Assessing Depth Anything V2 monocular depth estimation as a LiDAR alternative in robotics

Jakobus Murray Louw[1*], *Jaco* Verster[1], and *John* Dickens[1]

[1]Centre for Robotics and Future Production, CSIR, Pretoria, South Africa

Abstract. This paper evaluates the performance of Depth Anything V2, a deep learning-based monocular depth estimation model, as a low-cost alternative to LiDAR for robotic depth sensing. LiDAR, while widely used, is expensive, prompting the search for affordable solutions. Six datasets were recorded in indoor environments to assess the performance of the pre-trained metric depth model. Qualitative analysis showed that overall relative depth is well estimated, but fine details and close-range depths in feature-sparse areas are not represented well. Quantitative analysis revealed variability in performance across datasets, with mean errors ranging from 0.32 m to 0.66 m. Additionally, performance varies with distance. For objects within 2 m, 89.1% of errors are within ±0.5 m. This decreases to 77.0% for objects within 4 m and further drops to 70.8% for objects within 6 m. Depth Anything V2 demonstrates higher pixel resolution than LiDAR but with significantly reduced metric depth accuracy. While not suitable for high-precision applications like indoor navigation and obstacle avoidance, the model can still provide useful depth information in scenarios where fine-grained accuracy is less critical.

1 Introduction

1.1 Background on monocular depth estimation

Monocular depth estimation (MDE) represents a significant advancement in computer vision [1,2], aiming to predict the three-dimensional geometry of a scene using only a single image input. MDE has gained significant interest due to its cost-effective sensor requirements compared to traditional methods. Its applications extend across autonomous vehicles, robotics, AR, and 3D reconstruction, where depth accuracy is critical for system operation.

Despite its practical utility, extracting reliable depth information from a single image presents formidable challenges. The core challenge stems from the inherent ambiguity of 3D-to-2D projection. Environmental factors such as lighting variations, occlusions, and texture characteristics further complicate accurate depth prediction. Such factors create additional complexity in the already demanding task of inferring depth from monocular images.

Recent advances in deep learning have addressed some of these limitations through end-to-end trained models that generate dense depth maps. These developments have

* Corresponding author: mlouw2@csir.co.za
** Language editing to improve clarity and readability of this paper was assisted by ChatGPT.

substantially improved both the performance and accuracy of depth prediction systems. A notable example is the Depth Anything V2 [3] model, which demonstrates significant advances in depth prediction accuracy, efficiently handling complex scenes while preserving fine-grained details. This represents a notable advancement in addressing fundamental MDE limitations.

1.2 LiDAR technology in mobile robotics and the case for monocular alternatives

Light Detection and Range (LiDAR) technology represents a cornerstone technology in mobile robotics, using active time-of-flight (ToF) sensing principles to generate precise 3D depth measurements. LiDAR has become integral to Simultaneous Localization and Mapping (SLAM) systems [4]. The widespread adoption of LiDAR in robotics applications can be attributed to several key advantages which include high spatial precision, reliable performance in challenging conditions, and real-time processing capabilities, making it essential for autonomous robotic systems.

However, despite these significant advantages, the high cost of LiDAR limits its widespread adoption in robotics applications. This constraint has motivated the exploration of alternative approaches such as MDE, which presents a cost-effective alternative by leveraging standard Red-Green-Blue (RGB) cameras, making it particularly appealing when compared to expensive multi-beam LiDAR systems.

1.3 Objectives and contributions of this study

This study evaluates Monocular Depth Estimation as a LiDAR alternative in mobile robotics, focusing on the Depth Anything V2 model. The research examines practical implementation aspects and evaluates performance across critical tasks including obstacle avoidance, visual mapping, and navigation. Our analysis provides insights into deploying MDE on robotic platforms, with particular emphasis on real-world applications.

By thoroughly examining the limitations and constraints of substituting LiDAR with deep learning-based MDE, we identify crucial operational boundaries and performance characteristics. This research bridges the gap between theoretical capabilities and practical implementation, providing essential insights for roboticists and researchers working on vision-based navigation systems.

2 Related work

2.1 Recent progress in deep learning-based depth estimation

MDE has seen significant advancement in recent years, primarily driven by deep learning approaches and the availability of large depth-annotated datasets [5-11] has enabled significant progress in both indoor and outdoor scenes. Recent breakthrough approaches have focused on leveraging large-scale training datasets that combine both supervised and unsupervised learning paradigms. Notable contributions include MiDaS [12], Depth Anything (V1 and V2) [3,13], Metric3D [14], ZeroDepth [15], and ZoeDepth [16]. MiDaS initially established a benchmark by training on mixed labelled datasets, though its limited data coverage affected generalization capabilities. Subsequent developments by ZeroDepth and Metric3D V2 enhanced model robustness through extensive datasets of 15M and 16M labelled images, respectively.

A significant evolution occurred with Depth Anything V1, which expanded upon MiDaS's foundation by incorporating 62M unlabelled images from eight large-scale public datasets. This approach, combining data augmentation strategies with pseudo annotations, demonstrated exceptional zero-shot depth estimation capabilities. Depth Anything V2 further

refined this approach by exclusively using synthetic images for initial training, addressing the inherent limitations of real labelled data such as label noise and overlooked details.

The latest iteration, Depth Anything V2, represents the current state-of-the-art in zero-shot capability and metric depth estimation when fine-tuned. Its success has led to widespread adoption, particularly evident in its use as a backbone architecture by leading teams in the 2024 Monocular Depth Estimation Challenge [17]. This progression demonstrates the field's evolution toward more robust and generalizable depth estimation solutions, effectively combining synthetic data advantages with sophisticated training methodologies.

2.2 LiDAR alternatives in robotics

Alternative sensing technologies, including Time-of-Flight (ToF) [18], Ultrasonic sensors [19], depth cameras [20], and stereo cameras [21], offer lower-cost navigation solutions, often integrated with motor encoders to generate environmental maps. Camera-based systems leveraging historical LiDAR data have shown promise in enhancing monocular 3D detection [22]. However, these alternatives face significant limitations: ToF and Ultrasonic sensors suffer from resolution and range constraints, while camera-based solutions remain sensitive to environmental conditions. Depth cameras provide fast imaging and high lateral resolution, but they can be prone to noise in depth estimation. Stereo cameras, which use vision disparity for depth perception, can offer robust depth information but may be limited by computational demands and the need for precise calibration. Furthermore, the integration of multiple sensors to compensate for individual limitations often results in increased system complexity and can paradoxically lead to higher overall costs despite using lower-cost components.

LiDAR's continued dominance in the field can be attributed to its superior accuracy, environmental robustness, and real-time capabilities, supported by well-established infrastructure for its applications. While the search for cost-effective alternatives continues, the challenge remains to match LiDAR's comprehensive performance characteristics within a more economical package.

