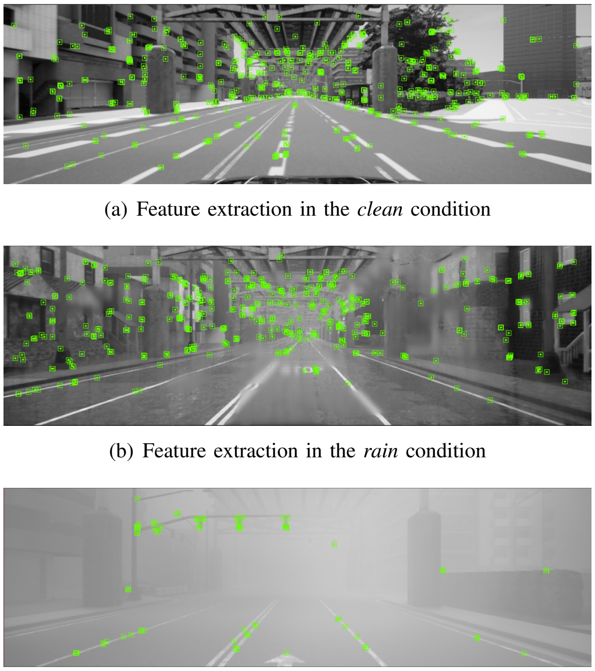
UMDrive: Towards Robust Outdoor SLAM under Any Condition

Ye Li, Caesar Guo, Yifeng Xu, Ruihan Xu
{yeyli,cesarguo,yifengxu,rhxu}@umich.edu

***Abstract*— Robust and reliable Simultaneous Localization and Mapping (SLAM) systems are crucial for safety-critical applications such as mobile robots and autonomous driving. The past few years have witnessed significant advancements in RGB-based SLAM. While most of the previous works have improved the performance of RGB-based SLAM through modern learning methods and novel optimization algorithms, notable degradation in SLAM performance under diverse conditions has been consistently observed. In addition, there exists a scarcity of datasets that cover diverse environments, which are crucial for assessing and enhancing the robustness of SLAM systems. To this end, we propose UMDrive, a comprehensive full-cycle pipeline encompassing data generation, SLAM optimization, and downstream evaluations. Our pipeline makes three appealing contributions. 1) To improve the robustness of SLAM in dynamic environments, we propose a novel generative in-painting method for dynamic objects. 2) We propose a novel assessment metric based on frame drop rates to identify the most reliable SLAM systems. 3) Centered around the theme of robust and reliable SLAM, we establish a systematic data generation platform in CALRA to synthesize RGB data under diverse conditions. Extensive experiments demonstrate that our framework can evaluate and optimize the performance of SLAM systems. Our code for the framework is publicly accessible at https://github.com/ywyeli/UMDrive.**



(a) Feature extraction in the *clean* condition

(b) Feature extraction in the *rain* condition

(c) Feature extraction in the *fog* condition

Fig. 1. Qualitative analysis of feature extraction process of SLAM in *clean*, *rain*, and *fog* conditions.

I. Introduction

The increasing deployment of mobile robots in environments characterized by complexity and dynamism [1], [2], referred to as *noisy worlds*, highlights the imperative need for enhanced robustness in robotic systems. This robustness is crucial for maintaining effective functionality amidst disruptive influences. As such, the evaluation of robustness in these scenarios has become a vital area of research [3]. At the heart of this research lies Simultaneous Localization and Mapping (SLAM) [4], [5], a cornerstone technology for robotic autonomy. The primary challenge is thus to develop a reliable and comprehensive framework for assessing the resilience of SLAM systems against diverse disturbances.

Recent progress in the evaluation of SLAM systems has been largely concentrated on the compilation of challenging datasets. These datasets subject SLAM systems to specific environmental conditions that deteriorate performance, thereby enriching our understanding of the operational challenges in real-world scenarios [6]–[9]. However, the practical challenges of data collection and labeling in natural settings restrict the size of these datasets, limiting a thorough and expansive evaluation. Furthermore, the complex interactions among environmental variables complicate the task of pinpointing the effects of specific disturbances on SLAM performance. To address these challenges, simulation-based benchmarks have gained traction as a valuable alternative [3], [10]–[13]. These simulations provide a platform for generating limitless 'battlefields' where data scalability and diversity enhance the 'survival testing' of SLAM models. They also allow for the creation of highly customizable and increasingly challenging scenarios, fostering ongoing improvements in SLAM robustness [12]. Although current simulators may lack complete real-world accuracy, advancements in visual content synthesis are progressively narrowing this fidelity gap [14], [15].

Despite the increasing availability of (nearly) photo-realistic 3D scene datasets and simulators [16]–[19] for SLAM evaluation, they often lack varied and controllable disturbances. As a result, these simulations typically represent idealized, perturbation-free environments, *i.e.*, *perfect world*, leaving the simulated perturbed environment, *i.e.*, *noisy world*, largely unexplored.

