Robustness Enhancement for Deep learning based Visual Odometry

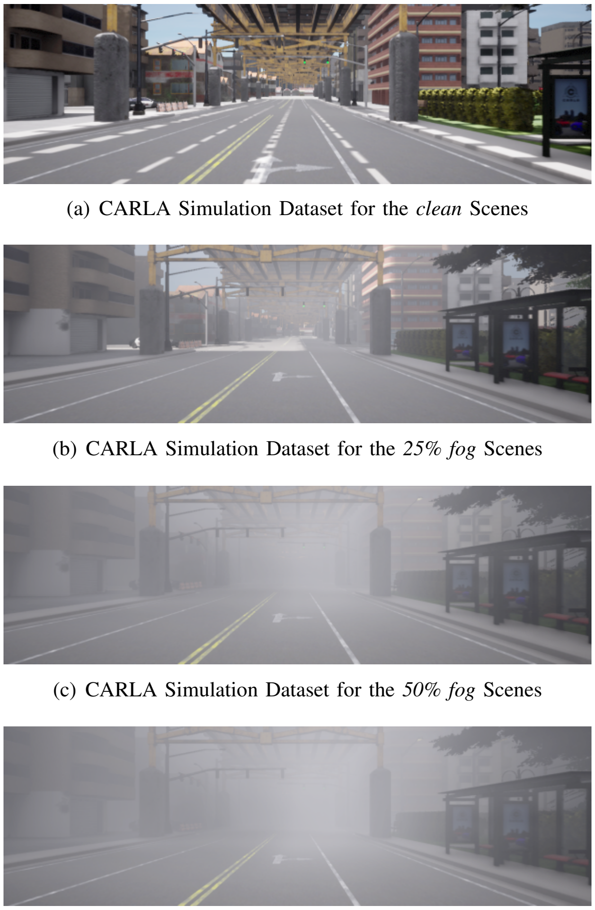
Yifeng Xu, Raghav Mishra, Diwen Zhu
{yifengxu,imraghav,diwenzhu}@umich.edu

***Abstract*— Robust and dependable Simultaneous Localization and Mapping (SLAM) systems are essential for critical applications in fields such as mobile robotics and autonomous vehicles. Recent years have seen considerable progress in SLAM technology that utilizes RGB data. While many studies have enhanced the capabilities of RGB-based SLAM through advanced learning techniques and innovative optimization methods, a consistent decline in SLAM efficacy under varied conditions remains a challenge. Moreover, there is a notable lack of comprehensive datasets that encompass a range of environments, which are vital for evaluating and improving the robustness of SLAM systems. Our extensive experimental work has shown that our framework effectively assesses and refines the performance of Deep-learning based Visual Odometry (VO) systems. In this project, the contribution can be concluded as: (1) we generate a KITTI-formatted virtual dataset for data augmentation towards robust learning for VO system(dataset can be accessed at: https://drive.google.com/drive/folders/1GWc_3JzKqdBCun9hw0ZAAk10-ay9vUNK?usp=drive_link). (2) we propose an evaluation framework that systematically examines the robustness of the DeepVO performance. (3) we implement a novel inpainting modules to address the dynamic objects issues for a leaning-based VO system. Our code for the framework is publicly accessible at https://github.com/IlikeSukiyaki/Enhanced-Leaning-based-Visual-Odometry.git.**

I. Introduction

The growing use of mobile robots in dynamic and complex settings, referred to as ”noisy worlds,” underscores the critical need for improved robustness in robotic technologies to ensure consistent functionality amidst disruptions. This has made the assessment of robustness an increasingly important focus in robotics research [1], [2]. Central to this area of study is Simultaneous Localization and Mapping (SLAM), which is fundamental to robotic autonomy [3], [4]. The main challenge involves developing a robust and detailed framework for evaluating the resilience of SLAM systems to various disturbances.

Recent advances in this field have primarily focused on assembling demanding datasets that expose SLAM systems to adverse environmental conditions, enhancing our understanding of their limitations in practical environments [5]–[8]. However, the complexities of data collection and annotation in natural environments limit the size and scope of these datasets, constraining a comprehensive evaluation. Additionally, the intricate interplay of environmental factors makes it difficult to isolate the impact of specific disturbances on SLAM performance. In response, simulation-based benchmarks have emerged as an effective alternative [9]–[13]. These simulations offer an environment for endless ’battlefields,’ where the scalability and diversity of data improve the ’survival testing’ for SLAM models. They also facilitate the creation of tailored and increasingly challenging scenarios, contributing to continuous enhancements in SLAM robustness [12]. While current simulation technologies may not perfectly replicate real-world conditions, ongoing improvements in visual content synthesis are gradually closing this fidelity gap [14], [15].