3 Methodology

3.1 Depth Anything V2 model overview

Depth Anything V2 advances the state-of-the-art in monocular depth estimation [17]. The model's training data combines 595K synthetic images and 62M unlabelled images [3], with synthetic data replacing previously used real images. This approach, supplemented by pseudo-labelled real images, enhances the model's cross-environment generalization capabilities.

The architecture of Depth Anything V2 focuses on enhancing the model's generalization capabilities by employing a teacher-student framework. The teacher model is scaled up to increase its capacity, allowing it to provide more accurate and robust depth predictions. The student models are trained using pseudo-labelled real images, which act as a bridge between synthetic and real-world data. This approach enables the model to perform well across diverse scenarios without requiring novel technical modules. Capabilities include:

- Efficiency and Accuracy: Depth Anything V2 models are significantly more efficient (over 10x faster) and accurate compared to models built on Stable Diffusion.
- Generalization: The models demonstrate strong generalization capabilities, allowing them to be fine-tuned for metric depth estimation tasks.
- Scalability: Models are available in various sizes (from 1.3B to 25M parameters), supporting a wide range of applications.

While Depth Anything V2 primarily focuses on relative depth estimation, pre-trained metric depth models are available for both indoor and outdoor environments. Our experiments utilized the indoor-tuned Depth-Anything-V2-Large model (335.3M parameters), the most accurate variant, without additional fine-tuning.

3.2 Hardware setup

The evaluation of Depth Anything V2 monocular depth estimation as a replacement for LiDAR was performed on the Voyager mobile robotic platform [23]. The robot was equipped with an Ouster OS0-32 high resolution imaging LiDAR with a depth accuracy of better than ±5 cm over a range of 35 m [24], and a FLIR Grasshopper3 4.1 Megapixel camera recording at the maximum resolution of 2048x2048.

The robot runs ROS2 Humble [25] and records the camera images and LiDAR point cloud to a rosbag [25] for offline processing. The camera and LiDAR are both mounted on the payload mount of the Voyager robot in such a way that their depth measurements can be easily compared using a simple coordinate transformation.

3.3 Camera calibration

The camera used in the study was calibrated to determine the intrinsic parameters for accurate data alignment [26], and to compute radial and tangential distortion coefficients for distortion correction [27]. The calibration process used a chessboard target and followed the plumb-bob camera model [28]. Using the computed calibration values, each image was rectified before further processing. The OpenCV functions *getOptimalNewCameraMatrix* and undistort [28] were applied to correct distortion and generate an optimized intrinsic matrix, which was used as the effective camera intrinsic matrix ($\overline{A}$) for all rectified images.

3.4 LiDAR-camera extrinsic calibration

The transformation between the camera and LiDAR coordinate frames was carefully calibrated to ensure precise alignment between the two sensors. The LiDAR measurements serve as ground-truth depth data, providing a reference for evaluating the performance of the monocular depth estimation model.

An open-source LiDAR-camera calibration toolbox [29] was used to perform the LiDAR-camera extrinsic calibration. Numerous rosbags were recorded in various settings, involving slow vertical movement of the robot to generate dense LiDAR point clouds through point cloud registration. Fig. 1 and Fig. 2 present examples of a corresponding camera image and a LiDAR intensity image used in the calibration.

Fig. 1: An example camera image used for the LiDAR-camera extrinsic calibration.

Fig. 2: An example corresponding image representing the LiDAR point intensities used for the LiDAR-camera extrinsic calibration.

3.5 Data collection

Datasets were generated by recording synchronized camera and LiDAR data in various environments. For each dataset, the robot was placed in 200 arbitrary positions within a room. Each camera-LiDAR data pair was captured while the robot was completely stationary to ensure precise synchronization between the camera and LiDAR and eliminating motion-induced misalignments. Objects in the robot's field of view remained stationary throughout the recording to maintain consistency.

Six datasets were recorded in different indoor locations. Fig. 3 shows labelled photos of the dataset locations. These locations feature diverse object types across various depth ranges. Depth points beyond 10 m are excluded, as they are sparse and typically irrelevant for the given scenes.

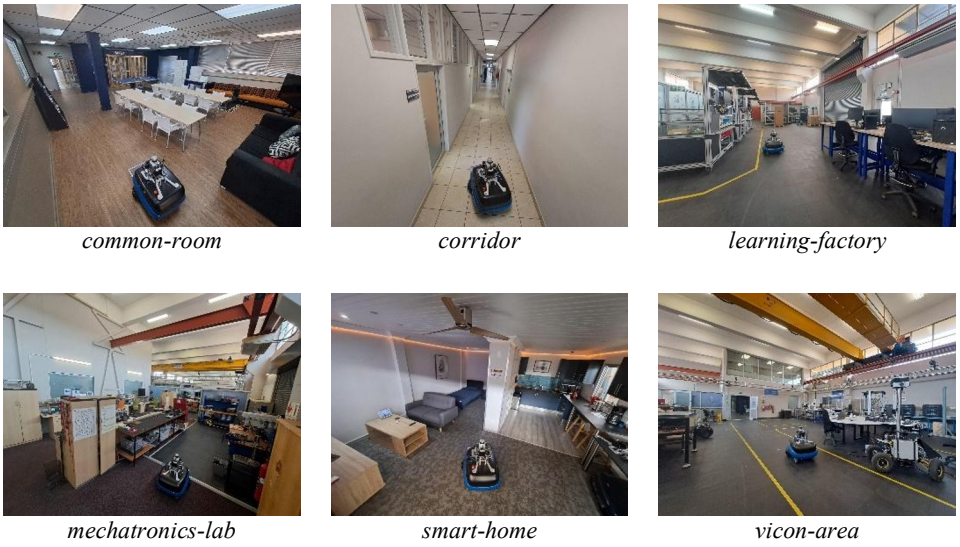
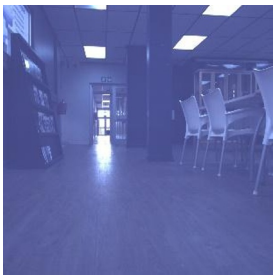


Fig. 3: Photos for the different locations used to record the datasets.

3.6 Colour histogram equalization

The FLIR Grasshopper camera used in this study does not apply any automatic colour balancing, which produces blue-dominated images. This has an unnatural appearance and is not representative of the data used to train the Depth Anything V2 model. Colour histogram equalization is a technique that redistributes the intensity values in an image's colour channels to achieve a more balanced and natural appearance. Fig. 4 compares an image from the NYU-D [30] dataset with an image from our common-room dataset, with and without colour histogram equalization applied. Our colour histogram equalized image, and the NYU-D image, both have a more natural colour balance than the blue-dominated raw image. All images from our FLIR Grasshopper camera are first colour histogram equalised before being processed.



