Pose Estimation from Camera Images for Underwater Inspection

Luyuan Peng, *Member, IEEE,* Hari Vishnu, *Senior Member, IEEE,* Mandar Chitre, *Senior Member, IEEE,* Yuen Min Too, *Member, IEEE,* Bharath Kalyan, *Senior Member, IEEE,* Rajat Mishra, *Member, IEEE* and Soo Pieng Tan

Abstract

High-accuracy localization is essential for underwater reinspection missions, which often require revisiting sites with complex structures for inspection and maintenance. Traditional localization systems—such as inertial navigation, Doppler velocity logs, and acoustic positioning—frequently fall short in accuracy or cost-effectiveness for these tasks. Machine learning-based visual relocalization presents a promising alternative, estimating poses from monocular images captured by onboard cameras using models trained on data from previous deployments. In this work, we evaluate the performance of such learning-based estimators in both clear and turbid water environments, examining the effects of color information, model architecture, and training data diversity. We further propose a novel view synthesis-based strategy to augment training data, enhancing pose estimation at previously unseen viewpoints. Finally, we improve localization robustness by fusing pose estimates with additional sensor inputs via an extended Kalman filter, resulting in smoother and more accurate trajectories.

Index Terms

underwater, localization, marine robotics, novel view synthesis, deep learning, sensor fusion.

I. Introduction

LOCALIZATION plays a crucial role in underwater reinspection missions [1]. These are tasks carried out by underwater vehicles to examine the health and function of submerged structures like pipelines, offshore platforms, and ship hulls, required to ensure the safety and durability of infrastructure vital to industries like oil and gas, renewable energy, and maritime transport [2], [3]. They stand apart from many underwater navigation tasks in their complexity and the precision required. Unlike general

L. Peng and M. Chitre are with the Acoustic Research Laboratory, Department of Electrical and Computer Engineering, National University of Singapore, Singapore 119227 (e-mail: luyuan@nus.edu.sg; mandar@nus.edu.sg).

H. Vishnu, B. Kalyan, R. Mishra and S. P. Tan are with the Acoustic Research Laboratory, Tropical Marine Science Institute, National University of Singapore, Singapore 119222 (e-mail: harivishnu@gmail.com; bharath@nus.edu.sg; rajat@nus.edu.sg; soopieng@nus.edu.sg).

Y. M. Too is with Subnero Pte. Ltd, Singapore 159836 (email: too@subnero.com).

underwater navigation that involves moving from place to place, often prioritizing pathfinding and obstacle avoidance, reinspection missions demand detailed, close-range examination of often complicated underwater structures [3]. As such, reinspection missions require the precise positioning and orientation of underwater vehicles to ensure thorough coverage, accurate data collection and the safety of the vehicles and the structures themselves.

In underwater environments, the use of global positioning systems is hindered due to the rapid dissipation of electromagnetic waves in water [4]. Traditionally, underwater localization has relied on inertial navigation systems (INS), Doppler velocity loggers (DVL) and acoustic positioning systems. However, these methods face significant challenges in the context of inspection missions. Acoustic navigation is often compromised by shadowing effects and multipath interference near marine structures, which can severely distort signal paths and reduce accuracy. Consequently, achieving precise acoustic navigation requires complex and costly setups [5]. Furthermore, INS and DVL, despite their widespread use, suffer from an accumulation of errors over time [5]. This limits their ability to provide the positioning accuracy required for detailed inspection of underwater structures. Although high-grade INS and DVL may be able to provide sufficient accuracy, they, too, come with high costs.

In recent years, advancements in underwater localization have explored the use of optical sensors, such as cameras [6]. Some of these approaches necessitate the deployment of active markers [6], [7] or elaborate setups by divers [8], [9], adding complexity and expense. In contrast, visual localization methods—estimating camera poses from images of the surrounding scene—present a more cost-effective solution. Since inspection vehicles typically come equipped with cameras, visual-based localization can be implemented without the need for extra hardware. Moreover, visual-based localization methods, such as simultaneous localization and mapping (SLAM) [5], visual odometry [10]–[14] and visual relocalization [15], [16], have shown promise in navigating terrestrial and underwater environments.

Underwater reinspection missions typically involve the vehicle returning to the same sites for routine monitoring, assessment and/or maintenance. In this sense, these missions have another difference from normal underwater navigation tasks in that they have prior information of the scene or environment available, i.e., the environment is “known” to some degree after the first mission. We can use this available prior information to perform relocalization. This approach can be made effective if in the initial mapping run, we collect positioning information as accurately as possible using precise (and typically expensive and complex) positioning infrastructure such as ultra-short baseline acoustic positioning to characterize the

environment. Using these data collected, visual relocalization methods can directly estimate poses from camera images in the following runs, significantly reducing the cost and complexity of reinspection. While SLAM and visual odometry are effective for general navigation, they do not utilize the additional prior information available in reinspection missions. In contrast, visual relocalization uses prior information and thus allows us to use more affordable vehicles and setups for localization in subsequent reinspection missions, significantly simplifying operations.

Visual relocalization techniques are categorized into feature-based methods such as Active Search [17], and deep-learning methods like PoseNet [15]. Active search achieves image-based localization by systematically identifying and matching 2D features in query images with 3D points in a scene model. In contrast, PoseNet is a deep learning model that utilizes a pretrained convolutional neural network (CNN) to directly regress the 6-degree-of-freedom (6-DOF) camera pose from images, bypassing the need for feature extraction and matching.

While Active Search has demonstrated state-of-the-art results in structured terrestrial environments, its reliance on salient features [15] and high computational cost [18] limits its suitability for underwater scenes, which are often characterized by sparse features and obscured textures due to poor visibility. Learning-based regressors inspired by PoseNet offer an appealing alternative for these environments by enabling fast and robust estimations.

Numerous methods have since been developed by building upon the PoseNet framework. For example, DSAC [19] combines learning-based regression with differentiable RANSAC for robust and accurate localization, though at significantly higher computational cost. VidLoc [20] extends PoseNet by modeling temporal dependencies across video sequences using recurrent networks. EffLoc [21], in contrast, retains a single-image input at inference but uses a transformer-based encoder to capture long-range spatial dependencies. MapNet [22] introduces additional geometric constraints during training by incorporating sensor-derived relative poses to improve consistency and robustness.

While these methods achieve impressive results in terrestrial applications, they all share PoseNet as a foundational architecture. In this work, we focus on PoseNet-based architectures for underwater relocalization due to their lower computational cost and simpler training requirements. Previous research has demonstrated PoseNet's efficacy in conducting inspection tasks within tanks with toy structures and simulated underwater environments [16], [23], [24]. However, the performance of machine learning-based pose estimators with realistic structures and in at-sea environments has not been thoroughly investigated.

A key limitation of learning-based pose estimators is their reliance on diverse and comprehensive training data. In underwater environments, collecting such data is costly and labor-intensive. To address this, we propose the use of Novel View Synthesis (NVS) models to generate augmented training data from limited original samples. Recent advancement in NVS models, such as Neural Radiance Fields (NeRF) [25] and 3D Gaussian Splatting (3DGS) [26], can synthesize photorealistic views of complex 3D scenes from a sparse set of input views by optimizing an underlying continuous volumetric scene function. When provided with a camera pose, NVS models utilize classical volumetric rendering techniques to project synthesized colors and densities into an image [25]. Using a trained NVS model, we can render images from any viewpoint within the boundary, allowing us to bypass the need for extensive physical data collection. We can then use these rendered images to augment our training data.

In this paper, our contributions are as follows:

1) We examine the performance of neural-network based pose estimators with different configurations in inspection missions in confined waters. We investigate the effects of different parameters, such as using RGB information versus grayscale, on the performance. We present the dataset collected, methods employed and results obtained in Section II.
2) We propose a new loss function, d-loss, incorporating the geometry of the inspection missions for training the pose estimators. The d-loss provides interpretability, and improves computation efficiency and estimation performance. We present the method and results in Section II.
3) We utilize underwater 3D NVS techniques to generate augmented training data. We demonstrate the performance improvement due to this in Section III.
4) We enhance the localization performance by integrating our pose estimation model with data from additional sensors, such as altimeters and compasses. We use an extended Kalman filter (EKF) for tracking and fusion. In Section IV, we present these methods and results showing improved robustness and accuracy of this approach.
5) We evaluate the performance of our proposed methods in at-sea environments. We present these results and discuss the overall performance of the entire pipeline in Section V.

Finally, we conclude this paper in Section VI.

II. Pose Estimation

Nielsen et al. [23] evaluated the performance of PoseNet in a small tank, inspecting a subsea connector attached to a metal stick. In our previous work, we assessed the performance of various pretrained CNNs

as pose estimators in a simulated underwater environment inspecting a subsea pipe [24]. In this section, we evaluate the performance of visual localization using two neural-network model architectures inspired from PoseNet [15]. The data for training and testing were collected from an artificial ocean basin at the Technology Center for Offshore and Marine, Singapore (TCOMS) [27]. The originally presented PoseNet [15] works on RGB images. Here, we also evaluate the visual localization performance using grayscale images instead of RGB images to determine if similar accuracy can be achieved with higher efficiency, based on the intuition that underwater images typically have limited color information. Finally, we investigate the models' capability for (1) estimating pose on test images from the same dataset (i.e., capability to interpolate within same dataset), and (2) their capability to generalize to datasets outside that used for training, by using data from different runs for training and testing, which have different paths and conditions during acquisition.

A. Methods

1) Architecture: The objective of PoseNet is to estimate a 6-DOF pose from a single monocular RGB image given as input to a neural network. The pose consists of the position (in 3D coordinates, x-y-z) and the orientation, which is represented in terms of a quaternion. Thus, the model outputs a 7-dimensional (7D) estimated pose vector $\mathbf{y} = [\hat{\mathbf{p}}, \hat{\mathbf{q}}]$ containing a position vector estimate $\hat{\mathbf{p}}$ and an orientation vector estimate $\hat{\mathbf{q}}$, where $\hat{}$ represents an estimate.

The PoseNet model originally presented by Kendall et al [15] was a CNN, a modified version of the GoogLeNet architecture [28] pretrained on the ImageNet dataset [29], with the softmax classifiers changed to affine regressors, and another fully connected (FC) layer of feature size 2048 inserted before the final regressor. However, regressing a 7D pose vector from a high dimensional output of the FC layer is not optimal [30]. A later work aimed to tackle this by modifying PoseNet by reshaping the FC layer of size 2048 to a 32×64 matrix and applying four long-short-term-memory networks (LSTMs) to perform structured dimensionality reduction [30]. This algorithm, which we refer to as CNN+LSTM, showed a performance improvement compared to PoseNet in terrestrial environments [30], and also in an underwater tank environment [16]. We implement and evaluate both model architectures – the CNN (shown in Fig. 1) and the CNN+LSTM (shown in Fig. 2). Additionally, we assess the performance of these using a pretrained ResNet50 [31] as the backbone.

2) Loss Function: Kendall et al [15] used a composite loss function that is a weighted sum of the (1) L2 loss $\mathcal{L}_{\mathbf{p}}$ between the predicted positions and the true positions, and the (2) L2 loss $\mathcal{L}_{\mathbf{q}}$ between the

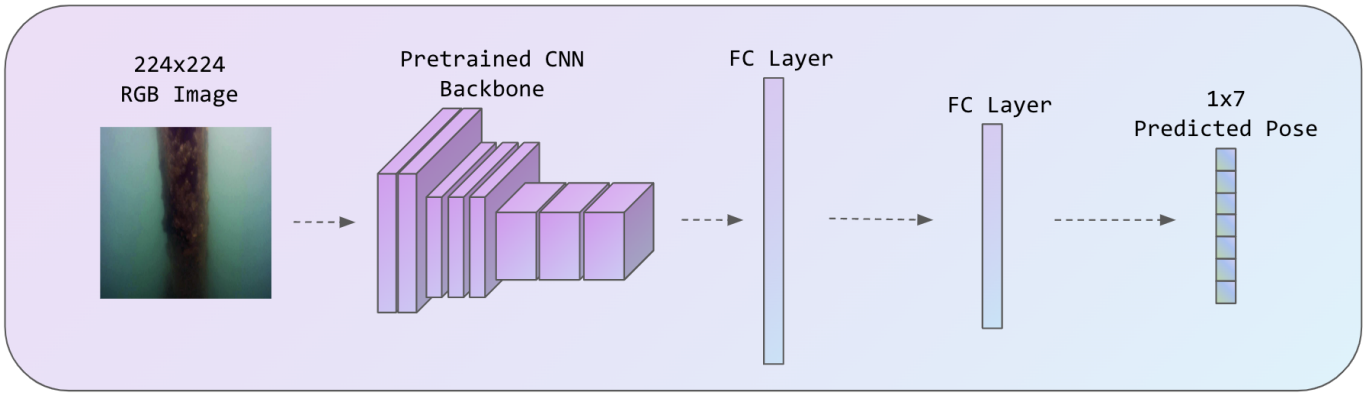


Fig. 1: Overview of the CNN-based architecture for visual localization.