In this work, we propose an innovative simulation framework to emulate a broad spectrum of environmental challenges, such as rain and fog weather, thereby enhancing the assessment and development of SLAM systems. Incorporating dynamic simulations and generative inpainting techniques, the framework addresses issues arising from oc-

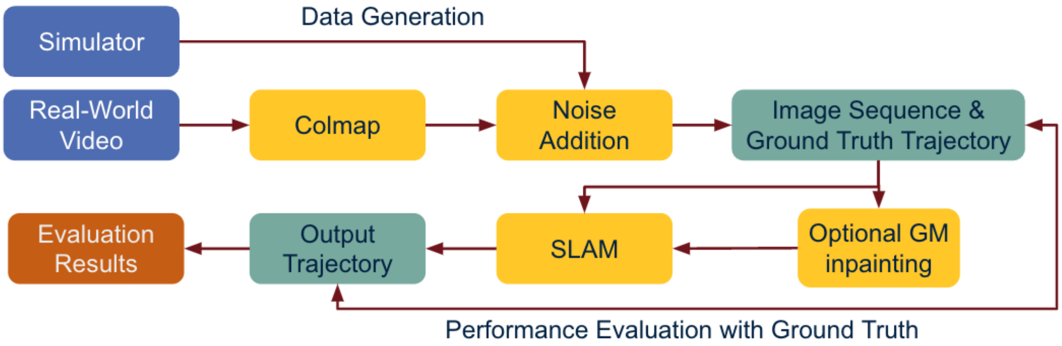


Fig. 2. Overview of our evaluation pipeline for Outdoor SLAM.

clusions and the presence of dynamic objects. Such simulations allow for testing SLAM performance where traditional static mapping approaches fail due to rapid environmental changes. To summarize, this work makes the following key contributions:

- We propose **UMDrive**, a comprehensive pipeline encompassing data generation, SLAM system optimization, and downstream evaluations.
- To improve the robustness of SLAM in dynamic environments, we propose a novel generative in-painting method for dynamic objects.
- We propose a novel assessment metric based on frame drop rates to identify the most reliable SLAM systems.
- Centered around robust and reliable SLAM, we establish a systematic data generation platform in CALRA to synthesize RGB data under diverse conditions.

II. Related Work

A. SLAM Methods

This section focuses on visual-related SLAM systems, though comprehensive reviews of SLAM systems are available in various sources, including [4], [20], [21]. Classical single-modal SLAM methods such as the visual-only ORB-SLAM [22] have achieved remarkable accuracy in clean benchmark environments like TUM VI [8] and Replica [16]. Addressing the challenges of real-world environments, research has explored techniques [23]–[26] that integrate multi-view sensors and fuse different data modalities, including visual-inertia and RGBD. The development of multi-agent SLAM models [27]–[29] has enabled collaborative localization and mapping among diverse robots, improving robustness in navigation. Furthermore, approaches leveraging neural networks and neural representations [12], [30]–[35] have enhanced generalization capabilities and 3D map reconstruction quality. However, the robustness of these models against sensor corruption and motion perturbations needs further exploration.

B. Robustness Benchmark

For mobile robots, perception modules must demonstrate resilience against shifts in natural distributions [36]. The benchmark ImageNet-C [37] has been pivotal in studying image corruption robustness by evaluating image classification methods against typical corruptions and perturbations. Subsequent studies have broadened this investigation to include other perception tasks such as object detection [38]–[40], segmentation [41]–[43], and embodied navigation [3], [44]. In SLAM, challenges include not only image-level corruptions from camera malfunctions but also dynamic variations in sensor corruption and sensor transformation deviations over time due to time-variant environmental effects and robots' diverse movements. This study introduces a perturbation taxonomy for RGBD SLAM in dynamic environments (e.g., varying illumination) and unstructured environments (e.g., uneven terrains causing vibrations for mobile robots).

C. Robustness Evaluation for SLAM

The robustness of SLAM systems is crucial for their reliable and accurate operation in dynamic and challenging real-world environments [4]. This robustness is vital for managing sensor faults and ensuring sustained performance. To facilitate robustness evaluation, several datasets have been collected in degraded environments with challenges like low illumination or motion blur [2], [8], [9], [45], [46]. SLAMBench [47] has compared the performance of various classical SLAM models across multiple challenging datasets, highlighting their vulnerabilities. Considering the difficulties and limitations of creating real-world datasets via robot platforms, Wang et al. [48] have employed photo-realistic simulation environments to develop TartanAir, a simulated SLAM benchmark for robustness evaluation. This study extends the evaluation scope to include the robustness of multi-modal SLAM models—covering both classical and neural methods—against a wider array of sensor corruptions and motion patterns (e.g., varying speed and motion-induced sensor trajectory deviations).

III. Methodology

A. Evaluation Pipeline

Our primary objective is to generate a challenging image sequence capable of capturing and replicating extreme real-world scenarios. Figure 2 illustrates the overall workflow of our methodology. We can incorporate simulator data generated by CARLA Simulator [11] as well as real-world data obtained from videos. Since arbitrary real-world video sequences lack ground truth camera pose, we employ Colmap for structure-from-motion [49] with loop closure to estimate a reliable camera trajectory, serving as the ground truth trajectory. Subsequently, various types of noise, such as rain, fog, and dynamic objects, are introduced into the image sequence to heighten the level of difficulty. Optionally, a generative inpainting method may be applied to enhance SLAM performance when dynamic objects are present; further details are discussed in Section III-B. Finally, we employ state-of-the-art algorithms, such as ORB-SLAM 2/3 and DeepVO, to estimate the trajectory and calculate errors using the Absolute Trajectory Error (ATE).

B. *Generative inpainting for dynamic objects*

As indicated by our experimental results in Section IV and supported by findings in [50], dynamic objects in uncertain environments pose significant challenges to existing SLAM models. Traditionally, SLAM algorithms rely on building maps of static environments with stationary objects to accurately localize the robot and update the map. However, the presence of dynamic objects disrupts this assumption, introducing considerable noise to the model. Specifically, observations of dynamic objects may yield a set of point clouds, features, etc., in frame $\mathbf{t}$, and a different set of point clouds and features in frame $\mathbf{t+1}$. But, because the object has changed its position in space, matching these features between frames becomes more challenging due to these inconsistent observations.