(a) CARLA Simulation Dataset for the *clean* Scenes

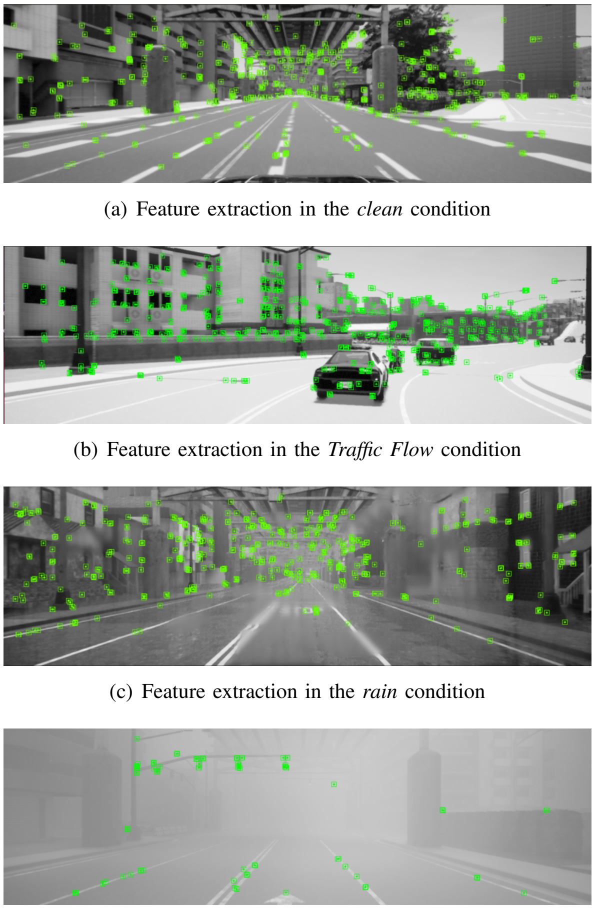
(b) CARLA Simulation Dataset for the *25% fog* Scenes

(c) CARLA Simulation Dataset for the *50% fog* Scenes

(d) CARLA Simulation Dataset for the *75% fog* Scenes

Fig. 1. CARLA Dataset Synthesis for Various Weather Conditions

In our research, we introduce a novel simulation framework designed to replicate a wide array of environmental challenges, including adverse weather conditions like rain and fog, thus advancing the evaluation and development of SLAM systems. This framework utilizes dynamic simulations combined with generative inpainting techniques to tackle challenges associated with occlusions and the presence of dynamic objects. Such simulations are crucial for evaluat-



(a) Feature extraction in the *clean* condition

(b) Feature extraction in the *Traffic Flow* condition

(c) Feature extraction in the *rain* condition

(d) Feature extraction in the *fog* condition

Fig. 2. Feature Extraction from CARLA Synthetic Dataset

ing SLAM performance in scenarios where traditional static mapping approaches are ineffective due to rapid environmental shifts. In summary, this work delivers the following significant contributions:

- Generate Kitti-formatted virtual dataset for data augmentation
- Propose Evaluation Framework to examine robustness of DeepVO Performance
- Implement Novel impainting module to address dynamic objects issues for learning based VO systems.

By sharing the code to research community we hope to enable more research along these lines. Finally using our evaluation framework can be used with nuscenes data by converting to SemanticKitti[]

II. Related Work

A. *Supervised Learning of Visual Odometry*

Supervised learning approaches in visual odometry (VO) focus on training a deep neural network (DNN) using labeled datasets to map consecutive image pairs to their respective motion transformations. This method contrasts with traditional VO techniques, which rely on analyzing the geometric properties of images. Typically, the DNN receives two sequential images as input and outputs the calculated translation and rotation between these image frames [17].

Early contributions in this field include the work by Konda et al. [18], which treats visual odometry (VO) as a classification task. In their model, a convolutional neural network (ConvNet) is employed to deduce discrete shifts in direction and velocity based on the input imagery. However, this approach is constrained by its reliance on discrete motion predictions and its inability to accurately delineate the complete camera trajectory. Addressing these limitations, Costante et al. [19] advanced this field by integrating dense optical flow for feature extraction and a ConvNet for assessing the motion between successive frames. Their methodology offers enhanced accuracy and smoother camera trajectories compared to the earlier work by Konda et al. [18]. Despite these advancements, both techniques stop short of providing a fully integrated end-to-end learning model from images to motion predictions and continue to underperform compared to traditional VO algorithms, such as VISO2 [20], especially in terms of accuracy and reliability. Moreover, both methods lack in exploiting the comprehensive geometric data inherent in the images, which is vital for precise motion detection. The training and evaluation datasets used also lack diversity, which may compromise their effectiveness in varied settings.

DeepVO [16] facilitates end-to-end learning in visual odometry by employing a synergistic approach that integrates a convolutional neural network (ConvNet) with a recurrent neural network (RNN). As illustrated in Figure 4, this conventional RNN+ConvNet based VO model captures visual features from image pairs using a ConvNet and leverages RNNs to manage the temporal association of these features. The architecture of its ConvNet encoder draws on the [21] design, which is optimized for extracting visual features critical for optical flow and autonomous motion determination. The recurrent component of the model consolidates historical data within its hidden states, allowing the system to generate outputs that reflect both prior knowledge and recent observations from ConvNet features. DeepVO is trained using datasets that include precisely verified poses to serve as training labels.

