TOLIE, H.F., REN, J., HASAN, M.J. and KANNAN, S. 2024. Enhancing underwater situational awareness: RealSense camera integration with deep learning for improved depth perception and distance measurement. In Bouma, H., Prabhu, R., Yitzhaky, Y. and Kuijf, H.J. (eds.) *Artificial intelligence for security and defence applications II: proceedings of the 2024 SPIE Security + defence*, 16-20 September 2024, Edinburgh, UK. Proceedings of SPIE, 13206. Bellingham, WA; SPIE [online], paper 1320605. Available from: https://doi.org/10.1117/12.3030972

Enhancing underwater situational awareness: RealSense camera integration with deep learning for improved depth perception and distance measurement.

TOLIE, H.F., REN, J., HASAN, M.J. and KANNAN, S.

2024

Enhancing Underwater Situational Awareness: RealSense Camera Integration with Deep Learning for Improved Depth Perception and Distance Measurement

Hamidreza Farhadi Tolie[a], Jinchang Ren[a], Md. Junayed Hasan[a], and Somasundar Kannan[b]

[a]National Subsea Centre, Robert Gordon University, Aberdeen, UK
[b]School of Computing, Engineering, and Technology, Robert Gordon University, Aberdeen, UK

ABSTRACT

This work presents a depth image refinement technique designed to enhance the usability of a commercial camera in underwater environments. Stereo vision-based depth cameras offer dense data that is well-suited for accurate environmental understanding. However, light attenuation in water introduces challenges such as missing regions, outliers, and noise in the captured depth images, which can degrade performance in computer vision tasks. Using the Intel RealSense D455 camera, we captured data in a controlled water tank and proposed a refinement technique leveraging the state-of-the-art Depth-Anything model. Our approach involves first capturing a depth image with the Intel RealSense camera and generating a relative depth image using the Depth-Anything model based on the recorded color image. We then apply a mapping between the Depth-Anything generated relative depth data and the RealSense depth image to produce a visually appealing and accurate depth image. Our results demonstrate that this technique enables precise depth measurement at distances of up to 1.2 meters underwater.

Keywords: Depth image, depth refinement, RealSense cameras, Depth-Anything

1. INTRODUCTION

With technological advancements and the availability of affordable depth cameras, 3D measurement has become significantly more straightforward. Unlike 2D vision, 3D vision provides accurate information about environments, particularly regarding object shapes, dimensions, and distances, which is crucial for real-world control. This capability is especially important for robotics, including underwater robotics for manipulation tasks. In the subsea world, where visibility and control are limited, utilizing depth and distance information from various sources is beneficial. Depth cameras offer precise environmental data, essential for robotic tasks such as object picking and obstacle avoidance.

Depth cameras have been widely used in various domains,[1–3] including manufacturing, robotics, healthcare, and agriculture,[4] to detect, recognize, track, and avoid objects/obstacles. To this end, depth cameras from various companies have been employed, but Intel cameras are particularly popular due to their affordability, compact size, and high performance, providing high frame rates and accurate depth perception.

In this study, our objective is to use an underwater depth camera to obtain precise 3D measurements of the subsea environment, thereby facilitating the guidance of remotely operated vehicles (ROVs) in object retrieval tasks that require centimeter-level accuracy. We focus on the utilization of Intel RealSense cameras, known for their high accuracy in distance measurement by emitting infrared (IR) light and capturing reflections. However, these cameras face limitations, such as the appearance of IR shadows, which create black regions in the depth images where distance measurement is not possible. Additionally, light attenuation in water further complicates the production of precise and dense depth maps, limiting the usability of depth cameras for constructing accurate 3D maps of the environment. To address these challenges, we explore a refinement technique to improve the

Further author information: (Send correspondence to J.R.)
H.F.T.: E-mail: h.farhadi-tolie@rgu.ac.uk
J.R.: E-mail: j.ren@rgu.ac.uk

