A Review of Visual SLAM Systems and Single-Image Depth Estimation for Challenging Scenarios

Brianna A. Balam-Velasco, L. Oyuki Rojas-Perez, and Jose Martinez-Carranza *
Computer Science Department, Instituto Nacional de Astrofísica, Óptica y Electrónica (INAOE), Puebla 72840, Mexico.

ABSTRACT

This article reviews monocular SLAM systems operating in challenging scenarios, with a particular emphasis on agricultural environments. Key metrics such as pose accuracy, processing time, and trajectory fidelity are examined across complex datasets. To address the known limitations of monocular systems in textureless or repetitive scenes, we explore the integration of single-image depth estimation models, including Depth Anything and ZoeDepth. We evaluate a hybrid SLAM pipeline based on ORB-SLAM2 enhanced with synthetic depth maps and dynamic keypoint adaptation, tested on the Rosario Dataset. The findings demonstrate improvements in localisation accuracy, robustness, and global trajectory consistency under real-world agricultural conditions, establishing a foundation for future research in SLAM enhancement via monocular depth priors.

1 INTRODUCTION

Visual SLAM (Simultaneous Localisation and Mapping) plays an important role in enabling autonomous navigation in environments where GPS is unreliable or unavailable. Among the many SLAM modalities, monocular SLAM systems have gained attention due to their minimal hardware requirements and potential for deployment on lightweight platforms such as drones. However, traditional monocular SLAM faces persistent challenges in environments with low texture, repetitive structures, or dynamic lighting conditions-scenarios common in real-world agricultural and extraterrestrial settings.

To address these challenges, recent research has explored augmenting monocular systems with single-image depth estimation models. These approaches aim to bridge the gap between RGB-only and RGB-D SLAM by generating synthetic depth information that can improve tracking robustness and reduce scale drift. This paper presents a comprehensive review of monocular SLAM systems under difficult environmental conditions and explores how the integration of learning-based depth estimation methods, such as Depth Anything, ZoeDepth, and MiDaS, can enhance system performance.

*Corresponding author: carranza@inaoep.mx

By using ORB-SLAM2 with real-time synthetic depth maps and adaptive feature modulation, we evaluate the proposed hybrid strategy on the Rosario Dataset [1], a benchmark for structured and unstructured fields, and across the six sequences, we report qualitative trajectory alignments and quantitative metrics, highlighting the feasibility and limits of image-based depth priors for strengthening monocular SLAM where conventional methods often fail.

2 RELATED WORK

The development of visual SLAM has been fundamental for autonomous navigation in GPS-denied environments. Early monocular approaches, such as PTAM [2] and MonoSLAM [3], enabled real-time performance with a single camera but were limited by scale ambiguity and poor performance in low-texture scenes. ORB-SLAM and ORB-SLAM2 [4] addressed these issues by incorporating loop closure, map reuse, and relocalisation. These systems later evolved into visual-inertial frameworks like OpenVINS [5], which improved robustness in dynamic and GPS-denied scenarios [6]. More recent efforts have integrated monocular and multi-view depth estimation to improve scale recovery, as demonstrated by Zhang et al. [7] with a tightly coupled SLAM system for indoor environments.

In parallel, semantic information has been explored to enhance localisation in visually ambiguous scenes. For instance, Martinez-Carranza et al.[8] combined natural language with YOLOv8 to support metric SLAM. Similarly, UAV-based studies have applied CNN-based RGB-D perception for GPS-denied localisation[9], and monocular cues for landing detection [10] and obstacle avoidance [11]. These contributions illustrate how combining geometric and semantic data can increase robustness in complex environments.

To further address challenges like visual degradation and dynamic motion, hybrid SLAM frameworks have emerged. Gallagher et al.[12] proposed a sparse-dense system that merges traditional feature tracking with dense depth prediction for increased resilience outdoors. Scherer[13] developed a GPS-free solution for MAVs using RGB-D constraints in low-altitude flight. However, RGB-D sensors remain limited by short range, lighting sensitivity, and vulnerability to fast motion. To overcome these constraints, Ghosh and Gallego [14] introduced a stereo depth method for event cameras, suitable for high-speed or low-light conditions. Gao et al. [15] provide a comprehensive review of hybrid stereo systems, highlighting the benefits of combining geometry with

http://www.imavs.org/

http://www.imavs.org/

learned depth.