In [50], the authors propose a novel framework called DynaSLAM, which effectively addresses the challenge posed by dynamic objects. They employ the state-of-the-art Mask R-CNN model [51] to initially detect segments of dynamic objects such as humans and cars. Utilizing the binary masks generated by Mask R-CNN, the SLAM algorithms then disregard the features detected within regions masked out by dynamic objects. To enhance the algorithm's robustness, if the ground truth static background of the mask area is observed in previous frames, they replace the area covered by dynamic objects with the ground truth background projected to the current camera position. This substitution of dynamic objects with static ones ensures more consistent feature detection and enhances the overall robustness of the algorithm.

However, the framework has a limitation. If there are areas consistently not observed, such as in heavy traffic scenarios where the background behind a stream of moving cars is never observed, then a large number of areas are discarded by the algorithm and do not contribute to feature detection and matching. In an attempt to address this problem, we propose a simple extension involving generative inpainting. This extension allows the model to infer the background behind dynamic objects when ground truth information is insufficient to replace all dynamic objects with static backgrounds.

To achieve the above idea, we adopt a similar Mask R-CNN model provided by YOLOv8 [52] and a generative inpainting model provided by LaMa inpainting model [53] (Other choices of generative inpainting models are discussed in IV).

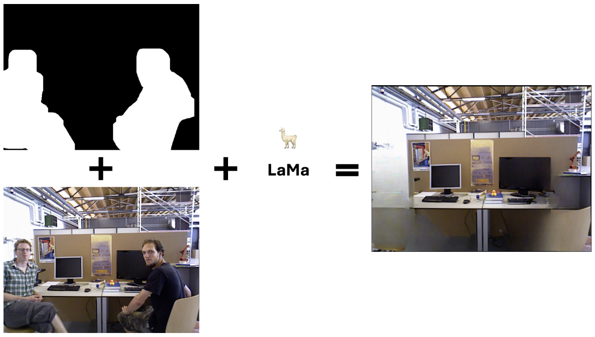


Fig. 3. Generative inpainting model LaMa is able to take in RGB image and the binary mask of target inpainting area to paint the inferred background

LaMa is robust and generalizable due to its ability to encode global and local contexts with high receptive fields. Specifically, they use Fast Fourier Convolution (FFC) [54] that uses Fast Fourier Transform (FFT) to retain this global context in the early stage. In FFC, we have

1) apply Real FFT2d to the input tensor (image + mask)

$$Real\ FFT2d : \mathbb{R}^{H\times W\times C} \to \mathbb{C}^{H\times \frac{W}{2}\times C}$$

2) combine the real and imaginary parts

$$ComplexToReal : \mathbb{C}^{H\times \frac{W}{2}\times C} \to \mathbb{R}^{H\times \frac{W}{2}\times 2C}$$

3) frequency domain convolution

$$ReLU \circ BN \circ Conv_{1\times 1} : \mathbb{R}^{H\times \frac{W}{2}\times C} \to \mathbb{R}^{H\times \frac{W}{2}\times 2C}$$

4) apply inverse transform to recover a spatial structure

$$RealToComplex : \mathbb{R}^{H\times \frac{W}{2}\times 2C} \to \mathbb{C}^{H\times \frac{W}{2}\times C}$$

$$Inverse\ Real\ FFT2d : \mathbb{C}^{H\times \frac{W}{2}\times C} \to \mathbb{R}^{H\times W\times 2C}$$

Finally, by combining the global contexts from FFC and local contexts from conventional convolutions, LaMa is able to create better generative backgrounds to the masked areas.

IV. Experiments

A. *Experiment Setup*

We tested all SLAM algorithms used in our work on the generated synthetic dataset. In addition, to bridge the gap of the effects of our improving methods between simulation and the real world, we test our generative inpainting method on the real-world TUM dataset [55].

B. *Experimental Results*

Supervised Learning-based Visual Odometry. In Figure 5(b), the DeepVO model's trajectory estimation under incremental fog densities is appraised, revealing its proficiency in clear weather and the ensuing decrement in accuracy with intensifying fog. Notably, the model's performance exhibits a marked fidelity to the ground truth in optimal visual conditions, signifying the robustness of its feature

