GS image. This measures preservation of geometric structure after RS correction rather than absolute metric depth accuracy. The teacher provides a consistent reference across all methods, isolating errors introduced by rolling shutter distortions.

Pipeline Details. The stages are trained sequentially: first, the Marigold U-Net is trained for RS correction; second, the corrected outputs are used to fine-tune the depth model.

Training Setup. Both stages are trained using our RS-GS paired datasets (**SYN1**, **SYN2**, **REAL**). Standard image augmentation is applied, and losses include MSE loss for RS correction and scale invariant loss for Depth-Anything-V2.

C. Comparisons and Baselines

All trainable baselines were fine-tuned on the training splits used by our method whenever training code and model weights were publicly available [1]. For methods whose temporal reference is fixed by design, results are reported using the authors' original formulation. Evaluation was performed on identical testing data. Baselines of the author's original implementation are included alongside fine-tuned implementations.

V. Experimental Analyses

Our experiments were designed to validate the theoretical motivation for anchor points, evaluate VFM's on RSC, and analyze how different design choices affect rolling shutter correction and depth estimation quality.

A. Anchor Points for Controlled Synthetic Scenes

We conduct single-scene experiments on the synthetic dataset **SYN1** to isolate the effect of anchor-point selection. Scenes exhibit rotation about the image center, suggesting an optimal anchor of $t_a = \frac{N}{2}$. Six diffusion models were trained across three scenes, comparing optimal ($t_a = \frac{N}{2}$) and sub-optimal anchor points. Only models trained with the optimal anchor successfully corrected all three scenes, confirming our theoretical predictions (Fig. 4, Table I). We further compare against Manhattan World [5] and RS-Diffusion [1], which struggle under extreme distortions.

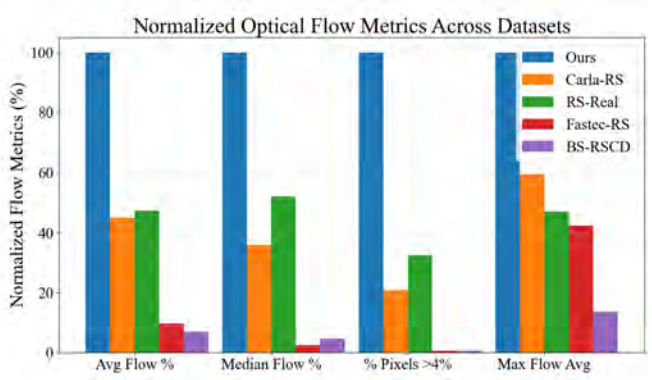

Fig. 5: Optical-flow motion statistics for RS datasets. The percentage of pixels greater than 2%, 6%, and 10% for our dataset was also over double other datasets compared. **No other dataset matches our egocentric motion, allowing us to model legged robots and other mobile systems.**

TABLE I: Comparisons of different anchor points with the first stage of our two-stage architecture on single-scene datasets (Ceiling Fan, Table Fan, Plane). The models were evaluated using PSNR and SSIM on intermediate RGB images. $t_a = \frac{N}{2}$ achieves the highest performance across all scenes.

Anchor	Ceiling Fan		Table Fan		Plane	
	PSNR	SSIM	PSNR	SSIM	PSNR	SSIM
$t_a = 0$	34.93	0.971	26.82	0.831	24.21	0.808
$t_a = \frac{N}{2}$	**36.21**	**0.986**	**27.57**	**0.842**	**25.02**	**0.841**

B. Anchor Points in General Realistic Scenes

Next, we evaluated the full two-stage pipeline, RS correction followed by monocular depth estimation.

Synthetic Results. On the X4K-based dataset [37], we trained models using different anchor-point choices spanning the exposure window. Performance consistently peaked near the temporal midpoint, with $t_a = \frac{N}{2}$ yielding the highest PSNR across training (Fig. 8). Visual comparisons show cleaner reconstructions and more consistent depth than prior RS correction methods (Fig. 6).