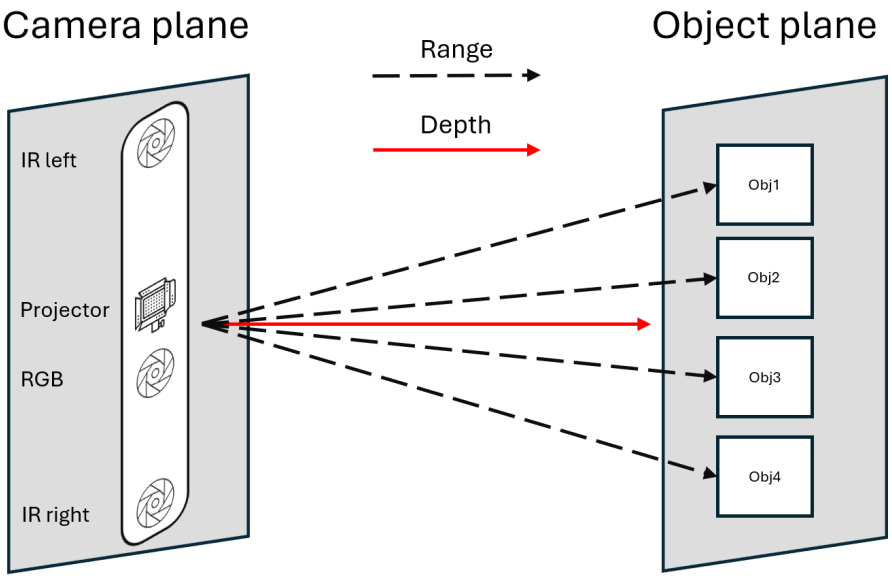

Figure 1. Schematic diagram of the depth and range (diagonal distance) from the camera plane to the objects.

captured depth images by filling in missing regions and removing noise. For our study, we use the Intel RealSense D455 camera,[5] which has an in air operating range of 0.6 to 6 meters.

The remainder of this paper is organized as follows: In Section 2, we explain the working pipeline of the RealSense camera and discuss current refinement methods developed for these cameras. Section 3 introduces our proposed approach. The experimental environment and results are discussed in Section 4.

2. RELATED WORK

In this section, we briefly overview the setting of the utilized RealSense camera and discuss the current state-of-the-art in depth image refinement both airborne and underwater.

2.1 Intel RealSense D455 camera

RealSense cameras include Stereoscopic depth sensors, red green blue (RGB) sensor with a resolution of 1280 × 800 and frame rate of 30 per second, and an IR projector. Using these three sensors, a depth map is generated by detecting the IR lights reflected from the objects in the scene. Compared with traditional stereo vision, using the IR lights enables the RealSense camera to even operate in low-light conditions, which fits the underwater purposes.

Technically, the depth image is formed using the stereo vision algorithms. The IR projector, emits invisible structured IR rays to the scene to improve the depth image calculation. Then using the stereo vision algorithms the correlation between each pixel of the left- and right-camera recorded images is calculated via a processor places in the camera module. Note that, the formed depth image determines the distance between the camera plane and the object plane not the diagonal distance to the objects.[6] As shown in Figure 1, although the diagonal distances to the objects vary, the depth distance to all four objects is same.

RealSense D455 camera has a very good accuracy for both indoor and outdoor environments, however, for underwater the performance drops significantly. This is mainly because the IR lights refracted in the water. In addition, as the camera is not water proof by itself, using an additional water housing adds extra refraction rate to the IR lights all resulting in noisy depth images with lots of missing regions.[7] As example, we have illustrated depth images captured from two scenes in our experimental water tank in Figure 2. As seen, the depth image contains outliers, noise, and missing regions making it hard to distinguish the shape and dimension of the objects.