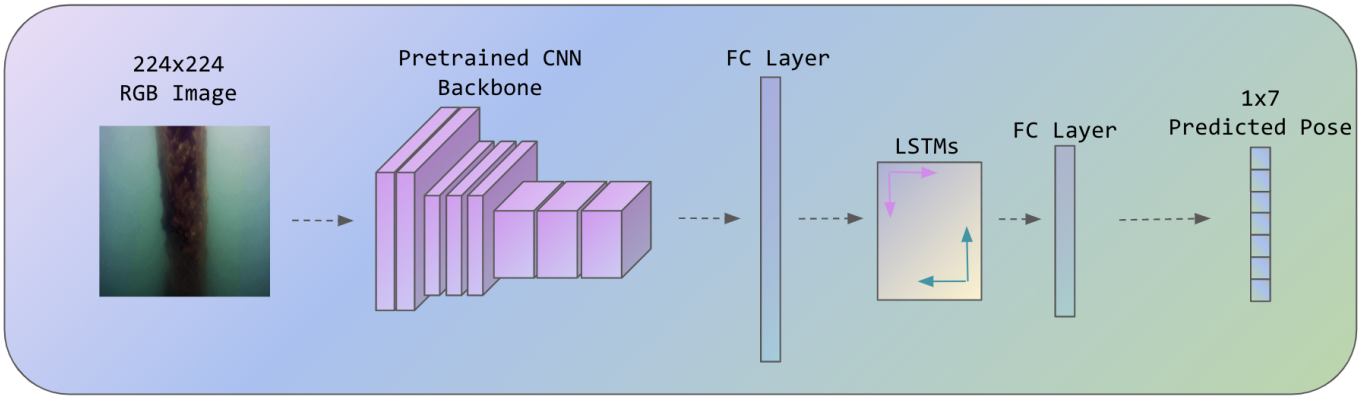


Fig. 2: Overview of the CNN+LSTM-based architecture for visual localization

predicted quaternions and the true quaternions:

$$\mathcal{L} = \mathcal{L}_{\mathbf{p}} + \beta \mathcal{L}_{\mathbf{q}}, \tag{1}$$

where $\mathcal{L}_{\mathbf{p}} = ||\mathbf{p} - \hat{\mathbf{p}}||_2$ and $\mathcal{L}_{\mathbf{q}} = ||\mathbf{q} - \hat{\mathbf{q}}/\|\hat{\mathbf{q}}\||_2$, and $\mathbf{p}$ and $\mathbf{q}$ represent the true pose. β is a free parameter that determines the trade-off between the desired accuracy in translation and orientation. In PoseNet and CNN+LSTM, the value of β is fine-tuned using a grid search to ensure the expected value of position and orientation errors are approximately equal, which the authors suggest lead to overall optimal performance. We refer to this loss function as the β-loss.

We argue that the β-loss is not the optimal approach to our problem, due to three reasons. Firstly, we argue that optimal performance is not necessarily achieved when position and orientation errors are roughly equal. Instead, the performance criteria and loss should incorporate geometry and physics relevant to the inspection task at hand. Secondly, the L2 loss between the predicted and true quaternions does not directly translate to an orientation error interpretable in degrees or radians, and thus, it does not accurately reflect

the geometric distance between the predicted and true orientations. Thirdly, searching for the optimal β value often involves extensive computational resources. This search can become a significant bottleneck, especially in scenarios where training needs to be done fast.

To overcome these shortcomings, we propose a new loss function more relevant to our problem, the d-loss, to improve the training effectiveness, interpretability and efficiency. The d-loss is defined as:

$$\mathcal{L} = \mathcal{L}_{\mathbf{p}} + d\mathcal{L}_{\theta}. \tag{2}$$

Note that we have replaced the quaternion loss in (1) with a loss based on the Eulerian angular difference, $\mathcal{L}_{\theta}$, which is calculated as follows. We first determine the rotation between the estimated and ground truth quaternions through quaternion multiplication, $\mathbf{\Delta q} = \mathbf{q}\,(\hat{\mathbf{q}}/\|\hat{\mathbf{q}}\|)^{*}$, where * denotes the conjugate of the quaternion. $\mathbf{\Delta q}$ is a unit quaternion which can be expressed as $(r, \vec{v})$ where r is the scalar part of the quaternion, and $\vec{v}$ is the vector part. r is related to a spatial rotation around a fixed point of $\mathcal{L}_{\theta}$ radians about a unit axis by the relation $r = \cos(\mathcal{L}_{\theta}/2)$ [32], thus $\mathcal{L}_{\theta} = 2\cos^{-1}(r)$. We approximate $\mathcal{L}_{\theta} \approx \frac{\pi}{2}(1 - r)$, using a Taylor series approximation, which is valid for small rotation angles where $r \approx 1$. For large angle deviations, this approximation becomes less accurate, and using the exact formula $L_{\theta} = 2\cos^{-1}(r)$ is recommended. In our use case, however, the model is trained to minimize pose errors, and angular deviations remain small in practice. The approximation also offers computational efficiency for model training. The Eulerian angular difference loss provides a more intuitive and direct measure of orientation error.

Additionally, we replace the hyperparameter weight factor β in (1) which required tuning, with the average distance d between the camera and the object of interest. In our experiments, d is computed based on prior knowledge of the inspection setup. In more typical deployments, d can be estimated using onboard sensors such as forward-looking sonars, or stereo depth estimation. The intuition here is that this factor translates the rotational error to an equivalent "average" translational error (attributed to the orientation difference). Thus, the overall loss can be interpreted as the "total positional error" in meters, including contributions from translational and orientation error components. Note that this formulation relies on several assumptions typical of underwater reinspection scenarios. It assumes that the vehicle maintains a relatively constant distance d from the structure during inspection, and that the camera is generally oriented toward the target (i.e., the bearing is aligned). We also assume the orientation errors are small enough for the small-angle approximation to hold. These conditions are commonly met in many

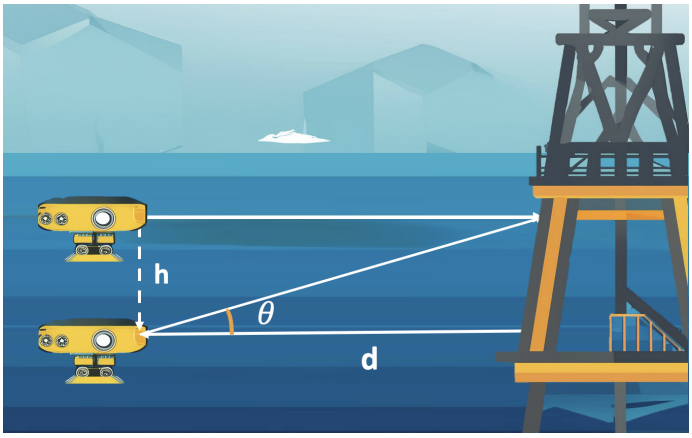


Fig. 3: Schematic showing the interpretation of the orientation error in terms of equivalent translational error. This assumes slow motion, small angles, and constant range.

kinds of underwater inspection tasks, where deliberate movements are required for safety and image quality.

The translation between rotational error and the "average" translational error is described as follows. As illustrated in the example in Fig. 3, if the camera has a pitch orientation error $\mathcal{L}_\theta$ of θ, the point it observes on the structure remains roughly the same as if the camera had an equivalent translational error $\mathcal{L}_\mathbf{p}$ of h (i.e., moves up by h) for small values of h and θ. Based on the geometry, equivalent translational error can be expressed in terms of orientation error $\mathcal{L}_\theta$ and the average horizontal range between the camera and the structure as:

$$\mathcal{L}_\mathbf{p} = d\tan(\mathcal{L}_\theta). \tag{3}$$

Assuming the case when the rotational error is small, we approximate $\tan(\mathcal{L}_\theta) \approx \mathcal{L}_\theta$. Thus, we obtain:

$$\mathcal{L}_\mathbf{p} \approx d\mathcal{L}_\theta. \tag{4}$$

This modified loss function (2) leverages the inherent geometric relationship between positional and rotational errors in inspection missions. By converting orientation error into an equivalent translational error using the physical distance d, both terms are expressed in the same unit (meters). This provides a more interpretable loss function with physical meaning, and avoids the need for manually tuning trade-off weights like β, simplifying the scaling challenge in pose estimation.

3) Implementation: To evaluate the effectiveness of deeper backbones, additional LSTM layers, the proposed d-loss, and the color information in images, we tested multiple configurations of the two visual

localization network architectures. The details of these configurations are summarized in Table I.

TABLE I
DESCRIPTION OF CONFIGURATIONS

ID	Architecture	Backbone	Loss	Color
C1	CNN	GoogLeNet	β-loss	RGB
C2	CNN	GoogLeNet	d-loss	Grayscale
C3	CNN	GoogLeNet	d-loss	RGB
C4	CNN	ResNet50	d-loss	RGB
C5	CNN+LSTM	GoogLeNet	d-loss	RGB
C6	CNN+LSTM	ResNet50	d-loss	RGB

During both training and testing for all configurations, we rescaled input images directly into a 224×224 pixels input, deviating from PoseNet's approach of resizing the images to 256×256 before cropping into 224×224. This adjustment was made to minimize the loss of image information, a concern particularly acute in underwater images where available information is inherently more limited compared to terrestrial settings. To speed up training, we normalized the images against the ImageNet dataset's mean and standard deviation. Additionally, poses are normalized to lie within the range [-1, 1].

As part of our investigation into color information, we explored the use of grayscale input to reduce input dimensionality and potentially improve computational efficiency, under the assumption that underwater images often contain limited useful color information due to turbidity and poor lighting. To preserve the benefits of transfer learning, we adapted a pretrained GoogLeNet model, which is originally designed for RGB input, to accept grayscale input. This was done by modifying the first convolutional layer to accept a single-channel input instead of three channels. The weights were initialized by summing across the RGB channels of the pretrained filters, and the modified layer was fine-tuned during training. This modification reduced the number of parameters in the first convolutional layer by a factor of three, since it now operates on a single channel instead of three.

We used the PyTorch deep learning framework to implement and train the models. The experiments were conducted using an RTX 6000 Ada GPU. For training, we used the stochastic gradient descent optimizer for configurations C1, C2, and C3. For the remaining configurations, we used the Adam optimizer. A batch size of 32 was used. Hyperparameters, including the learning rate, weight decay, and β for C1, were tuned using grid search strategy over a predefined set of values. The best set of hyperparameters was selected based on validation performance. Training continued until early stopping was triggered.

B. Testing in Controlled Environment

The artificial ocean basin at TCOMS is an indoor pool measuring 60 m $\times$ 48 m $\times$ 12 m. As illustrated in Fig. 4, a structure was placed in the basin, which consisted of six piles interconnected by metallic pipes, with each pile comprising three metallic oil barrels. The overall dimensions of the structure were approximately 3.9 m $\times$ 4.6 m $\times$ 3.0 m. The whole structure was yellow in color. To better differentiate the barrels, duct tape strips of various colors with different patterns were stuck on each barrel to create uniquely identifiable features.

