Underwater marker-based pose-estimation with associated uncertainty

Petter Risholm*[1], Peter Ørnulf Ivarsen [1], Karl Henrik Haugholt[1], and Ahmed Mohammed[1]

[1]SINTEF Digital, Smart Sensor Systems, Oslo, Norway

Abstract

We propose a system for 6-DoF estimation of Aruco markers with associated uncertainties in the challenging underwater environment. A state-of-the-art object detection framework (EfficientDet) was adapted to predict the corner locations of Aruco markers, while dropout sampling at inference time is used to estimate the predictive 6-DoF pose uncertainty. A dataset of Aruco markers captured in a wide variety of turbidities, with ground truth position of the corner locations, was gathered and used to train the network to robustly predict the 6-DoF pose. We report median translational errors of 2.6cm at low turbidity (8.5m attenuation length) and up to 10.5cm at high turbidities (0.3m attenuation length). The respective uncertainty, reported as interquartile ranges (IQRs), range from 3.2cm up to 27.9cm. The rotational median errors varied from 5.6° to 10.7° with IQRs of 6.4° to 26.2°. We also discuss how the pose uncertainty can be applied to reduce the risk in a subsea intervention operation.

1. Introduction

Autonomous underwater vehicles (AUVs) can only operate reliably if they have robust 6-degrees of freedom (6-DoF) localization capabilities. Basing critical autonomous decisions on highly uncertain localization information can potentially lead to catastrophic outcomes that not only risk the success of the AUVs mission, but also endanger human lives.

Deep learning networks have shown impressive accuracy, even surpassing human performance, on several challenging vision tasks [1]. Object detection with deep learning, which is crucial for scene understanding and generally a backbone for 6-DoF localization methods, has over the last years seen strong improvements on both detection accuracy and efficiency. Some of the most prominent networks are two-stage region-based CNNs [4, 9, 10] and one-stage detectors such as YOLO [18]. Recently, EfficientDet [21] has improved the efficiency of one-stage detectors by including a bi-directional feature pyramid network on top of an EfficientNet [20] backbone, and they use AutoML to optimize for size, resolution and depth of the model structures.

Pose estimation of relevant objects is an important prerequisite to enable vehicles to safely operate autonomously. There are two main approaches to estimate the 6-DoF – direct and indirect. The direct approach regresses the translational and rotational parameters directly, such as in PoseNet [13] and EfficientPose [3]. A challenge with these approaches is to define the loss function over the complex pose space in a stable manner. In the indirect approach, object keypoints are regressed in 2D, and a Perspective-N-Point [16] (PnP) algorithm is used to estimate the 6-DoF pose. Because the predictions are performed in 2D image space, robust loss functions are easily constructed [22]. A relatively simple approach to 6-DoF estimation is to detect artificial markers with known geometry in the scene, e.g. Aruco/Charuco markers where the four corners are detected and used to compute the 6-DoF pose [8]. In DeepCharuco [11] they showed that a deep learning approach can outperform the classical Charuco approach to detect the markers, especially under challenging light situations. Some works have studied the effect of the underwater environment on the detection of Aruco markers [5], [6] and proposed methods to improve the detection rate by e.g. applying dehazing algorithms [24]. They are generally evaluated according to the rate of detection of the markers – not the pose accuracy. In [17] they evaluate the accuracy of the Aruco detection in an underwater environment with very clear water. They report a mean translational error of 11.8cm and a mean rotational error of 4.2 degrees.

A pertinent question is how much a robot can trust the 6-DoF pose predictions of the deep learning (DL) networks? Current deep neural network (DNN) methodology tends to make overconfident decisions based on point-predictions. DL systems typically returns softmax scores that are proportional to the systems confidence of the prediction - not calibrated probabilities [14].