Blue-dominated raw image from common-room

Colour histogram equalised image from common-room

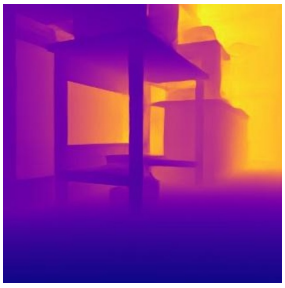
Image from the NYU-D dataset

Fig. 4: Comparison of an image from the NYU-D [30] dataset and images from our common-room dataset, with and without colour histogram equalisation.

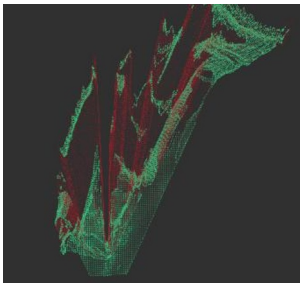
3.7 Depth image edge filter

In depth images produced by Depth Anything V2, edges exhibit a depth gradient instead of a discontinuity. This occurs due to a gradual transition in depth values over a small number of pixels, which may not be clearly visible in the depth image but become pronounced when projected into a 3D point cloud. These artifacts manifest as elongated lines radiating from the camera's principal point, leading to inaccurate 3D representations, and posing challenges for robotic perception. Fig. 5 illustrates this artifact, showing a depth image alongside two different viewing angles of the corresponding point cloud.

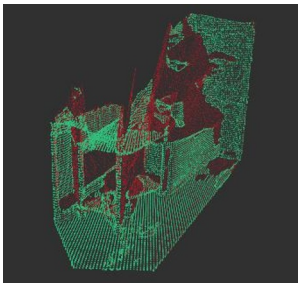
To address this, an edge filter is applied to remove affected pixels from the depth images. A pixel mask is generated using the Sobel operator in the OpenCV library [28] to detect edges in the depth images with a kernel size of 5 and magnitude threshold of 95%. The mask is dilated using a square kernel of size 20×20 to provide better coverage of the affected regions. Fig. 5 shows removed points in red and remaining points in green after applying this filter.



Depth image



Top-view of point cloud



Isometric view of point cloud

Fig. 5: Depth cloud with different views of the corresponding point cloud showing filtered points in red and remaining points in green

3.8 Performance evaluation

The performance of the depth estimations is evaluated by comparing them to ground-truth data. Many monocular depth estimation studies use depth images from depth cameras as ground truth [3,13,16]. With this approach, both RGB and depth images share the same coordinate system, making it straightforward to compare corresponding pixel values.

In this study, LiDAR data is used as the ground-truth source. Compared to depth cameras, LiDAR offers the advantage of greater accuracy over long distances and superior

performance in challenging lighting conditions. However, LiDAR has the disadvantage of lower spatial resolution compared to a camera. Despite this limitation, the multi-beam LiDAR used in this study provides an adequate number of ground-truth depth values to effectively evaluate depth estimation performance in a robotic context.

Another disadvantage is that LiDAR requires extrinsic calibration with the RGB camera to determine the transformation between their respective coordinate systems before comparisons can be made. Inaccuracies in this extrinsic calibration can lead to errors in aligning the LiDAR data with the RGB images. A method we use to account for misalignments is discussed below.

Points in the world frame (W) are initially measured in the LiDAR frame (L), and represented as a homogeneous coordinate, $\boldsymbol{p_L} = [x_L, y_L, z_L, 1]^T$. Each measured point is then transformed into the camera frame (C) using the extrinsic LiDAR-camera transformation matrix, $\boldsymbol{p_C} = T_{LC}\, \boldsymbol{p_L}$. In the camera frame, any measured point that falls outside the camera's field of view (FOV) can be disregarded, as it will not correspond to any pixels in the image.

The point is then projected into the image frame using the rectified camera intrinsic matrix $\boldsymbol{p_I} = [x_I, y_I, z_I] = \bar{A}\, \boldsymbol{p_C}$. Next, the point coordinates are normalized to obtain the pixel coordinates on the image plane, $u = x_I/z_I$ and $v = y_I/z_I$. The ground-truth depth value z_C can then be compared to the estimated depth value at the pixel coordinates $[u, v]$ in the image.

As previously discussed, slight misalignments between the ground-truth depth values and corresponding pixel coordinates may arise due to small errors in the extrinsic LiDAR-camera calibration. To mitigate this, each ground-truth depth value is compared to the estimated depths within a 7x7 pixel window, with the value closest to the ground-truth being selected as the corresponding depth estimation. The depth estimation error for each pixel corresponding to a LiDAR measurement can now be calculated as:

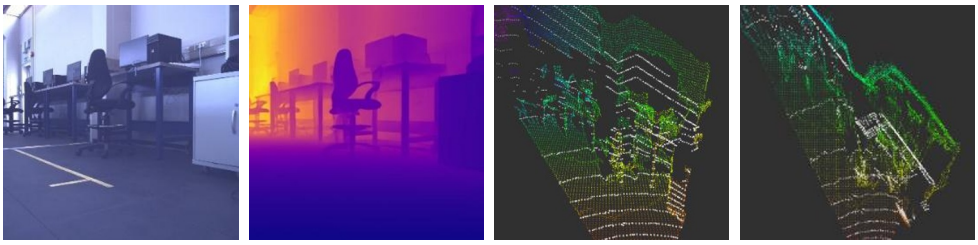
$$error = estimated\ depth\ -\ ground\text{-}truth\ depth$$

4 Qualitative results

This section presents qualitative analyses of the performance of Depth Anything V2 as a potential LiDAR alternative. Key strengths and limitations are highlighted through visual comparisons and analysis of depth consistency, structural similarity, and depth accuracy of different feature. The focus is on assessing the model's ability to estimate depth in different scenes, including edge-case conditions. The qualitative visual assessment aims to provide a comprehensive understanding of how the monocular depth estimations compare to LiDAR measurements in various real-world robotics scenarios.

4.1 Various types of features in scene

Qualitative comparisons provide valuable insights into the monocular depth estimation performance by visually demonstrating how the inferred depth maps align with the actual scene structures. This highlights strengths and limitations that could affect real-world robotic applications and cannot be fully captured by quantitative metrics alone.