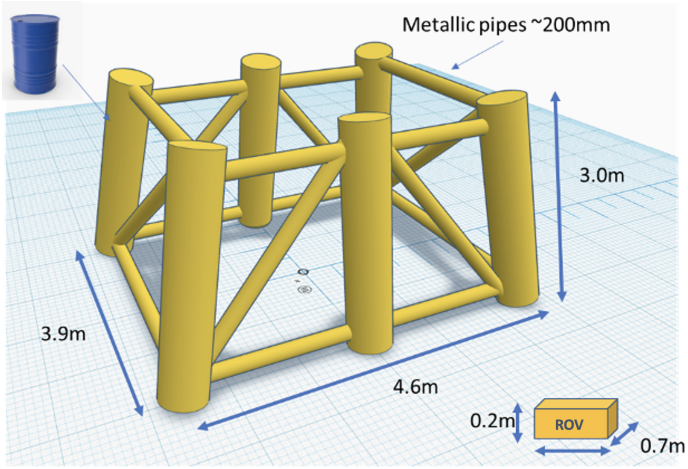
We used a customized remotely operated vehicle (ROV) based on the BlueROV2 platform. The ROV was equipped with a 1080p monocular camera provided by Blue Robotics for collecting RGB image data. For navigation and sensor integration, the vehicle used the BlueROV2 Navigator Flight Controller (NFC) mounted on a Raspberry Pi 4. The NFC includes an onboard IMU, compass, depth sensor, and Analog-to-Digital Converter sensors. The compass and depth sensor were used to provide orientation and depth information, respectively. To estimate the horizontal (x-y) position of the ROV, we employed a customized ultra-short baseline (USBL) positioning system. This setup was based on a Subnero high-speed acoustic modem (model WNC-S40HSS4+xCh) configured with four receivers and deployed near the operating region, as illustrated in Fig. 5. The USBL system enabled accurate localization of the vehicle during the trials in the TCOMS basin.

We executed three trials within the environment at different depths to gather data while the ROV surveyed the structure. Each trial features a roughly similar lawnmower trajectory around the structure, with a total path length of approximately 37 meters per trial. The trials were conducted at average depth levels of -1.5 m, -3 m, and -4 m, respectively.

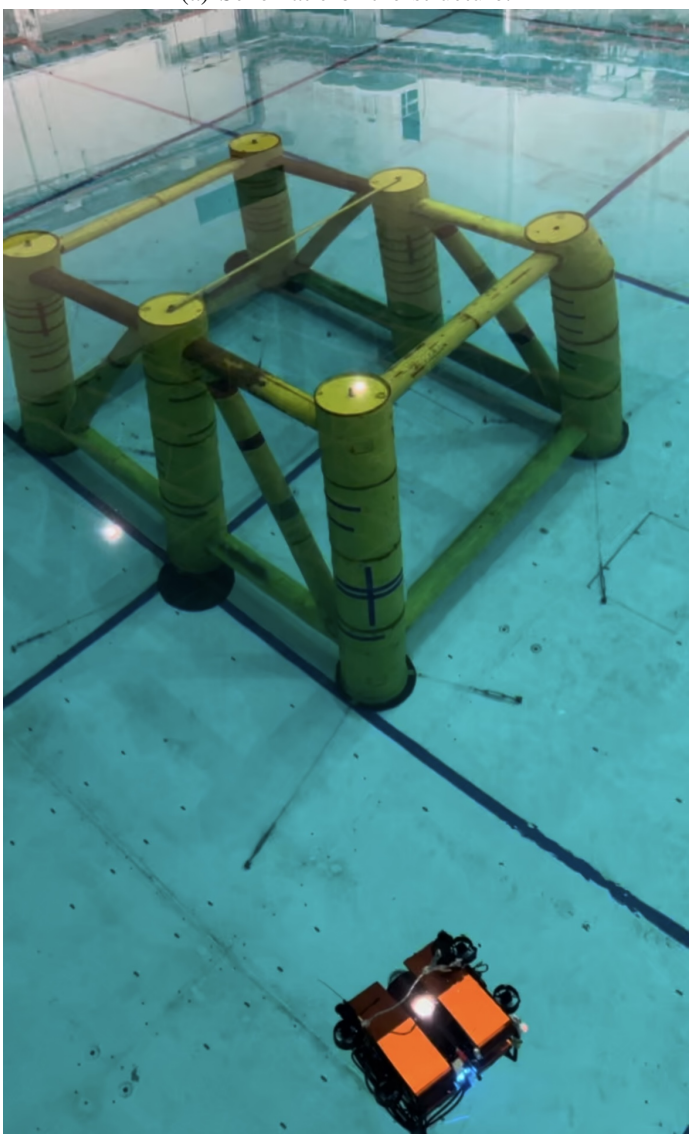
The sensor data from the vehicle was captured using ROS (Robot Operating System) and sampled at a frequency of 5 Hz. We synchronized the sampled data with the USBL position estimates based on timestamps and interpolated where necessary. For ground truth, we used the x and y coordinates from the USBL, and the z coordinate and the orientation data from the NFC.

From the recorded data collected during these trials, we curated three datasets, referred to as D1, D2, and D3. These datasets vary in depth and size, as summarized in Table II. Notably, D3 was constructed by downsampling the raw data from the third trial to create a more challenging dataset for testing purposes.

We use D1 as the primary dataset to evaluate the models' capability of interpolation. We randomly select 60% points from the data for training, 20% for validation and 20% for testing. We further assess



(a) Schematic of the structure.



(b) Topview of ROV surveying the structure.

Fig. 4: The structure surveyed in the TCOMS facility.

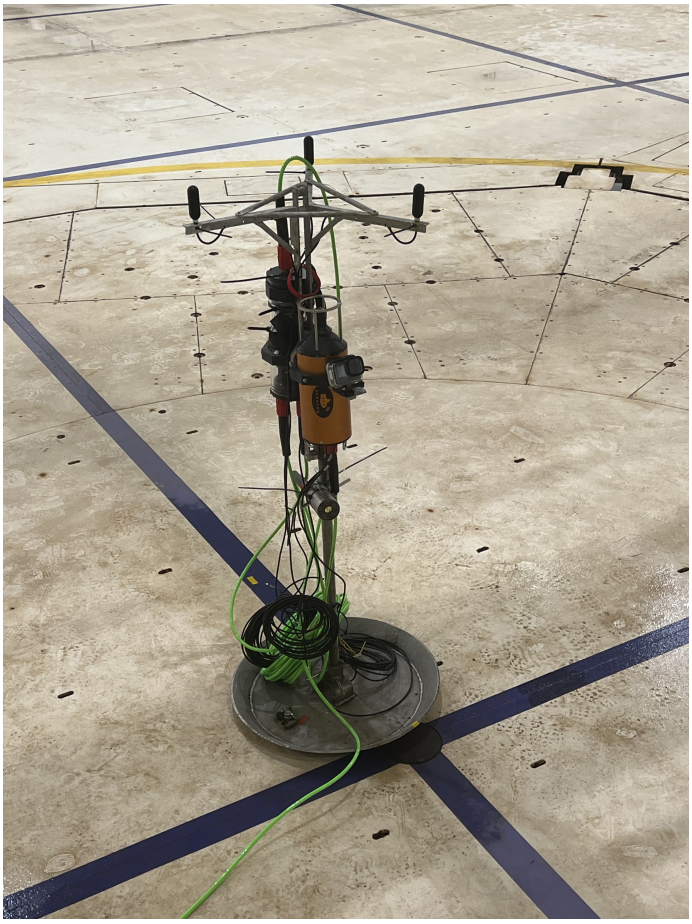


Fig. 5: The USBL setup at TCOMS to estimate the location of the ROV (shown).

TABLE II
DESCRIPTION OF DATASETS

ID	Dataset Name	Dataset Size
D1	Clear Water-Deep	2165
D2	Clear Water-Shallow	2956
D3	Clear Water-Mid	933
D4	Clear Water-NVS	4193
D5	Sea Water-1	2360
D6	Sea Water-2	735
D7	Sea Water-NVS	18918

the models' ability to generalize to new depths by employing D1 as the training dataset and D3 for validation and testing. Additionally, we investigate the impact of incorporating data from diverse depths on the models' generalization performance by using D1 and D2 together as the training data and D3 as the validation and test data.

C. Results & Discussions

1) Model performance: We present the performance of different configurations in Table III. The benchmark for our evaluation is the performance of C1.

TABLE III
PERFORMANCE OF ALL CONFIGURATIONS TRAINED AND TESTED ON DATASET D1. $\mathcal{L}_{\mathbf{p}}$ AND $\mathcal{L}_\theta$ TABULATED ARE THE MEDIAN OF ESTIMATES ACROSS THE TEST DATA. $\mathcal{L}$ WAS CALCULATED USING 2 WITH $d = 3$ M. THE BEST PERFORMANCE FOR EACH METRIC IS HIGHLIGHTED IN BOLD.

ID	$\mathcal{L}$ **(m)**	$\mathcal{L}_{\mathbf{p}}$ **(m)**	$\mathcal{L}_\theta$ (°)	**Inference time (ms)**
C1	2.41	2.36	**0.86**	2.20
C2	0.61	0.53	1.50	1.65
C3	0.41	0.36	0.99	1.62
C4	0.34	0.29	0.88	1.16
C5	0.30	0.22	1.51	0.78
C6	**0.19**	**0.12**	1.34	**0.77**

We observe the following:

1. Comparing the performance of C3 against C1, our results demonstrate that training with our proposed d-loss significantly enhances model performance, especially in terms of the overall performance metric $\mathcal{L}$. This improvement can be attributed to the simplicity and ease of use of the d-loss. Unlike the β-loss, which requires extensive hyperparameter tuning through grid search to identify an appropriate trade-off between translation and rotation errors, the d-loss eliminates this need by expressing both components in the same unit. While an optimally tuned β may achieve comparable results, the d-loss performs well out-of-the-box, reducing the tuning burden and yielding stable, reliable performance.

2. Comparing the performance of C2 against C3, it can be observed that using grayscale images shows significantly worse performance and too little an improvement in inference time, contrary to our initial expectation. The worse performance of grayscale images can be attributed to the fact that since D1 was collected in a non-turbid fresh water environment, the color information in the underwater images is not as limited as one might anticipate in an image taken in a sea environment. As shown in Fig. 6, the underwater RGB images in D1 retain valuable color information that may provide distinguishing features in these environments. Thus, the grayscale images have much less information than RGB images and thus lead to poorer performance. The lack of improvement in inference time is due to the fact that we only reduce the number of channels in the first CNN layer of the pretrained model, resulting in a minimal reduction

in computational load. To achieve more substantial computational savings, the entire model architecture would need to be better streamlined for grayscale images, not just the initial layer.

3. Comparing the performance of C6 to C5 and C4 to C3 shows that using ResNet50, a deeper network, as the backbone, improves performance for both CNN and CNN+LSTM. This is likely due to ResNet50's higher representational capacity and its residual connections, which facilitate better feature extraction and gradient flow during training. These benefits are especially useful in underwater scenes where discriminative features may be subtle or degraded. The observed improvements were consistent across several configurations, indicating that the choice of backbone architecture plays a substantial role in pose estimation accuracy.

4. Comparing the performance of C6 to C4 and C5 to C3 shows that the CNN+LSTM architecture consistently outperforms the CNN architecture. This improvement can be attributed to the LSTM layers' ability to perform structured dimensionality reduction, helping the network learn more meaningful and stable pose representations.

Among all the configurations, C6, which uses the CNN+LSTM architecture with the ResNet50 backbone and is trained using the proposed d-loss, performs the best, achieving 0.12 m of positional accuracy and 1.34°of orientation accuracy with an inference time of 0.77 ms.

2) Generalization performance: We test the performance of generalization using the model with the best configuration, C6. We first trained the model on D1 and tested on D3. A significant performance degradation is observed, as shown in the first row of Table IV. This is on expected lines because the test data is sampled from a different distribution than the training data with possibly different paths and conditions, and deep-learning models often fail to extrapolate beyond the bounds of the training data.

To address this issue, we evaluate the use of a larger and more diverse training dataset, by expanding the training data to include both D1 and D2. This augmentation introduces a wider distribution of data, notably enhancing the diversity in depth information. This leads to a 49% improvement in model performance in overall loss, as shown in the second row in Table IV.