As an alternative to more complex sensor setups, monocular depth estimation has proven effective for lightweight robotic platforms. Pellegrin et al.[16] proposed a single-image model tailored to aerial views, while Alquisiris et al.[17] employed optical flow for real-time depth on mobile robots. Rojas-Perez et al. [18] introduced a decentralised monocular SLAM strategy for coordinating UAVs in GPS-denied environments using learned priors. These works highlight the adaptability of monocular estimation in constrained systems.

Recent advances have focused on transformer-based and zero-shot models to improve generalisation in depth prediction. AdaBins [19] introduced adaptive binning for depth regression, ZoeDepth [20] combined metric and relative depth for zero-shot transfer, and Depth Anything [21] leveraged large-scale pretraining to estimate depth in novel environments. Integrated into SLAM pipelines, these models help reduce scale drift and mapping errors. Additionally, synthetic depth priors have proven useful in low-texture or poorly lit conditions [22]. Bladh [23] evaluated AdaBins, MiDaS, and Monodepth2 within ORB-SLAM3, reporting improved rotational consistency and suggesting that synthetic depth can substitute conventional sensors in some cases.

Despite these advances, monocular SLAM still faces challenges in repetitive or texture-sparse environments. In this paper, we propose a hybrid approach that combines monocular depth priors with adaptive feature modulation to simulate RGB-D input from a single RGB stream. This system is evaluated in real-world agricultural environments, where traditional SLAM pipelines often fail due to texture scarcity and scene variability.

3 METHODOLOGY

We propose a theoretical strategy to improve the localisation performance of ORB-SLAM2 in unstructured environments by generating synthetic depth images from monocular images and using them to simulate RGB-D operation. The aim is to overcome the scale ambiguity and reduce tracking robustness typically associated with monocular SLAM, particularly in environments characterised by low texture, inconsistent illumination, or repetitive structures.

The proposed approach comprises two main conceptual contributions; first, we introduce a method to infer and adapt scene depth using spatial statistics, enabling monocular SLAM to approximate the metric scale, and we propose a dynamic visual feature modulation mechanism that adjusts the number of extracted keypoints based on interframe motion, improving tracking robustness under varying conditions.

3.1 Scale recovery through depth compensation

Monocular SLAM lacks inherent metric information; to overcome this, we propose a depth compensation method that generates synthetic metric depth maps by fusing predictions from *Depth Anything* and *ZoeDepth* estimators.

Our method employs a hybrid strategy, where *Depth Anything* generates dense, relative depth maps from each RGB frame, and while these maps preserve structural consistency, they lack a fixed metric scale. To address this, we periodically apply *ZoeDepth* to a cropped, structurally grounded region of the image to obtain a metric reference, which then allows us to compute a scaling factor by comparing the local depth variation from *Depth Anything* with the global depth range from *ZoeDepth*.

To ensure temporal coherence between consecutive frames, we apply a smoothing function to the maximum depth estimation using an exponentially weighted moving average:

$$\text{depth_max}_t = \alpha \cdot \text{depth_max}_{t-1} + (1-\alpha) \cdot r_{\text{avg}} \quad (1)$$

where α is a smoothing coefficient and r_{avg} is the average of recent scaling factors. This formula maintains a stable and consistent depth representation over time, supporting pose estimation and trajectory reconstruction in the simulated RGB-D configuration of ORB-SLAM2.

And, to enable the integration of synthetic depth images into the RGB-D pipeline of ORB-SLAM2 using monocular input, it was defined a scaling strategy that transforms relative depth predictions into scale-consistent metric maps. We first periodically apply *ZoeDepth* to a cropped image region to compute the global depth range:

$$dr = \max(\text{depth}) - \min(\text{depth}) \quad (2)$$

This value serves as a reference for calibrating the relative predictions. Next, for every frame, we use *Depth Anything* to generate a dense, relative depth map; this map is then normalized, inverted, and scaled using the global range dr, for each pixel p in the region of interest, we calculate a local scaling factor:

$$r = \frac{(p - \min)}{ds} \cdot dr \quad (3)$$

where $ds = \max - \min$ is the local depth variation.

Finally, it is rescaled the entire depth map using the smoothed maximum depth value:

$$\text{depth}_{\text{metric}} = \left(1.0 - \frac{\text{depth}_{\text{norm}}}{255.0}\right) \cdot \text{depth_max} \quad (4)$$

This produces a synthetic depth image that is both dense and scale-consistent, enabling monocular input to be treated as if it were obtained from a depth sensor.