B. *Hybrid Visual Odometry*

In contrast to end-to-end models that depend exclusively on deep neural networks for pose interpretation from data, hybrid models merge traditional geometric approaches with deep learning techniques. These models employ deep neural networks to enhance parts of geometric models, offering more sophisticated representations. A prominent challenge in conventional monocular visual odometry (VO) is the issue of scale ambiguity, wherein monocular VOs are restricted to estimating relative scale only. This limitation becomes critical in scenarios demanding absolute scale measurements. A viable solution to this problem involves incorporating learned depth estimates into traditional visual odometry frameworks to facilitate the recovery of absolute scale metrics for poses. Depth estimation has long been a focal area of research

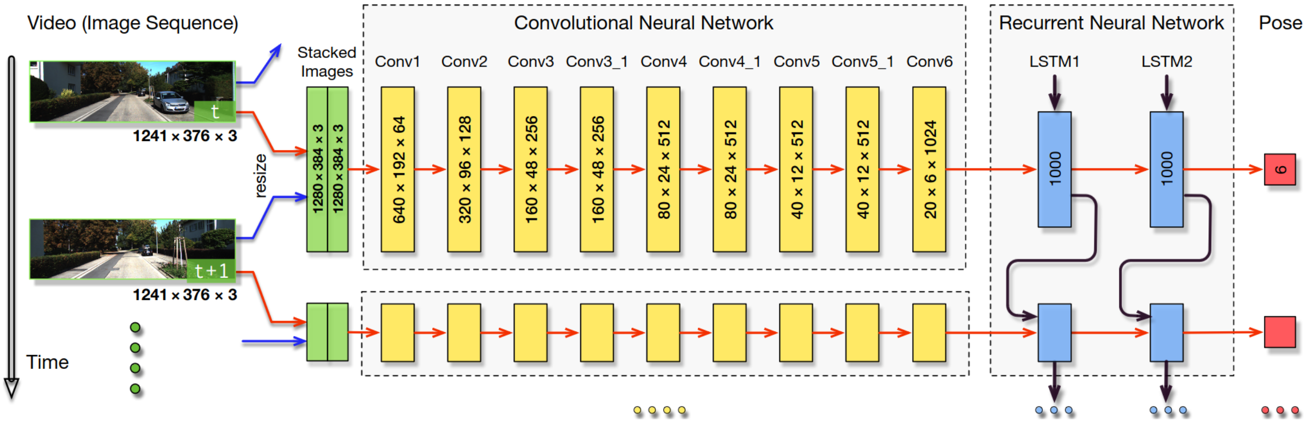


Fig. 3. Architecture of the proposed RCNN based monocular VO system. [16]

in computer vision, with numerous strategies proposed to address this challenge. For example, Godard et al. [22] have developed a deep neural architecture capable of predicting per-pixel depths on an absolute scale.

In the study by [23], a convolutional neural network (ConvNet) generates initial coarse depth estimates from raw images, which are subsequently refined using conditional random fields. The scale factor is determined by correlating these depth predictions with observed point positions. Once this factor is established, the calculated scale is applied to the estimated translations from a monocular visual odometry (VO) algorithm to produce ego-motions with absolute scale, thereby addressing the scale ambiguity issue through the integration of depth data. Further advancements are reported in [24], which introduces the use of predicted ephemeral masks (areas representing moving objects) alongside depth maps within a traditional VO framework to enhance its resilience to dynamic objects. This technique allows for the generation of metric-scale pose estimations using a single camera, even in instances where dynamic objects obscure large areas of the image. Additionally, [24] suggests combining classical VO with learned pose corrections to significantly reduce error drifts found in traditional VOs. Unlike purely learning-based VO systems that regress inter-frame pose changes directly, this approach adjusts pose based on regressions from data, eliminating the requirement for pose ground truth in training. Similarly, [25] enhances classical monocular VO with learned depth estimates, introducing a depth estimation module that operates in two distinct modes to support localization and mapping. This approach demonstrates a high level of adaptability to various environments, surpassing other learning-based VOs. Moreover, [23] integrates learned depth and optical flow predictions into a conventional VO framework. This method utilizes optical flow and single-view depth outputs from deep ConvNets to establish 2D-2D and 3D-2D correspondences, with consistent scale depth estimates helping to alleviate scale drift challenges in monocular VO/SLAM systems. By fusing deep learning predictions with geometric-based techniques, these studies illustrate how deep VO models can enhance traditional VO/SLAM systems.

In summary, hybrid models that incorporate geometric or physical priors alongside deep learning strategies typically surpass the accuracy of end-to-end VO/SLAM systems and often exceed the performance of traditional monocular VO systems in standard benchmarks. These geometry-based models enhance VO/SLAM frameworks by integrating deep neural networks to refine depth and egomotion predictions, and to bolster resilience against dynamic entities. Additionally, models that utilize physical motion principles merge deep learning with established motion models, such as Kalman or particle filters, to seamlessly incorporate these traditional motion models into the learning algorithms of VO/SLAM systems. The advantages of melding geometric or physical priors with advanced learning techniques generally result in hybrid models achieving superior precision compared to end-to-end VO systems, as documented in Table II. It is noteworthy that recent hybrid models have not only advanced rapidly but also outperformed several well-known conventional monocular VO systems in prevalent benchmarks [26], underscoring swift progress in this field.