These findings underscore the importance of comprehensive baseline mapping to collect sufficiently diverse training data. This is essential for training models that are robust enough to perform accurate localization during reinspection tasks.

TABLE IV
PERFORMANCE OF CONFIGURATION C6 ON DATASET D3. $\mathcal{L}_{\mathbf{p}}$ AND $\mathcal{L}_{\theta}$ ARE MEDIAN VALUES ACROSS THE TEST DATA. $\mathcal{L}$ WAS CALCULATED USING 2 WITH THE AVERAGE DISTANCE $d = 3$ M. THE BEST PERFORMANCE FOR EACH METRIC IS HIGHLIGHTED IN BOLD.

Training Dataset	EKF	Color Jittering	Performance Metrics		
			$\mathcal{L}$ (m)	$\mathcal{L}_{\mathbf{p}}$ (m)	$\mathcal{L}_{\theta}$ (°)
D1			1.45	1.34	2.09
D1+D2			0.75	0.58	3.20
D1+D2	✓		0.47	0.47	0.00
D1+D2+D4			0.52	0.40	2.28
D1+D2+D4		✓	0.20	0.15	0.93
D1+D2+D4	✓	✓	**0.11**	**0.11**	**0.00**

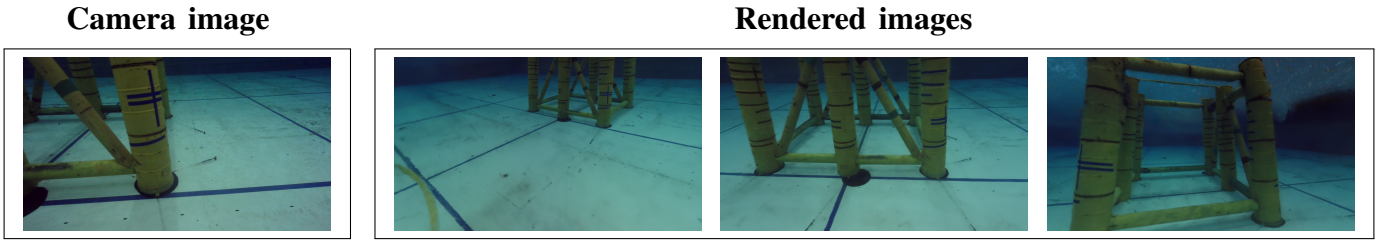


Fig. 6: Camera images and NVS rendered images in the controlled experiment in TCOMS. The rendered images produce photorealistic views of the structure but exhibit discrepancies in brightness. Some of the rendered views have artifacts in the background as shown in the right-most image.

III. Augmented Training with Novel View Synthesis

The previous section demonstrated the importance of diverse training data with good coverage of the surveyed location. Although it may sometimes be possible to collect such data by extensively covering areas during the baseline mapping run, the practical constraints of cost and labor often limit this approach or render it infeasible. We explore alternative approaches to improve model performance in such data-limited scenarios. We propose to use NVS techniques to create models of the 3D scene, and then use these to generate more images from new aspects to augment the training data. In this section, we present the methods of augmenting training data using NVS models and the results of this approach.

A. Methods

We first select 540 images from D1 and D2 to train an NVS model for the TCOMS scene. For this, we employ *COLMAP* [33], [34], an open-source Structure-from-Motion computation software, to compute the camera pose associated with each image within an arbitrary reference coordinate.

We employed the nerfacto pipeline from nerfstudio, an open-source library that provides a modular and user-friendly framework for training, and evaluating NVS-based 3D scene representations [35], as our

NVS model to render views for training data augmentation. Nerfacto is a simplistic modular NeRF implementation that adopts recent advancements to improve computational efficiency and handle unbounded scenes [35].

To train the model, we used 540 images from the original trials, along with their corresponding poses estimated via COLMAP. Inspired by the RobustNeRF variant [36], we replaced the default nerfacto loss with a robust photometric loss that down-weights inconsistent or noisy regions during training. This improves rendering quality in scenes with transient features or non-uniform illumination. The details of training the model are presented in our previous work [37].

To generate novel camera poses for rendering, we applied controlled perturbations to the original COLMAP-estimated poses. For each pose, we randomly sampled a new depth value within the feasible range, defined by the minimum and maximum depths observed in the collected data, and replaced only the z-coordinate to preserve the viewing direction. Additionally, we perturbed the x and y positions by scaling the vector from the pose to the structure using a random factor sampled from the range [0.8, 1.2], effectively varying the lateral distance while maintaining orientation toward the target. These synthesized poses were kept within the scene bounds to ensure rendering consistency. The trained NVS model was then used to render photorealistic images at these new viewpoints, which were added to the pose estimator's training set to improve generalization. In total, we generate 4193 images, and we refer to this dataset as D4. We then use D1, D2, and D4 for training, and D3 for validation and testing to test the improvement provided by using the NVS-based augmentation.

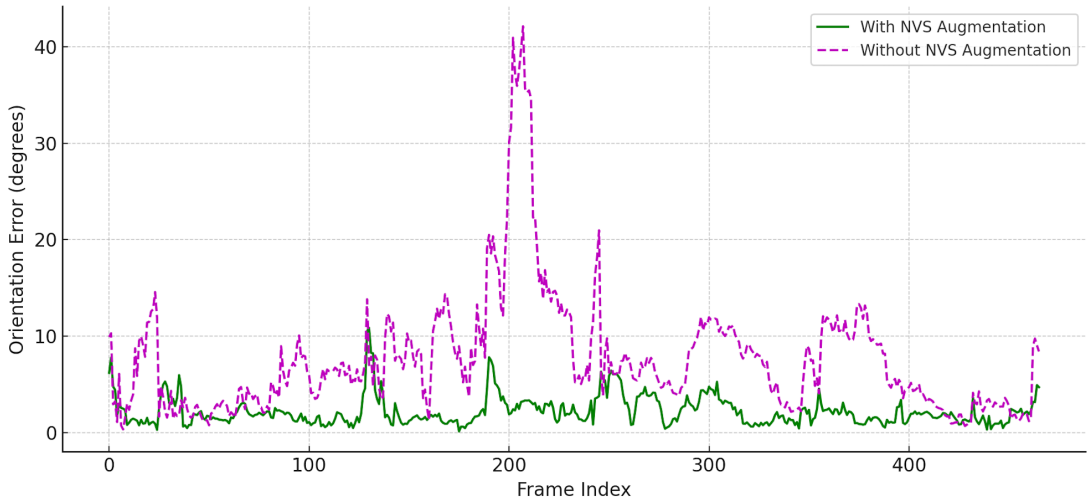
Additionally, it is noted that the images in D4 exhibited different brightness levels and background noise as compared to the original data, introduced during the NVS model reconstruction. To address the potential degradation due to this, we further augment the data by jittering the color of each image during training, thus making the pose estimator robust to minute color and lighting changes. For evaluation, we use the same GPU, framework, and hyperparameter tuning methods as described in the previous section.

B. Results & Discussion

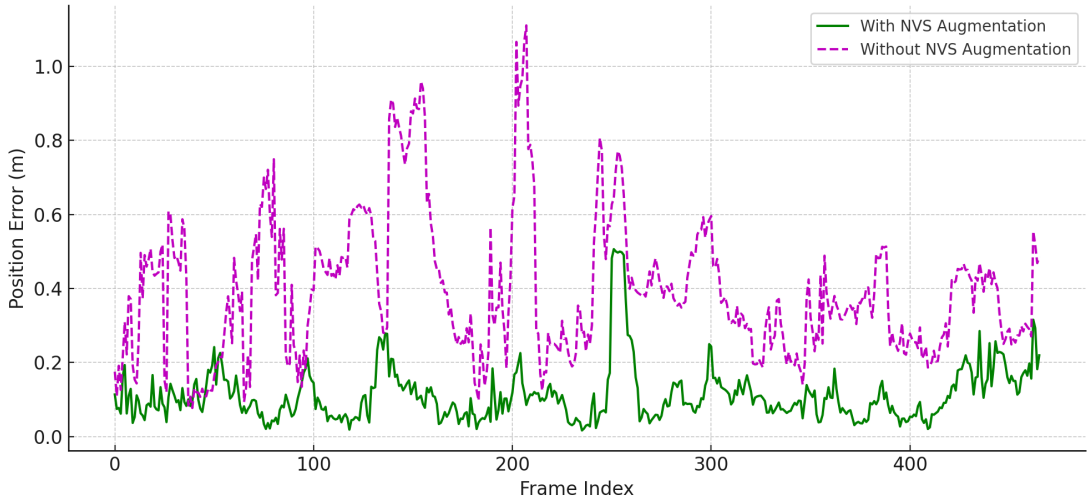
Our results show that utilizing augmented training data generated by a NVS model leads to a significant enhancement in localization accuracy. Comparing row 2 and row 4 in Table IV, we find that by augmenting the training data with D4, the overall localization error can be reduced by 30%.

Color jittering augmentation is also highly effective in further improving the model performance, further reducing the error by an additional 61.5%. We compare the performance of the augmented training with

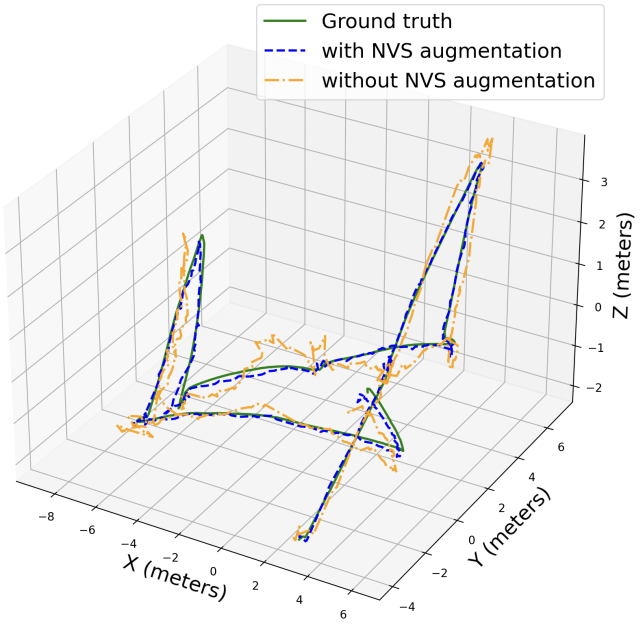
color jittering with the performance without augmented training in Fig. 7 and Fig. 8. These plots show that the proposed augmented training with NVS significantly improves the pose estimator's accuracy and reliability in terms of both position and orientation.



(a) Orientation error over time.



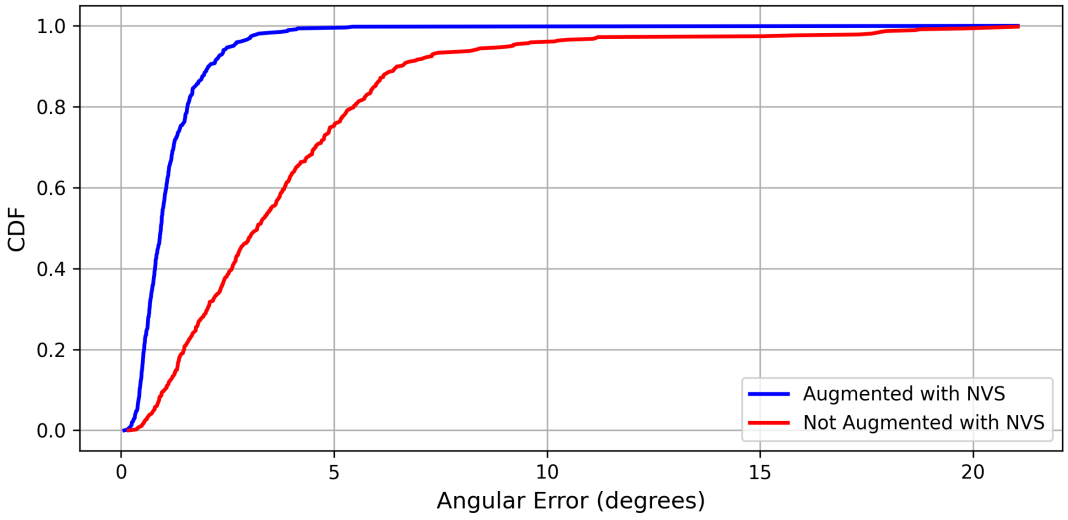
(b) Positional error over time.