3.2 Adaptive feature modulation based on motion analysis

In addition to spatial enhancement, a dynamic visual tracking mechanism is proposed that adjusts the number of keypoints extracted per frame. Traditional SLAM systems typically employ a fixed feature count, which may lead to inefficiencies or instability when operating in dynamic or visually complex environments.

http://www.imavs.org/

To address this, a motion-aware adjustment strategy is defined that monitors translation and rotation between consecutive camera poses:

$$\Delta t = \|t_1 - t_0\|, \quad \Delta R = \|R_1 - R_0\| \tag{5}$$

where t_0, t_1 represent positions and R_0, R_1 are the corresponding orientations. In cases of significant motion or sudden changes in direction, the number of features is increased to preserve tracking accuracy; in contrast, during slow or stable movements, the feature count is reduced to improve computational efficiency.

This modulation is conducted incrementally and constrained by predefined thresholds to ensure smooth transitions and avoid performance oscillations.

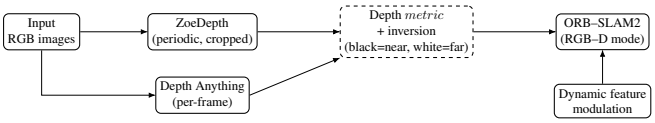


Figure 1: Proposed system pipeline, ZoeDepth periodically estimates dr on cropped frames; Depth Anything provides dense depth. Scaled, inverted maps (black=near, white=far) feed ORB–SLAM2 (RGB–D) with dynamic feature modulation.

4 EXPERIMENTS

By using sequences from the Rosario Dataset, we evaluate the system's performance under realistic agricultural conditions, focusing on how depth augmentation affects trajectory accuracy and robustness; each experiment is analyzed both visually and through quantitative error metrics to assess tracking consistency and global alignment.

4.1 Rosario Dataset

The Rosario Dataset [1] is specifically designed to assess visual localisation and mapping algorithms in realistic agricultural settings; it was captured under natural outdoor conditions, offering challenging scenarios typical of agricultural fields, including repetitive visual patterns, low-textured surfaces, dynamic illumination, and environmental variability.

Comprised of six distinct sequences, this dataset provides synchronized multi-sensor data for evaluating visual and visual-inertial SLAM systems. Key characteristics include high-resolution stereo imagery at 672×376 pixels and 15 Hz, 6-axis IMU data at 140 Hz, and a ground truth reference from wheel odometry and precise GPS-RTK positional information at 5 Hz.

These characteristics make the Rosario Dataset especially suitable for testing SLAM performance in challenging, real-world scenarios, allowing for a comprehensive benchmarking of the robustness, accuracy, and adaptability of monocular SLAM and depth estimation techniques.

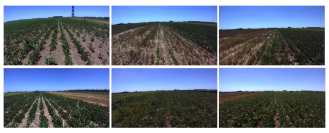


Figure 2: The Rosario dataset samples.

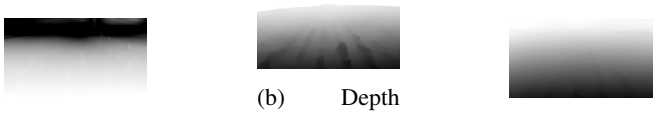
4.2 Monocular depth estimation

Depth estimation from single images has evolved rapidly, transitioning from classical convolutional neural networks to transformer-based and zero-shot learning. Early efforts, such as Eigen et al.'s multi-scale CNN [24], struggled to generalise and lacked metric accuracy, while later approaches like AdaBins [19] introduced improvements in both generalisation and precision by using an adaptive binning strategy.

ZoeDepth [20] represents a recent advance in zero-shot depth estimation; by combining relative and metric training signals, it can predict metric depths without requiring fine-tuning on each new domain. The model performs particularly well in out-of-distribution settings, but due to its high computational cost, it was used to generate global metric depth references at a lower frequency rather than for real-time processing, so depth map inversion was not implemented.

In contrast, Depth Anything [21] adopts a vision-language pretraining strategy and is trained on over 62 million images; its lightweight nature allowed for real-time synchronisation. The model produces inherently relative, highly generalisable, and visually consistent dense depth maps, so the inversion of depth values was applied, significantly facilitating clearer visual interpretation and integration with the RGB-D pipeline of ORB-SLAM2.