C. *Robustness Benchmark*

To guarantee the dependable operation of mobile robots, their perception systems are required to be robust against natural distribution shifts [27]. A foundational benchmark for this purpose, ImageNet-C [28], was established to rigorously assess the robustness of image classification techniques against various common image corruptions and disturbances. Building on this foundation, further research has broadened the examination to cover additional perception tasks such as object detection [29]–[31], segmentation [32]–[34], and embodied navigation [11], [35]. These investigations highlight the critical importance of testing the corruption resilience of models. Within the realm of SLAM, the challenges are not only confined to image-level corruptions, such as those resulting from camera failures, but also include managing dynamic sensor corruption and shifts in sensor transformations over time. These variations typically stem from environmental changes that occur over time and the varied movements of robots. In our research, we introduce a perturbation taxonomy specifically designed for RGBD SLAM operating in dynamic environments (e.g., changing light conditions) and unstructured settings (e.g., rough terrains that may induce vibrations in mobile robots).

D. *Robustness Evaluation for SLAM*

The reliability and accuracy of SLAM systems in dynamic and complex real-world settings are crucial, necessitating robust systems capable of enduring sensor malfunctions and sustaining long-term performance [3]. To assess the robustness of SLAM models effectively, various datasets have been compiled in challenging conditions featuring disturbances such as low light or motion blur [2], [7], [8], [36], [37]. Additionally, SLAMBench [38] evaluates the performance of multiple traditional SLAM models against these tough datasets, highlighting their susceptibilities. Given the logistical and scalability challenges of developing real-world datasets through robotic platforms for SLAM, Wang *et al.* [39] have pioneered the creation of a simulated SLAM benchmark known as TartanAir using photo-realistic simulation environments, designed specifically for robustness testing. In our research, we broaden the evaluation framework to include the robustness of multi-modal SLAM models, covering both traditional and neural SLAM approaches, against a wider range of sensor corruptions and motion variables (*e.g.*, changes in speed and motion-induced deviations in sensor trajectories).

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 4 illustrates the overall workflow of our methodology. We can incorporate simulator data generated by CARLA Simulator [10] 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 [40] 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 DeepVO with incorporated dataset for training, to estimate the trajectory and calculate errors using the Absolute Trajectory Error (ATE).

B. *Generative inpainting for dynamic objects*

Our experimental findings, as detailed in Section IV and supported by literature in [41], demonstrate that dynamic objects in unpredictable environments significantly challenge current SLAM models. Traditional SLAM algorithms are predicated on the creation of maps from static environments with immobile objects, which is essential for precise robot localization and map updates. Yet, dynamic objects disrupt these foundational assumptions by introducing significant variability. Specifically, observations can differ dramatically from one frame to the next, as dynamic objects shift positions, complicating the task of matching features across frames due to these inconsistent observations.

The DynaSLAM framework, proposed in [41], effectively mitigates issues posed by dynamic objects. This method integrates the advanced Mask R-CNN model [42] to first identify dynamic objects such as humans and vehicles. Leveraging the binary masks produced by Mask R-CNN, the SLAM algorithms then exclude features within the masked regions from consideration. To bolster the system's robustness, areas obscured by dynamic objects and visible in previous frames are reconstructed using the ground truth static background, reprojected to align with the current camera view. This process of replacing dynamic elements with static images aids in maintaining consistent feature detection and improves the resilience of the algorithm.

Nevertheless, this framework has limitations, especially in scenarios where the background is consistently obscured, such as in dense traffic conditions where the background behind a continuous flow of moving vehicles remains unseen. Large sections of the background thus get excluded, reducing the potential for feature detection and matching. To address this gap, we suggest an enhancement using generative inpainting. This adaptation enables the model to hypothesize the background behind dynamic objects when actual background data is lacking.

For implementing this solution, we utilize a similar Mask R-CNN model as employed in YOLOv8 [43], coupled with the LaMa inpainting model [44] for background prediction (alternative inpainting models are explored in IV).

LaMa demonstrates robustness and a strong ability to generalize by effectively encoding both global and local contexts, facilitated by its extensive receptive fields. This capability is largely due to the utilization of Fast Fourier Convolution (FFC) [45], which incorporates the Fast Fourier Transform (FFT) in the initial stages to preserve global context. In the FFC process, we observe that

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

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

2) combine the real and imaginary parts

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

3) frequency domain convolution

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

4) apply inverse transform to recover a spatial structure

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

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

In the final stage of the proposed framework, the LaMa (Large Mask Inpainting) model significantly enhances the