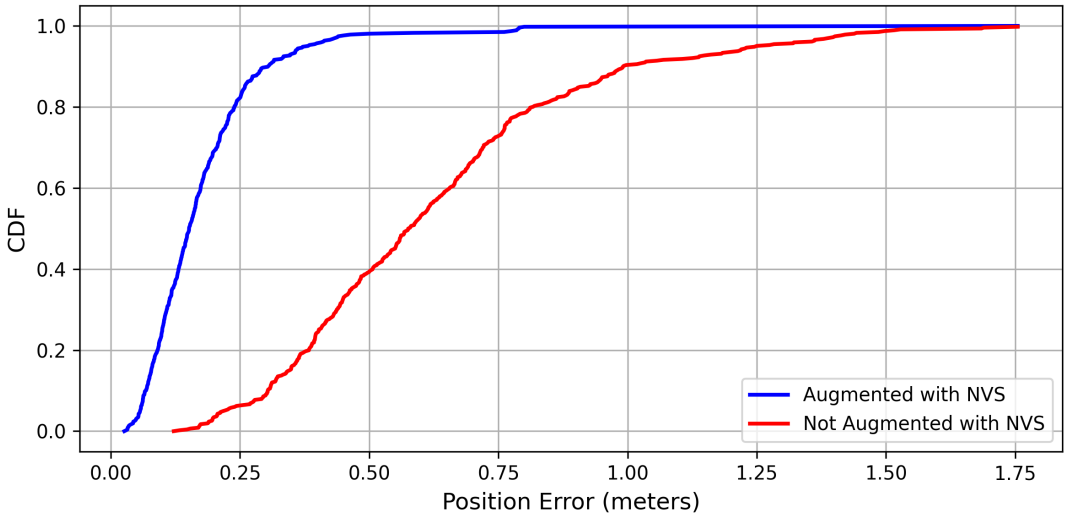
(c) 3D trajectory comparison between the ground truth and model predictions.

Fig. 7: Comparison of pose estimation results with and without NVS-based training augmentation in a controlled environment.

Nonetheless, we observed the presence of outliers. Upon examining the data, we found that these outliers were caused by transient objects, such as the tether shown in Fig. 9(b), which were not present in the training data.



(a)

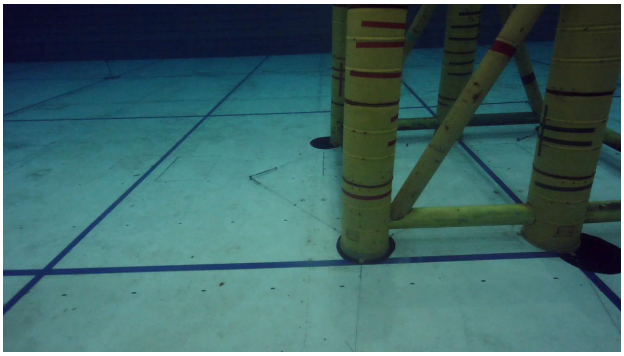


(b)

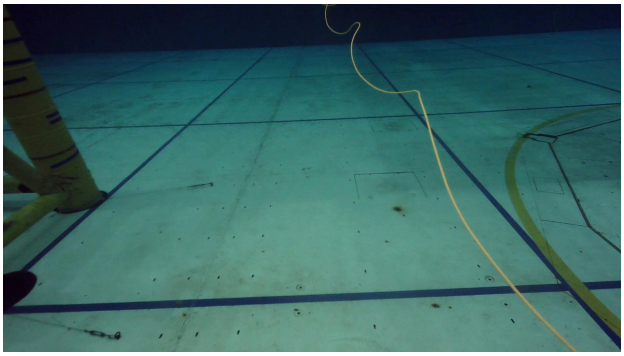
Fig. 8: Cumulative distribution function (CDF) of (a) orientation and (b) position errors for models trained with and without NVS augmentation in controlled environment. The plots show that augmented training with NVS yields significantly lower errors for both orientation and position compared to training without augmentation.

IV. Localization Enhancement via Sensor Data Fusion

While the trained pose estimators yield small median orientation and position errors, their estimates exhibit some volatility. Our model currently treats each sample independently, ignoring temporal context, and utilizes only the camera inputs during deployment. However, additional information, such as temporal information and other sensor inputs from the ROV, is available. To enhance localization accuracy and achieve a more stable trajectory estimation, we propose sensor fusion using an EKF. This section details the integration of the pose estimator with additional sensor data and presents the results of the sensor



(a)



(b)

Fig. 9: Test images from Clear Water-Mid with amongst the best and worst pose estimation accuracy. Panel (a) is the image with one of the best pose estimation accuracy and panel (b) is the image with one of the worst pose estimation accuracy.

fusion.

A. Methods

Given the sequential nature of data in reinspection missions and the availability of additional sensors, incorporating temporal information and other sensor data presents a viable strategy for improving the model's estimation stability and accuracy. Currently, the visual localization model without sensor fusion occasionally results in estimation of poses that are physically implausible or outliers, in context of the dynamics from previous poses. By integrating knowledge of the ROV's physics model and leveraging previous pose estimates, we can enhance pose accuracy and stability.

Furthermore, during reinspection missions, ROVs are commonly equipped with depth sensors and compasses, which have a reasonable accuracy. As such, we could use these reliable depth and orientation measurements during reinspection to further improve the overall localization accuracy.

We assume that the vehicle moves with a constant translational velocity and constant angular velocity since the vehicle normally moves slowly during inspection missions. Our EKF fuses measurements from three sources: the pose estimator (x, y, z position and orientation in quaternion form), compass (orientation), and depth sensor (z position). The filter maintains a 13-dimensional state vector comprising position, velocity, quaternion orientation, and angular velocity. The structure of the EKF, including its iterative prediction-update loop, is illustrated in Fig. 10.

The EKF maintains and updates three covariance matrices: the state covariance $\mathbf{P}$, the process noise covariance $\mathbf{Q}$, and the measurement noise covariance $\mathbf{R}$. The state covariance $\mathbf{P} \in \mathbb{R}^{13\times13}$ reflects the uncertainty in the estimated state and is propagated and corrected at each timestep. The process noise

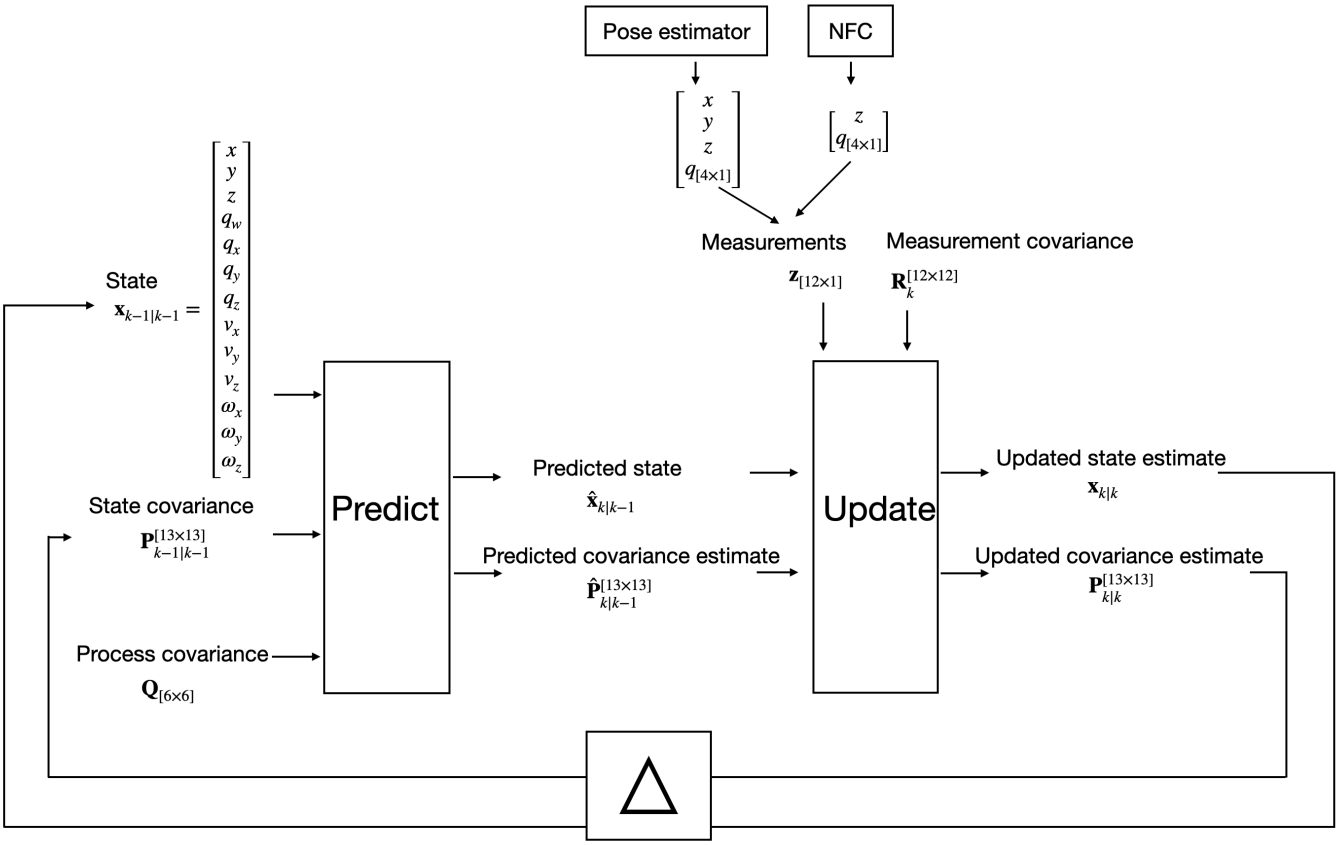


Fig. 10: EKF schematic for sensor fusion.

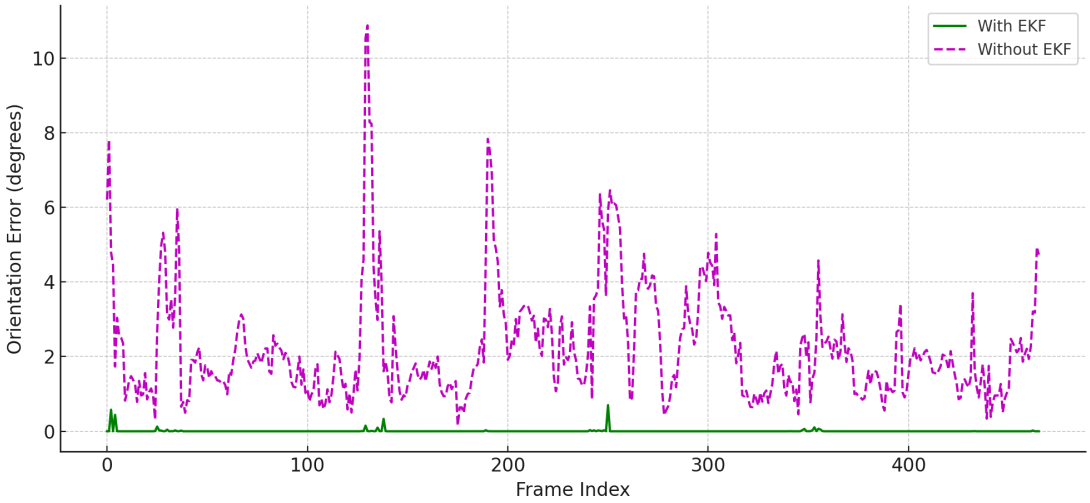
covariance $\mathbf{Q} \in \mathbb{R}^{6\times6}$ is treated as a tunable hyperparameter and models uncertainty in the velocity and angular velocity components. The measurement noise covariance $\mathbf{R} \in \mathbb{R}^{12\times12}$ incorporates nominal noise levels from manufacturer specifications for the compass and depth sensor.