MiDaS [25], another monocular depth estimation model, leverages vision transformer architectures to enhance generalisation. The Swin2-Tiny variant provided a good balance of computational efficiency and depth estimation quality, achieving stable real-time synchronisation; thus, depth inversion was also applied; however, MiDaS produced slightly blurrier textures compared to Depth Anything, impacting feature matching quality and localisation accuracy.



(a) ZoeDepth (b) Depth Anything (c) MiDaS

Figure 3: Comparison of depth maps generated from the three models (Sequence 01). (a) ZoeDepth (original convention), (b) Depth Anything (inverted convention), (c) MiDaS (Swin2-Tiny, inverted convention).

Depth map inversion was strategically implemented for Depth Anything and MiDaS to facilitate visual interpretation and streamline integration with ORB-SLAM2's RGB-D

http://www.imavs.org/

pipeline; conversely, ZoeDepth was used without inversion due to computational limitations, serving as a periodic global depth reference.

4.3 Rosario Dataset Experiments

The effectiveness of the proposed methodology was evaluated using sequences from the Rosario Dataset [1]; the ORB-SLAM2 system was operated in RGB-D mode, with monocular RGB images augmented in real-time by synthetic depth maps from the Depth Anything model. These maps were produced at 15Hz, matching the frame rate of the `rosbag` stream to ensure temporal alignment.

The first sequence involved a semi-structured agricultural path with long linear stretches and sparse visual variation. The estimated trajectory (Fig. 4) successfully maintained global consistency along the main linear portions, but noticeable deviations appeared at the initial segment and sharp turns; vertical consistency was well preserved.

Figure 4: Comparison of ground truth and ORB-SLAM2 trajectory for Sequence 01.

Complementary error analysis (Fig. 5) indicates that the Absolute Position Error (APE) increases at the tight turns and endpoints, but drift is kept low on long linear extensions.

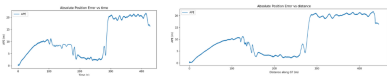


Figure 5: Sequence 01 absolute position error (APE) time (left) and distance (right).

Sequence 02 introduced increased variability in vegetation and texture, which initially enhanced SLAM tracking, but repetitive patterns later reappeared, challenging continuous feature association. The estimated trajectory (Fig. 6) successfully reproduced the global structure, including extended straight sections, but discrepancies became more pronounced at turning points and curved portions; the side view, however, showed strong consistency.

Complementary error analysis (Fig. 7) shows that the APE increases almost monotonically, with larger increments around turns, indicating cumulative drift along the long straight sections.

Sequence 03 exhibited linear motion with denser and more structured weed patterns, providing robust and stable

Figure 6: Comparison of ground truth and ORB-SLAM2 trajectory for Sequence 02.

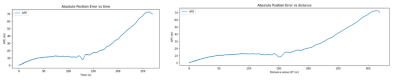


Figure 7: Sequence 02 absolute position error (APE) time (left) and distance (right).

visual features that significantly enhanced tracking. The SLAM-estimated path (Fig. 8) aligns closely with the ground truth along most linear sections, particularly in the early part of the sequence; however, noticeable deviations arise in the latter half, especially around the long curved descent and at the endpoints. The side view reveals a strong vertical alignment between both trajectories.

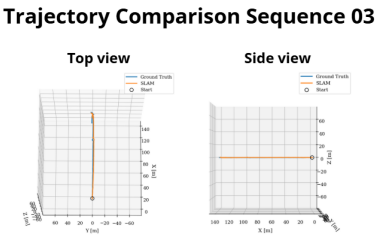


Figure 8: Comparison of ground truth and ORB-SLAM2 trajectory for Sequence 03.

As Figure 9 shows, the APE remains low and stable initially but increases in stages with local peaks appearing around the curved segment.

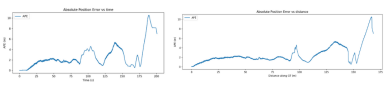


Figure 9: Sequence 03 absolute position error (APE) time (left) and distance (right).