*petter.risholm@sintef.no

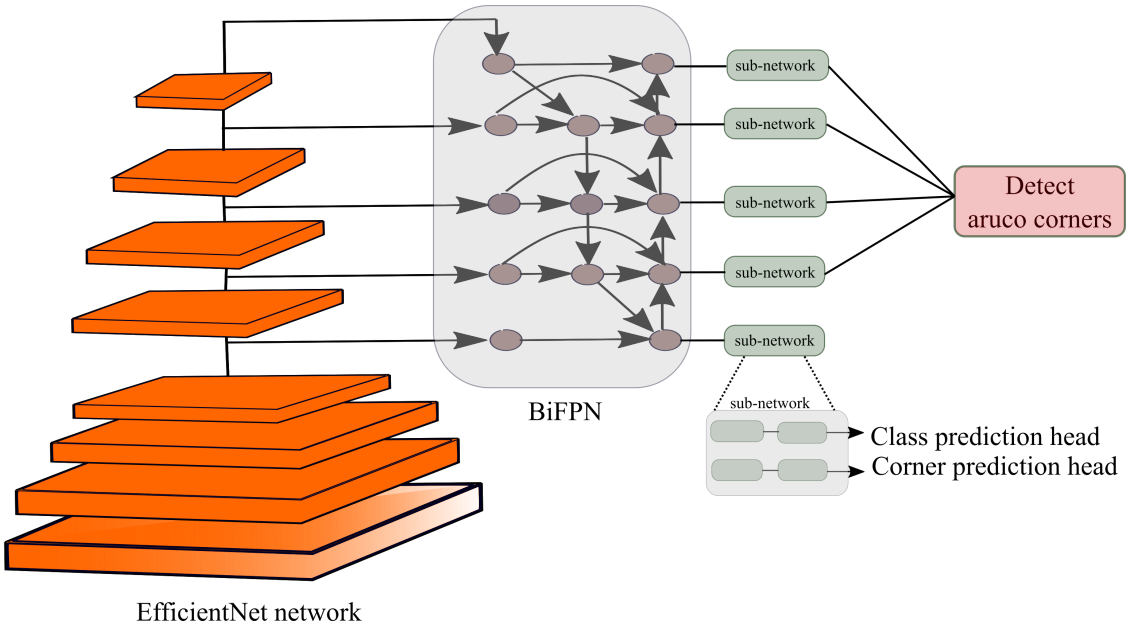


Figure 1: Network structure. We use an EfficentNet backbone network for feature extraction, a bi-directional Feature Pyramid Network (BiFPN) for efficient feature fusion, and separate class and corner prediction heads.

Recent advances have shown potential for combining Bayesian methodology with DNNs to effectively reason about model and data uncertainty to characterize the distribution over outcomes [12]. These methods can incorporate data uncertainties in the prior and characterize the full predictive uncertainty. These methods are often based on computationally expensive sampling schemes such as Markov Chain Monte Carlo (MCMC), however approximations such as Dropout sampling [7] and ensemble methods [14] have been proposed to alleviate this.

We propose a system for 6-DoF estimation of Aruco markers with associated uncertainties in the challenging underwater environment. A state-of-the-art object detection framework (EfficientDet) was adapted to predict the corner locations of Aruco markers, while dropout sampling at inference time is used to estimate the predictive 6-DoF uncertainty. A dataset of Aruco markers captured in a wide range of turbidities, with ground truth position of the corner locations, was gathered and used to train the network to robustly predict the 6-DoF pose.

We present results, both in terms of rotational and translational errors with associated robust uncertainty measures and how this may be useful to reduce the operational risk in an autonomous subsea intervention procedure.

2. Methods

2.1. Prediction of 6-DoF using EfficientDet

We build our Aruco regressor on top of EfficientDet [21] which is a state-of-the-art object detector. EfficientDet uses an EfficientNet backbone and a Bi-directional Feature Pyramid Network (BiFPN) for feature fusion which also shares features across scales. The classification and corner prediction heads consists of three-layer convolutional networks the same width as the output of the BiFPN layer. EfficientDet uses anchor bounding boxes to regress object bounding boxes. We have adapted the 4-component $(x, y, width, height)$ box prediction of the original EfficientDet to instead predict the 8-component vector of the four corner coordinates: $\mathbf{c}_i = [x_1, y_1, x_2, y_2, x_3, y_3, x_4, y_4]$. The loss function of the regression head then becomes: $L_r(\mathbf{c}; \hat{\mathbf{c}}, \delta) = \frac{1}{N} \sum_{i=1}^{N} E(\mathbf{c}_i, \hat{\mathbf{c}}_i; \delta)$, where $\hat{\mathbf{c}}$ is the ground truth corner locations and E is a Huber loss. Each regression difference (predicted corner position versus the annotation difference) is scaled by the respective anchor box width / height. A standard focal loss [15] is used for the classification head.

2.2. Uncertainty characterization using MC-dropout

Given a predicted set of four corner points $\mathbf{c}_i$, we use a perspective 4-point algorithm to estimate the 6-DoF of the Aruco marker in the camera coordinate system. It has been shown that a neural network with dropout applied before

each weight layer is equivalent to a (variational) approximation to the deep Gaussian process (marginalised over the covariance function parameters) [7]. In practice, standard dropout is not applied before each weight layer, but rather in the last layers in the model, and this has been shown to produce equivalent results. We apply dropout on the second and third to last convolutional layers in the regression head.

In a Bayesian perspective, dropout is equivalent to drawing samples from: $\mathbf{c}_i \sim \int p(\mathbf{c}_i|I_i,\theta)p(\theta)d\theta$, where I_i is an image of an Aruco marker and θ describes the neural weight parameters. The predictions from a standard Gaussian Process is by definition a Gaussian. However, in our case, the Markov Chain (MC) dropout approach is only an approximation to a Gaussian process and may be multi-modal and potentially severely skewed as can be seen in Figure 6. Hence, we cannot assume that the mean and standard deviation are robust and accurate summaries of the MC-dropout samples. The primary mode of the distribution is the most likely outcome, but can be difficult to robustly estimate in multi-modal settings. Consequently, we choose to use the median as a proxy for the mode, and use interquartile range (IQR) as robust uncertainty summaries.

2.3. Dataset

A dataset was established of Aruco markers imaged in a small aquarium under a large span of turbidities. The aquarium measured 100cm x 50cm x 60cm and is shown in Figure 2. A Toshiba BU238M machine vision camera was placed at the short end of the aquarium. A halogen lamp and a LightCrafter 4500 DMD was used for illumination of the scene. Three Aruco markers (7,9 and 10 from DICT_4x4_250) with sides of 10cm were placed in five different geometrical configurations as shown in Figure 3. The markers were rigidly attached to a wood plate laying on top of the aquarium. They were located approximately 0.7m-0.9m from the camera.