Characterization of the pose estimator's measurement noise requires a more involved process. The noise primarily stem from the fact that network estimations are inconsistent and can sometimes exhibit substantial errors. As such, setting a static value for the pose estimator's measurement noise, such as the standard deviation of localization error derived from validation performance, is inadequate. To more accurately represent the dynamic noise in the pose estimator, we employ dropout techniques at test time for Monte Carlo sampling from the model output's posterior distribution. Dropout is a technique commonly used as a regularizer in training neural networks to prevent overfitting. Recent works have shown that using dropout during inference can be used to approximate Bayesian inference over the distribution of the network's weights at test time, without requiring any additional model parameters [38]. Here, we apply Monte Carlo dropout at inference - specifically, we enable dropout in the second-to-last fully connected layer of the pose estimator using a dropout rate of 0.1. At test time, we perform 100 forward passes

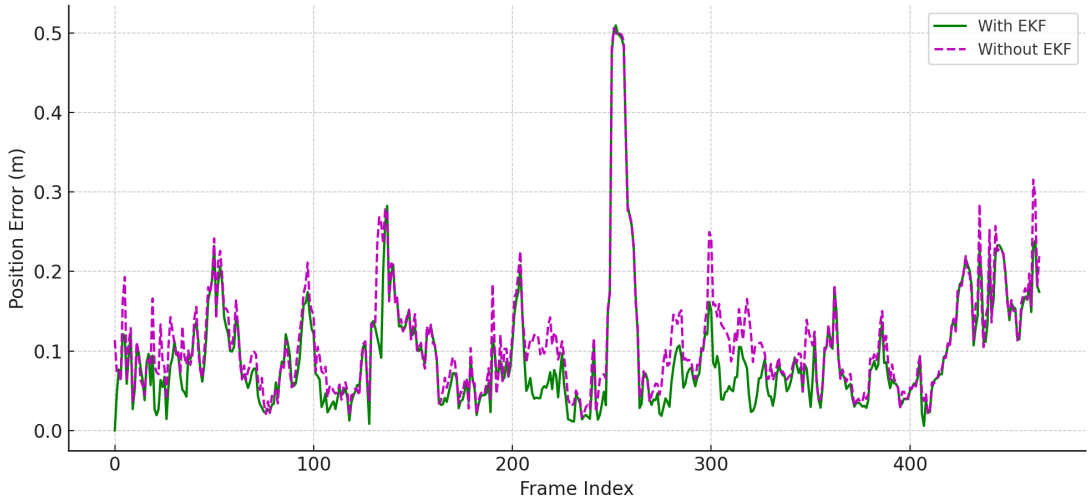
per image and compute the variance across pose predictions. This variance is then used to populate the relevant entries in $\mathbf{R}$, allowing the EKF to down-weight lower-confidence visual estimates and improve robustness in uncertain conditions. As sensor biases were minimized through careful calibration prior to data collection, we did not observe any consistent bias in the compass and depth sensor measurements during the trials. As such, we assume the measurement noise is zero-mean and unbiased.

B. Results & Discussion

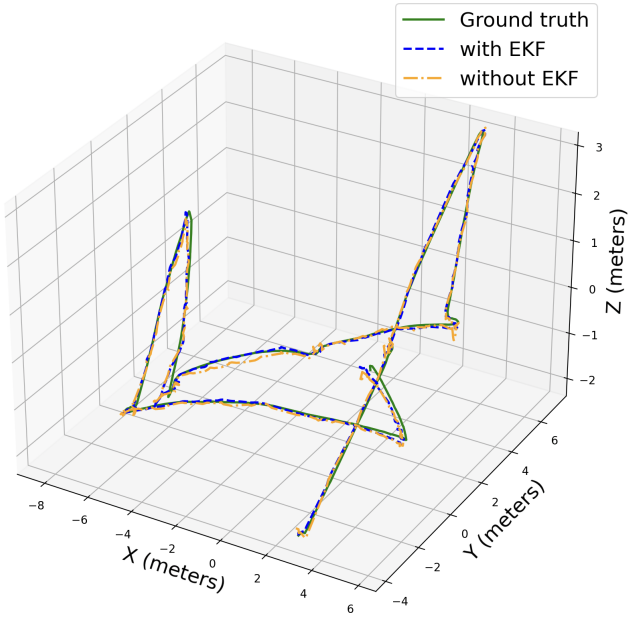
As shown in Table IV, sensor fusion with the EKF consistently improves pose estimation accuracy across different training setups. For configuration C6 trained on D1+D2 and tested on D3 (see rows 2 and 3), applying EKF reduces the median position error $\mathcal{L}p$ from **0.58 m** to **0.47 m**, and the orientation error $\mathcal{L}\theta$ from **3.20**° to **0.00**°. Similarly, for C6 trained with the NVS-augmented dataset (see rows 5 and 6), EKF reduces the position error from **0.15 m** to **0.11 m**, and orientation error from **0.93**° to **0.00**°. This consistent improvement demonstrates the robustness of the EKF-based fusion method in filtering noisy frame-level predictions and leveraging inertial priors. As also illustrated in Fig. 11, the estimated trajectory becomes noticeably smoother and more aligned with ground truth. While the inference time increases by approximately 10 times due to Monte Carlo sampling, this trade-off may be acceptable in scenarios where pose stability and accuracy are critical.



(a) Orientation error over time.



(b) Positional error over time.



(c) 3D trajectory comparison between the ground truth and model predictions.

Fig. 11: Comparison of pose estimation results with and without EKF in a controlled environment.

V. Field trials at sea

To further validate our proposed methods, we conducted field trials in a bay near St. John's Island, Singapore (SJI). In this section, we present the methods, results and challenges encountered in using our proposed methods from the previous section in a real-world setting.

A. Methods

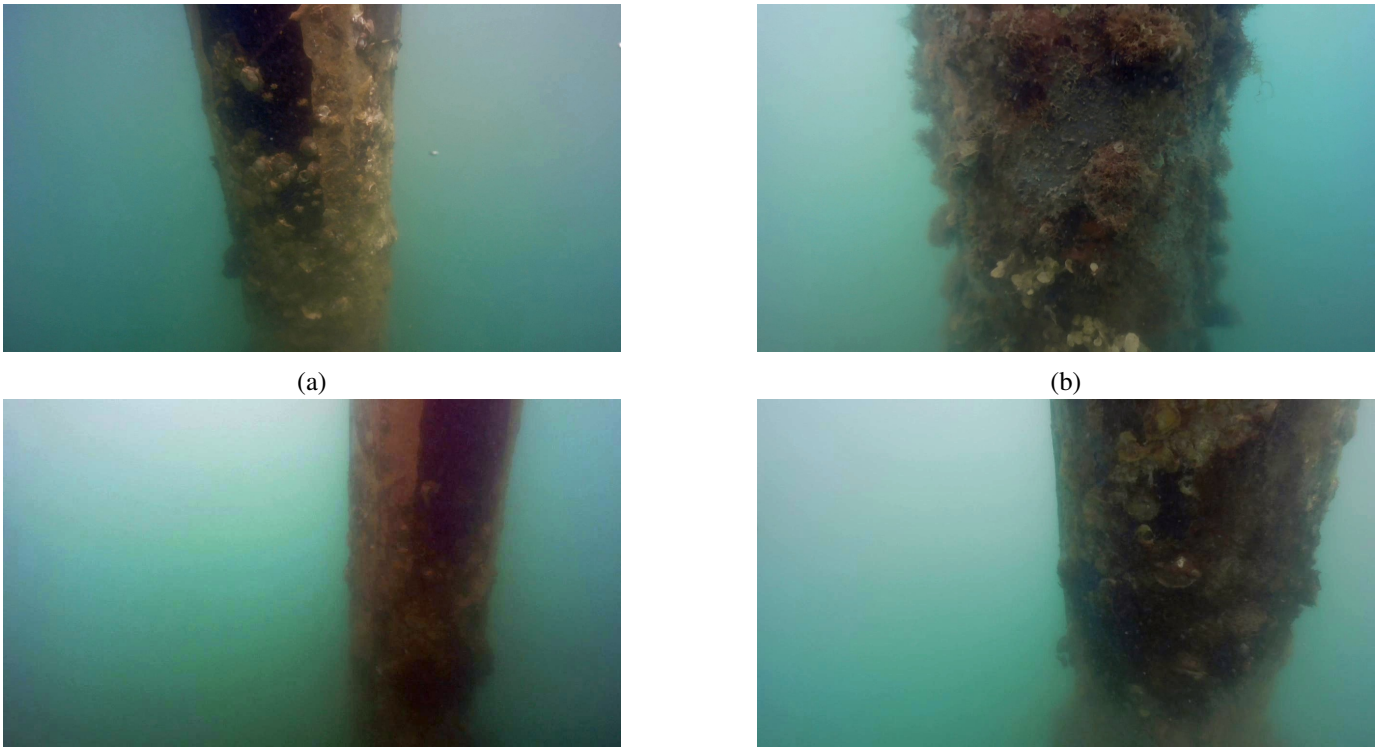
We used the ROV to collect data in an at-sea environment, inspecting a submerged pillar. The pillar selected was approximately 5 m tall and 0.5 m in diameter. Although the pillar was a simple black structure, the barnacles and algae growing on its surface provided visual features that could be used for pose estimation. We drove the ROV following a vertical lawnmower path around the pillar, while recording the video from the camera. Due to the high turbidity in the water, we operated the ROV in close proximity to the structure with the average distance being 1 m.

The accuracy of USBL in our at-sea experiments was compromised due to high measurement noise and the absence of detailed information about the deployment geometry. In contrast, COLMAP was able to produce camera pose estimates with centimeter-level accuracy using structure-from-motion on the collected images. We therefore used *COLMAP* to estimate camera poses using the collected image data. Although these poses are not ground-truth in the absolute sense, they provide a consistent reference trajectory suitable for evaluating relative pose estimation performance in the field setting.

We collected two datasets, named as D5 and D6, on two different days. Although the inspection was carried out on the same structure with similar trajectories, there were noticeable differences in environmental conditions between the two runs. D6 was collected under higher turbidity compared to D5, resulting in fewer visual features and noisier images. This variability reflects typical challenges encountered in real-world underwater inspections, where it is difficult to guarantee the same visibility, lighting, or exact path between mapping and reinspection runs. Samples of images collected in these datasets are shown in Fig. 12.

We use D5 to train an NVS model following the method described in Section III. New camera poses are generated using the same approach. The NVS model is then utilized to create an augmented training dataset, named D7. Samples of images generated at new poses using the NVS model are shown in Fig. 13.

We train the best visual localization architecture configuration, C6, both with augmented training data (datasets D5+D7) and without any augmentation (only D5). Dataset D6 is used for validation and testing.



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

Fig. 12: Sample images from Sea Water-1 and Sea Water-2 dataset. Panels (a) and (b) are from Sea Water-1 dataset, and (c) and (d) are from Sea Water-2 dataset. The images from Sea Water-2 dataset show higher turbidity and thus fewer features than images from Sea Water-1.