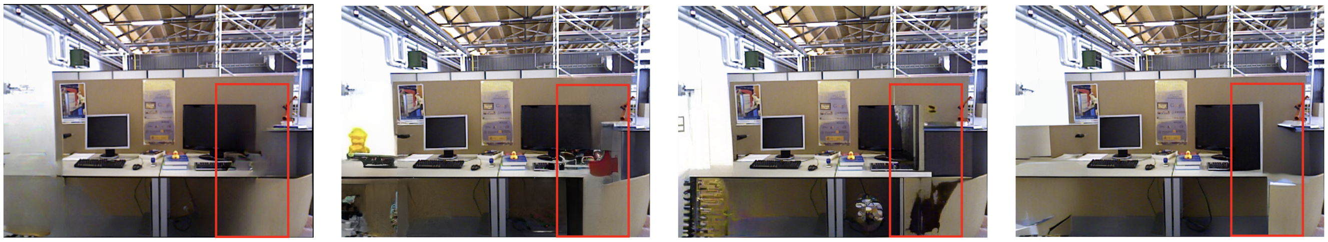
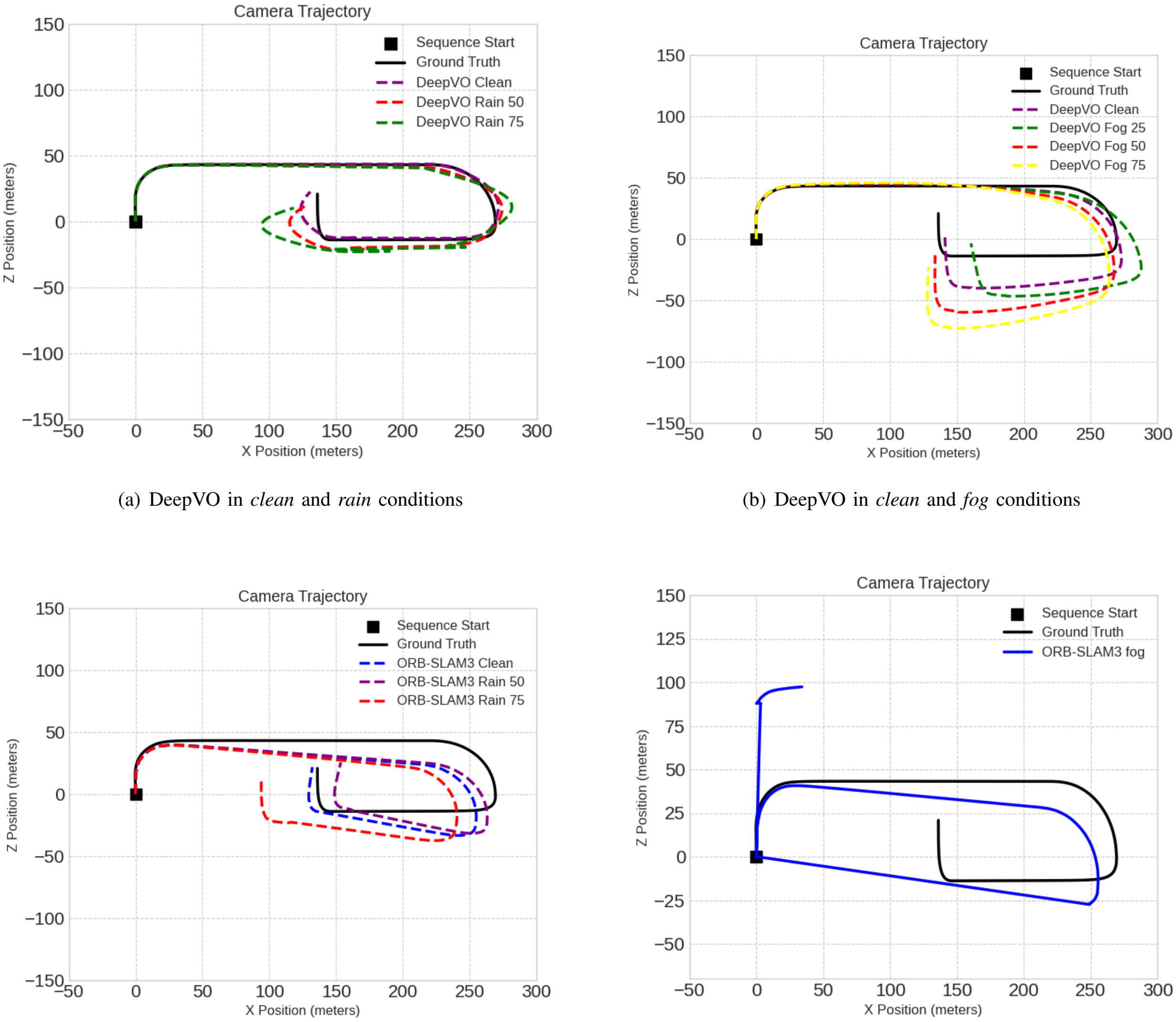

(a) (b) (c) (d)

Fig. 4. **Qualitative comparison** among inpainting models reveals that **LaMa** (top-left) outperforms all other models, including **MIGAN** (top-right), **MAT** (bottom-left), and **LDM** (bottom-right), by generating fewer artifacts. The ground truth image is identical to that shown in 3



(a) DeepVO in *clean* and *rain* conditions

(b) DeepVO in *clean* and *fog* conditions

(c) ORB-SLAM3 in *clean* and *rain* conditions

(d) ORB-SLAM3 in *clean* and *fog* conditions

Fig. 5. **Comparative analysis** of selected visual odometry in this work for *clean*, *rain*, and *fog* conditions.

extraction capabilities. However, the augmenting fog conditions precipitate a decline in trajectory precision, with the most pronounced divergence observed during intricate navigational maneuvers, indicative of the perturbed feature extraction by the convolutional neural network.

Correspondingly, Figure 5(a) demonstrates the DeepVO model's performance on simulated rainfall. It becomes evident that as precipitation intensifies, the model's trajectory estimates progressively deviate from the ground truth. This deviation, attributed to the rain-induced visual disruptions, underscores the susceptibility of the model's pre-trained parameters to the stochastic visual patterns presented by rainfall.

In essence, while DeepVO is adept at navigating clear environmental conditions, the vicissitudes presented by fog and rain elucidate the need for an enriched training paradigm. The incorporation of a panoply of adverse weather scenarios is imperative for the advancement of the model's generalization capabilities. Such enhancement is essential for the realization of robust autonomous navigation systems capable of operating reliably in the multifaceted and unpredictable dynamics of real-world settings.

Featured-based Visual Odometry. Figures 5(c) and 5(d) demonstrate the trajectory estimation challenges encountered by ORB-SLAM3, a traditional feature-based SLAM method, under adverse weather conditions. Notably, ORB-SLAM3's performance is significantly impeded by foggy scenarios, as visualized in Figure 5(d), where the deviation from the ground truth trajectory is pronounced. This performance impediment can be attributed to the fog's obscuration of critical features within the environment, which are essential for the SLAM algorithm to maintain spatial awareness and localization accuracy.

The extracted features, integral to ORB-SLAM3's operation as illustrated in Figures 1(a), 1(b), and 1(c), become increasingly sparse and unreliable in fog, leading to trajectory estimations that substantially diverge from the ground truth. Moreover, the method's susceptibility to rapid turns and other swift motions exacerbates this divergence, as these dynamic conditions further challenge the feature tracking capabilities of the algorithm.