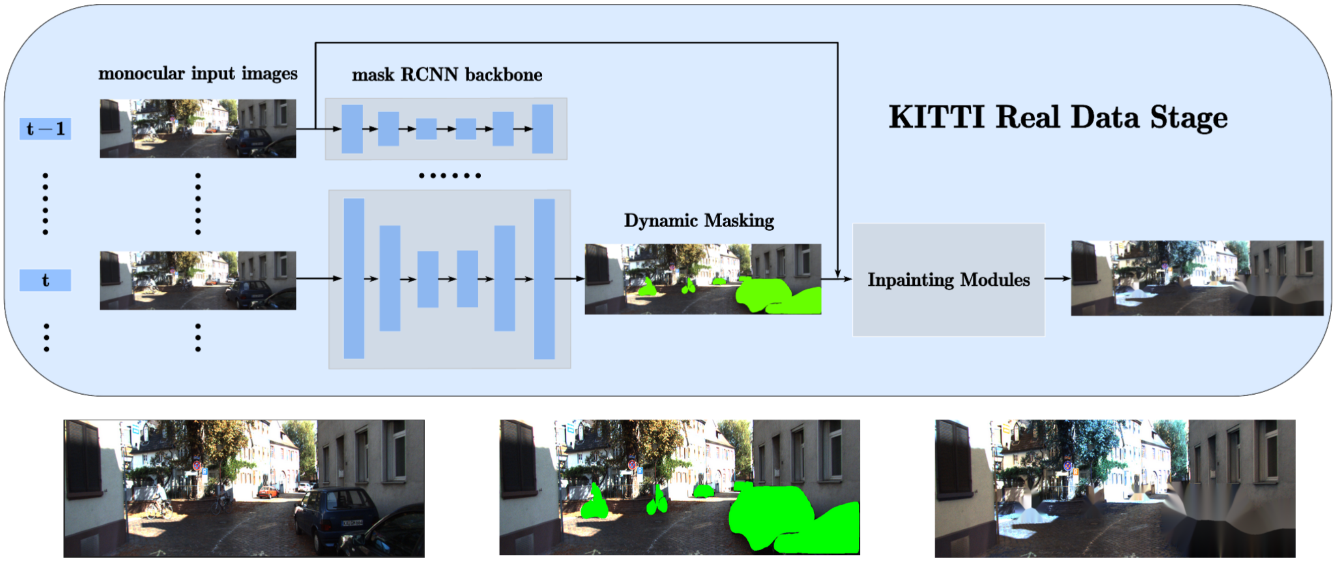


Fig. 4. Architecture of the RCNN based monocular VO system combined with the inpainting modules. [16]

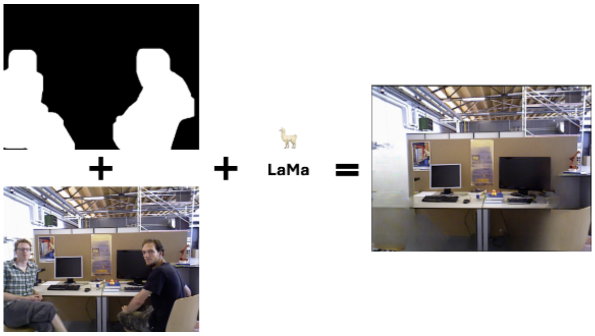


Fig. 5. Generative inpainting model LaMa is able to take in RGB image and the binary mask of target inpainting area to paint the inferred background

quality of the generated backgrounds for masked regions by effectively integrating global and local contextual information. To capture the global context, LaMa employs Fast Fourier Convolution (FFC) [45], which incorporates the Fast Fourier Transform (FFT) in the initial stages. This ensures that the generated background is coherent and consistent with the overall structure and layout of the scene, preventing abrupt or unrealistic changes. Simultaneously, LaMa utilizes traditional convolution techniques to obtain local contextual information from the regions surrounding the masked areas. This enables the generation of backgrounds with realistic textures, patterns, and details that seamlessly blend with the adjacent areas. The combination of global and local contexts empowers LaMa to generate visually coherent and convincing background inpainting results, effectively filling in the masked regions while preserving the overall structure and details of the scene. This robust background inpainting capability significantly contributes to the resilience of the proposed monocular visual odometry system, enabling reliable operation in challenging real-world environments.

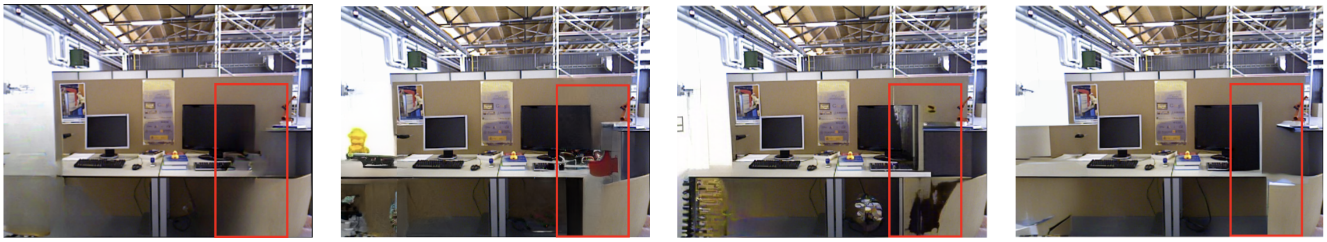
IV. Experiments

A. Experiment Setup

The monocular VO model utilizes a deep Recurrent Convolutional Neural Network (RCNN), streamlining input video by standardizing image sequences and feeding them into a ConvNet inspired by FlowNet architecture, optimized to detect fine features. The processed data is then analyzed by a dual LSTM network within the RNN stage, which interprets both recent and historical information to determine object poses.

To tackle the issue of vanishing or exploding gradients that RNNs typically face, the model integrates LSTM nodes that maintain the efficacy of the gradient through depth and time via three specialized gates. This enables the capture of complex dynamic patterns essential for visual odometry. Furthermore, bidirectional LSTMs are incorporated to improve accuracy by considering information from both past and future timeframes, although this demands additional computational power. This approach aims to address some limitations associated with a monocular viewpoint, expanding the potential for more robust visual odometry.