For each static geometrical configuration of the markers, we acquired a set of 30 images for each permutation of pre-defined lighting conditions, exposures and turbidities. Images were acquired at a resolution of 1900x1200, but scaled down to 512x512 with fixed aspect ratio to be suitable for input in the CNN network. All external lighting was turned off for the acquisition. Lighting configurations included: 1) Halogen lamp behind zero, one and two sheets of diffusive paper, 2) LightCrafter projector. Fifteen exposure times were used within the range 30μs to 16 400μs. In total, we acquired 4 950 images for each configuration.

The pool was filled with fresh tap-water before adding 1g of blue modelling clay for each increase of turbidity. We acquired a full set of images for five different turbidities. Figure 4 shows example images across turbidities for one of the configurations. We use the attenuation length of the water as a proxy for the turbidity and measured it with a camera and blinking light source as shown in Figure 2a. The attenuation length is defined as the distance at which $\frac{1}{e} \approx 37\%$ amount of light remains. The estimation procedure we employ is described in [19]. We estimated the attenuation to be 8.6m in clear tap water (turbidity 0), and down to 0.3m after adding a total of 4g of dissolved blue clay (turbidity 4).

We used the images acquired at turbidity 0 to establish a ground truth location of the Aruco corners which would be valid across the turbidities, lighting and exposure variations. We used the standard Aruco library to compute the mean ground truth corner locations of the markers using the 10 exposures of the lighting and exposure configuration of the marker which provided the best signal to noise ratio (SNR) over the marker. The SNR was computed by taking the signal of the pixels representing the white parts of the marker over the noise of the pixels representing the dark parts of the marker.

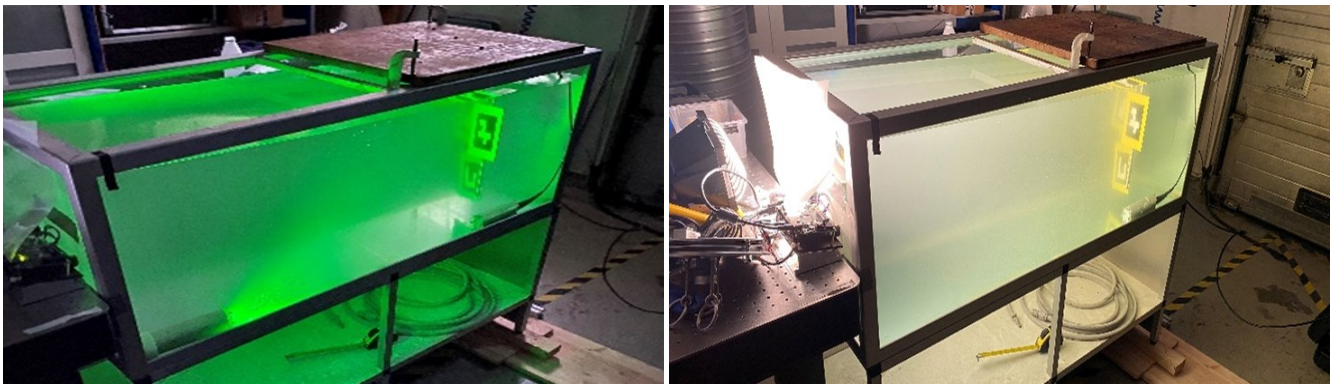
The intrinsic calibration matrix was established through a standard checkerboard calibration [23].

2.3.1 Dataset augmentations

The static dataset only includes the markers at a small set of geometrical configurations but includes a diverse range of lighting and turbidity configurations. Hence, we add augmentations that varies the geometrical configuration of the markers as seen in Figure 5. We apply a random affine transformation to simulate variations in distance (scaling), position (translation) and orientation (shearing, rotation). Random crops were added to the image (see Figure 5b) to make sure that the network learns the marker and not its location in relation to the background. We also added random noise over markers that ended up on the image boundary as shown in Figure 5c.

2.4. Training and inference

The dataset was divided into a training set using images from configurations 1-4, while images from configuration 5 was used for the test set. We trained the network for 200 epochs, using mini-batches of 16 images, and observed that the loss converged after approximately 150 epochs. We used a drop-out rate of 30% during training for regularization purposes, and 0.0% during test time when we evaluate the loss function. During inference, post training, we used a dropout rate of 30% and drew 1 000 samples to characterize the distributions. For the classification (focal) loss we used $\alpha = 0.25$ and $\gamma = 2.0$, while the Huber regression loss function used $\delta = 1/9$.



(a) (b)

Figure 2: Acquisition setup. 2a The attenuation measurement setup. A blinking spotlight is located at the bottom left of the aquarium, while the camera is located on the bottom right of the aquarium. 2b Acquisition of the Aruco markers. The camera and light sources are located on the left side of the aquarium.