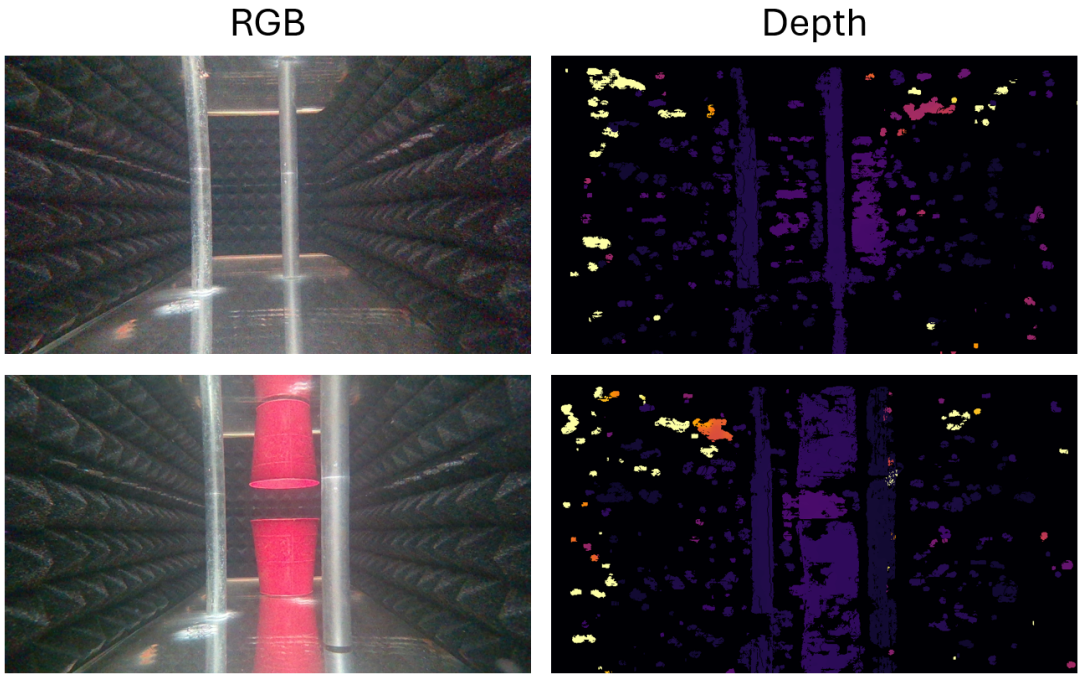

Figure 2. Sample RGB and depth images captured within the water tank.

2.2 Background

To eliminate noise and fill missing regions in captured depth images, various efforts have been made in recent years for both air and underwater applications. Most of these approaches correct depth images using information from adjacent pixels. However, since information in shadow regions is already lost, the accuracy of such refinements is often unreliable. Given the severity of these issues in underwater scenarios, adjacent pixel-based refinements are particularly ineffective. Therefore, we propose using both the RGB image and the captured depth image to achieve more accurate measurements. In this section, we discuss current depth image refinement techniques and RGB-based depth image estimation.

In an early study, Chen *et al.*[8] explored methods to improve the depth images captured by the Microsoft Kinect camera, addressing issues such as invalid pixels, noise, and mismatched edges. They proposed a region-growing method to first identify and remove erroneous depth values using the corresponding color image. Subsequently, they estimated the missing regions by incorporating a joint bilateral filter. Additionally, they employed an adaptive bilateral filter to reduce noise in the Kinect's depth images. In another attempt, Min *et al.*[9] introduced using weighted mode filtering based on a joint histogram for depth video enhancement. This method computes weights from color similarities to generate the histogram and determine a global mode, extended to temporally neighboring frames for consistency, addressing flickering issues. Later, in 2015, Matsuo *et al.*[10] proposed to use the local tangent planes instead of the local pixel-coordinates from the color image as they cannot effectively represent the measured distances from the surfaces. Matsuo *et al.* calculate the local tangents using a color heuristic and orientation correction, followed by surface reconstruction via ray-tracing to these tangents. More recently, Jun-Park and Baek-Kim[11] proposed to use the adjacent pixel information with conventional image processing techniques to fill the missing regions in Intel Realsense camera's depth images. In particular, this method refines the depth values by computing the average recorded depth value of the adjacent same-color objects.

However, applying these methods in underwater scenarios faces several challenges. These include issues with light scattering, color distortion, and the unique noise characteristics of underwater environments, which degrade the quality of both color and depth images. These conditions result in non-uniform color distributions and large missing regions, making the use of adjacent pixel information impractical.

With the emergence of automatic and hierarchical feature extraction in deep learning models,[12] several studies have focused on designing deep neural networks to optimize the mapping between input color and depth images for noise removal and filling in missing regions. Zhang and Wu[13] developed a convolutional deep neural network to train a pixel-wise generative model for depth image refinement and introduced a pre-processing step to enhance important edge areas, as missing regions often occur around edges. While this method performs well for depth image refinement in air, it struggles with underwater images due to the large portions of missing regions in depth images and the requirement for well-correlated color and depth inputs.