Real-World Results. On the novel **REAL** dataset, our pipeline consistently outperformed baseline RS correction models and Depth-Anything-V2 in both RGB fidelity and depth accuracy (Table. II). PSNR improvements are most pronounced in high-motion sequences, demonstrating robustness to large RS distortions (Fig. 7).

C. Ablation Study: Model Configuration

We evaluated three configurations: **one-stage**, **parallel two-stage**, and **orthogonal two-stage** (Table II). **One-stage** models (Marigold [38], Depth-Anything [32], EcoDepth [39]) jointly learn de-rolling and depth but struggle with structural consistency. **Parallel two-stage** models decouple the tasks sequentially using identical architectures; this improves stability but provides limited gains. **Orthogonal two-stage** models combine distinct architectures (Marigold U-Net for de-rolling and Depth-Anything-V2 for depth) and achieve the highest PSNR, SSIM, AbsRel, and δ_1. Even untuned orthogonal baselines outperform parallel two-stage designs, indicating that sequential specialization of tasks is a pivotal factor behind performance improvements. Fine-tuning further improves results, confirming that the pipeline effectively leverages VFMs for RSC and Monocular Depth Estimation.

The effects of anchor-point selection and architecture choice were studied independently. Table II isolates anchor-point selection while holding the underlying network architecture fixed. Table III evaluates different architectural configurations while using the optimal anchor point identified in Table II. The results indicate that both factors contribute to performance. Anchor-point selection consistently improves photometric re-

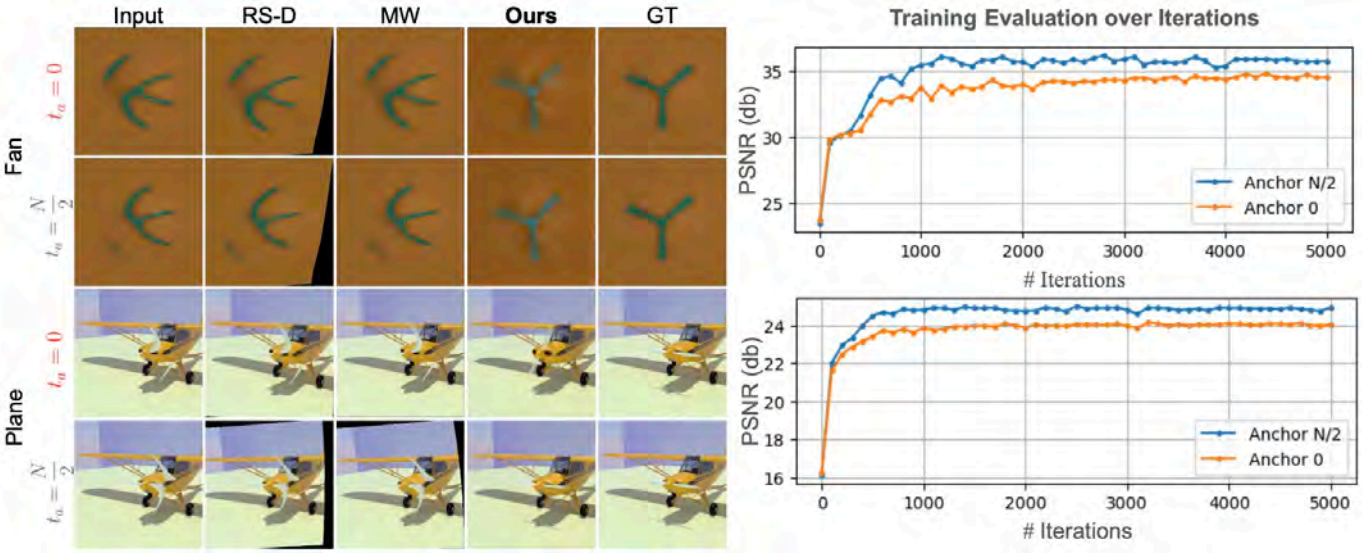

Fig. 4: Results on SYN1 comparing RS-Diffusion [1], Manhattan World [5], and our method under two anchor-point choices ($t_a = 0$, $t_a = \frac{N}{2}$). PSNR curves show validation performance per anchor-point.