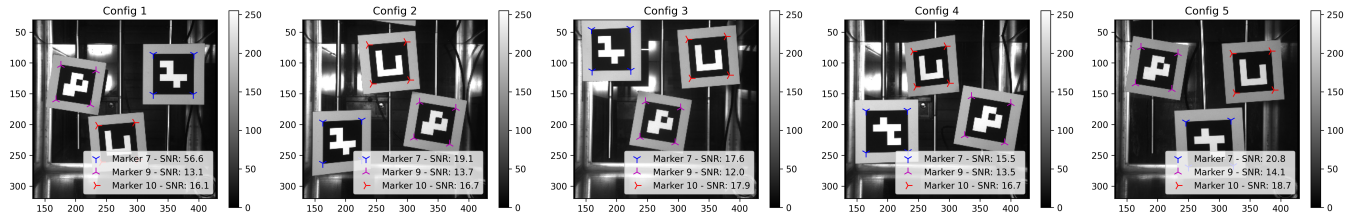


Figure 3: Aruco marker configurations. The legend shows the marker id and the corresponding marker SNR.

3. Results

3.1. 6-DoF estimation in turbid waters

This section summarizes the results on the test set (i.e. configuration 5). In Figure 6, we show the inferred distributions for each corner location of marker 9 across turbidities at the same lighting and camera exposure of 1ms. At attenuation length 0.3m the SNR for the marker is 0.4. It is very difficult to discern the marker visually, and the network produces wider distributions where some of them are multimodal.

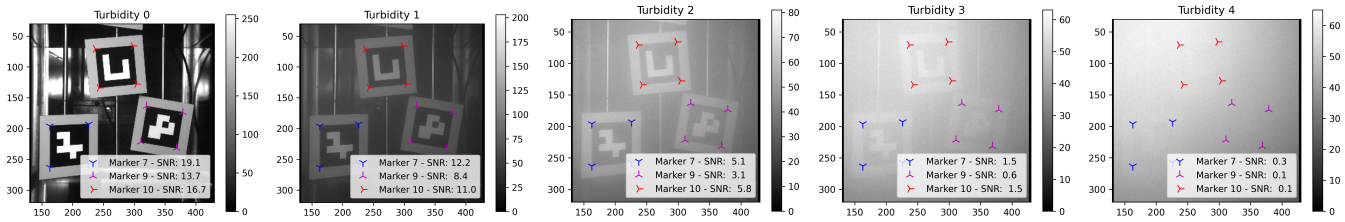
In Figure 7 (left) we show results of the estimated translational error in centimetres in relation to the turbidity for the three translational components. In Figure 7 (right) we show the translational error for the three translational components in relation to the SNR of the marker. Since these results aggregate information across exposures and lighting conditions, there is not a direct relationship between the SNR and the turbidity. The uncertainty (IQR) is stable at approximately 0.4cm, 0.5cm and 3cm for the three components down to an SNR of about 3, where the uncertainty increases rapidly. The z-uncertainty is considerably higher because of the PnP computation which effectively aggregates the x- and y- uncertainty.

The rotational error estimates are shown in Figure 8. The error was computed as the single axis rotation between the ground truth and estimated coordinate frames. The figure shows the rotational error in relation to both the attenuation length and the SNR. We observe the same effect as for the translation, that the rotational error and uncertainty increases at low SNRs/high turbidity.

We summarize robust measures of the translational and rotational errors in Table 1. The accuracy and detection rate of the OpenCV Aruco library using standard detection parameters is also reported. We observe that the detection rate falls off rapidly with higher turbidities, while the proposed method provides a detection rate of 100%. However, one should probably characterize detections with a high uncertainty as a missed detection, and consequently the detection rate would be reduced accordingly.

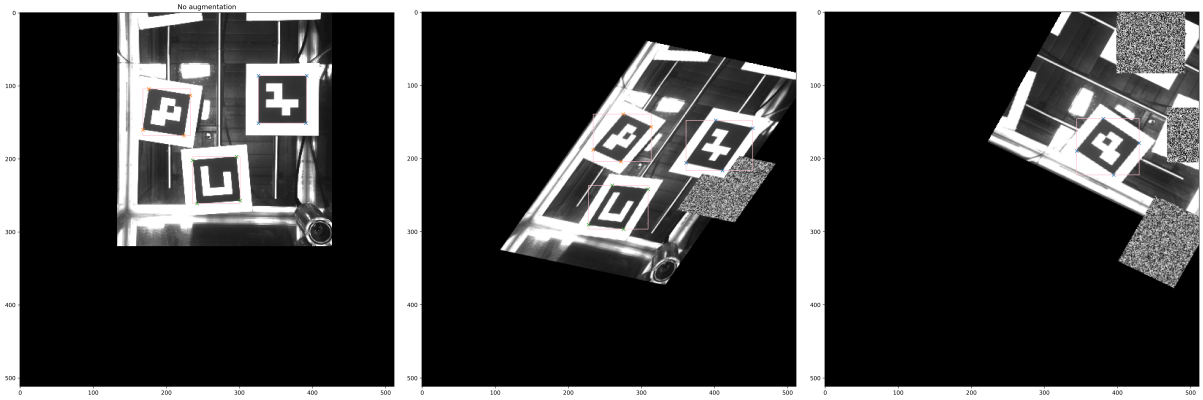
3.2. Subsea intervention: gripping of a fish-tail handle

One direct application of the proposed approach to 6-DoF pose estimation is for autonomous interventions subsea. If an Aruco marker is rigidly placed in relation to a



(a) Turbidity 0: 8.6m (b) Turbidity 1: 1.2m (c) Turbidity 2: 0.7m (d) Turbidity 3: 0.4m (e) Turbidity 4: 0.3m

Figure 4: Example images from across turbidities. Images were acquired of the same configuration under five different turbidites ranging from 8.6m down to 0.3m attenuation length. The legend shows the marker id and the corresponding marker signal to noise ratio.