Comparison. For quantitative analysis of the methods in this experiment, the error metric we adopt is the absolute trajectory error (ATE) proposed by Sturm *et al.* [56]. Using the sequences of estimated trajectory $P_{1:n}$ and ground truth trajectory $Q_{1:n}$, the ATE at a certain time step i can be computed as:

$$E_i = Q_i^{-1} S P_i$$

where S is a rigid body transformation matrix that maps the estimated trajectory onto the ground truth trajectory. The rooted mean squared error (RMSE) over all time indices is computed as:

$$RMSE(E_{i:n}) = (\frac{1}{n}\sum_{i=1}^{n} ||trans(E_i)||^2)^{1/2}$$

The ATE for ORB-SLAM3 and deepVO are shown in table I. Notice how DeepVO achieved a much lower error compared to ORB-SLAM3 in the fog environment. It is due to the fact that well-trained DeepVO networks tend to accumulate larger errors with noisier environment input, but does not have the problem of totally losing track, which can be a problem of feature-based visual odometries including ORB-SLAM. It is also evident that ORB-SLAM3 largely outperformed DeepVO in rain. It implies the ability of ORB-SLAM to restore orientation based on scale-invariant features from different levels in the pyramid (and thus different distances to the camera). On the other hand, rain can be attached to the camera lens in our dataset. The random large areas of blurred pixels could have an impact on the CNN in DeepVO.

TABLE I

ABSOLUTE TRAJECTORY ERROR FOR DEEPVO AND ORBSLAM3

Conditions	**ORB-SLAM3** (m) ↓	**DeepVO** (m) ↓
Clean	13.989	14.693
Rain	16.816	37.249
Fog	104.852	26.327
Dynamic	20.066	–

TABLE II

ABSOLUTE TRAJECTORY ERROR FOR DIFFERENT FRAME DROP RATES

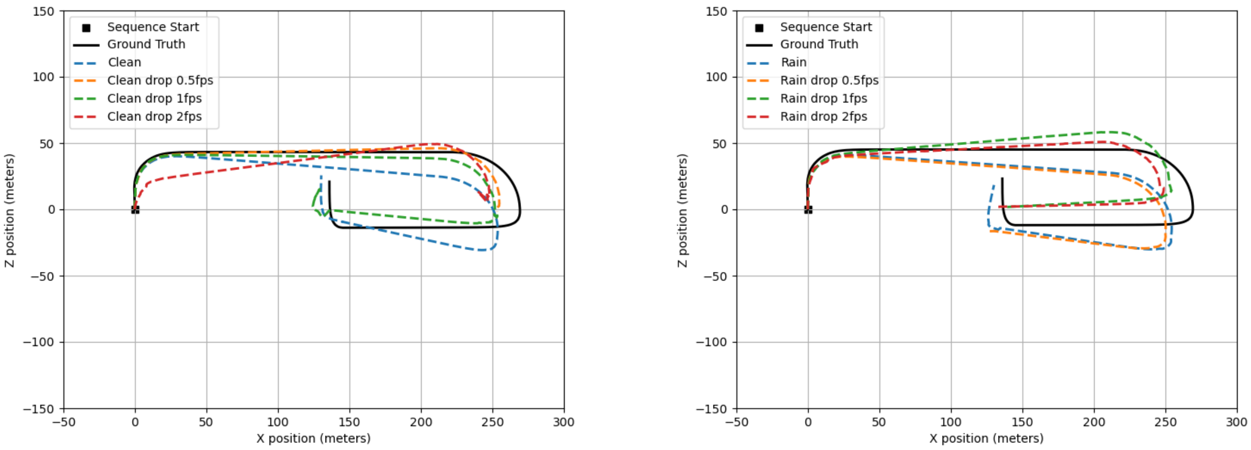
Conditions	0.5fps ↓	1fps ↓	2fps ↓
Clean	40.403	26.075	30.156
Rain	25.208	26.159	50.186

A comparative analysis reveals that both DeepVO and ORB-SLAM3 maintain high fidelity to the ground truth in clear conditions, implying robustness in normal environmental settings. Under rainfall, both methods suffer performance losses, yet the extent and nature of the deviations differ. DeepVO shows a tendency for smoother trajectory deviations, possibly due to its reliance on learned features which may generalize better in varied conditions. In contrast, ORB-SLAM3 exhibits sharper departures from the ground truth, which could be due to the loss of reliable feature points in the rain, a challenge for feature-based methods that rely heavily on the distinctiveness and matchability of features across frames.

In addition, in clear conditions and normal operational scenarios, as reflected in Figure 5(c), ORB-SLAM3 tends to outperform learning-based methods like DeepVO, displaying a trajectory closely aligned with the ground truth. This superiority in nominal cases suggests that in environments with abundant and stable visual features, ORB-SLAM3 is well-equipped to provide accurate and reliable pose estimation. It underscores the potential for feature-based methods in applications where environmental conditions can be controlled or predicted with a high degree of certainty.

Generative inpainting model. we conduct comparative analyses of the state-of-the-art generative inpainting models. We include LaMa, MAT [57], MIGAN [58], and LDM [59] in our comparison. These models are showcased in Fig. 4, where each is assessed for its ability to seamlessly inpainting images. This comparison helps in identifying which model performs best under various conditions and contributes to advancements in the field of Visual Odometry.