The experimental setup also involved training and testing on both real-world and synthetic datasets. For the real-world data, the KITTI dataset, which includes 10 sequences with pose ground truth, was used: sequences 00 to 08 for training and sequences 09 to 10 for testing. A synthetic dataset, created to simulate various weather conditions, consisted of 630 frames; 500 frames were designated for training and the remaining 130 for testing. The models were trained with the Adagrad optimizer with a learning rate of 0.0005. For feature extraction in the ConvNet, a pretrained FlowNet was utilized, which is indicative of leveraging existing architectures for improved feature extraction in complex tasks like VO.



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

Fig. 6. **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 5

TABLE I
ABSOLUTE TRAJECTORY ERROR FOR CARLA SYNTHETIC DATASET

Sequences	Rotational RMSE	Translational RMSE
Clean Dataset	7.593	4.158
Rain 25%	9.825	10.067
Rain 50%	10.365	16.327
Rain 75%	20.066	15.553
Fog 25%	60.717	26.357
Fog 50%	79.027	42.327
Fog 75%	85.342	66.743
Traffic Flow	8.664	11.527

B. Experimental Results

Camera Trajectory Evaluation under Disturbances: As shown in the KITTI dataset trajectories in Figure 7, Robust DeepVO generally adheres closer to the ground truth compared to DeepVO, especially in complex sequences like 00, 07, and 09. Sequence 00's intricate path shows Robust DeepVO's better alignment despite environmental complexities, while DeepVO diverges more noticeably. In extended paths such as Sequence 01, Robust DeepVO indicates superior error management over distance. Sequence 03's performance suggests Robust DeepVO's adeptness at sequential frame processing.

Sequence 07 demonstrates both algorithms' struggle with complexity, yet Robust DeepVO shows better recovery toward the sequence's end. The complexity of Sequence 09 illustrates Robust DeepVO's ability to approximate the ground truth more closely, highlighting its resilience. In Sequence 10, both algorithms start off similarly, but Robust DeepVO pulls ahead with greater accuracy in the latter half.

Overall, Robust DeepVO exhibits a trend of more accurate and resilient trajectory tracking across various environmental conditions, indicating advanced algorithmic capabilities in handling complex routes and dynamic changes.

Robustness Evaluation Metrics: 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.* [46]. 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:

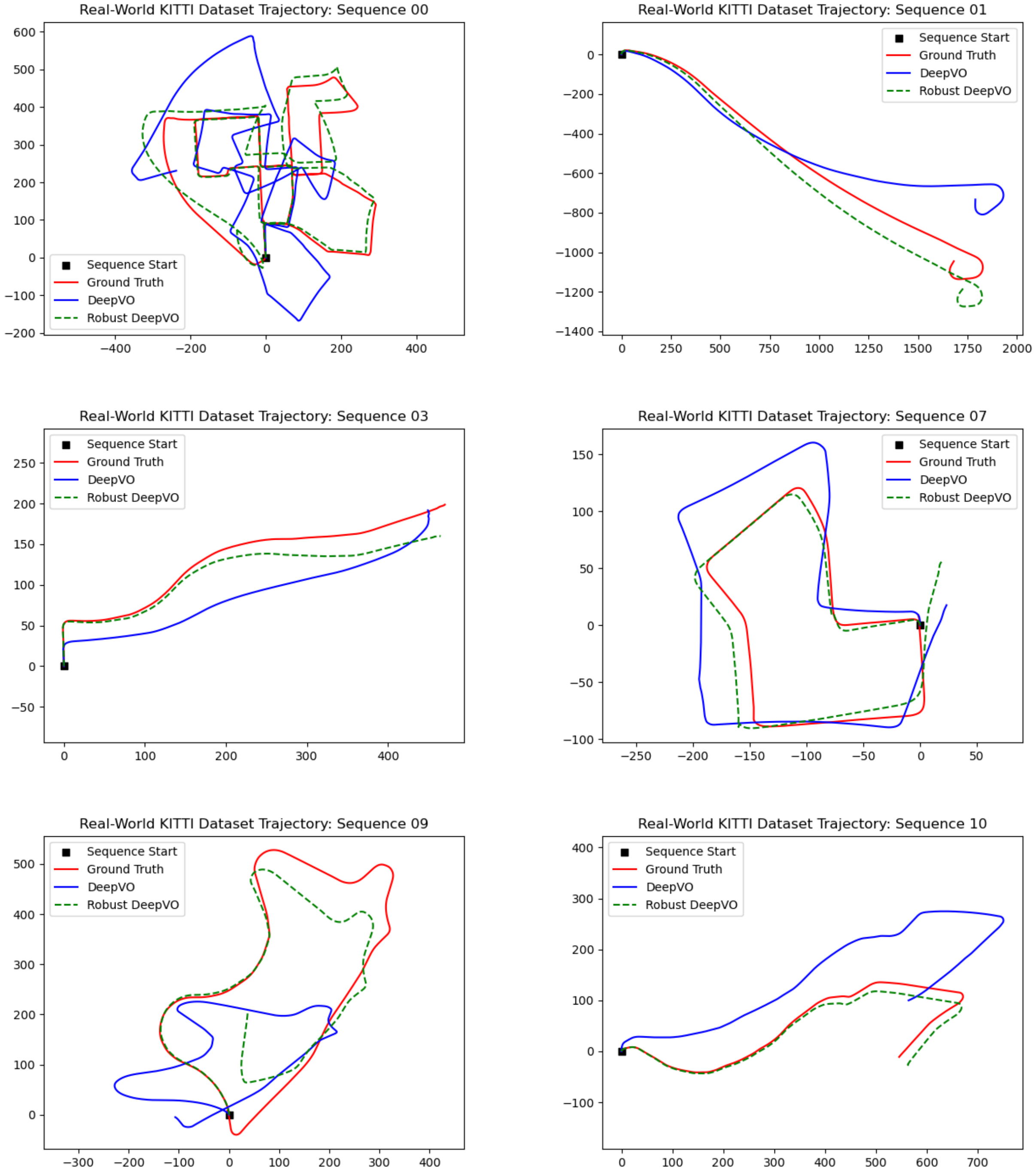
$$RMSE(E_{i:n}) = (\frac{1}{n}\sum_{i=1}^{n} ||trans(E_i)||^2)^{1/2}$$