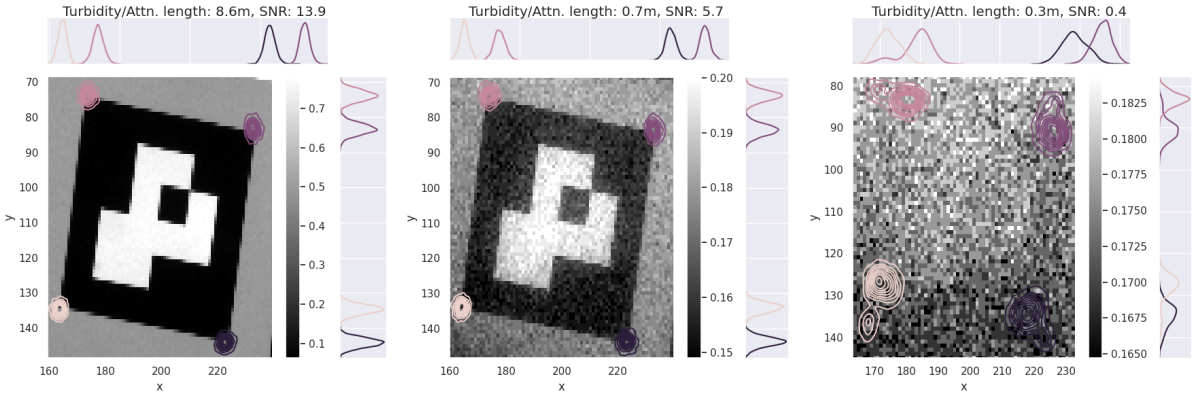
(a) (b) (c)

Figure 5: Augmentations. Figure 5a An original image without augmentation. The marker corners are overlaid. To span the space of possible configurations we applied random affine transformations (translation, rotation, shear and scale) to the images. Two examples are shown in Figure 5b and Figure 5c. We also added random crops filled with random values to the images to make sure that the network learns the marker and not the location of the marker on the background (see 5b). Markers which end up on the image boundary after augmentation are removed by filling the bounding box with random values. This can be seen in Figure 5c.

Turbidity[m]	T%25[cm]	T%50[cm]	T%75[cm]	R%25[deg]	R%50[deg]	R%75[deg]	Aruco T%25	Aruco T%50	Aruco T%75	ArucoDetectionRate	DetectionRate
8.6	1.0	2.0	3.2	3.1	4.6	7.8	0.0	0.0	12.3	100.0	100.0
1.1	1.1	2.3	3.9	3.3	5.0	8.1	0.0	0.0	0.0	51.7	100.0
0.7	1.2	2.4	4.1	3.6	5.4	8.5	0.0	0.0	6.1	11.7	100.0
0.4	1.3	2.5	4.0	4.5	6.5	8.8	NaN	NaN	NaN	0.0	100.0
0.3	1.8	3.8	10.6	5.7	8.1	10.9	NaN	NaN	NaN	0.0	100.0

Table 1: Quantitative results. Robust estimates of the translational (T) and rotational (R) errors. We report the median (%50) and the lower (%25) and upper (%75) IQRs. Notice that none of these error estimates can go negative, while the component-wise measures in Figure 7 can go negative. We also report the detection rate based on the OpenCV Aruco library and the proposed method, where NaN means that no markers were detected. One may argue that the detections with the proposed method should be classified according to the uncertainty, which would result in a decline in detection rate according to the user specified uncertainty threshold.

fish-tail handle which the AUV should intervene with, the AUV can automatically position itself and the gripper in relation to the fish-tail. The gripping procedure can be adjusted according to the uncertainty of the pose estimate of the fish-tail. With high uncertainty, the movements can be slower, and the gripper can open up more before closing up the gripper. This will help reduce the risk of damaging the gripper and the fish-tail. This is of course a bit simplis-



(a) (b) (c)

Figure 6: Corner prediction overlaid on image for attenuation lengths 8.6m (6a), 0.7m (6b) and 0.3m (6c). We have plotted iso-contours of the marginal probability of the different corner locations. The prediction is relatively stable across turbidities, except for at the highest turbidity where there is limited signal. At the highest turbidity we observe that the distributions are multimodal and considerably wider than at lower turbidities.