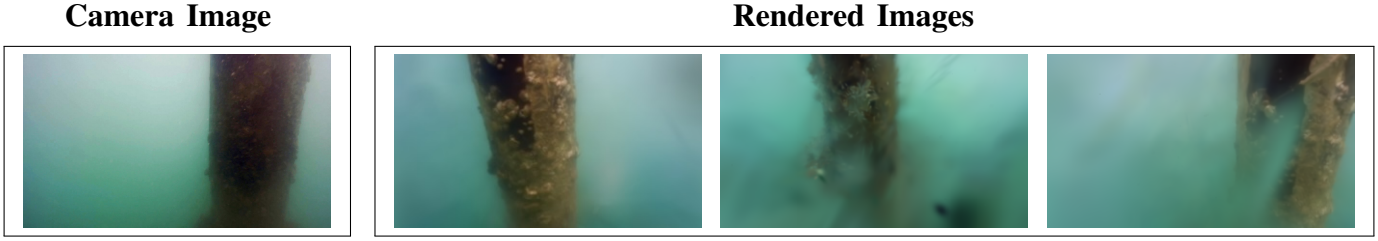


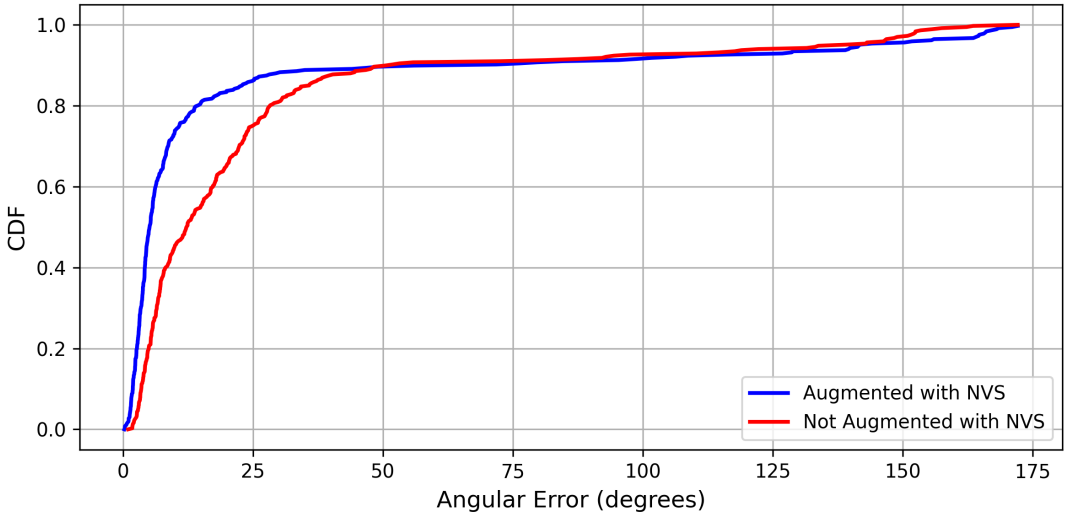
Fig. 13: Camera image and NVS rendered images in the bay near SJI. The rendered images produce photorealistic views of the structure but exhibit some artifacts and noise depending on the camera pose.

The training methods are similar to those described in Section II.

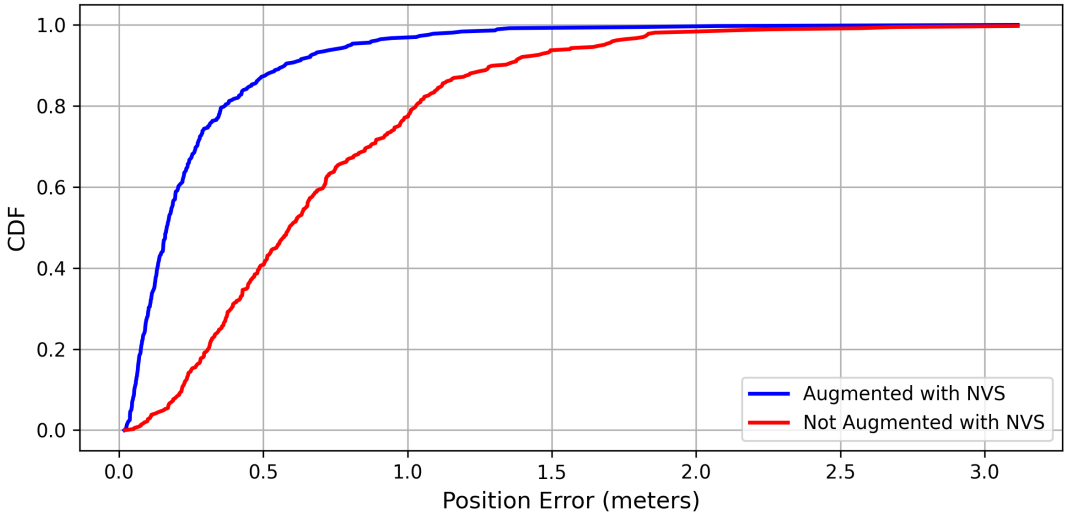
B. Results & Discussion

As shown in Fig. 14, augmented training with NVS yields significant improvement in both position and orientation accuracy compared to training without NVS augmentation. The improvements brought by NVS can be attributed not only to an increase in training samples, but more importantly to the expanded coverage of viewpoints, especially those that may be underrepresented or missing due to inevitable variations in inspection trajectories. This highlights the strength of NVS augmentation in realistic

underwater applications, where achieving complete and repeatable scene coverage is inherently difficult. With configuration C6 and augmented training, we are able to achieve a position accuracy of 0.17 m and orientation accuracy of 5.09°. We present the performance of C6 on D6 in Table V. While the median accuracy is comparable to the performance in the controlled environment, we note that the standard deviation in the errors are much larger at sea.



(a)



(b)

Fig. 14: CDF of (a) orientation and (b) position errors for models trained with and without NVS augmentation in at-sea environment. The plots show that augmented training with NVS yields significantly lower errors for both orientation and position compared to training without augmentation.

Clearly, the real-world setting at sea presents several challenges that are not present in controlled environments. The biggest challenge is the turbidity of the water, which significantly affects the quality of the images. Moreover, lighting is inconsistent at different camera poses and on different days, causing

TABLE V
PERFORMANCE OF CONFIGURATION C6 ON DATASET D6. $\mathcal{L}_{\mathbf{p}}$ AND $\mathcal{L}_{\theta}$ ARE MEDIAN VALUES ACROSS THE TEST DATA. $\mathcal{L}$ WAS CALCULATED USING 2 WITH THE AVERAGE DISTANCE $d = 1$ M.

Training Dataset	**Color Jittering**	**Performance Metrics**		
		$\mathcal{L}$ (m)	$\mathcal{L}_{\mathbf{p}}$ (m)	$\mathcal{L}_{\theta}$ (°)
D5		0.80	0.59	12.15
D5+D7	✓	**0.26**	**0.17**	**5.09**

(a) (b)

Fig. 15: Test images from Sea Water-2 with amongst the best and worst pose estimation accuracy. Panel (a) is the image with one of the best pose estimation accuracy and (b) is the image with one of the worst pose estimation accuracy.

high variablity in the image quality. This introduced three new challenges. First, the noisy images make it challenging to compute camera poses in *COLMAP*, resulting in a sparse number of registered images. Consequently, the EKF model could not be used for performance improvement since it would not be feasible to assume constant velocity and angular velocity in the vehicle model. Second, the turbidity and inconsistent lighting in the training data introduced artifacts in the NVS model. Thus, the rendered images are more noisy compared to images in clear waters, as shown in Fig. 13. Third, the high variability in image quality can lead to more estimation outliers and large errors during inference. As illustrated in Fig. 15, some test images contain rich textures and clear structure boundaries, which are favorable for accurate pose estimation. In contrast, others suffer from turbidity and challenging environment lighting, resulting in severely degraded visual features and consequently poor pose estimates. All of these factors contribute to a decrease in the model's performance.

VI. Conclusion

In this paper, we addressed the challenge of localization in underwater inspection missions with a neural-network based pose estimator. We proposed a new loss function to train the pose estimator, and

demonstrated that training with d-loss significantly improved the model's performance in pose estimation tasks. This improvement is attributed to the incorporation of domain-specific physics, as the d-loss accounts for the relevant geometric considerations in the inspection mission. Furthermore, this loss function also lends more interpretability to the loss. Employing the ResNet50 backbone with a CNN+LSTM architecture allows us to efficiently use the available visual information to estimate the pose, and yielded improvements in the localization performance as compared to benchmark architectures.

In terms of the generalization, using more diverse data with a wider distribution significantly enhances the localization performance on test data that lies outside the training distribution. We additionally investigated the use of NVS techniques to augment training data and showed that this significantly improves the estimator's performance with previously unsurveyed poses. Thus, this provides a cost-effective and information-efficient method to improve the generalization performance without having to undertake expensive field trials to collect additional data. Further integrating the pose estimator with an EKF allows us to fuse sensor data with the visual-based estimates, and we demonstrated that this further improved the performance and stability. We validated our proposed methods in both controlled environments in a clear water tank and real-world settings at sea.

Overall, our results show that our proposed methods significantly improve the visual localization performance in both controlled underwater environments and real-world settings and achieve good localization accuracy to within desired limits, providing a cost-effective alternative or complement to existing localization solutions. Real-world challenges such as turbidity and noise limit the performance achievable, but the proposed method still performs reasonably, especially when data augmentation using color-based augmentation is used to robustify the technique against color distortion.

Potential improvements to this technique may include utilizing temporal information (i.e., more than one image at a time) to improve the accuracy of pose estimates, fusing more data such as control input information and sonar data, exploring better sensor fusion techniques such as particle filters and using per-pixel loss together with NVS rendered images to fine-tune the model.

The algorithm developed in this work is also utilized as part of a model-based image compression technique for low bandwidth scenarios. The details of this approach and the preliminary results are presented in our previous work [39].

ACKNOWLEDGMENT

This research project is supported by A*STAR under its RIE2020 Advanced Manufacturing and Engineering (AME) Industry Alignment Fund - Pre-Positioning (IAF-PP) Grant No. A20H8a0241.

REFERENCES

[1] E. Vargas, R. Scona, J. S. Willners, T. Luczynski, Y. Cao, S. Wang, and Y. R. Petillot, "Robust underwater visual SLAM fusing acoustic sensing," in *2021 IEEE International Conference on Robotics and Automation (ICRA)*. IEEE, May 2021, pp. 2140–2146.

[2] B. Bingham, B. Foley, H. Singh, R. Camilli, K. Delaporta, R. Eustice, A. Mallios, D. Mindell, C. Roman, and D. Sakellariou, "Robotic tools for deep water archaeology: Surveying an ancient shipwreck with an autonomous underwater vehicle," *Journal of Field Robotics*, vol. 27, no. 6, pp. 702–717, Nov. 2010.

[3] M. Carreras, J. D. Hernandez, E. Vidal, N. Palomeras, and P. Ridao, "Online motion planning for underwater inspection," in *2016 IEEE/OES Autonomous Underwater Vehicles (AUV)*. IEEE, Nov. 2016, pp. 336–341.

[4] H.-P. Tan, R. Diamant, W. K. G. Seah, and M. Waldmeyer, "A survey of techniques and challenges in underwater localization," *Ocean Engineering*, vol. 38, no. 14, pp. 1663–1676, Oct. 2011.

[5] S. Zhang, S. Zhao, D. An, J. Liu, H. Wang, Y. Feng, D. Li, and R. Zhao, "Visual SLAM for underwater vehicles: A survey," *Computer Science Review*, vol. 46, p. 100510, Nov. 2022.

[6] S. Kuutti, S. Fallah, K. Katsaros, M. Dianati, F. Mccullough, and A. Mouzakitis, "A survey of the state-of-the-art localization techniques and their potentials for autonomous vehicle applications," *IEEE Internet of Things Journal*, vol. 5, no. 2, pp. 829–846, Apr. 2018.

[7] A. D. Buchan, E. Solowjow, D.-A. Duecker, and E. Kreuzer, "Low-cost monocular localization with active markers for micro autonomous underwater vehicles," in *2017 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*. IEEE, Sep. 2017, pp. 4181–4188.

[8] A. Gomez Chavez, C. Mueller, T. Doernbach, D. Chiarella, and A. Birk, "Robust gesture-based communication for underwater human-robot interaction in the context of search and rescue diver missions," Oct. 2018.

[9] D. Chiarella, M. Bibuli, G. Bruzzone, M. Caccia, A. Ranieri, E. Zereik, L. Marconi, and P. Cutugno, "Gesture-based language for diver-robot underwater interaction," in *OCEANS 2015 - Genova*, May 2015, pp. 1–9.

[10] B. Teixeira, H. Silva, A. Matos, and E. Silva, "Deep learning for underwater visual odometry estimation," *IEEE Access*, vol. 8, pp. 44 687–44 701.

[11] A. Bucci, L. Zacchini, M. Franchi, A. Ridolfi, and B. Allotta, "Comparison of feature detection and outlier removal strategies in a mono visual odometry algorithm for underwater navigation," *Applied Ocean Research*, 2022.