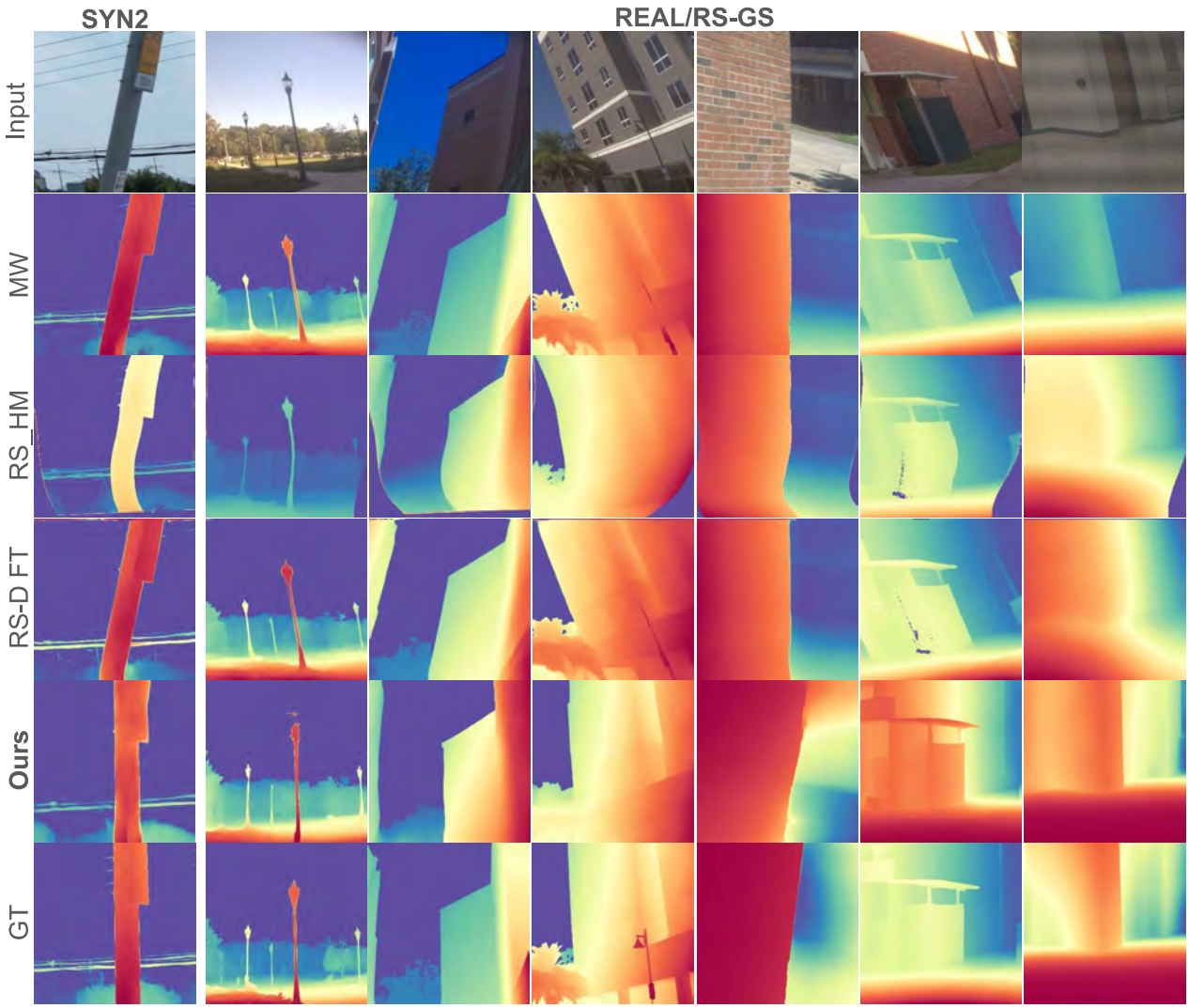

Fig. 6: Qualitative results on real scenes. From top to bottom: input image, two baseline comparison outputs, fine-tuned comparison output, our model's output, and ground truth. Only our method produces accurate depth, demonstrating the two-stage approach preserves scene structure. **Note Column 5**, where naive affine-based rectification does not work, since the ground truth has a slanted structure. Additional visualizations and comparisons can be seen in the supplementary video.