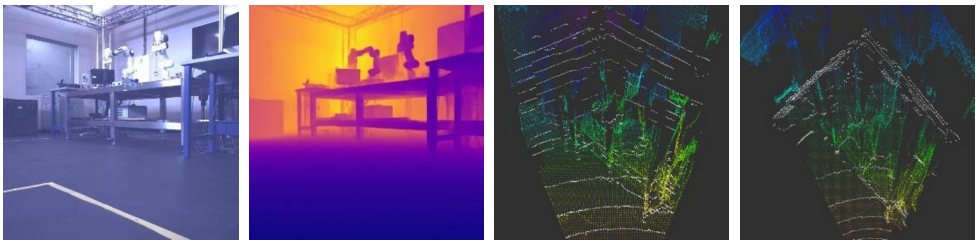


a. Rectified image *b. Depth estimation* *c. LiDAR compared to estimated depth cloud* *d. LiDAR compared to estimated depth cloud*

Fig. 6: Comparison of LiDAR and estimated depth data for a scene from learning-factory dataset.

Fig. 6 contains image tiles comparing LiDAR data and depth estimation data from a scene in the learning-factory dataset. The image in Fig. 6a shows the distortion-free image that was rectified using the camera calibration parameters, and Fig. 6b shows the resulting depth image estimated by the Depth Anything v2 model. Fig. 6c and Fig. 6d show different view angles of the LiDAR and depth estimation point cloud data in the same coordinate frame. The white dots represent the LiDAR point readings, and the colourful dots represent the estimated depth image pixels projected as a 3D point cloud with a colour scale based on depth values.

Fig. 6 shows how the estimated depth cloud consistently aligns with the general structure of the real-world scene. Note from Fig. 6d that the pillar against the wall in the centre of the image frame is wrongly estimated as being further away than the LiDAR readings, even though the structure is correctly represented. However, the estimation accurately represents the rest of the large geometric features like the floor, wall, and cabinet, with depth values that follow the contours of the LiDAR data. Conversely, the structure of the smaller features like the chairs and tables are not captured well enough to identify them in the depth cloud, even though each of these features are still represented by a few estimated points which should be enough for a robot to register their position as an obstacle.



a. Rectified image *b. Depth estimation* *c. LiDAR compared to estimated depth cloud* *d. LiDAR compared to estimated depth cloud*

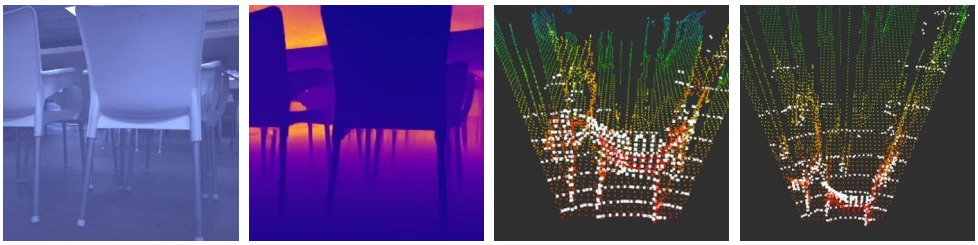
Fig. 7: Comparison of LiDAR and estimated depth data for a scene from vicon-area dataset.

Fig. 7 shows data from a scene in the vicon-area dataset. In this comparison the structure of the walls is clearly represented by the estimated depth cloud, even though the metric depth is overestimated. However, the smaller features like the table surface and legs are estimated more accurately. Furthermore, it can be noted the complex objects like the desktop robot arms and items on the table, are represented badly in the estimated depth cloud with inconsistent and wavy depth values, even though they can be recognised in the depth image. This could also be seen in Fig. 6c where the features on the chairs and computers were inaccurately estimated. These inconsistencies suggest that Depth Anything V2 prioritizes

global scene understanding over precise fine-grained depth estimation. Large, regular structures like walls, tables, and floors are well-represented in the model's training data, leading to better inference for similar shapes. In contrast, unique and intricate objects are less common in the training data, resulting in poorer depth estimation for these structures.

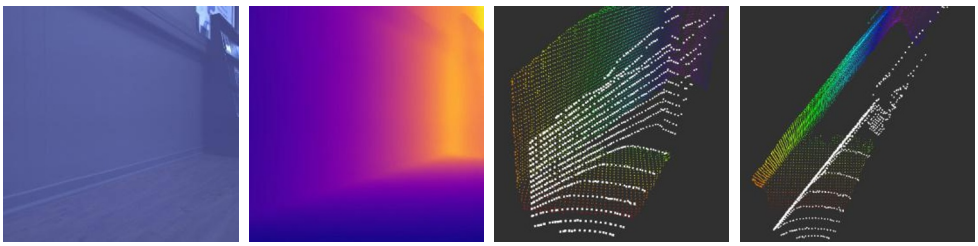
4.2 Close-range images

An edge-case to consider with depth estimation for robotics is the inference performance on close-range images. In robotics, close range depth estimation is import because it directly affects the robot's ability to navigate and avoid obstacles in indoor or cluttered environments. Fig. 8 shows the depth estimation performance of a scene from the common-room dataset. Fig. 8c and Fig. 8d show that the nearest two chairs are estimated accurately, and even the thin legs of the chairs are represented.



a. Rectified image *b. Depth estimation* *c. LiDAR compared to estimated depth cloud* *d. LiDAR compared to estimated depth cloud*

Fig. 8: Comparison of LiDAR and estimated depth data for a scene from common-room dataset.



a. Rectified image *b. Depth estimation* *c. LiDAR compared to estimated depth cloud* *d. LiDAR compared to estimated depth cloud*

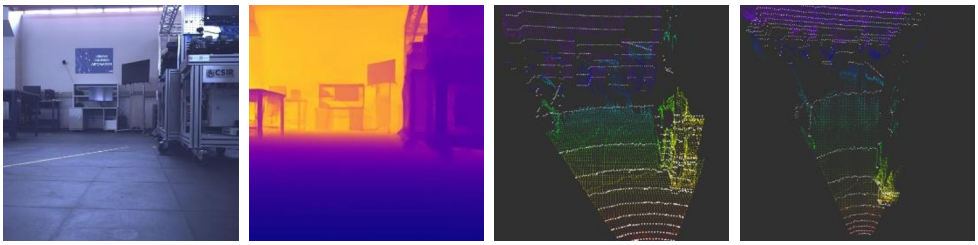
Fig. 9: Comparison of LiDAR and estimated depth data for a scene from common-room dataset.

In scenarios like approaching a wall, close-range images may contain fewer features, reducing the contextual information needed for accurate depth prediction. To mitigate this, robots may need feature detection mechanisms to recognize such situations or account for lower confidence in depth estimates. Fig. 9 shows depth data from a scene in the common-room dataset where the robot approaches a wall with limited features in the image frame. While the overall structure of the scene is captured, the depth estimation is notably inaccurate at close range.