[12] M. Ferrera, J. Moras, P. Trouvé-Peloux, and V. Creuze, "Real-time monocular visual odometry for turbid and dynamic underwater environments," *Sensors*, vol. 19, no. 3, 2019.

[13] S. Wirth, P. L. N. Carrasco, and G. Codina, "Visual odometry for autonomous underwater vehicles," *2013 MTS/IEEE OCEANS - Bergen*, pp. 1–6, 2013.

[14] A. Burguera, F. Bonin-Font, and G. Oliver, "Trajectory-based visual localization in underwater surveying missions," *Sensors (Basel, Switzerland)*, vol. 15, pp. 1708 – 1735, 2015.

[15] A. Kendall, M. Grimes, and R. Cipolla, "PoseNet: A convolutional network for real-time 6-DOF camera relocalization," in *2015 IEEE International Conference on Computer Vision (ICCV)*, Dec. 2015, pp. 2938–2946, ISSN: 2380-7504.

[16] L. Peng, H. Vishnu, M. Chitre, Y. M. Too, B. Kalyan, and R. Mishra, "Improved image-based pose regressor models for underwater environments," Mar. 2024. [Online]. Available: http://arxiv.org/abs/2403.08360

[17] T. Sattler, B. Leibe, and L. Kobbelt, "Efficient & effective prioritized matching for large-scale image-based localization," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 39, no. 9, pp. 1744–1756, Sept. 2017.

[18] Y. Shavit and R. Ferens, "Introduction to camera pose estimation with deep learning," *arXiv*, Jul. 2019.

[19] E. Brachmann, A. Krull, S. Nowozin, J. Shotton, F. Michel, S. Gumhold, and C. Rother, "Dsac — differentiable ransac for camera localization," in *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2017, pp. 2492–2500.

[20] R. Clark, S. Wang, A. Markham, N. Trigoni, and H. Wen, "Vidloc: A deep spatio-temporal model for 6-dof video-clip relocalization," in *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2017, pp. 2652–2660.

[21] Z. Xiao, C. Chen, S. Yang, and W. Wei, "Effloc: Lightweight vision transformer for efficient 6-dof camera relocalization," in *2024 IEEE International Conference on Robotics and Automation (ICRA)*, 2024, pp. 8529–8536.

[22] S. Brahmbhatt, J. Gu, K. Kim, J. Hays, and J. Kautz, "Geometry-aware learning of maps for camera localization," *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 2616–2625, 2017.

[23] M. C. Nielsen, M. H. Leonhardsen, and I. Schjolberg, "Evaluation of PoseNet for 6-DOF underwater pose estimation," in *OCEANS 2019 MTS/IEEE SEATTLE*. IEEE, Oct 2019, pp. 1–6.

[24] L. Peng and M. Chitre, "Regressing poses from monocular images in an underwater environment," in *OCEANS 2022 - Chennai*, Feb. 2022, pp. 1–4.

[25] B. Mildenhall, P. P. Srinivasan, M. Tancik, J. T. Barron, R. Ramamoorthi, and R. Ng, "NeRF: representing scenes as neural radiance fields for view synthesis," *Commun. ACM*, vol. 65, no. 1, p. 99–106, Dec. 2021.

[26] B. Kerbl, G. Kopanas, T. Leimkuehler, and G. Drettakis, "3d gaussian splatting for real-time radiance field rendering," *ACM Transactions on Graphics*, vol. 42, no. 4, pp. 1–14, Aug. 2023.

[27] "TCOMS Research & Development," Oct. 2023. [Online]. Available: https://www.tcoms.sg/research-development/

[28] C. Szegedy, W. Liu, Y. Jia, P. Sermanet, S. Reed, D. Anguelov, D. Erhan, V. Vanhoucke, and A. Rabinovich, " Going deeper with convolutions ," in *2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. Los Alamitos, CA, USA: IEEE Computer Society, Jun. 2015, pp. 1–9. [Online]. Available: https://doi.ieeecomputersociety.org/10.1109/CVPR.2015.7298594

[29] J. Deng, W. Dong, R. Socher, L.-J. Li, K. Li, and L. Fei-Fei, "Imagenet: A large-scale hierarchical image database," in *2009 IEEE Conference on Computer Vision and Pattern Recognition*, 2009, pp. 248–255.

[30] W. F., C. Hazırbaş, L. Leal-Taixé, T. Sattler, S. Hilsenbeck, and D. Cremers, "Image-based localization using lstms for structured feature correlation," in *2017 IEEE International Conference on Computer Vision (ICCV)*. Piscataway, NJ: IEEE, 2017, Conference Paper, pp. 627 – 637, 16th IEEE International Conference on Computer Vision (ICCV 2017); Conference Location: Venice, Italy; Conference Date: October 22-29, 2017.

[31] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2016, pp. 770–778.

[32] E. Bernardes and S. Viollet, "Quaternion to Euler angles conversion: A direct, general and computationally efficient method," *PLOS ONE*, vol. 17, no. 11, p. e0276302, Nov. 2022.

[33] J. L. Schönberger, E. Zheng, J.-M. Frahm, and M. Pollefeys, "Pixelwise view selection for unstructured multi-view stereo," in *Computer Vision – ECCV 2016*, B. Leibe, J. Matas, N. Sebe, and M. Welling, Eds. Springer International Publishing, pp. 501–518.

[34] J. L. Schonberger and J.-M. Frahm, "Structure-from-motion revisited," in *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. IEEE, Jun. 2016, pp. 4104–4113.

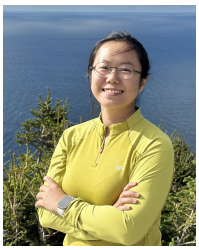
[35] M. Tancik, E. Weber, E. Ng, R. Li, B. Yi, J. Kerr, T. Wang, A. Kristoffersen, J. Austin, K. Salahi, A. Ahuja, D. McAllister, and A. Kanazawa, "Nerfstudio: A modular framework for neural radiance field development," in *ACM SIGGRAPH 2023 Conference Proceedings*, ser. SIGGRAPH '23, 2023.

[36] S. Sabour, S. Vora, D. Duckworth, I. Krasin, D. J. Fleet, and A. Tagliasacchi, "Robustnerf: Ignoring distractors with robust losses," in *2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2023, pp. 20 626–20 636.

[37] Y. M. Too, H. Vishnu, M. Chitre, B. Kalyan, L. Peng, and R. Mishra, "Feasibility study on novel view synthesis of underwater structures using neural radiance fields," in *OCEANS 2024 MTS/IEEE Singapore*. IEEE, 2024.

[38] A. Kendall and R. Cipolla, "Modelling uncertainty in deep learning for camera relocalization," in *2016 IEEE International Conference on Robotics and Automation (ICRA)*, 2016, pp. 4762–4769.

[39] R. Mishra, M. Chitre, B. Kalyan, Y. M. Too, H. Vishnu, and L. Peng, "An architecture for virtual tethering of ROVs," in *OCEANS 2024 MTS/IEEE Singapore*. IEEE, 2024.



Luyuan Peng is a Ph.D. candidate in the Department of Electrical and Computer Engineering at the National University of Singapore (NUS) and a research engineer at the Acoustic Research Laboratory (ARL), NUS. She received her B.Eng. degree in Aerospace Engineering from Nanyang Technological University (NTU), Singapore, in 2020. Her research interests include marine robotics, perception, computer vision, and multi-sensor fusion.

Luyuan serves as the Editor of the IEEE Oceanic Engineering Society (OES) Student Newsletter. She was the Local Arrangements Chair for IEEE OCEANS 2024 Singapore and the Logistics Chair for the IEEE Singapore AUV Challenge 2025. In 2024, she was selected as an IEEE OES Women in Engineering (WIE) Propel Laureate.

Hari Vishnu is a Senior Research Fellow at the Acoustic Research Laboratory, National University of Singapore. His interests include machine learning and processing in impulsive noise. These are used in a wide range of underwater applications ranging from biodiversity or defense-related scenarios in shallow tropical waters infested with snapping shrimp noise, to polar ice sheets where glacier melt noise dominates the soundscape.

From 2019, Hari is focusing on the acoustics of melting ice, machine-learning based marine-mammal quantification, and distributed acoustic sensing. He obtained his Ph.D from Nanyang Technological University, Singapore, in Computer Engineering on underwater signal processing including robust detection and localization.

He is the Chief Editor on the IEEE OES Science outreach magazine Earthzine and serves on the IEEE Oceanic Engineering society Executive committee as Deputy Secretary. In 2019, he was awarded the IEEE OES YP-BOOST award which aims to encourage young professionals to participate in the society leadership.

Mandar Chitre (S'04–M'05–SM'11–F'25) received the B.Eng. and M.Eng. degrees in electrical engineering from the National University of Singapore (NUS), Singapore, in 1997 and 2000, respectively, the M.Sc. degree in bioinformatics from the Nanyang Technological University (NTU), Singapore, in 2004, and the Ph.D. degree in underwater communications from NUS, in 2006.

From 1997 to 1998, he was with the Acoustic Research Laboratory (ARL), NUS. From 1998 to 2002, he headed the technology division of a regional telecommunications solutions company. In 2003, he rejoined ARL, initially as the Deputy Head (Research) and is currently the Head of the laboratory. He also holds a joint appointment with the Department of Electrical and Computer Engineering, NUS as an Associate Professor. His current research interests include underwater communications, acoustic propagation modeling, marine robotics, model-based signal processing, and machine learning.

Dr. Chitre served as the Editor-in-Chief for the IEEE JOURNAL OF OCEANIC ENGINEERING from 2018 to 2023. He was the Technical Co-Chair for IEEE OCEANS 2020 Singapore – U.S. Gulf Coast and the Technical Chair for IEEE OCEANS 2024 Singapore. He was the Chairman of the student poster committee for the IEEE OCEANS 2006 Singapore and the founding Chairman for the IEEE Singapore AUV Challenge in 2013. He was awarded the Distinguished Technical Achievement Award by the IEEE Oceanic Engineering Society in 2020 for his work on underwater communications & networking.

Yuen Min Too (M'15) received the Bachelor of Engineering degree in Biomedical from the Universiti Teknologi Malaysia (UTM), Malaysia in 2010, and the Ph.D. degree in underwater acoustics and signal processing from the Department of Electrical and Computer Engineering, National University of Singapore (NUS), Singapore in 2017.

From 2015 to 2025, he was with the Acoustic Research Laboratory at NUS, where he served as a Research Engineer and later as a Research Fellow. His research focused on underwater acoustics and signal processing, with particular interest in compressed sensing and array signal processing in complex underwater noise environments. He also contributed to the development of machine learning techniques for underwater sound event detection and acoustic signal denoising.

After a decade in academia, he transitioned to industry to pursue a more practical role, motivated by the desire to see his research applied in real-world settings. In 2025, he joined Subnero Pte. Ltd. as a Software Engineer, developing advanced solutions for underwater wireless communication networks.

Bharath Kalyan (SM'19) is a Senior Research Fellow at the Acoustic Research Laboratory, National University of Singapore, specializing in underwater robotics and navigation. With over two decades of experience, he has led programs on AUVs, hybrid underwater vehicles, and deepsea mineral exploration, including efforts under Keppel-NUS Corporate Lab and A*STAR Marine & Offshore Program Office. He received his PhD from Nanyang Technological University and has held visiting positions at the University of Tokyo. An active member of IEEE OES for over 15 years, he has served in various leadership roles and organized major events such as SAUVC and OCEANS conferences.

Rajat Mishra (M'15) is a Research Fellow at the Acoustic Research Laboratory, National University of Singapore, specializing in marine autonomy and underwater robotics. His research focuses on informative path planning for autonomous surface and underwater vehicles, hybrid ROV/AUV systems, and underwater perception. He received the PhD degree from the National University of Singapore in 2020 and the B.Tech degree in Electronics and Electrical Engineering from VIT University, India.

Soo Pieng Tan holds a Bachelor of Electrical & Electronic Engineering from Nottingham Trent University. Prior to her graduation, she was attached to ARL, and she joined the company as an Engineer in 2001. Her work focuses on supporting hardware prototyping, system integration, and equipment maintenance. In her free time, she enjoys playing Sudoku and hiking.