Qualitative Analysis: The performance of the Robust DeepVO visual odometry system is analyzed under fog and rain conditions using the CARLA synthetic dataset. The system's adaptability to environmental conditions is critical for autonomous navigation, as evidenced by the trajectories plotted against the ground truth.

In Figure 8(a), which displays performance in fog, the trajectory remains close to the ground truth at the lowest density of 25. At fog densities of 50 and 75, the deviation from the ground truth increases progressively, with the density of 75 leading to the largest discrepancies. This pattern indicates a deterioration in the system's ability to localize as conditions become less visually clear, suggesting that visual feature extraction and sensor performance are hampered by higher fog densities.

Figure 8(b) illustrates performance in rain conditions. Here again, the system accurately follows the ground truth in the absence of rain. However, as the rain density escalates to 25, and further to 50 and 75, we witness a parallel escalation in trajectory deviation. The highest rain density of 75 exhibits the most substantial deviation, signifying that the system's performance is notably impacted by heavy rain, likely due to the impairment of visual cues and sensor noise.

Overall, the Robust DeepVO system shows resilience to low-density environmental changes but struggles under higher density conditions of both fog and rain. The performance under high-density conditions underscores the need for the system to incorporate more sophisticated means of handling adverse weather, possibly through enhanced sensor fusion or more advanced feature detection that is less affected by visual impediments. These improvements are critical for the system's utility in real-world autonomous navigation where variable weather is a common challenge. **Quantitative Analysis:** The data in Table I reflect that the visual odometry system maintains high precision under clear conditions and experiences a deterioration in accuracy as environmental complexity increases due to weather factors. There is a notable trend where errors in both rotation and translation incrementally escalate with rising densities of rain and fog,


Real-World KITTI Dataset Trajectory: Sequence 00

Real-World KITTI Dataset Trajectory: Sequence 01

Real-World KITTI Dataset Trajectory: Sequence 03

Real-World KITTI Dataset Trajectory: Sequence 07

Real-World KITTI Dataset Trajectory: Sequence 09

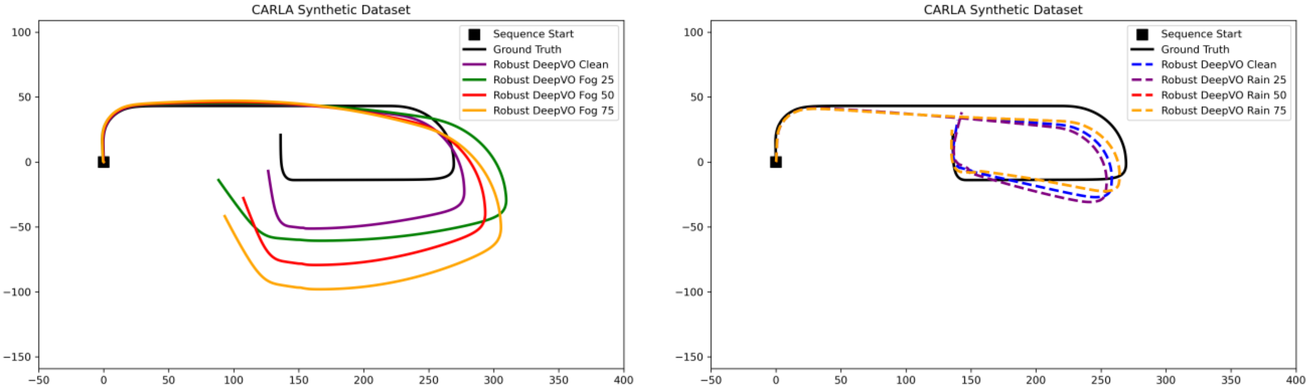
Real-World KITTI Dataset Trajectory: Sequence 10

Fig. 7. **Comparative analysis** of DeepVO and Enhanced Robust DeepVO on Real-world KITTI Dataset

suggesting that adverse weather significantly challenges the system's sensory and processing capabilities.

For the real-world KITTI dataset in Table II, the system's performance exhibits variability across different sequences, which may be attributed to the varying complexities and characteristics inherent to each real-world scenario. This suggests that while the system is capable of handling real-world environments to a degree, its robustness is influenced by the specific dynamics and environmental features encountered in each sequence.