More recently, employing deep neural networks, various researchers have proposed new methodologies for improved stereo matching performance, resulting in more accurate depth images. Notable examples include StereoNet[14] and PSMNet.[15] StereoNet introduces a cost volume that encodes the matching cost of pixels from stereo images and utilizes a lightweight network architecture to efficiently produce disparity maps. PSMNet employs a pyramid structure to process images at different scales and a 3D convolutional network to capture spatial dependencies, refining depth estimation through both local and global contexts.

In recent works, to eliminate the need for stereo cameras, monocular depth estimation has become a hot topic, leading to the development of methods such as DenseDepth,[16] NDDepth,[17] and Depth-Anything.[18] DenseDepth uses an encoder-decoder network architecture with a hierarchical structure and multi-scale feature aggregation to refine depth predictions. NDDepth leverages a deep neural network to jointly learn depth and surface normals from single RGB images, using the relationship between normals and depth to improve overall estimation. Depth-Anything introduces a self-supervised learning approach, using geometric and photometric constraints to learn depth from unannotated images.

Although these monocular methods have been successful in producing high-quality depth images, they cannot provide absolute distance information, which is crucial for robotic manipulations. This limitation is particularly problematic in underwater environments, where these methods often fail to generate accurate depth images and frequently miss some objects. Therefore, we propose integrating these depth estimation methodologies with real recorded depth images to provide accurate, refined depth images capable of delivering absolute distance information.

3. METHODOLOGY

To explore the capabilities of the Realsense camera, we conducted experiments in a controlled underwater environment using the water tank shown in Figure 3. Since the Realsense camera is not waterproof, we attached it to the outside glass wall of the tank and placed target objects, such as a metal pipe and a bucket, at various distances inside the tank to study the camera's properties. As shown in Figure 2, the captured depth images suffered from noise and missing regions. To address this, we proposed utilizing the state-of-the-art depth estimation module, Depth-Anything, to generate a relative depth map. By using sample points from both the

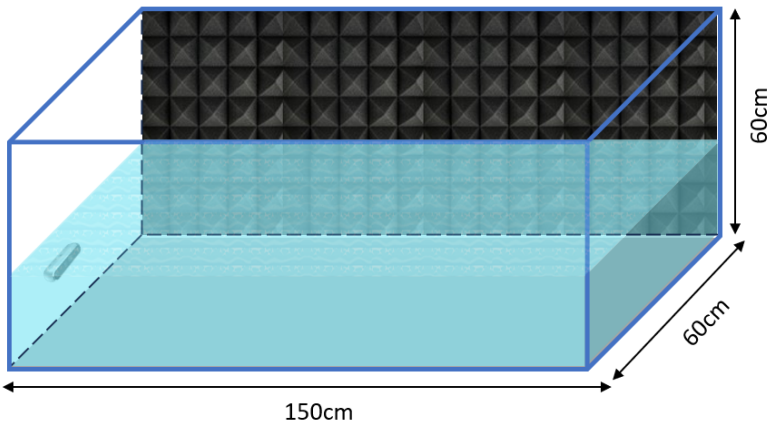

Figure 3. A schematic diagram of the testing environment within a water tank.