TABLE II: Comparisons of PSNR, SSIM, AbsRel, and δ_1 on the testing dataset used in Sect. 5.2. Our fine-tuned, two-stage architecture (U-Net, Transformer) performs better on validation than any other combination of the architectures. Values that are not possible to record (intermediate RGB metrics for one-stage architecture) are represented with NA. Comparisons of two-stage architectures with shared first stages will have repeated metrics for intermediate RGB.

Architecture	**Method**	Intermediate RGB		Depth Map	
		PSNR ↑	SSIM ↑	AbsRel ↓	δ_1 ↑
One Stage	Marigold [34]	NA	NA	6.047	0.257
	Depth-Anything [32]	NA	NA	0.655	0.757
	Eco-Depth [39]	NA	NA	0.488	0.676
Comparisons	RS-Diff [1]	19.47	0.433	0.904	0.190
	RS-Diff Fine-tuned [1]	21.52	0.469	0.802	0.214
	Deep RS-HM [40]	20.75	0.456	0.788	0.446
	Manhattan [5]	21.41	0.454	0.621	0.623
Parallel Two Stage	[34]- [34]	22.87	0.494	3.22	0.382
Orthogonal Two Stage	[34]- [39]	22.87	0.494	0.506	0.649
	Baseline [34]- [32]	22.87	0.494	0.377	0.773
	Fine Tuned (ours)	**22.87**	**0.494**	**0.191**	**0.842**

TABLE III: Comparisons of anchor points with the first stage of our architecture on **SYN2** and **REAL**. These comparisons were done on the intermediate RGB images, and $t_a = \frac{N}{2}$ consistently achieves the best or comparable performance.

Anchor	X4K		Collocated	
	PSNR	SSIM	PSNR	SSIM
$t_a = 0$	15.594	0.578	20.238	0.639
$t_a = \frac{N}{4}$	17.291	**0.630**	20.292	0.640
$t_a = \frac{N}{2}$	**17.733**	0.629	**21.778**	**0.674**

construction quality, while architectural specialization further improves downstream depth estimation metrics.

D. Ablation Study: Anchor Points in Other RSC Algorithms

We evaluated RS-Diffusion [1] across a sweep of anchor points and found that centering the predicted flow at the temporal midpoint ($t_{\text{anch}} = \frac{N}{2}$) consistently yields the best reconstruction on synthetic and real RS images. RS-Diffusion was chosen for ablation due to it's ability to change predicted flow easily without retraining. This confirms that mid-exposure alignment improves performance across different RSC methods without retraining, highlighting the generality of the anchor-point insight (Fig. 1).

E. Effectiveness on Legged Robots

To evaluate our method in this real-world setting, we collected 1500 images across three outdoor locations using the collocated sensor system mounted on Puppy-Pi. Runtime measurements were performed on an NVIDIA RTX A6000 GPU using FP16 inference at a resolution of 300×300.

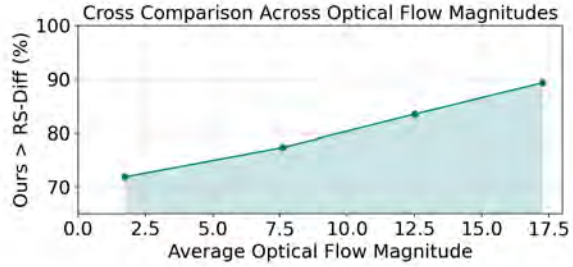

Fig. 7: Relative performance of our model vs. RS-Diffusion across optical flow magnitudes. Each bin represents the percentage of images where our model achieves higher PSNR.