4.3 Inconsistent performances for different scenes

Understandably, the Depth Anything V2 performance differs significantly for different scenes. This may be due to variations in the types of objects and lighting in the scene, as well as the distribution of similar data in the training set. Fig. 10c and Fig. 10d show an example

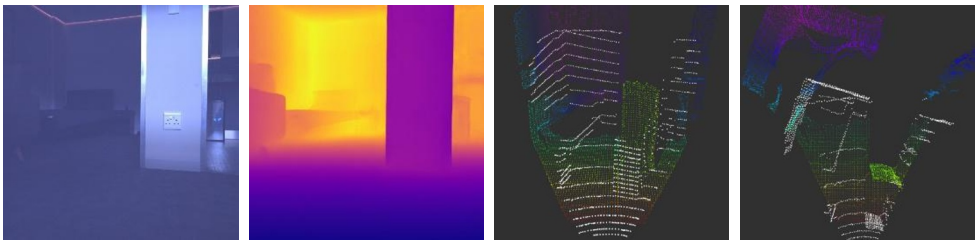
scene where the estimated depth cloud clearly matches the LiDAR scan for large and small structures.



a. Rectified image *b. Depth estimation* *c. LiDAR compared to estimated depth cloud* *d. LiDAR compared to estimated depth cloud*

Fig. 10: Comparison of LiDAR and estimated depth data for a scene from learning-factory dataset.

In contrast, Fig. 11 show a scene where the estimated depth cloud has a clear offset of scaling error compared to the LiDAR scan. Although the overall structure of the room is captured, the metric depth values are significantly overestimated. This could stem from factors such as low or high-contrast lighting. Regardless, the inconsistency in depth estimations across different scenes highlights the need for robots to adapt to variations in accuracy, if depth estimation models are to be used as LiDAR substitutes. Since Depth Anything V2 does not provide confidence values for individual inferences, robots also need to rely on alternative methods to assess depth accuracy for decision making algorithms.



a. Rectified image *b. Depth estimation* *c. LiDAR compared to estimated depth cloud* *d. LiDAR compared to estimated depth cloud*

Fig. 11: Comparison of LiDAR and estimated depth data for a scene from smart-home dataset.

5 Quantitative results

This section presents quantitative analyses of Depth Anything V2's performance as a LiDAR alternative, focusing on depth accuracy and consistency across different scenes. The results provide numerical insights into the model's reliability and suitability for robotics applications.

5.1 Error distribution

The normalized frequency distribution of depth estimation errors is a key consideration for practical robotics applications. Fig. 12 presents a histogram of depth estimation errors across all datasets, where each pixel associated with a ground-truth depth value constitutes a data point. This figure plots the normalized frequency of these errors as a probability density distribution such that the area under the curve equates to 1. The x-axis is divided into bins, each with a width of 0.1 m. The probability distribution has a positive skew, with 66.5% of the data on the positive side of the x-axis. This suggests a bias towards overestimating depths,

which poses a challenge for obstacle sensing applications, as objects may be measured as farther away than they are, potentially leading to collisions.

The analyses indicates that 16.5% of the data points fall within the [-0.1, 0.1] m error range, 66.5% within [-0.5, 0.5] m, and 87.0% within [-1, 1] m. The results suggest that the accuracy is sufficient for robotic navigation in expansive environments, where coarse accuracy is sufficient; however, it may be inadequate for high-precision navigation in confined spaces.

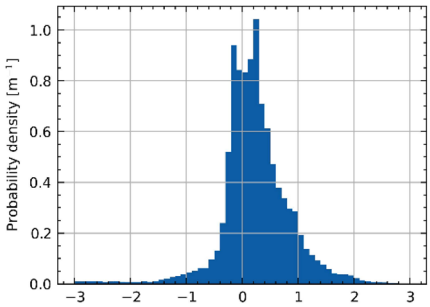


Fig. 12: Normalized frequency distribution of depth estimation errors across all datasets.

5.2 Accuracy at different depths

The accuracy of the metric depth estimation is influenced by the distance to a feature in the scene. Fig. 13 presents box-and-whisker plots comparing depth estimation errors to ground-truth depths for each dataset. The plots aggregate all estimation errors across the various scenes within each dataset, including outliers. The x-axis represents discrete ranges of ground-truth depths, while the y-axis shows the corresponding depth estimation errors ($error = estimated - actual$).

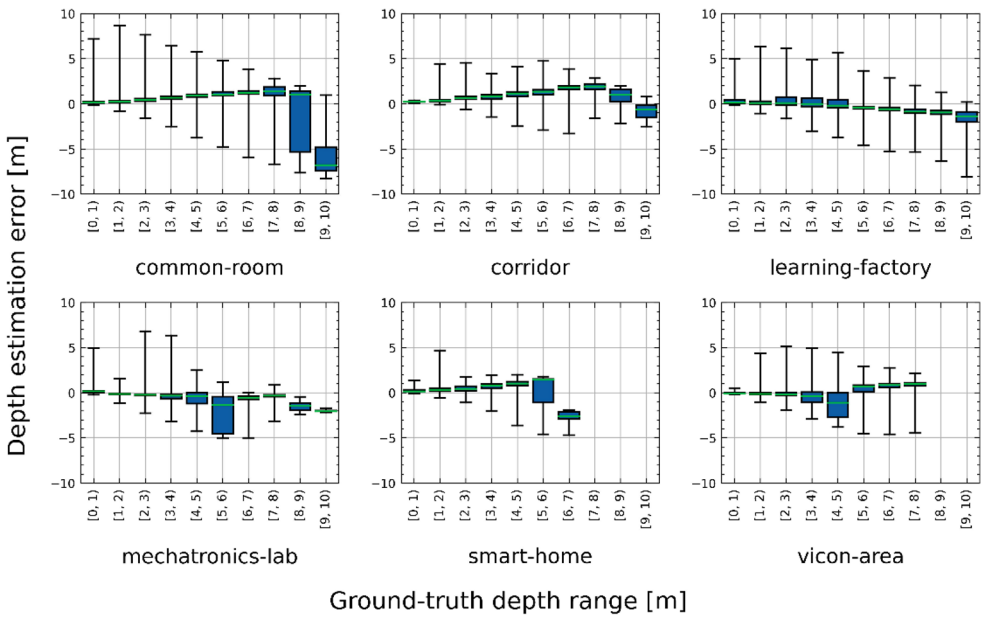


Fig. 13: Box-and-whisker plots of depth estimation errors for each dataset, grouped by ground-truth depth ranges.

Fig. 13 illustrates consistently small interquartile ranges (IQRs) across various depths and datasets, highlighting reliable performance. Exceptions are observed in certain box plots, where large IQRs appear at specific depths, likely due to specific features which are poorly estimated. The smallest median depth estimation errors are associated with low ground-truth distances, while both the median error and max-min range generally increase with larger ground-truth depths. Despite a few outliers, the median error remains close to zero across varying ground-truth depths, demonstrating consistent depth estimation accuracy.