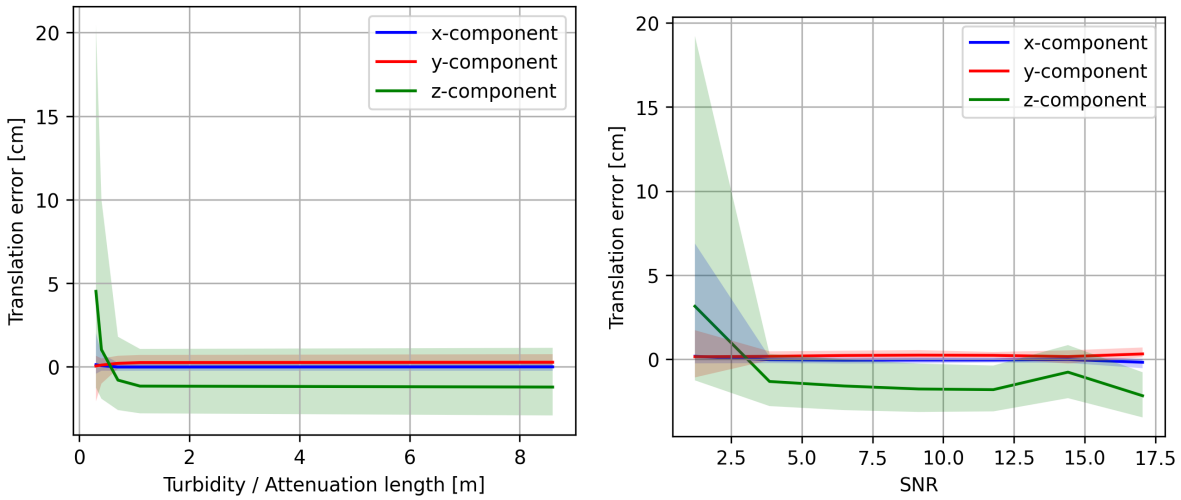


Figure 7: Translational errors across turbidities (left) and across signal-to-noise ratios (right).

tic picture of the challenging intervention procedure which involves a number of complex tasks such as floating base-control [2].

In Figure 9, we show an example where we have used the pose distribution (1000 samples) given by the proposed algorithm when detecting an Aruco marker at an attenuation length of 0.7m at 90cm distance. We create a marginal probability volume around the gripper which tells us the probability of the fish-tail being present in that particular voxel. This is done by transforming the fish-tail handle with the 1000 pose samples and for each time a voxel is inside the transformed model an accumulator is incremented for that voxel. When the gripper (the yellow model to the right in the figure) is closing its grip, we can report the probability of whether it is now gripping the fish-tail. The snapshot of the gripping process in the figure shows that the probability of the grip is 99%.

4. Discussion

Underwater vehicles operating autonomously are dependent on having localization systems that can provide 6-DoF pose estimates that are both accurate and where the associated uncertainty of the prediction is characterized. Making high-risk decisions without knowing the full predictive distribution over the 6-DoF pose can lead to catastrophic out-

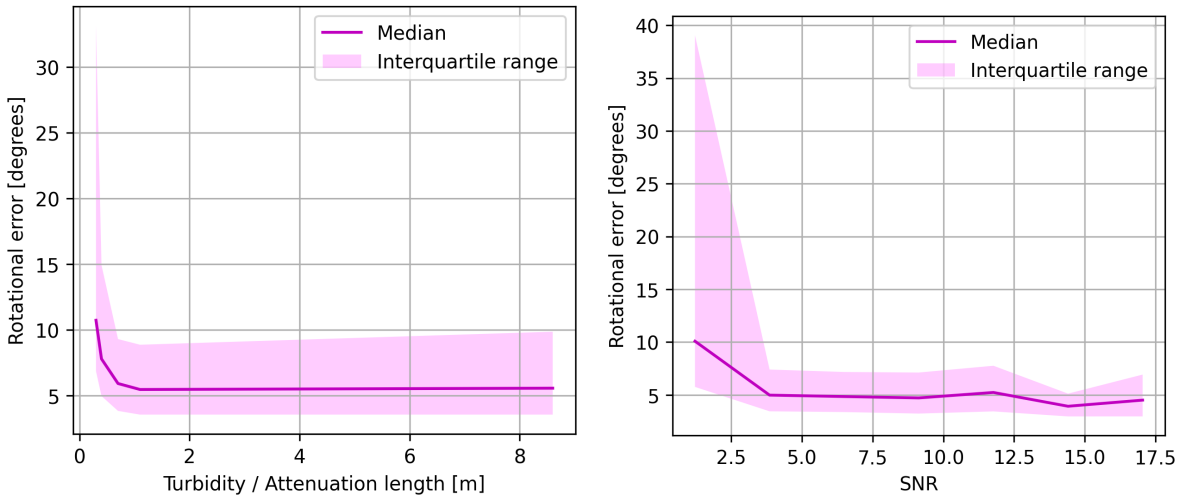
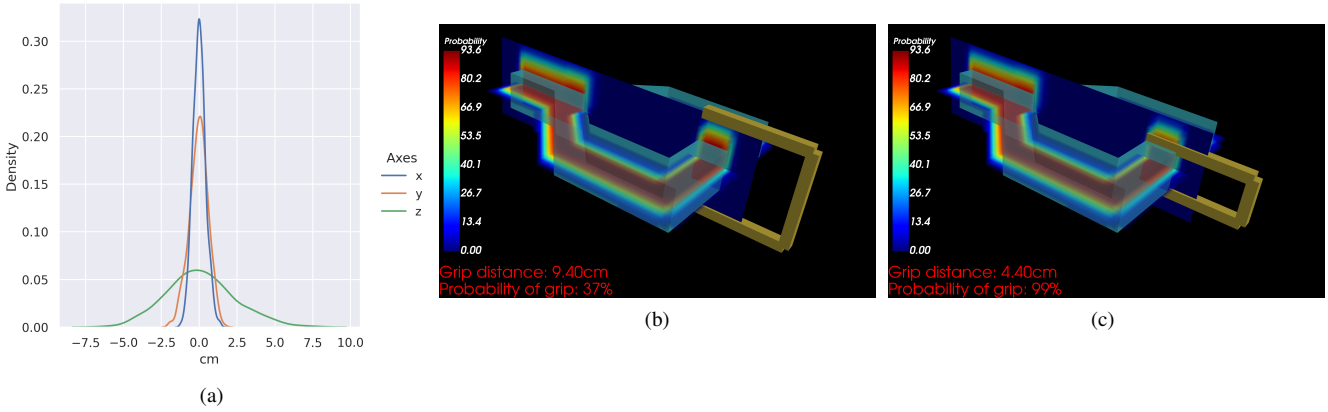