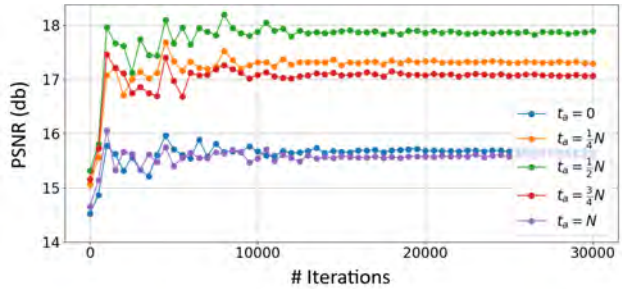

Fig. 8: Comparison of PSNR in the five models described in Section-5B. Performing a grid search on the possible anchor point location yielded different networks with varying accuracies. This experiments supported the use of the anchor point $\frac{N}{2}$ for the real-world scenes.

TABLE IV: Depth estimation performance in a legged robot scenario comparing the rolling shutter depth against our method. AbsRel and δ_1 are averaged over 1500 images collected across three outdoor locations.

Method	absRel ↓	δ_1 ↑
Raw RS Image	0.112	0.895
Our Algorithm	**0.096**	**0.903**

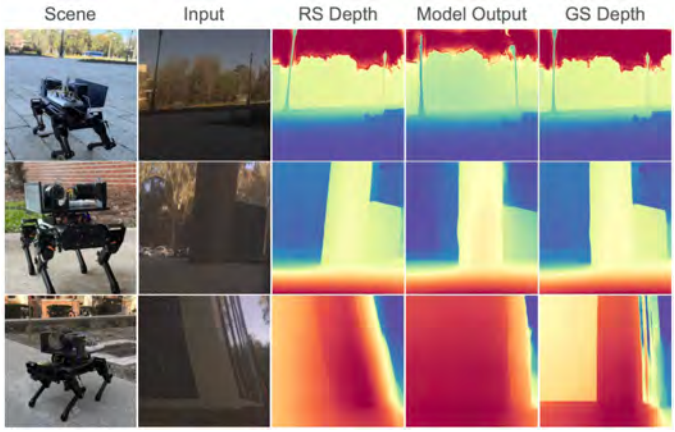

Fig. 9: Visualizations of our model on RS distortions in legged robots. The qualitative results demonstrate effective rectification of rolling shutter artifacts in legged robotic scenarios.

To improve deployment efficiency, the diffusion model was accelerated by reducing the number of denoising steps from 50 to 10, reducing the average inference time from 2850 ms to 349 ms.

As shown in Table IV, our method reduces AbsRel by 14.3% and improves δ_1 accuracy by 0.9% over raw RS input without additional fine-tuning beyond training on the REAL dataset. The qualitative results in Fig. 9 are consistent with these improvements, where rolling shutter–induced warping is visibly mitigated. While the proposed pipeline demonstrates promising low-latency performance, deployment at higher image resolutions or alongside additional perception tasks presents a natural computational trade-off due to increased diffusion and image-processing costs. Improving scalability through more efficient diffusion strategies, model compression, or shared-backbone architectures remains an important direction for future work.

VI. Conclusion

In this paper, we investigate monocular depth estimation under severe rolling shutter distortions induced by robot ego-motion and identify anchor-point selection as a key factor governing geometric consistency and depth accuracy. By explicitly modeling this temporal reference, we demonstrate that existing rolling shutter correction methods and vision foundation models can be substantially improved without architectural modifications or retraining. Using a new collocated RS-GS dataset and extensive synthetic and real-world experiments, we demonstrated robust depth estimation under large distortions

and aggressive motion, including deployment on a legged robot with an inference latency under 400 ms. These findings highlight the importance of temporal alignment in rolling shutter perception and suggest that anchor-based modeling can serve as a simple yet effective design principle for improving visual perception in mobile robotic systems.

ACKNOWLEDGMENTS

The authors thank Amazon Robotics for their support of this work with AGR DTD 10-06-2025. We also thank the National Science Foundation and the Office of Naval Research for their support. The authors were partially supported by NSF grants 1942444 and d2330416 as well as ONR grants N000142312429 and N000142312363.