Influence of Frame Drops. Frame drops are also a problem for visual SLAM systems that have a large impact on robustness and accuracy. It can not only appear due to hardware or software limitations and issues, but also arise from motion blur, occlusions, and varying lighting conditions. Those environmental factors are present in different weathers and can cause data to be unavailable for the system. We run an extra experiment to test the influence of frame



(a) Frame drops in *clean* condition

(b) Frame drops in *rain* condition

Fig. 6. Comparison of different frame drop rates on *Clean* and *Rain* datasets

drop on the traditional feature-based ORB-SLAM2. Two conditions where ORB-SLAM performs the best – clean and rain – were selected as experiment groups. We tried three frame drop rates: 0.5fps, 1fps, 2fps, equivalent to dropping one frame in every 20, 10, or 5 frames, with our camera frequency at 10 fps. A sliding time window of those lengths was applied to the collected data, and a random frame was dropped within every window. ORB-SLAM2 was then run on the processed data respectively.

Figure 6 shows estimated and ground truth trajectories in the two scenarios. Absolute trajectory errors are computed and listed in Table II. In general, as frame drops become more frequent, the quality and accuracy of trajectory estimation decrease. However, it is not always the case, as is evident in the clean condition, where frame drop rates of both 2fps and 0.5fps cause the system to lose track at a corner while 1fps gives the best tracking result. During running, we observed a significant drop in the number of matched features when the system processed frames adjacent to the dropped frames. Given also the fact that loss of track happens during turns at sharp corners, we conclude ORB-SLAM is not robust when handling large camera motions or frequent drop of frames. They cause a large difference between consecutive frames, and make it hard for feature tracking, especially matching feature points between frames. With the feature-learning abilities and temporal information, learning-based methods are likely more robust in the presence of dropped frames and large motions.

V. Conclusion

This work presents a new framework designed to assess the robustness of visual odometry systems using diverse environmental simulations. It addresses current challenges in SLAM technologies by providing a customizable way to generate virtual datasets, which include dynamic obstacles and varying weather conditions for comprehensive system evaluations. Key contributions include a rigorous testing environment and the integration of advanced models like DeepVO and ORB-SLAM2/3. The research highlights both the potential and limitations of learning-based methods in adapting to environmental changes and underscores the need for ongoing research to enhance the resilience and reliability of SLAM systems for autonomous navigation in complex real-world scenarios.

References

[1] E. Kaufmann, L. Bauersfeld, A. Loquercio, M. Müller, V. Koltun, and D. Scaramuzza, “Champion-level drone racing using deep reinforcement learning,” *Nature*, vol. 620, no. 7976, pp. 982–987, 2023.

[2] K. Ebadi, L. Bernreiter, H. Biggie, G. Catt, Y. Chang, A. Chatterjee, C. E. Denniston, S.-P. Deschênes, K. Harlow, S. Khattak, L. Nogueira, M. Palieri, P. Petráček, M. Petrlík, A. Reinke, V. Krátký, S. Zhao, A.-a. Agha-mohammadi, K. Alexis, C. Heckman, K. Khosoussi, N. Kottege, B. Morrell, M. Hutter, F. Pauling, F. Pomerleau, M. Saska, S. Scherer, R. Siegwart, J. L. Williams, and L. Carlone, “Present and future of slam in extreme environments: The darpa subt challenge,” *IEEE Transactions on Robotics*, pp. 1–20, 2023.

[3] P. Chattopadhyay, J. Hoffman, R. Mottaghi, and A. Kembhavi, “Robustnav: Towards benchmarking robustness in embodied navigation,” in *IEEE/CVF International Conference on Computer Vision*, 2021, pp. 15 691–15 700.

[4] C. Cadena, L. Carlone, H. Carrillo, Y. Latif, D. Scaramuzza, J. Neira, I. Reid, and J. J. Leonard, “Past, present, and future of simultaneous localization and mapping: Toward the robust-perception age,” *IEEE Transactions on Robotics*, vol. 32, no. 6, pp. 1309–1332, 2016.

[5] C. Chen, B. Wang, C. X. Lu, N. Trigoni, and A. Markham, “Deep learning for visual localization and mapping: A survey,” *IEEE Transactions on Neural Networks and Learning Systems*, pp. 1–21, 2023.

[6] W. Maddern, G. Pascoe, C. Linegar, and P. Newman, “1 year, 1000 km: The oxford robotcar dataset,” *The International Journal of Robotics Research*, vol. 36, no. 1, pp. 3–15, 2017.

[7] Y. Choi, N. Kim, S. Hwang, K. Park, J. S. Yoon, K. An, and I. S. Kweon, “Kaist multi-spectral day/night data set for autonomous and assisted driving,” *IEEE Transactions on Intelligent Transportation Systems*, vol. 19, no. 3, pp. 934–948, 2018.

[8] D. Schubert, T. Goll, N. Demmel, V. Usenko, J. Stückler, and D. Cremers, “The tum vi benchmark for evaluating visual-inertial odometry,” in *2018 IEEE/RSJ International Conference on Intelligent Robots and Systems*, 2018, pp. 1680–1687.

[9] S. Zhao, D. Singh, H. Sun, R. Jiang, Y. Gao, T. Wu, J. Karhade, C. Whittaker, I. Higgins, J. Xu *et al.*, "Subt-mrs: A subterranean, multi-robot, multi-spectral and multi-degraded dataset for robust slam," *arXiv preprint arXiv:2307.07607*, 2023.

[10] T. Sayre-McCord, W. Guerra, A. Antonini, J. Arneberg, A. Brown, G. Cavalheiro, Y. Fang, A. Gorodetsky, D. McCoy, S. Quilter, F. Riether, E. Tal, Y. Terzioglu, L. Carlone, and S. Karaman, "Visual-inertial navigation algorithm development using photorealistic camera simulation in the loop," in *2018 IEEE International Conference on Robotics and Automation (ICRA)*, 2018, pp. 2566–2573.

[11] A. Dosovitskiy, G. Ros, F. Codevilla, A. Lopez, and V. Koltun, "Carla: An open urban driving simulator," in *Conference on robot learning*. PMLR, 2017, pp. 1–16.