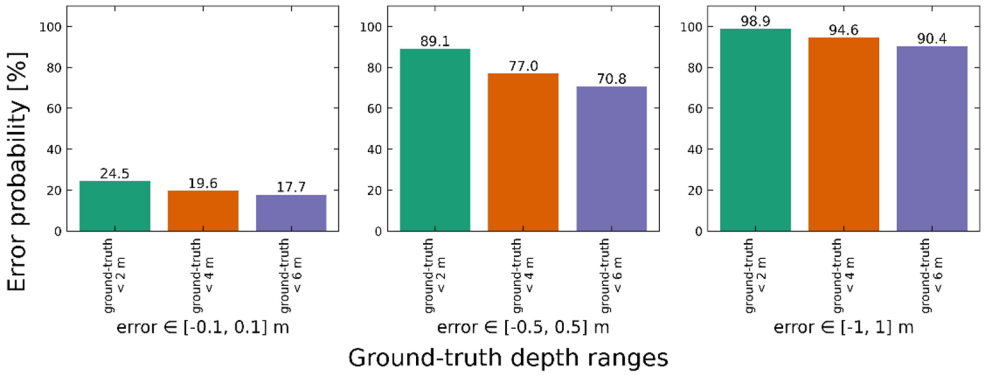


Fig. 14: Error probability distributions of depth estimation errors for different ground-truth depth ranges across all datasets.

Fig. 14 shows three bar graph subplots, with error probability on the y-axis and ground-truth depth ranges on the x-axis. From left to right, the subplots display error probability data for depth estimation errors in the ranges [-0.1, 0.1], [-0.5, 0.5], and [-1, 1] m, respectively. For instance, the bar in the leftmost subplot for the depth estimation range "ground-truth < 2 m" indicates that 24.5% of the error distribution falls within the [-0.1, 0.1] m range for points with a ground-truth distance of less than 2 m.

The figure shows that near objects have higher error probabilities across all error ranges compared to farther objects. For robotics applications requiring precise environmental sensing, the error probabilities are suboptimal, with only 24.5% confidence that an obstacle 2 m away is estimated within a ±0.1 m error range. However, for coarse environments, the accuracy may prove adequate, as there is 89.1% confidence that an obstacle 2 m away is observed within a ±0.5 m error range. For objects within 4 m, this decreases to 77.0% and for objects within 6 m, it decreases further to 70.8%. Improving these confidence values may be possible by combining depth estimations from multiple monocular cameras with different view angles, though this would introduce additional complexities.

5.3 Absolute relative error

The Absolute Relative Error (AbsRel) is a commonly used metric to evaluate the accuracy of depth estimation models, including Depth Anything. AbsRel, expressed as:

$$AbsRel = \frac{1}{N}\sum_{N}\frac{|\tilde{D}_i - D_i|}{D_i}, \tag{8}$$

measures the relative absolute difference between estimated ($\tilde{D}_i$) and ground-truth (D_i) depths, normalized by the ground-truth depth and averaged over all valid pixels [31]. By normalizing the error by ground-truth depth, AbsRel accounts for the tendency that depth estimation errors increase at larger distances. The AbsRel is calculated as 0.20 for the depth estimation performance across all the datasets recorded in this study.

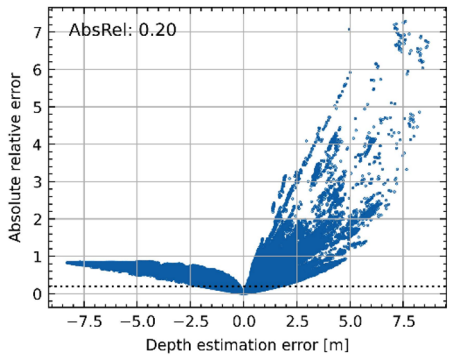


Fig. 15: Comparison of absolute relative error and depth estimation error across datasets.

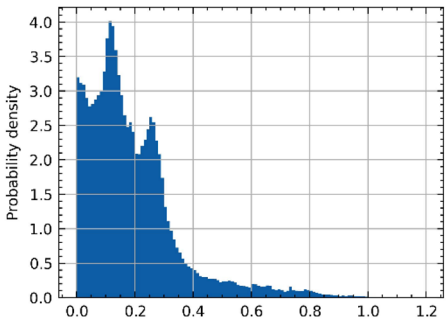


Fig. 16: Normalized frequency histogram of absolute relative error across datasets.

Fig. 15 presents a scatter plot comparing the absolute relative error on the y-axis with the depth estimation error on the x-axis for all measured points across each dataset. The AbsRel value, representing the mean of all absolute relative errors, is indicated by a horizontal line on the plot. In the negative x-direction, the absolute relative errors converge to 1. This behaviour stems from an inherent limitation in the AbsRel metric, where the numerator cannot exceed the denominator for negative depth estimation errors because the smallest possible depth estimation value is 0. Because overestimates can grow indefinitely, they can dominate the AbsRel mean metric, leading to a skewed representation of depth estimation performance.

The corresponding normalized frequency histogram in Fig. 16 shows the distribution of absolute relative error values for the same data. The y-axis represents the probability density, and the x-axis represents the absolute relative error values. This plot shows that overestimates do not dominate the AbsRel metric, since a very small proportion of the data lies beyond past AbsRel=1 on the x-axis.

5.4 Performance metrics

Fig. 17 presents a series of bar graph subplots, where the x-axis represents the datasets, and the y-axis of each subplot corresponds to a different error metric. Each subplot shows significant variation in the error metrics across datasets, highlighting the influence of different scene characteristics on depth estimation accuracy. The error standard deviation plot further emphasizes this variation, with the largest error standard deviation being 1.12 m, observed in the corridor dataset. The large standard deviation suggests that depth estimation is not consistently repeatable and not suitable for high-precision applications such as mapping or delicate navigation.