Generative inpainting model. we conduct comparative analyses of the state-of-the-art generative inpainting models. We include LaMa, MAT [47], MIGAN [48], and LDM [49] in our comparison. These models are showcased in Fig. 6, 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.



(a) Performance on CARLA synthetic dataset in *Fog* condition

(b) Performance on CARLA synthetic dataset in *Rain* condition

Fig. 8. Comparison of Robust DeepVO Performance on *Clean* and *Rain* datasets

TABLE II
ABSOLUTE TRAJECTORY ERROR FOR REAL-WORLD KITTI DATASET

Sequences	**Rotational RMSE**	**Translational RMSE**
Sequence 00	13.989	14.693
Sequence 01	13.162	25.454
Sequence 02	17.928	9.127
Sequence 03	11.525	10.327
Sequence 04	29.625	15.953
Sequence 05	13.552	8.240
Sequence 06	44.263	11.449
Sequence 07	25.193	26.327
Sequence 08	22.645	75.923
Sequence 09	33.715	55.210
Sequence 10	6.224	7.015

V. CONCLUSION

In conclusion, this project introduces a novel and comprehensive framework that combines deep learning techniques with traditional geometric methods to significantly enhance the robustness and performance of monocular visual odometry systems in complex, dynamic real-world environments. The key contributions include: (1) KITTI-inspired simulation pipeline that generates diverse synthetic datasets for effective training and evaluation of the proposed models. (2) Integration of the Mask R-CNN model and LaMa inpainting algorithm for robust handling of dynamic objects and occlusions. (3) Utilization of generative inpainting techniques, such as Fast Fourier Convolutions, to preserve global context and ensure consistent scene representation. Extensive experiments on both synthetic and real-world datasets demonstrate the superior performance of the proposed system in terms of camera trajectory tracking accuracy and resilience, particularly in challenging scenarios with dynamic objects and occlusions. The modular architecture allows for future enhancements and integration of additional components, such as semantic information and object priors.Future directions of research include, but not limited to incorporating incremental differentiable slam using the synthetic data generated using Carla with particle filters.

https://github.com/RaghavM11/Diff-Slam

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] 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.

[4] 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.

[5] 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.

[6] 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.

[7] 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.

[8] 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.

[9] 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.

[10] 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.

[11] 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.

[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] H. W. S. Wang, R. Clark and N. Trigoni, "Deepvo : Towards end-to-end visual odometry with deep recurrent convolutional neural networks," *The IEEE International Conference on Robotics and Automation (ICRA)*, vol. 39, no. 3, pp. 2429–2447, 2017.

[17] D. Scaramuzza and F. Fraundorfer, "Visual odometry [tutorial]," 2011.

[18] N. S. T. Zhou, M. Brown and D. G. Lowe, "Learning visual odometry with a convolutional network," *IEEE/CVF In- ternational Conference on Computer Vision and Pattern Recognition (CVPR)*, vol. 39, no. 3, pp. 2429–2447, 2015.

[19] P. V. G. Costante, M. Mancini and T. A. Ciarfuglia, "Exploring representation learning with cnns for frame-to-frame ego-motion estimation," *IEEE Robotics and Automation Letters*, vol. 39, no. 3, pp. 2429–2447, 2023.

[20] N. S. T. Zhou, M. Brown and D. G. Lowe, "Close the optical sensing domain gap by physics-grounded active stereo sensor simulation," *IEEE/CVF In- ternational Conference on Computer Vision and Pattern Recognition (CVPR)*, vol. 39, no. 3, pp. 2429–2447, 2023.

[21] H. P. C. H. P. V. D. S. D. C. P. Fischer, E. Ilg and T. Brox, "Flownet: Learning optical flow with convolutional net- works," *The International Conference on Computer Vision (ICCV)*, vol. 39, no. 3, pp. 2429–2447, 2016.

[22] O. M. A. C. Godard and G. J. Brostow, "Unsupervised monoc- ular depth estimation with left-right consistenc," *IEEE/CVF In- ternational Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 33–37, 2017.

[23] X. D. X. Yin, X. Wang and Q. Chen, "Scale recovery for monocular visual odometry using depth estimated with deep convolutional neural field," *The International Conference on Computer Vision (ICCV)*, p. 5870–5878, 2017.

[24] V. P. B. Wagstaff and J. Kelly, "Self-supervised deep pose corrections for robust visual odometry," *The IEEE International Conference on Robotics and Automation (ICRA)*, p. 2331–2337, 2020.

[25] E. X. Z. L. C. S. L. Sun, W. Yin and C. Shen, "Improving monocular visual odometry using learned depth," *IEEE Transactions on Robotics*, p. 3173–3186, 2020.

[26] R. W. N. Yang, L. von Stumberg and D. Cremers, "D3vo: Deep depth, deep pose and deep uncertainty for monocular visual odometry," *IEEE/CVF International Conference on Computer Vision and Pattern Recognition (CVPR)*, 2020.

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

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

[29] 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.

[30] 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.

[31] 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.

[32] 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.

[33] 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.

[34] 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.

[35] 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.

[36] 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.

[37] 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.

[38] 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.

[39] 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.

[40] 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.

[41] 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.

[42] K. He, G. Gkioxari, P. Dollár, and R. Girshick, "Mask r-cnn," 2018.

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

[44] 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.

[45] 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.

[46] 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.

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

[48] 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.

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