Sequence 04 presented an environment similar to Sequence 03, characterized by repetitive and uniform textures; despite these challenges, monocular depth augmentation allowed ORB-SLAM2 to maintain trajectory estimation throughout the sequence. The estimated trajectory (Fig. 10) shows an almost perfect alignment with the ground truth, with

http://www.imavs.org/

a near overlap of both paths along the long straight segment and no noticeable vertical drift.

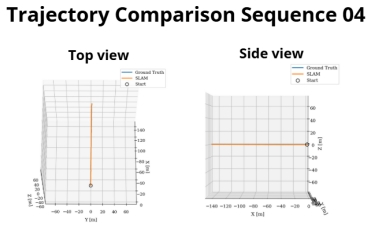


Figure 10: Comparison of ground truth and ORB-SLAM2 trajectory for Sequence 04.

Error analysis (Fig. 11) shows that the APE remains at the centimeter level throughout the entire run, indicating insignificant drift.

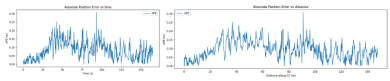


Figure 11: Sequence 04 absolute position error (APE) time (left) and distance (right).

Sequence 05 presented substantial difficulties due to an increased presence of repetitive crop patterns interspersed with abrupt changes in weed texture, which significantly challenged depth estimation and keypoint matching reliability. The estimated trajectory (Fig. 12) initially aligns with the ground truth but rapidly diverges, with significant drift evident in both top and side views, particularly during transitions between repetitive rows and abrupt texture changes.

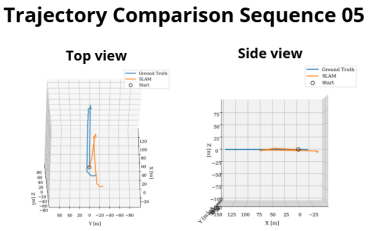


Figure 12: Comparison of ground truth and ORB-SLAM2 trajectories for Sequence 05.

Figure 13 shows that the APE grows steadily, peaking around the middle of the run and remaining large at all times, coinciding with strong flat drift and noticeable vertical bias.

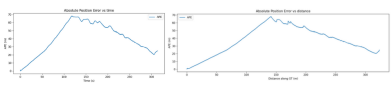


Figure 13: Sequence 05 absolute position error (APE) time (left) and distance (right).

Sequence 06 represents one of the most challenging trials, featuring long, repetitive crop rows with minimal textural distinction, which significantly hinders both keypoint extraction and depth inference. The ORB-SLAM2 system was unable to maintain trajectory accuracy; the estimated trajectory (Fig. 14) fails to capture the loop structure and deviates progressively from the ground truth, while the side view reveals a consistent vertical alignment.

Figure 14: Comparison of ground truth and ORB-SLAM2 trajectory for Sequence 06.

Complementary error analysis (Fig. 15) shows that the APE increases sharply in the final stretch, with the distance view following an almost linear trend on long straights and steepening towards the end.

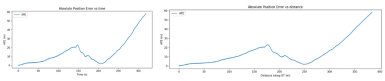


Figure 15: Sequence 06 absolute position error (APE) time (left) and distance (right).

4.4 Discussion

The experiments performed on the Rosario Dataset reveal valuable insights into integrating monocular depth estimation into a visual SLAM pipeline under challenging agricultural conditions. Qualitatively, the proposed hybrid approach, which uses Depth Anything for real-time relative estimation and ZoeDepth for periodic metric correction, enables the system to better preserve spatial coherence and trajectory fidelity than traditional monocular SLAM. The quantitative results in Table 1 confirm that while residual drift is still present, the system maintains consistent alignment in structured segments; this improvement stems from the adaptive modulation of feature extraction and dynamic scaling of depth maps.

Sequences 04 (RMSE of 0.05 m) and 03 (RMSE of 3.20 m) achieved the lowest errors, reflecting the system's ability to track accurately in stable, well-structured environments, while Sequence 01 (RMSE of 7.33 m) showed moderate accuracy. In contrast, sequences 02, 05, and 06 exhibited substantially higher errors due to texture-scarce or highly repetitive segments and re-localization challenges. Despite these errors, the overall shape and orientation of the trajec-

http://www.imavs.org/

tories were generally maintained, demonstrating the system's ability to preserve route consistency even when local accuracy was degraded.

Table 1: Trajectory error metrics on the Rosario Dataset using the proposed monocular depth-augmentation approach.