Figure 8: Rotational errors across turbidites (left) and across signal-to-noise ratios (right).



(a) (b) (c)

Figure 9: Qualitative example of gripping a fish-tail. The pose of the fish-tail is computed based on the marker shown in Figure 6b. 9a shows the distribution of the translational components centered around the median value. A probability volume of the location of the fish-tail is constructed based on the pose distribution. In 9b and 9c we show example visualizations of the gripper versus the probability volume of the fish-tail at two different stages of the grip process. This probabilistic information can help the AUV to decide how to approach and grip the fish-tail with minimum risk of damaging the gripper or fish-tail.

comes, such as loss of costly equipment or human lives. Consequently, the uncertainty threshold which is applied during operations should be adjusted according to the operational risk.

We have introduced a method for estimating the uncertainty of 6-DoF estimates based on Aruco markers. The network was trained on a realistic dataset of three Aruco markers acquired underwater in a wide range of turbidities and lighting conditions. We used the Aruco detections in clear water as a proxy for the ground truth detections, and acknowledge that there may be minor sub-pixel deviations from the real pixel positions. We observed median translational errors ranging from 2.6cm at low turbidity to 10.5cm at high turbidities. The respective IQRs (uncertainties) are 3.2cm up to 27.9cm. The rotational median errors varied from 5.6° to 10.7° with IQRs of 6.4° to 26.2°. Our translational errors across all turbidities are lower than the 11.8cm that is reported in clear water in [17], while their rotational error of 4.2° is approximately the same as we report for the clear water case.

Blue clay was used to generate a turbid environment which exhibits absorption and scattering. It may be that the optical effects (absorption, back- and forward-scattering) of other sediments is different. We have not verified how well

the results generalize to other turbid environments. We only applied the augmentations to the training dataset, hence we did not have any strongly inclined markers in the test set. We expect that the accuracy will decrease and uncertainty will increase with more inclined markers. This will be investigated in future work.

Markers associated with a 250-word dictionary was used in this study. We have not studied the sensitivity of the detections to the size of the dictionary. However, we observed minimal mis-classification of the markers with the 250-word dictionary, even at high turbidities, so it is unlikely that a smaller dictionary would lead to more reliable detections.

One interesting question is whether the generated distributions are well calibrated, i.e. do they accurately encapsulate the real uncertainty of the model? There are a number of approximations involved when establishing the identity between the MC-dropout sampling of a general-purpose CNN and sampling from a Gaussian process, which may cause the distributions to be uncalibrated. The drop-out rate as well as the location where the drop-out is applied may also affect the spread of the distributions. We chose a drop-out rate in line with what other authors have reported. We have not performed any checks to validate the sensitivity of the results to the drop-out rate. One approach to validate the distributions may be to acquire N images of the same Aruco marker and generate corner predictions for each of them. For a well calibrated distribution, the ground truth position of the corner should, for a given x, be within the x-percentile range $x\%$ of times.

Inference time (50ms per sample on an RTX 2080, so if 100 samples is adequate to characterize the distribution it would take about 5s) is a limiting factor for applying the proposed uncertainty estimation scheme in a practical setting. In future work we will evaluate the trade-off between the number of samples and the accuracy of the characterization of the distribution. We will also investigate further the implications of having the full 6-DoF distribution on decision making during high-risk underwater operations.

5. Acknowledgements

This research was funded by the Norwegian Research Council, grant number 280934. The work was carried out in the SEAVENTION project (www.sintef.no/SEAVENTION). The authors acknowledge the valuable input from the project partners Equinor, Norwegian University of Science and Technology (NTNU), FMC Technologies, IKM and Oceaneering.