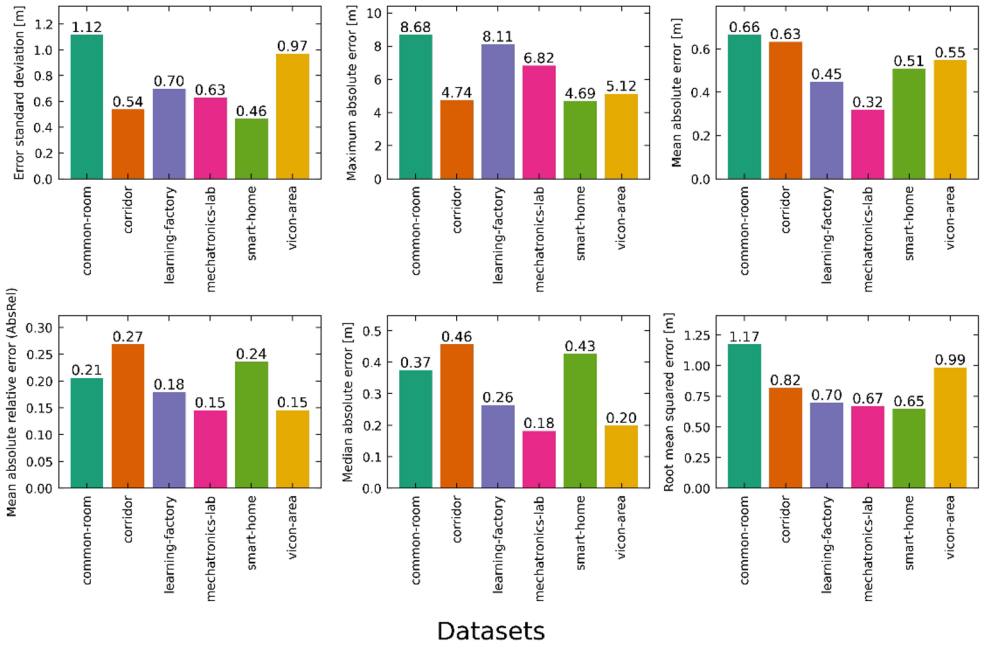


Fig. 17: Bar graph comparisons of depth estimation performance metrics across datasets.

The AbsRel metric also exhibits variation, ranging from a maximum of 0.27 for the corridor dataset to a minimum of 0.15 for both the mechatronics-lab and vicon-area datasets. In comparison, the original Depth Anything V2 paper reports an AbsRel of 0.045 [3] for the NYU-D [30] indoor dataset, which is significantly better than our results. While these results are not directly comparable due to differences in ground-truth sources (depth cameras versus LiDAR projections onto the image plane) the comparison offers insight into the relative performance of Depth Anything V2 on our datasets.

The common-room dataset exhibited the largest mean absolute error with a value of 0.66 m, a standard deviation of 1.12 m, and a maximum absolute error of 8.68 m. This is likely due to the complex and close-up images captured, including arrangements of multiple chairs and table legs, which pose a challenge for accurate depth estimation. The corridor dataset, on the other hand, showed the largest median absolute error of 0.46 m and a standard deviation of only 0.53 m. The low standard deviation may be attributed to the relatively uniform nature of the images in the long, narrow corridor, leading to less variation in depth estimation.

The mechatronics-lab dataset demonstrated the best performance, with a mean absolute error of 0.32 m, a median error of 0.18 m, and a standard deviation of 0.63 m. This is likely due to the simplicity of the scene, with predominantly large geometric objects such as cupboards and counters, making depth estimation more accurate compared to the intricate structures in the common-room dataset. Similarly, the vicon-area presents mostly simple and large features in the dataset and achieves comparative results with a median error of 0.20 m.

6 Conclusion

This study presents a comprehensive evaluation of Depth Anything V2 as a potential LiDAR alternative for robotics applications. Six datasets were recorded across diverse indoor scenes, and the performance of the pre-trained indoor metric depth model was assessed on this data. Both qualitative and quantitative analyses revealed key strengths and limitations of the model for depth sensing in robotics applications.

Qualitatively, Depth Anything V2 performs well in estimating the overall depth structure of large environments, accurately representing major features such as walls, floors, and tables. However, depth estimation accuracy varies significantly across scenes. The model generally captures the relative structure of a scene well but does not accurately represent the metric depth of features. Performance decreases for smaller or more intricate objects, resulting in less precise depth estimation for fine-grained structures. Close-range depth estimation also presents challenges, particularly in environments with fewer features, affecting the model's ability to generate accurate depth maps in confined spaces.

Quantitatively, the model's performance varies significantly across datasets, with simpler scenes featuring large geometric features yielding more consistent and accurate results compared to complex scenes. The best performance was observed in the mechatronics-lab dataset, with a mean absolute error of 0.32 m and a median absolute error of 0.18 m. In contrast, the worst-performing datasets had a mean absolute error of 0.66 m and a median absolute error of 0.46 m, highlighting the considerable impact of the environment on depth estimation accuracy.

While the depth estimation may not be accurate enough for robot navigation in confined indoor environments, it can still offer useful predictions in expansive scenes, where coarse accuracy is sufficient. Notably, 89% of errors for objects within 2 m are within a ±0.5 m range, indicating that the model can reliably estimate distances in such scenarios.

However, traditional cameras have a much smaller field of view than 360° LiDARs comprehensive spatial coverage. Moreover, the model’s metric depth accuracy in real-world datasets remains insufficient for applications requiring precise measurements, such as mapping or fined-grained navigation. Therefore, while Depth Anything V2 can serve as an alternative for coarse accuracy tasks, it cannot replace LiDAR where high-accuracy depth measurements are required.

Future work could explore using Depth Anything V2 to enhance, rather than replace, LiDAR in robotic 3D perception. By integrating Depth Anything V2 with LiDAR, the system could improve the resolution of local regions within the LiDAR point cloud, providing finer details where LiDAR's resolution is limited. In this hybrid approach, a lower-cost, lower-resolution 3D LiDAR could be employed, with Depth Anything V2 enhancing the effective resolution. This would offer a more comprehensive and accurate environmental representation, capitalizing on the strengths of both technologies. Furthermore, fine-tuning the Depth Anything V2 model with scene-specific data could improve its performance, increasing its potential as a LiDAR alternative in familiar environments.

References

1. U. Rajapaksha, F. Sohel, H. Laga, D. Diepeveen, M. Bennamoun, *Deep learning-based depth estimation methods from monocular image and videos: A comprehensive survey*, ACM Comput. Surv. **56**, 1-51 (2024).
2. Y. Ming, X. Meng, C. Fan, H. Yu, *Deep learning for monocular depth estimation: A review*, Neurocomputing **438**, 14-33 (2021).
3. L. Yang, B. Kang, Z. Huang, Z. Zhao, X. Xu, J. Feng, H. Zhao, *Depth anything V2*, arXiv preprint arXiv:2406.09414v2, (2024).
4. B. Alsadik, S. Karam, *The simultaneous localization and mapping (SLAM)-An overview*, J. Appl. Sci. Technol. Trends 2, 147-158 (2021).
5. A. Geiger, P. Lenz, C. Stiller, R. Urtasun, Vision meets robotics: *The kitti dataset*, Int. J. Robot. Res. **32**, 1231-1237 (2013).