REFERENCES

[1] Z. Yang, H. Li, M. Hong, B. Zeng, and S. Liu, "Single image rolling shutter removal with diffusion models," 2024. [Online]. Available: https://arxiv.org/abs/2407.02906

[2] C.-K. Liang, L.-W. Chang, and H. H. Chen, "Analysis and compensation of rolling shutter effect," *IEEE Transactions on Image Processing*, vol. 17, no. 8, pp. 1323–1330, 2008.

[3] P. Liu, Z. Cui, V. Larsson, and M. Pollefeys, "Deep shutter unrolling network," in *Proc. IEEE Conf. on Computer Vision and Pattern Recognition (CVPR)*, 2020.

[4] V. Rengarajan, A. N. Rajagopalan, and R. Aravind, "From bows to arrows: Rolling shutter rectification of urban scenes," in *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2016, pp. 2773–2781.

[5] P. Purkait, C. Zach, and A. Leonardis, "Rolling shutter correction in manhattan world," in *2017 IEEE International Conference on Computer Vision (ICCV)*, 2017, pp. 882–890.

[6] S. Su and W. Heidrich, "Rolling shutter motion deblurring," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2015.

[7] C.-K. Liang, L.-W. Chang, and H. H. Chen, "Analysis and compensation of rolling shutter effect," *IEEE Transactions on Image Processing*, vol. 17, no. 8, pp. 1323–1330, 2008.

[8] S. Baker, E. Bennett, S. B. Kang, and R. Szeliski, "Removing rolling shutter wobble," in *2010 IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, 2010, pp. 2392–2399.

[9] M. Grundmann, V. Kwatra, D. Castro, and I. Essa, "Calibration-free rolling shutter removal," in *2012 IEEE International Conference on Computational Photography (ICCP)*, 2012, pp. 1–8.

[10] J. Hedborg, M. Felsberg, P. E. Forssén, and E. Ringaby, "Rolling shutter bundle adjustment," in *CVPR Workshops*, 2012.

[11] M. Meingast, C. Geyer, and S. S. Sastry, "Geometric models of rolling-shutter cameras," in *OMNIVIS Workshop, ICCV*, 2005.

[12] J. Mo, M. J. Islam, and J. Sattar, "Imu-assisted learning of single-view rolling shutter correction," in *Conference on Robot Learning*. PMLR, 2022, pp. 861–870.

[13] P. Furgale, T. D. Barfoot, and G. Sibley, "Continuous-time batch estimation using temporal basis functions," in *ICRA*, 2012.

[14] C. Albl, Z. Kukelova, and T. Pajdla, "Rolling shutter relative pose," in *ICCV*, 2015.

[15] R. Saurer, C. Zhou, and M. Pollefeys, "A minimal solution to the rolling shutter pose estimation problem," in *IROS*, 2015.

[16] S. Leutenegger, S. Lynen, M. Bosse, R. Siegwart, and P. Furgale, "Okvis: Open keyframe-based visual-inertial slam," in *IROS*, 2015.

[17] A. Rosinol, M. Abate, Y. Chang, and L. Carlone, "Kimera: An open-source library for real-time metric-semantic localization and mapping," in *RSS*, 2020.

[18] D. Schubert, N. Demmel, L. von Stumberg, V. Usenko, and D. Cremers, "Rolling-shutter modelling for direct visual-inertial odometry," in *IROS*, 2019.

[19] Z. Zhong, Y. Zheng, and I. Sato, "Towards rolling shutter correction and deblurring in dynamic scenes," 2021. [Online]. Available: https://arxiv.org/abs/2104.01601

[20] A. Veeraraghavan, D. Reddy, and R. Raskar, "Coded strobing photography: Compressive sensing of high speed periodic videos," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 33, no. 4, pp. 671–686, 2010.

[21] Y. Hitomi, J. Gu, M. Gupta, T. Mitsunaga, and S. K. Nayar, "Video from a single coded exposure photograph using a learned over-complete dictionary," in *2011 International Conference on Computer Vision*. IEEE, 2011, pp. 287–294.

[22] J. Wang, M. Gupta, and A. C. Sankaranarayanan, "Lisens-a scalable architecture for video compressive sensing," in *2015 IEEE International Conference on Computational Photography (ICCP)*. IEEE, 2015, pp. 1–9.