[12] W. Wang, Y. Hu, and S. Scherer, "Tartanvo: A generalizable learning-based vo," in *Conference on Robot Learning*. PMLR, 2021, pp. 1761–1772.

[13] M. Johnson-Roberson, C. Barto, R. Mehta, S. N. Sridhar, K. Rosaen, and R. Vasudevan, "Driving in the matrix: Can virtual worlds replace human-generated annotations for real world tasks?" in *2017 IEEE International Conference on Robotics and Automation (ICRA)*, 2017, pp. 746–753.

[14] R. Rombach, A. Blattmann, D. Lorenz, P. Esser, and B. Ommer, "High-resolution image synthesis with latent diffusion models," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2022, pp. 10 684–10 695.

[15] A. Raistrick, L. Lipson, Z. Ma, L. Mei, M. Wang, Y. Zuo, K. Kayan, H. Wen, B. Han, Y. Wang *et al.*, "Infinite photorealistic worlds using procedural generation," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2023, pp. 12 630–12 641.

[16] J. Straub, T. Whelan, L. Ma, Y. Chen, E. Wijmans, S. Green, J. J. Engel, R. Mur-Artal, C. Ren, S. Verma *et al.*, "The replica dataset: A digital replica of indoor spaces," *arXiv preprint arXiv:1906.05797*, 2019.

[17] A. Dai, A. X. Chang, M. Savva, M. Halber, T. Funkhouser, and M. Nießner, "Scannet: Richly-annotated 3d reconstructions of indoor scenes," in *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2017, pp. 5828–5839.

[18] M. Deitke, E. VanderBilt, A. Herrasti, L. Weihs, K. Ehsani, J. Salvador, W. Han, E. Kolve, A. Kembhavi, and R. Mottaghi, "Procthor: Large-scale embodied ai using procedural generation," *Advances in Neural Information Processing Systems*, vol. 35, pp. 5982–5994, 2022.

[19] J. Zheng, J. Zhang, J. Li, R. Tang, S. Gao, and Z. Zhou, "Structured3d: A large photo-realistic dataset for structured 3d modeling," in *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part IX 16*. Springer, 2020, pp. 519–535.

[20] A. Macario Barros, M. Michel, Y. Moline, G. Corre, and F. Carrel, "A comprehensive survey of visual slam algorithms," *Robotics*, vol. 11, no. 1, p. 24, 2022.

[21] I. A. Kazerouni, L. Fitzgerald, G. Dooly, and D. Toal, "A survey of state-of-the-art on visual slam," *Expert Systems with Applications*, vol. 205, p. 117734, 2022.

[22] R. Mur-Artal, J. M. M. Montiel, and J. D. Tardos, "Orb-slam: a versatile and accurate monocular slam system," *IEEE transactions on robotics*, vol. 31, no. 5, pp. 1147–1163, 2015.

[23] R. Mur-Artal and J. D. Tardós, "Orb-slam2: An open-source slam system for monocular, stereo, and rgb-d cameras," *IEEE Transactions on Robotics*, vol. 33, no. 5, pp. 1255–1262, 2017.

[24] C. Campos, R. Elvira, J. J. G. Rodríguez, J. M. M. Montiel, and J. D. Tardós, "Orb-slam3: An accurate open-source library for visual, visual–inertial, and multimap slam," *IEEE Transactions on Robotics*, vol. 37, no. 6, pp. 1874–1890, 2021.

[25] N. Khedekar, M. Kulkarni, and K. Alexis, "Mimosa: A multi-modal slam framework for resilient autonomy against sensor degradation," in *2022 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, 2022, pp. 7153–7159.

[26] A. Rosinol, A. Violette, M. Abate, N. Hughes, Y. Chang, J. Shi, A. Gupta, and L. Carlone, "Kimera: From slam to spatial perception with 3d dynamic scene graphs," *The International Journal of Robotics Research*, vol. 40, no. 12-14, pp. 1510–1546, 2021.

[27] Y. Tian, Y. Chang, F. Herrera Arias, C. Nieto-Granda, J. P. How, and L. Carlone, "Kimera-multi: Robust, distributed, dense metric-semantic slam for multi-robot systems," *IEEE Transactions on Robotics*, vol. 38, no. 4, pp. 2022–2038, 2022.

[28] M. Patel, M. Karrer, P. Bänninger, and M. Chli, "Covins-g: A generic back-end for collaborative visual-inertial slam," in *2023 IEEE International Conference on Robotics and Automation (ICRA)*, 2023, pp. 2076–2082.

[29] H. Xu, P. Liu, X. Chen, and S. Shen, "D2 slam: Decentralized and distributed collaborative visual-inertial slam system for aerial swarm," *arXiv preprint arXiv:2211.01538*, 2022.

[30] Z. Teed and J. Deng, "Droid-slam: Deep visual slam for monocular, stereo, and rgb-d cameras," *Advances in neural information processing systems*, vol. 34, pp. 16 558–16 569, 2021.

[31] H. Wang, J. Wang, and L. Agapito, "Co-slam: Joint coordinate and sparse parametric encodings for neural real-time slam," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2023, pp. 13 293–13 302.

[32] E. Sucar, S. Liu, J. Ortiz, and A. J. Davison, "imap: Implicit mapping and positioning in real-time," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2021, pp. 6229–6238.

[33] Z. Zhu, S. Peng, V. Larsson, W. Xu, H. Bao, Z. Cui, M. R. Oswald, and M. Pollefeys, "Nice-slam: Neural implicit scalable encoding for slam," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2022, pp. 12 786–12 796.

[34] Y. Zhang, F. Tosi, S. Mattoccia, and M. Poggi, "Go-slam: Global optimization for consistent 3d instant reconstruction," in *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, October 2023.