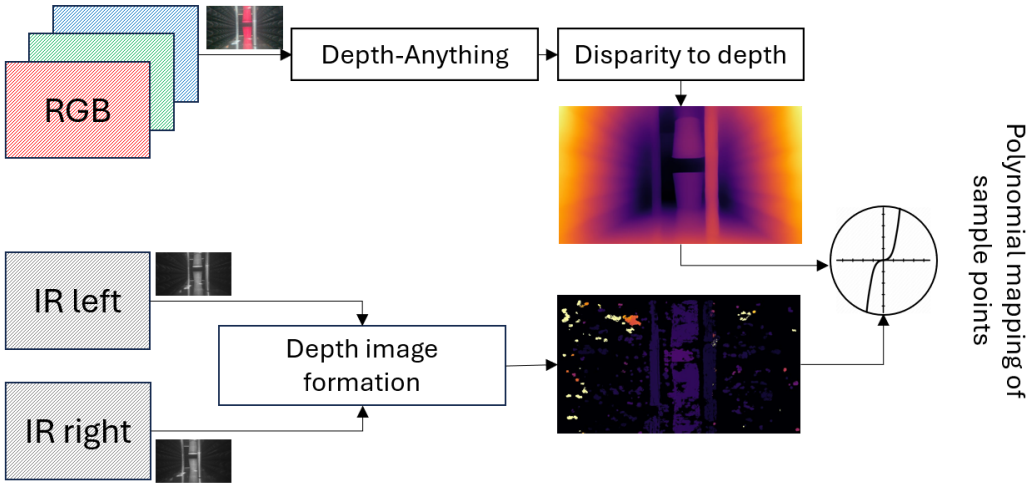

Figure 4. Framework of the proposed refinement strategy.

Realsense-recorded and Depth-Anything-generated images, we then found an appropriate mapping to estimate the missing regions and remove the noise.

The framework of the proposed refinement strategy is illustrated in Figure 4. As shown, we use the left and right IR cameras to record the scene and generate a depth image using the Realsense camera's technology. Additionally, we employ the recently proposed Depth-Anything method to generate a pseudo disparity image containing relative disparity information. However, since our objective is to refine the Realsense-generated depth image, we use the camera's baseline and focal length information to convert the pseudo disparity image into a relative depth image as follows.

$$\text{relative depth} = \frac{baseline * \text{focal length}}{\text{Pseudo disparity}} \tag{1}$$

where *baseline* and focal length equal to 95mm and 1.88mm, respectively, for the Realsense D455 camera.

As seen in Figure 4, the generated relative depth image is free from noise, and the shapes of the objects are clear. By establishing a proper mapping between the relative and absolute distance values from the generated relative depth image and the Realsense depth image, we can estimate the corresponding absolute depth values. To achieve this, we selected sample points with different intensity values from the relative depth image and their respective absolute values from the Realsense depth image. To minimize the error between the estimated and absolute depth values, we fitted a curve to the sample points, estimating the curve parameters using a polynomial function as follows.

$$f(x) = a_6x^6 + a_5x^5 + a_4x^4 + a_3x^3 + a_2x^2 + a_1x + a_0 \tag{2}$$

where $\sum_{i=1}^{6} a_i$ are the coefficient terms determining the influence of x^i on the formed curve, and a_0 is a constant term representing the value of the curve when x is 0. These coefficients are determined through data points collected from different sample images by solving a least squares problem.

4. IN LAB EXPERIMENT

To validate our refinement strategy, we conducted experiments in our water tank using two metal pipes and a bucket to capture depth images. These images were then used to evaluate the accuracy of our distance predictions compared to ground-truth distances. The reported distance values represent depth, which corresponds to the straight-line distance from the camera plane to the object plane, as illustrated in Figure 1.

As shown in Figure 3, the experiments took place within a glass water tank measuring 60 × 60 × 150 cm (height × width × length), with tank walls 1 cm thick. The water level reached 28 cm in height. To enhance reflection quality attributable to the tank's glass structure, acoustic foams were affixed to its interior walls and