[23] Z. Zhong, Y. Zheng, and I. Sato, "Towards rolling shutter correction and deblurring in dynamic scenes," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2021, pp. 9219–9228.

[24] O. Seiskari, J. Ylilammi, V. Kaatrasalo, P. Rantalankila, M. Turkulainen, J. Kannala, E. Rahtu, and A. Solin, "Gaussian splatting on the move: Blur and rolling shutter compensation for natural camera motion," in *European conference on computer vision*. Springer, 2024, pp. 160–177.

[25] G. Woo, A. Lippman, and R. Raskar, "Vrcodes: Unobtrusive and active visual codes for interaction by exploiting rolling shutter," in *2012 IEEE International Symposium on Mixed and Augmented Reality (ISMAR)*. IEEE, 2012, pp. 59–64.

[26] M. O'Toole, S. Achar, S. G. Narasimhan, and K. N. Kutulakos, "Homogeneous codes for energy-efficient illumination and imaging," *ACM Transactions on Graphics (ToG)*, vol. 34, no. 4, pp. 1–13, 2015.

[27] J. Wang, J. Bartels, W. Whittaker, A. C. Sankaranarayanan, and S. G. Narasimhan, "Programmable triangulation light curtains," in *Proceedings of the European Conference on Computer Vision (ECCV)*, 2018, pp. 19–34.

[28] M. Sheinin, D. Chan, M. O'Toole, and S. G. Narasimhan, "Dual-shutter optical vibration sensing," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2022, pp. 16 324–16 333.

[29] R. Ranftl, K. Lasinger, D. Hafner, K. Schindler, and V. Koltun, "Towards robust monocular depth estimation: Mixing datasets for zero-shot cross-dataset transfer," 2020. [Online]. Available: https://arxiv.org/abs/1907.01341

[30] R. Ranftl, A. Bochkovskiy, and V. Koltun, "Vision transformers for dense prediction," 2021. [Online]. Available: https://arxiv.org/abs/2103.13413

[31] S. Farooq Bhat, I. Alhashim, and P. Wonka, "Adabins: Depth estimation using adaptive bins," in *2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. IEEE, Jun. 2021, p. 4008–4017. [Online]. Available: http://dx.doi.org/10.1109/CVPR46437.2021.00400

[32] L. Yang, B. Kang, Z. Huang, Z. Zhao, X. Xu, J. Feng, and H. Zhao, "Depth anything v2," 2024. [Online]. Available: https://arxiv.org/abs/2406.09414

[33] B. Yu, J. Wu, and M. J. Islam, "Udepth: Fast monocular depth estimation for visually-guided underwater robots," in *IEEE International Conference on Robotics and Automation (ICRA)*. IEEE, 2023.

[34] B. Ke, A. Obukhov, S. Huang, N. Metzger, R. C. Daudt, and K. Schindler, "Repurposing diffusion-based image generators for monocular depth estimation," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2024.

[35] J. Gu, Y. Hitomi, T. Mitsunaga, and S. Nayar, "Coded rolling shutter photography: Flexible space-time sampling," in *2010 IEEE International Conference on Computational Photography (ICCP)*, 2010, pp. 1–8.

[36] O. Villar, *Learning Blender*. Addison-Wesley Professional, 2021.

[37] H. Sim, J. Oh, and M. Kim, "Xvfi: extreme video frame interpolation," 2021. [Online]. Available: https://arxiv.org/abs/2103.16206

[38] R. Rombach, A. Blattmann, D. Lorenz, P. Esser, and B. Ommer, "High-resolution image synthesis with latent diffusion models," 2021.

[39] S. Patni, A. Agarwal, and C. Arora, "Ecodepth: Effective conditioning of diffusion models for monocular depth estimation," 2024. [Online]. Available: https://arxiv.org/abs/2403.18807

[40] W. Yan, R. T. Tan, B. Zeng, and S. Liu, "Deep homography mixture for single image rolling shutter correction," in *2023 IEEE/CVF International Conference on Computer Vision (ICCV)*, 2023, pp. 9834–9843.