Increasing diver safety for heavy underwater works by Sonar-to-Video Image Translation

Jan Lorscheidt
German Federal Agency for Technical Relief
Jan.lorscheidt@thw.de

Bilal Wehbe
DFKI - Robotics Innovation Center
bilal.wehbe@dfki.de

Diego Cesar
Kraken Robotik GmbH
dcesar@krakenrobotics.com

Tom Becker
DFKI - Robotics Innovation Center
tom.becker@dfki.de

Thomas Vögele
DFKI - Robotics Innovation Center
thomas.voegele@dfki.de

ABSTRACT

Supervision of technical dives is particularly important in emergency and disaster response operations to ensure the safety of divers in unexplored locations with uncertain conditions. Diver monitoring relies primarily on voice communication and a video stream that gives the operator a first-person view of the diver. However, in many cases underwater visibility can drop to just a few centimeters, leaving the diver only able to feel his way with his hands and the operator depended only on voice communication, making it very difficult for both of them to identify upcoming hazards. In the DeeperSense research project, we are attempting to reduce the limitations caused by poor underwater visibility by using a sonar in combination with an AI-based algorithm designed to translate the sonar signal into a visual image that is independent of the turbidity of the water and gives an overview of the situation where the eye can no longer see anything. Laboratory results show that visual information can be recovered from sonar data.

Keywords

diver monitoring, GANs, sonar sensor fusion, marine perception

INTRODUCTION

In many underwater applications, technical divers are still indispensable when it comes to maintenance and inspection. Industrial divers regularly have to work in dangerous situations, but with good planning of the dives, only a controllable residual risk remains. Technical diving in civil protection and emergency response operations, however, presents different challenges as the situation at the place of action is often unknown and the risk level sometimes has to be assessed during the dive. In contrast to recreational divers who primarily target environments with good visibility and lighting, industrial divers mostly face low visibility situations that make navigating, self-locating and finding objects or recognizing hazards difficult.

To alleviate these difficult circumstances, industrial divers are assisted by an on-land operator who monitors the diver's activities and helps anticipate upcoming dangers. The operator has a voice link with the diver and receives a 1st person view by a camera that is attached to the diver's helmet. By this means, the operator can discuss next work steps in consultation with the diver and alert him if he notices any discrepancies or emerging dangers, while the diver can focus on the task at hand. Technological advances in the past decades helped to further improve the monitoring of the diver through the use of small remotely operated vehicles (ROV), which have become

increasingly available on the market. In some cases, ROVs are used to get a 3rd person view of the diver and enable the operator to get a better overview of the situation and to support the diver more actively. However, while this strategy works well to enable productive work and mitigate risks, it is mainly based on visual input that depends on the turbidity of the water. Even if visibility is good at the beginning of the dive, activities performed by the divers often cause the turbidity of the water to increase and can lead to white-out situations in which the operator loses sight of the situation.

Thus, the lack of visibility is a major problem in reliable diver monitoring. In order to overcome these limitations, sound waves can be used instead of light, because sound waves are not affected by the turbidity of the water. Sonar imaging is a state-of-the-art technology that has been improved over the last decades and high-resolution devices are available on the market that can be carried by a ROV and provide the ability to obtain a sonar image of the scene for monitoring. However, sonar images are often very hard to interpret even to a trained human eye and suffer usually from low signal-to-noise ratios. Therefore, we aim to use deep learning methods to learn a consistent association between sonar and optical camera images observing the same underwater scene. By this learned association the sonar input shall be translated into a realistic visual-like image, that can be easily interpreted by a human operator monitoring technical divers working in low visibility environments.

Within the European research project DeeperSense, the described use case is being elaborated. The necessary training data for the deep-learning algorithm is generated in multiple diving sessions with technical divers of the German Federal Agency for Technical Relief (THW). The sessions are simultaneously recorded by an optical camera and a sonar by the German Research Center for Artificial Intelligence (DFKI) and industrial partner KRAKEN Robotik. As deep-learning methods require a huge amount of input data for training and testing of the algorithm, multiple realistic working situations will be staged by the divers in different locations and settings in order to prevent the algorithm from overfitting. After the training phase, the algorithm developed by the researchers of DFKI is being tested within application-oriented scenarios and it is evaluated whether the technology can add value for the application of diver monitoring in low visibility environments.

RELATED WORK

Underwater image enhancement and restoration has been an active field of research due to the difficulties marine environments pose for optical imaging. Phenomena such as marine backscatter reduces the contrast of an image and produces fogginess, floating organic matter known as “marine snow" could create occlusions, additionally the colors of light dissipate in the water with depth.

Classical methods to improve the quality of underwater images generally involve techniques such as (1) noise filtering (Arnold-Bos, et al., 2005) (Bazeille, et al., 2006) (Jia & Ge, 2012), (2) color correction (Akkaynak, et al., 2014) (Berman, et al., 2020), and (3) image dehazing (Treibitz & Schechner, 2009) (Berman, et al., 2016).