Seq.	RMSE (m)	MAE (m)	Median AE (m)	Max (m)	Min (m)
01	7.33	6.55	7.48	12.14	1.08
02	32.87	25.36	14.35	72.89	0.00
03	3.20	2.31	1.42	9.27	0.05
04	0.05	0.04	0.03	0.22	0.00
05	45.52	41.32	44.90	68.11	0.00
06	19.82	15.53	12.05	54.81	0.09
Avg.	**18.13**	**15.19**	**13.37**	**36.24**	**0.21**

Purely monocular ORB-SLAM2 did not initialize in any of the sequences, highlighting the difficulty of repetitive row crop patterns. In contrast, the proposed hybrid approach initialized in all sequences, demonstrating substantially greater operational robustness; this indicates that periodic metric cues and dense per-frame relative depth are crucial for consistent tracking in challenging scenarios.

While the proposed system improves localisation, practical trade-offs remain. **ZoeDepth** is applied selectively on cropped regions due to its high computational cost, limiting the frequency of global depth corrections. In contrast, **Depth Anything** enables real-time synchronisation and produces dense relative depth maps; however, its non-metric nature can cause local inconsistencies in visually repetitive or texture-sparse regions.

In summary, the combination of **Depth Anything** for dense real-time inference and **ZoeDepth** for metric scaling allows the pipeline to achieve a consistent, globally coherent trajectory; while residual drift remains a limitation, the approach significantly improves localization robustness over traditional monocular SLAM, which often fails in these environments.

5 CONCLUSIONS

A hybrid monocular SLAM framework is presented, which integrates synthetic depth estimation from *Depth Anything* and *ZoeDepth* with adaptive keypoint modulation, enhancing localization robustness in visually complex and repetitive environments. Using real-time relative depth inference and periodic metric corrections, the system enables ORB-SLAM2 to approximate RGB-D performance with monocular input.

Evaluations on the Rosario Dataset demonstrated consistent improvements in global trajectory alignment and tracking stability, particularly in sequences with structured textures and repeatable features; although residual drift remained a challenge, the system preserved the overall trajectory configuration. Quantitative metrics, including RMSE and absolute trajectory errors, corroborated these findings, highlighting the practical value of synthetic depth priors in extending monocular SLAM to environments where traditional pipelines often fail.

Future research will focus on incorporating loop closure based on learned descriptors, using lightweight and uncertainty-sensitive depth models, and optimizing the system for embedded applications. Other interesting directions include long-term pipeline operation and scalable mapping in dynamic, unstructured environments; the contribution lays a more solid foundation for applying monocular SLAM in practical scenarios such as field robots, planetary exploration, and agriculture.

REFERENCES

[1] Taihú Pire, Juan Corti, Guillermo L. Grinblat, Diego Di Lernia, Sebastián Bianchi, and Pablo De Cristóforis. The rosario dataset: Multisensor data for localization and mapping in agricultural environments. In *Proceedings of the XX Congreso Argentino de Ciencias de la Computación (CACIC)*, pages 1–10. RedUNCI, 2019.

[2] Georg Klein and David Murray. Parallel tracking and mapping for small ar workspaces. In *2007 6th IEEE and ACM International Symposium on Mixed and Augmented Reality*, pages 225–234. IEEE, 2007.

[3] Andrew J Davison, Ian D Reid, Nicholas D Molton, and Olivier Stasse. Monoslam: Real-time single camera slam. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 29(6):1052–1067, 2007.

[4] Raul Mur-Artal and Juan D Tardós. Orb-slam2: An open-source slam system for monocular, stereo, and rgb-d cameras. *IEEE Transactions on Robotics*, 33(5):1255–1262, 2017.

[5] Patrick Geneva, Kyle Eckenhoff, Yulin Lee, and Guoquan Huang. Openvins: A research platform for visual-inertial estimation. In *IEEE International Conference on Robotics and Automation (ICRA)*, pages 4666–4672. IEEE, 2020.

[6] Michael Burri, Jakob Nikolic, Philip Gohl, Thomas Schneider, Johannes Rehder, Sami Omari, Markus Achtelik, and Roland Siegwart. The euroc micro aerial vehicle datasets. *The International Journal of Robotics Research*, 35(10):1157–1163, 2016.

[7] X. Wang, Z. Zhang, and L. Li. A tightly-coupled dense monocular visual-inertial odometry system with lightweight depth estimation network. *Applied Soft Computing*, 171:112809, 2025.