6. M. Cordts, M. Omran, S. Ramos, T. Rehfeld, M. Enzweiler, R. Benenson, U. Franke, S. Roth, B. Schiele, *The cityscapes dataset for semantic urban scene understanding*, Proc. IEEE Conf. Comput. Vis. Pattern Recognit., 3213-3223 (2016).
7. N. Silberman, D. Hoiem, P. Kohli, R. Fergus, *Indoor segmentation and support inference from RGBD images*, in Comput. Vis. ECCV 2012, Lecture Notes in Computer Science, vol. **7576**, Springer, Berlin, Heidelberg, 746-760 (2012).
8. D. J. Butler, J. Wulff, G. B. Stanley, M. J. Black, *A naturalistic open source movie for optical flow evaluation*, in Comput. Vis. ECCV 2012, Lecture Notes in Computer Science, vol. **7577**, Springer, Berlin, Heidelberg, 611-625 (2012).
9. T. Schops, J. L. Schonberger, S. Galliani, T. Sattler, K. Schindler, M. Pollefeys, A. Geiger, *A multi-view stereo benchmark with high-resolution images and multi-camera videos*, Proc. IEEE Conf. Comput. Vis. Pattern Recognit., 3260-3269 (2017).
10. I. Vasiljevic, N. Kolkin, S. Zhang, R. Luo, H. Wang, F. Z. Dai, A. F. Daniele, M. Mostajabi, S. Basart, M. R. Walter, G. Shakhnarovich, *DIODE: A dense indoor and outdoor depth dataset*, arXiv preprint arXiv:1908.00463 (2019).
11. D. Barnes, M. Gadd, P. Murcutt, P. Newman, I. Posner, *The oxford radar robotcar dataset: A radar extension to the oxford robotcar dataset*, Proc. IEEE Int. Conf. Robot. Autom., 6433-6438 (2020).
12. R. Birkl, D. Wofk, M. Müller, *Midas v3.1--a model zoo for robust monocular relative depth estimation*, arXiv preprint arXiv:2307.14460 (2023).
13. L. Yang, B. Kang, Z. Huang, X. Xu, J. Feng, H. Zhao, *Depth anything: Unleashing the power of large-scale unlabeled data*, in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Seattle, WA, USA, pp. 10371-10381, (2024)
14. W. Yin, C. Zhang, H. Chen, Z. Cai, G. Yu, K. Wang, X. Chen, C. Shen, *Metric3d: Towards zero-shot metric 3d prediction from a single image*, Proc. IEEE/CVF ICCV, 9043-9053 (2023)
15. V. Guizilini, I. Vasiljevic, D. Chen, R. Ambruș, A. Gaidon, *Towards zero-shot scale-aware monocular depth estimation*, Proc. IEEE Int. Conf. Comput. Vis., 9233-9243 (2023).
16. S. Bhat, R. Birkl, D. Wofk, P. Wonka, M. Müller, *ZoeDepth: Zero-shot transfer by combining relative and metric depth*, arXiv preprint arXiv:2302.12288, (2023).
17. J. Spencer, F. Tosi, M. Poggi, R. S. Arora, C. Russell, S. Hadfield, R. Bowden, G. Zhou, Z. Li, Q. Rao, Y. Bao, *The third monocular depth estimation challenge*, Proc. IEEE Conf. Comput. Vis. Pattern Recognit., 1-14 (2024).
18. A. García, M. Díaz, F. Martínez, *From concept to application: building and testing a low-cost light detection and ranging system for small mobile robots using time-of-flight sensors*, Int. J. Electr. Comput. Eng. **15**, 1 (2025).
19. P. Venkatesan, N. Pavitra, R. Mohan, *Performing SLAM using Low-Cost Sensors for Autonomous Navigation in household environments*, Proc. IEEE Int. Conf. Converg. Technol., 1-5 (2019).
20. I. C. Condotta, T. M. Brown-Brandl, S. K. Pitla, J. P. Stinn, K. O. Silva-Miranda, *Evaluation of low-cost depth cameras for agricultural applications*, Comput. Electron. Agric. 173, 105394 (2020).
21. M. Kytö, M. Nuutinen, P. Oittinen, *Method for measuring stereo camera depth accuracy based on stereoscopic vision*, Proc. SPIE 7864, 168-176 (2011).

22. Y. You, C. P. Phoo, C. A. Diaz-Ruiz, K. Z. Luo, W. L. Chao, M. Campbell, B. Hariharan, K. Q. Weinberger, *Better Monocular 3D Detectors with LiDAR from the Past*, arXiv preprint arXiv:2404.05139 (2024).
23. K. Purdon, J. Dickens, W. de Ronde, K. Ramruthan, G. Crafford, *Voyager, a ground mobile robotic platform for research development*, in *Proc. 2023 RAPDASA-RobMech-PRASA-AMI Conf.*, (2023).
24. Ouster, *OS0: Ultra-Wide View High-Resolution Imaging Lidar Datasheet*, REV: 12/2024, (2024), Available: https://data.ouster.io/downloads/datasheets/datasheet-rev7-v3p1-os0.pdf [Accessed: 29 January 2025].
25. S. Macenski, T. Foote, B. Gerkey, C. Lalancette, W. Woodall, *Robot Operating System 2: Design, architecture, and uses in the wild*, Sci. Robot. **7** (2022).
26. Z. Zhang, *A flexible new technique for camera calibration*, IEEE Trans. Pattern Anal. Mach. Intell. **22**, 1330-1334 (2000).
27. J. Heikkilä, O. Silvén, *A four-step camera calibration procedure with implicit image correction*, in *Proc. IEEE Comput. Soc. Conf. Comput. Vis. Pattern Recognit.*, pp. 1106-1112, (1997)
28. G. Bradski, *The OpenCV Library*, Dr. Dobb's J. Softw. Tools (2000).
29. K. Koide, S. Oishi, M. Yokozuka, A. Banno, *General, single-shot, target-less, and automatic LiDAR-camera extrinsic calibration toolbox*, in *Proc. IEEE Int. Conf. Robotics Autom. (ICRA)*, pp. 11301-11307, (2023)
30. N. Silberman, D. Hoiem, P. Kohli, R. Fergus, *Indoor segmentation and support inference from RGBD images*, in *Computer Vision–ECCV 2012*, A. Fitzgibbon, S. Lazebnik, P. Perona, Y. Sato, C. Schmid, Eds., Lecture Notes in Computer Science, vol. 7576, Springer, Berlin, Heidelberg, pp. 746-760, (2012)
31. X. Dong, M. Garratt, H. Abbass, *Towards real-time monocular depth estimation for robotics: A survey*, IEEE Trans. Intell. Transp. Syst. **23**, 1-12 (2021).