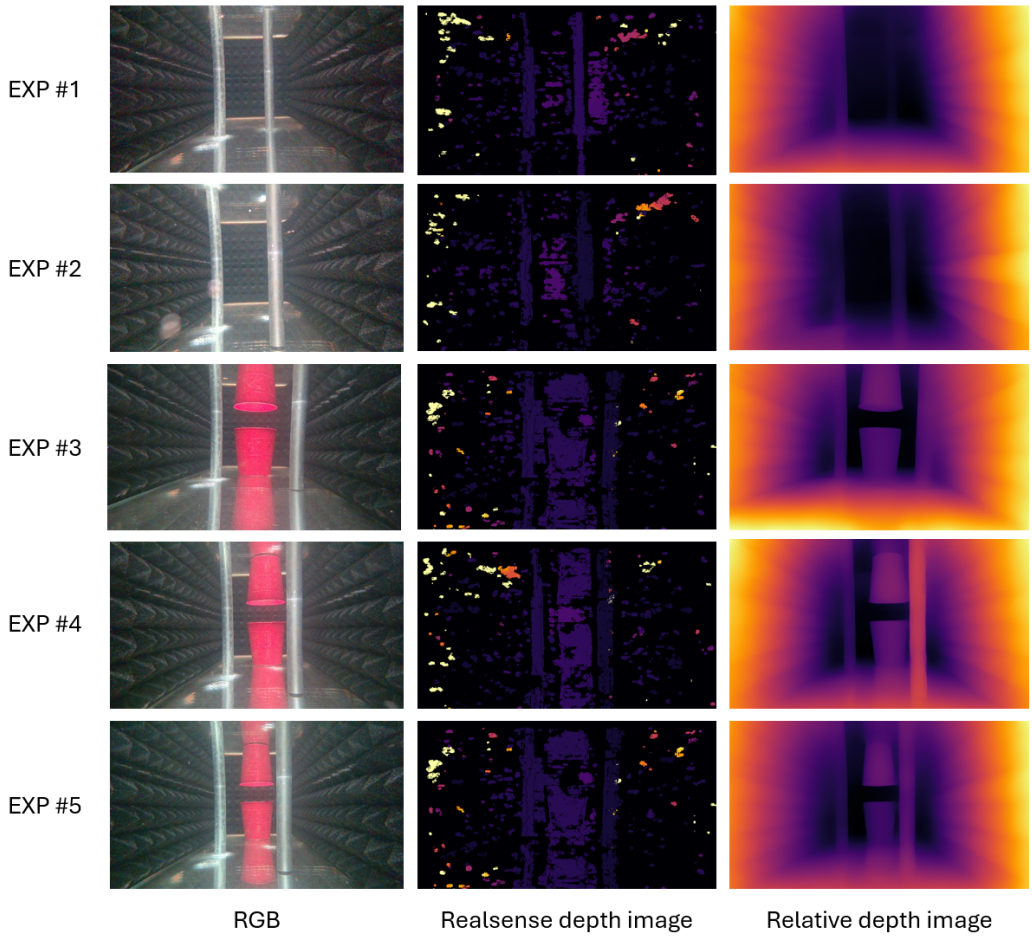

Figure 5. RGB, Realsense depth, and relative depth images taken through 5 experimental setting with objects at different distances.

the camera was attached to the exterior wall. Then, we have placed the objects at the distances of 60 cm to 120 cm. According to our experiments, we have observed the following:

- With the infrared technology used in the Realsense camera,it can capture depth images in low-light conditions.
- The minimum and maximum operational distance for the Intel Realsense camera in underwater setup is 60 cm and 120 cm, respectively.
- The accuracy in depth measurement in underwater environment drops approximately by 30%, thus it is required to compensate it by a factor of 1.30.
- According to our experiments, the error rate decreases when the objects is located in distant and gets to less than 1% of the ground-truth distance.
- The measured distances are dependant to shape of the object and the viewpoint of the camera.
- Using the refined depth images, more visually pleasent 3D images can be reconstructed from the underwater scene.

In Figure 5, we present the RGB image, the depth image captured by the Realsense camera, and the relative depth image generated using the Depth-Anything model. It can be observed that while the Realsense depth

image lacks detailed information about the objects in the environment, the generated relative depth image effectively highlights the target objects. Therefore, by applying our proposed refinement technique—mapping relative depth values to absolute depth values—and considering that depth refers to the distance from the camera plane to the object plane, we can provide detailed depth information about the target objects. We have reported the measured ground-truth and measured depth values to the objects in each experimental scenario in Table 1.

Table 1. Distance measurement results of uniform pipe at various distances.