[8] Jose Martinez-Carranza, David I. Hernandez-Farias, Leticia Oyuki Rojas-Perez, and A. A. Cabrera-Ponce. Language meets yolov8 for metric monocular slam. *Journal of Real-Time Image Processing*, 20(59), 2023.

http://www.imavs.org/

[9] Jose Martinez-Carranza, Leticia Oyuki Rojas-Perez, A. A. Cabrera-Ponce, and R. Munguia-Silva. Combining deep learning and rgbd slam for monocular indoor autonomous flight. In *15th Mexican International Conference on Artificial Intelligence (MICAI)*, Guadalajara, Jalisco, Mexico, 2018.

[10] Leticia Oyuki Rojas-Perez, R. Munguia-Silva, and Jose Martinez-Carranza. Real-time landing zone detection for uavs using single aerial images. In *10th International Micro Air Vehicle Conference (IMAV)*, Melbourne, Australia, 2018.

[11] Leticia Oyuki Rojas-Perez and Jose Martinez-Carranza. Metric monocular slam and colour segmentation for multiple obstacle avoidance in autonomous flight. In *IEEE 4th Workshop on Research, Education and Development of Unmanned Aerial Systems (RED-UAS)*, Linköping, Sweden, 2017.

[12] Liam Gallagher, Vageesh Kumar, Senthil Yogamani, and John McDonald. A hybrid sparse-dense monocular slam system for autonomous driving. In *2021 European Conference on Mobile Robots (ECMR)*, pages 1–8. IEEE, 2021.

[13] Sebastian A. Scherer. *Efficient Visual-Inertial SLAM for Micro Aerial Vehicles*. PhD thesis, University of Tübingen, Germany, 2016.

[14] Arnab Ghosh and Guillermo Gallego. Multi-event-camera depth estimation and outlier rejection by refocused events fusion. *arXiv preprint arXiv:2207.10494*, 2022.

[15] Boyu Gao, Haoxiang Lang, and Jing Ren. Stereo visual slam for autonomous vehicles: A review. In *2020 IEEE International Conference on Systems, Man, and Cybernetics (SMC)*, pages 1840–1847, Toronto, Canada, 2020. IEEE.

[16] Luca Pellegrin and Jose Martinez-Carranza. Towards depth estimation in a single aerial image. *International Journal of Remote Sensing*, 41(1):1–19, 2020.

[17] O. Alquisiris-Quecha and Jose Martinez-Carranza. Depth estimation using optical flow and cnn for the nao robot. In *18th Mexican International Conference on Artificial Intelligence (MICAI)*, Xalapa, Veracruz, Mexico, 2019.

[18] Leticia Oyuki Rojas-Perez and Jose Martinez-Carranza. Flight coordination of mavs in gps-denied environments using a metric visual slam. In *11th International Micro Air Vehicle Conference (IMAV)*, Madrid, Spain, 2019.

[19] Shariq Farooq Bhat, Ibraheem Alhashim, and Peter Wonka. Adabins: Depth estimation using adaptive bins. *arXiv preprint arXiv:2011.14141*, 2020.

[20] S. F. Bhat, R. Birkl, D. Wofk, P. Wonka, and M. Müller. Zoedepth: Zero-shot transfer by combining relative and metric depth. 2023.

[21] Zhen Li et al. Depth anything: Unleashing the power of large-scale unlabelled data. In *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*, 2023.

[22] Cesar Cadena, Luca Carlone, Henry Carrillo, Yasir Latif, Davide Scaramuzza, Jose Neira, Ian Reid, and John J. Leonard. Past, present, and future of simultaneous localization and mapping: Toward the robust-perception age. *IEEE Transactions on Robotics*, 32(6):1309–1332, 2016.

[23] Daniel Bladh. Deep learning-based depth estimation models with monocular slam: Impacts of pure rotational movements on scale drift and robustness. Master's thesis, Linköping University, 2023.

[24] David Eigen, Christian Puhrsch, and Rob Fergus. Depth map prediction from a single image using a multi-scale deep network. *arXiv preprint arXiv:1406.2283*, 2014.

[25] Rene Ranftl, Alexey Bochkovskiy, and Vladlen Koltun. Towards robust monocular depth estimation: Mixing datasets for zero-shot cross-dataset transfer. *arXiv preprint arXiv:1907.01341*, 2019.