[35] A. Rosinol, J. J. Leonard, and L. Carlone, "Nerf-slam: Real-time dense monocular slam with neural radiance fields," *arXiv preprint arXiv:2210.13641*, 2022.

[36] T. Zhang, W. Zhang, and M. M. Gupta, "Resilient robots: Concept, review, and future directions," *Robotics*, vol. 6, no. 4, p. 22, 2017.

[37] D. Hendrycks and T. Dietterich, "Benchmarking neural network robustness to common corruptions and perturbations," *Proceedings of the International Conference on Learning Representations*, 2019.

[38] C. Michaelis, B. Mitzkus, R. Geirhos, E. Rusak, O. Bringmann, A. S. Ecker, M. Bethge, and W. Brendel, "Benchmarking robustness in object detection: Autonomous driving when winter is coming," *arXiv preprint arXiv:1907.07484*, 2019.

[39] A. Carlson, K. A. Skinner, R. Vasudevan, and M. Johnson-Roberson, "Modeling camera effects to improve visual learning from synthetic data," in *Proceedings of the European Conference on Computer Vision (ECCV) Workshops*, September 2018.

[40] L. Kong, Y. Liu, X. Li, R. Chen, W. Zhang, J. Ren, L. Pan, K. Chen, and Z. Liu, "Robo3d: Towards robust and reliable 3d perception against corruptions," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2023, pp. 19 994–20 006.

[41] C. Kamann and C. Rother, "Benchmarking the robustness of semantic segmentation models," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2020, pp. 8828–8838.

[42] X. Xu, J. Wang, X. Ming, and Y. Lu, "Towards robust video object segmentation with adaptive object calibration," in *Proceedings of the 30th ACM International Conference on Multimedia*, 2022, pp. 2709–2718.

[43] X. Li, J. Wang, X. Xu, X. Li, B. Raj, and Y. Lu, "Robust referring video object segmentation with cyclic structural consensus," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2023, pp. 22 236–22 245.

[44] N. Yokoyama, Q. Luo, D. Batra, and S. Ha, "Benchmarking augmentation methods for learning robust navigation agents: the winning entry of the 2021 igibson challenge," in *2022 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*. IEEE, 2022, pp. 1748–1755.

[45] M. Helmberger, K. Morin, B. Berner, N. Kumar, G. Cioffi, and D. Scaramuzza, "The hilti slam challenge dataset," *IEEE Robotics and Automation Letters*, vol. 7, no. 3, pp. 7518–7525, 2022.

[46] Y. Tian, Y. Chang, L. Quang, A. Schang, C. Nieto-Granda, J. P. How, and L. Carlone, "Resilient and distributed multi-robot visual slam: Datasets, experiments, and lessons learned," *arXiv preprint arXiv:2304.04362*, 2023.

[47] M. Bujanca, X. Shi, M. Spear, P. Zhao, B. Lennox, and M. Luján, "Robust slam systems: Are we there yet?" in *2021 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*. IEEE, 2021, pp. 5320–5327.

[48] W. Wang, D. Zhu, X. Wang, Y. Hu, Y. Qiu, C. Wang, Y. Hu, A. Kapoor, and S. Scherer, "Tartanair: A dataset to push the limits of visual slam,"

in *2020 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*. IEEE, 2020, pp. 4909–4916.

[49] S. Agarwal, N. Snavely, I. Simon, S. M. Seitz, and R. Szeliski, “Building rome in a day,” in *2009 IEEE 12th International Conference on Computer Vision*, 2009, pp. 72–79.

[50] B. Bescos, J. M. Facil, J. Civera, and J. Neira, “Dynaslam: Tracking, mapping, and inpainting in dynamic scenes,” *IEEE Robotics and Automation Letters*, vol. 3, no. 4, p. 4076–4083, Oct. 2018.

[51] K. He, G. Gkioxari, P. Dollár, and R. Girshick, “Mask r-cnn,” 2018.

[52] J. Redmon, S. Divvala, R. Girshick, and A. Farhadi, “You only look once: Unified, real-time object detection,” 2016.

[53] R. Suvorov, E. Logacheva, A. Mashikhin, A. Remizova, A. Ashukha, A. Silvestrov, N. Kong, H. Goka, K. Park, and V. Lempitsky, “Resolution-robust large mask inpainting with fourier convolutions,” 2021.

[54] L. Chi, B. Jiang, and Y. Mu, “Fast fourier convolution,” in *Advances in Neural Information Processing Systems*, H. Larochelle, M. Ranzato, R. Hadsell, M. Balcan, and H. Lin, Eds., vol. 33. Curran Associates, Inc., 2020, pp. 4479–4488.

[55] D. Schubert, T. Goll, N. Demmel, V. Usenko, J. Stuckler, and D. Cremers, “The tum vi benchmark for evaluating visual-inertial odometry,” in *2018 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*. IEEE, Oct. 2018.

[56] J. Sturm, N. Engelhard, F. Endres, W. Burgard, and D. Cremers, “A benchmark for the evaluation of rgb-d slam systems,” in *2012 IEEE/RSJ International Conference on Intelligent Robots and Systems*, 2012, pp. 573–580.

[57] W. Li, Z. Lin, K. Zhou, L. Qi, Y. Wang, and J. Jia, “Mat: Mask-aware transformer for large hole image inpainting,” 2022.

[58] A. Sargsyan, S. Navasardyan, X. Xu, and H. Shi, “Mi-gan: A simple baseline for image inpainting on mobile devices,” in *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, October 2023, pp. 7335–7345.

[59] R. Rombach, A. Blattmann, D. Lorenz, P. Esser, and B. Ommer, “High-resolution image synthesis with latent diffusion models,” 2022.