Experiment	Object	Ground-truth depth	Compensated measured distance
EXP #1	Left pipe	84.0cm	83.7cm
EXP #1	Right pipe	97.5cm	98.2cm
EXP #2	Left pipe	84.0cm	83.7cm
EXP #2	Right pipe	61.5cm	62.5cm
EXP #3	Left pipe	84.0cm	83.7cm
EXP #3	Right pipe	64.8cm	63.8cm
EXP #3	Bucket	92.5cm	93.0cm
EXP #4	Left pipe	84.0cm	85.0cm
EXP #4	Right pipe	60.7cm	59.9cm
EXP #4	Bucket	104.3cm	105.3cm
EXP #5	Left pipe	84.0cm	83.7cm
EXP #5	Right pipe	60.7cm	59.9cm
EXP #5	Bucket	120.5cm	119.6cm

Based on the results, we observed that the Intel Realsense D455 camera can measure depth information with an absolute error of 1 cm in this experimental setup. To illustrate this, we have plotted the relative and absolute error values in Figure 6. The results show that the minimum error occurs for objects placed within the [65, 100] cm distance range, indicating that mounting the camera within this range on a robotic arm is ideal for reducing error in robotic manipulations. Additionally, the relative error plot shows that as the distance between the objects and the camera increases, the relative error compared to the ground truth decreases, demonstrating the camera's potential usability.

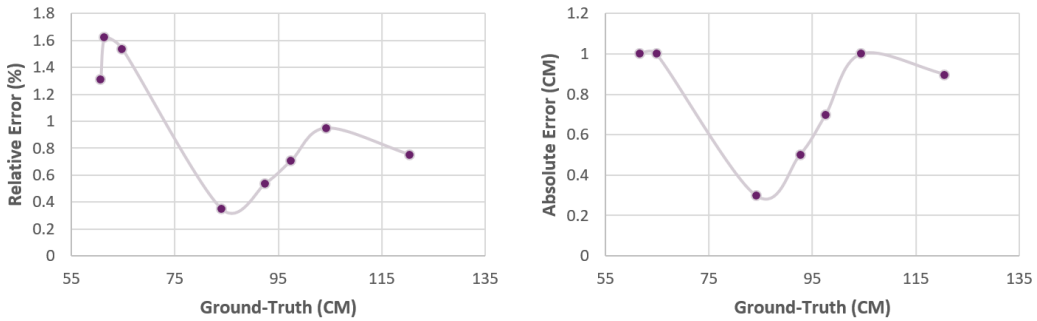

Figure 6. Relative and absolute depth measurement error diagrams.

5. CONCLUSION

In conclusion, we have evaluated the usability of depth cameras in underwater environments and explored the potential of the Intel Realsense D455 camera for depth/distance measurement in such scenarios. Given the limitations of stereo vision underwater, we proposed using a state-of-the-art depth estimation model, Depth-Anything, to generate relative depth images. These images were then refined using recorded depth values from

the Realsense camera through a polynomial curve fitting process. Our lab experiments demonstrated that with this refinement technique and by applying a compensation rate to the measured depth values, the Realsense camera can be effectively used for robotic manipulations within a range of 60 to 120 cm. The results validate the effectiveness of the proposed strategy in producing visually accurate depth images.

Looking ahead, we plan to explore more advanced refinement methodologies, taking into account that depth is defined as the distance from the camera plane to the object plane. Specifically, we aim to enhance[19] and segment the RGB image to identify objects and then use the recorded depth values to assign a consistent distance value to the pixels corresponding to each object. To achieve this, we can either utilize existing segmentation methods[20] or apply change detection approaches[21] that incorporate temporal information from the recorded images integrated with directional guided filters.[22]

ACKNOWLEDGMENTS

This work is partially supported by the SeaSense project (SPARK-2294), funded by the Net Zero Technology Centre, UK.

REFERENCES

[1] Carey, N., Werfel, J., and Nagpal, R., "Fast, accurate, small-scale 3d scene capture using a low-cost depth sensor," in [*2017 IEEE winter conference on applications of computer vision (WACV)*], 1268–1276, IEEE (2017).

[2] Rusu, R. B., Marton, Z. C., Blodow, N., Dolha, M., and Beetz, M., "Towards 3d point cloud based object maps for household environments," *Robotics and Autonomous Systems* **56**(11), 927–941 (2008).

[3] El-Sayed, E., Abdel-Kader, R. F., Nashaat, H., and Marei, M., "Plane detection in 3d point cloud using octree-balanced density down-sampling and iterative adaptive plane extraction," *IET Image Processing* **12**(9), 1595–1605 (2018).