References

[1] Yoshua Bengio, Yann LeCun, et al. Scaling learning algorithms towards ai. *Large-scale kernel machines*, 34(5):1–41, 2007. 1

[2] A. Birk, T. Fromm, C. A. Mueller, T. Luczynski, A. Gomez Chavez, D. Koehntopp, A. Kupcsik, S. Calinon, A. K. Tanwani, G. Antonelli, P. Di Lillo, E. Simetti, G. Casalino, G. Indiveri, L. Ostuni, A. Turetta, A. Caffaz, P. Weiss, T. Gobert, B. Chemisky, J. Gancet, T. Siedel, S. Govindaraj, X. Martinez, and P. Letier. Dexterous underwater manipulation from distant onshore locations. *IEEE Robotics and Automation Magazine*, 2018. 6

[3] Yannick Bukschat and Marcus Vetter. Efficientpose–an efficient, accurate and scalable end-to-end 6d multi object pose estimation approach. *arXiv preprint arXiv:2011.04307*, 2020. 1

[4] Zhaowei Cai and Nuno Vasconcelos. Cascade r-cnn: Delving into high quality object detection. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 6154–6162, 2018. 1

[5] Jan Čejka, Fabio Bruno, Dimitrios Skarlatos, and Fotis Liarokapis. Detecting square markers in underwater environments. *Remote Sensing*, 11(4):459, 2019. 1

[6] Diego Brito dos Santos Cesar, Christopher Gaudig, Martin Fritsche, Marco A dos Reis, and Frank Kirchner. An evaluation of artificial fiducial markers in underwater environments. In *OCEANS 2015-Genova*, pages 1–6. IEEE, 2015. 1

[7] Yarin Gal and Zoubin Ghahramani. Dropout as a bayesian approximation: Representing model uncertainty in deep learning. In *international conference on machine learning*, pages 1050–1059. PMLR, 2016. 2, 3

[8] Sergio Garrido-Jurado, Rafael Muñoz-Salinas, Francisco José Madrid-Cuevas, and Manuel Jesús Marín-Jiménez. Automatic generation and detection of highly reliable fiducial markers under occlusion. *Pattern Recognition*, 47(6):2280–2292, 2014. 1

[9] Ross Girshick. Fast r-cnn. In *Proceedings of the IEEE international conference on computer vision*, pages 1440–1448, 2015. 1

[10] Kaiming He, Georgia Gkioxari, Piotr Dollár, and Ross Girshick. Mask r-cnn. In *Proceedings of the IEEE international conference on computer vision*, pages 2961–2969, 2017. 1

[11] Danying Hu, Daniel DeTone, and Tomasz Malisiewicz. Deep charuco: Dark charuco marker pose estimation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 8436–8444, 2019. 1

[12] Alex Kendall and Yarin Gal. What uncertainties do we need in bayesian deep learning for computer vision? In I. Guyon, U. V. Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, and R. Garnett, editors, *Advances in Neural Information Processing Systems*, volume 30. Curran Associates, Inc., 2017. 2

[13] Alex Kendall, Matthew Grimes, and Roberto Cipolla. Posenet: A convolutional network for real-time 6-dof camera relocalization. In *Proceedings of the IEEE international conference on computer vision*, pages 2938–2946, 2015. 1

[14] Balaji Lakshminarayanan, Alexander Pritzel, and Charles Blundell. Simple and scalable predictive uncertainty estimation using deep ensembles. In I. Guyon, U. V. Luxburg,

S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, and R. Garnett, editors, *Advances in Neural Information Processing Systems*, volume 30. Curran Associates, Inc., 2017. 1, 2

[15] Tsung-Yi Lin, Priya Goyal, Ross Girshick, Kaiming He, and Piotr Dollár. Focal loss for dense object detection. In *Proceedings of the IEEE international conference on computer vision*, pages 2980–2988, 2017. 2

[16] Francesc Moreno-Noguer, Vincent Lepetit, and Pascal Fua. Accurate non-iterative o(n) solution to the pnp problem. In *2007 IEEE 11th International Conference on Computer Vision*, pages 1–8. IEEE, 2007. 1

[17] Christian A Mueller, Tobias Doernbach, Arturo Gomez Chavez, Daniel Koehntopp, and Andreas Birk. Robust continuous system integration for critical deep-sea robot operations using knowledge-enabled simulation in the loop. In *2018 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, pages 1892–1899. IEEE, 2018. 1, 7

[18] Joseph Redmon and Ali Farhadi. Yolo9000: better, faster, stronger. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 7263–7271, 2017. 1

[19] Petter Risholm, Jostein Thorstensen, Jens T Thielemann, Kristin Kaspersen, Jon Tschudi, Chris Yates, Chris Softley, Igor Abrosimov, Jonathan Alexander, and Karl Henrik Haugholt. Real-time super-resolved 3d in turbid water using a fast range-gated cmos camera. *Applied optics*, 57(14):3927–3937, 2018. 3

[20] Mingxing Tan and Quoc Le. Efficientnet: Rethinking model scaling for convolutional neural networks. In *International Conference on Machine Learning*, pages 6105–6114. PMLR, 2019. 1

[21] Mingxing Tan, Ruoming Pang, and Quoc V Le. Efficientdet: Scalable and efficient object detection. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 10781–10790, 2020. 1, 2

[22] Bugra Tekin, Sudipta N Sinha, and Pascal Fua. Real-time seamless single shot 6d object pose prediction. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 292–301, 2018. 1

[23] Zhengyou Zhang. A flexible new technique for camera calibration. *IEEE Transactions on pattern analysis and machine intelligence*, 22(11):1330–1334, 2000. 3

[24] Marek Žuži, Jan Čejka, Fabio Bruno, Dimitrios Skarlatos, and Fotis Liarokapis. Impact of dehazing on underwater marker detection for augmented reality. *Frontiers in Robotics and AI*, 5:92, 2018. 1