[4] Tadic, V., Toth, A., Vizvari, Z., Klincsik, M., Sari, Z., Sarcevic, P., Sarosi, J., and Biro, I., "Perspectives of realsense and zed depth sensors for robotic vision applications," *Machines* **10**(3), 183 (2022).

[5] Intel RealSense, "Intel realsense d455." https://www.intelrealsense.com/depth-camera-d455/.

[6] Tadic, V., Odry, A., Kecskes, I., Burkus, E., Kiraly, Z., and Odry, P., "Application of intel realsense cameras for depth image generation in robotics," *WSEAS Transac. Comput* **18**, 2224–2872 (2019).

[7] Digumarti, S. T., Chaurasia, G., Taneja, A., Siegwart, R., Thomas, A., and Beardsley, P., "Underwater 3d capture using a low-cost commercial depth camera," in [*2016 IEEE Winter Conference on Applications of Computer Vision (WACV)*], 1–9, IEEE (2016).

[8] Chen, L., Lin, H., and Li, S., "Depth image enhancement for kinect using region growing and bilateral filter," in [*Proceedings of the 21st International Conference on Pattern Recognition (ICPR2012)*], 3070–3073 (2012).

[9] Min, D., Lu, J., and Do, M. N., "Depth video enhancement based on weighted mode filtering," *IEEE Transactions on Image Processing* **21**(3), 1176–1190 (2012).

[10] Matsuo, K. and Aoki, Y., "Depth image enhancement using local tangent plane approximations," in [*Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*], 3574–3583 (2015).

[11] Park, H. J. and Kim, K. B., "Depth image correction for intel realsense depth camera," *Indones. J. Electr. Eng. Comput. Sci* **19**, 1021–1027 (2020).

[12] Nejad, A., de Haan, G. A., Heutink, J., and Cornelissen, F. W., "Ace-dnv: Automatic classification of gaze events in dynamic natural viewing," *Behavior Research Methods* , 1–15 (2024).

[13] Zhang, X. and Wu, R., "Fast depth image denoising and enhancement using a deep convolutional network," in [*2016 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*], 2499–2503 (2016).

[14] Khamis, S., Fanello, S., Rhemann, C., Kowdle, A., Valentin, J., and Izadi, S., "Stereonet: Guided hierarchical refinement for real-time edge-aware depth prediction," in [*Proceedings of the European conference on computer vision (ECCV)*], 573–590 (2018).

[15] Chang, J.-R. and Chen, Y.-S., "Pyramid stereo matching network," in [*Proceedings of the IEEE conference on computer vision and pattern recognition*], 5410–5418 (2018).

[16] Alhashim, I. and Wonka, P., "High quality monocular depth estimation via transfer learning," *arXiv e-prints* **abs/1812.11941** (2018).

[17] Shao, S., Pei, Z., Chen, W., Wu, X., and Li, Z., "Nddepth: Normal-distance assisted monocular depth estimation," in [*Proceedings of the IEEE/CVF International Conference on Computer Vision*], 7931–7940 (2023).

[18] Yang, L., Kang, B., Huang, Z., Xu, X., Feng, J., and Zhao, H., "Depth anything: Unleashing the power of large-scale unlabeled data," in [*Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*], 10371–10381 (2024).

[19] Tolie, H. F., Ren, J., and Elyan, E., "Dicam: Deep inception and channel-wise attention modules for underwater image enhancement," *Neurocomputing* **584**, 127585 (2024).

[20] Hasan, M. J., Elyan, E., Yan, Y., Ren, J., and Sarker, M. M. K., "Segmentation framework for heat loss identification in thermal images: Empowering scottish retrofitting and thermographic survey companies," in [*International Conference on Brain Inspired Cognitive Systems*], 220–228, Springer (2023).

[21] Li, Y., Ren, J., Yan, Y., Ma, P., Assaad, M., and Gao, Z., "Abbd: Accumulated band-wise binary distancing for unsupervised parameter-free hyperspectral change detection," *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing* **17**, 9880–9893 (2024).

[22] Tolie, H. F., Faraji, M. R., and Qi, X., "Blind quality assessment of screen content images via edge histogram descriptor and statistical moments," *The Visual Computer* **40**(8), 5341–5356 (2024).