In recent years, Deep Learning based methods for underwater image enhancement have been gaining more attention. A Generative Adversarial Network (GAN) was trained in (J. Li, et al., 2018) to generate underwater images from a pair of in-air optical and depth images. Their network was used then to generate a large dataset for training another image restoration network that predicts a color corrected underwater image. This method, however, only handles color correction and is not trained to remove haziness or restore images that are corrupted with floating particulate matter.

This work proposes the use of GANs in order to recover underwater images that have been corrupted by adding blur and darkness. The GAN is trained to fuse a sonar image and the corrupted camera image in order to recover the original clear color image.

MACHINE LEARNING CONCEPT

The core idea behind the machine learning (ML) concept is to learn an end-to-end association between an imaging sonar and visual camera images observing the same underwater scene. This learned association would then be used to generate realistic visual-like images, given only sonar images as input or a combination of a sonar image and a dark or turbid visual image. The purpose is to provide images that can be easily interpreted by a human operator, even in bad visibility conditions, for example to monitor the status of a diver working in turbid or dark waters.

In this work we propose the use of a class of ML algorithms known as Conditional Generative Adversarial Networks (C-GAN) (Mirza & Osindero, 2014), that has become popular in the literature for the tasks of image-to-image translation (P. Isola, et al., 2017) and (Wang, et al., 2018). This network architecture is mainly composed of an encoder-decoder style Generator G(x) network that maps an input image to a desired target image by

minimizing the pixel loss between the predicted and real image. Additionally, a Discriminator network D(x) is trained simultaneously which randomly takes a generated image or the corresponding real image as input and tries to predict whether the given image is fake or real. In this case the input image acts as a condition that is imposed on the generator and discriminator inputs to compel the network to perform image translation tasks.

For our use case, we use the sonar data and a camera image that has been blurred and darkened to simulate bad visibility conditions as input. Both inputs are first passed through a 3 x 3 convolutional layers and then concatenated. This is then passed into the generator network that is composed of a U-Net encoder-decoder architecture (Ronneberger, et al., 2015), with 7 up and downsampling CBR (Convolutional-BatchNorm-ReLU) layers. The discriminator network is composed of a PatchGAN (P. Isola, et al., 2017), see Figure 1:

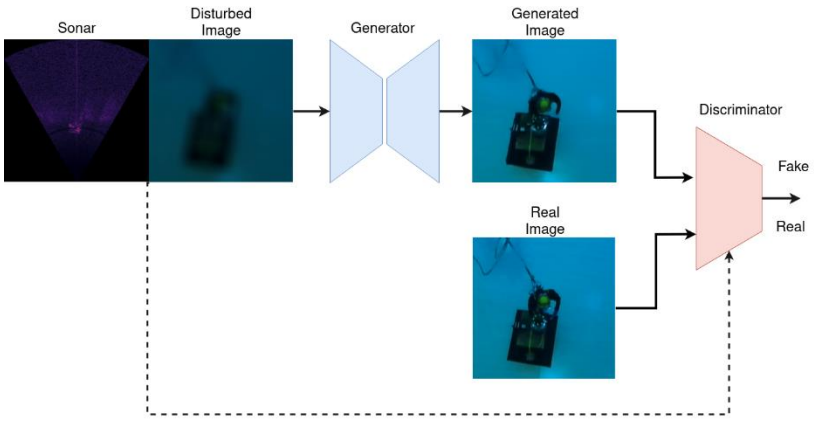

Figure 1: Diagram representing the CGAN architecture

Sensor Setup and Data Collection Sessions

In the data collection sessions, the training data for the algorithm training is being generated. In every session there is a scene consisting of a diver performing technical underwater tasks and a set of sensors that are aligned to record the same field of view of the scene. The sensor set comprises a multi-beam imaging sonar, an optical stereo camera and an underwater laser scanner. The sensors are synchronized and record data simultaneously in order to obtain matching images of the scene from different modalities. The image of the optical camera is used as a ground truth for the visual-like image that shall be generated by the algorithm. In the initial phase of the algorithm development, no depth information from the stereo camera nor 3D laser scanner have been used for training, but data is collected for future iterations. A typical setup for data generation is shown in Figure 2:

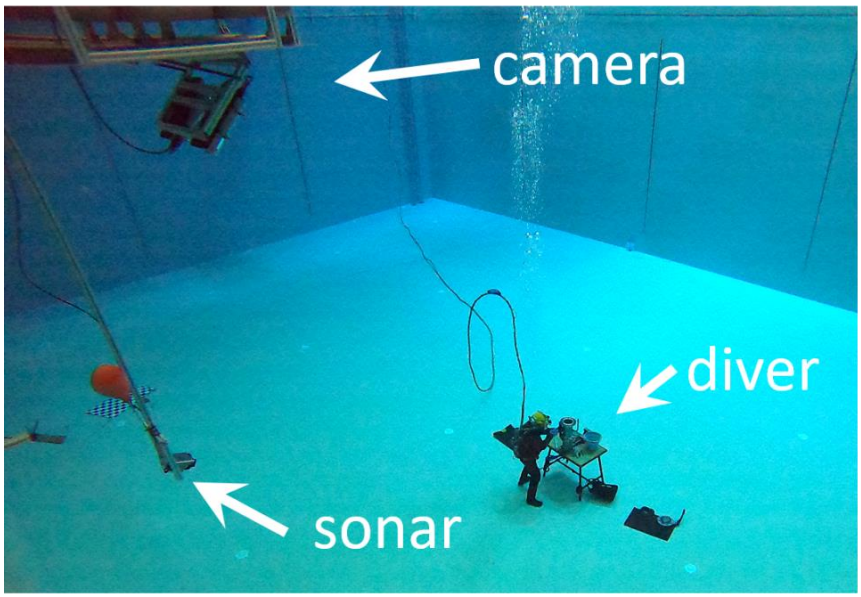

Figure 2: Data Generation Sensor Setup

This sensor setup was used in all data acquisition sessions with slight variations and it was either fixed on a floating platform or fixed at the basin wall. Between August 2021 and July 2022 four sessions were conducted, at three different locations in Germany: Two sessions under laboratory conditions at the Maritime Exploration Hall at DFKI-RIC in Bremen, one field session at the Kreidesee in Hemmoor, and one field session at a training basin in Neu-Ulm. In all cases, a typical underwater work environment for the divers was simulated, including workbench and tools. Divers performed various tasks such as assembling pipe flanges, tightening nuts and bolts, or building fittings.

Laboratory sessions 1 and 2 in The Maritime Exploration Hall

The Maritime Exploration Hall contains a saltwater basin measuring 23m × 19m × 8m, Figure 3 shows a diver entering the basin. The environment conditions in the Maritime Exploration Hall are stable and independent of the outside weather, the visibility is greater than 50 m.

Figure 3: Maritime Exploration Hall

The main goal of the laboratory sessions is to generate very clear optical images of the divers performing their work in order to obtain a high-quality ground truth. Therefore, it is important to have unrestricted visibility given by a minimum turbidity of the water. Furthermore, we aim for a homogeneous background to ensure that the diver and his equipment are the only visible objects in the scene. These optimal circumstances we encounter in the Maritime exploration hall, which is why it was chosen for the laboratory sessions. The only limitation is that we cannot carry out tasks that would pollute the pool and impair the visibility.

For the first session the focus was to verify the overall setup of the data collection and therefore it was decided to only perform simple works like mounting pipe flanges, wrenching nuts and bolts and moving objects. Figure 4 shows an example of the collected data as a matching pair of images that were taken by an optical camera and a Gemini 720i Sonar.

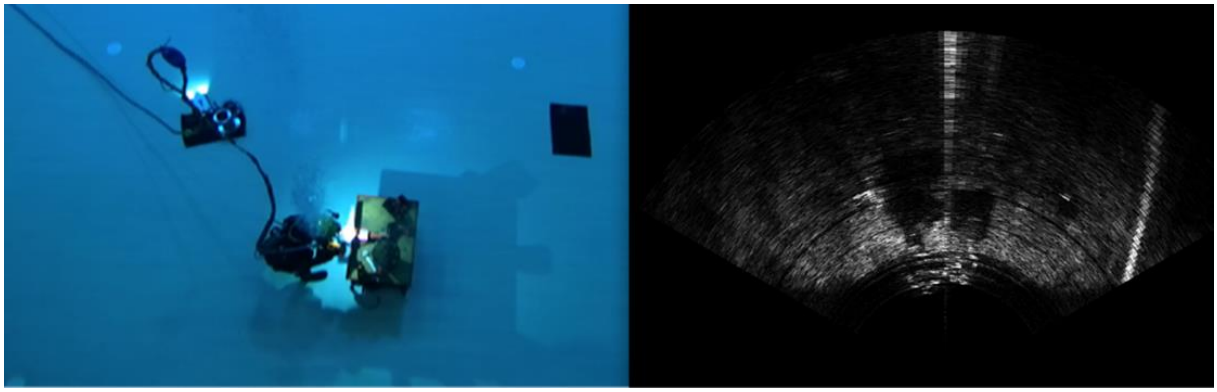

Figure 4: Example of data collected during the first session. Left: RGB image from optical Camera. Right: Sonar image from Gemini 720i

A total of 12 hours of divers in action was captured during the first session. Looking at the results, it was found that the resolution of the sonar was too low for the purpose of identifying objects and thus an Oculus M1200d sonar was used for the second session yielding more detailed sonar images as shown in Figure 5:

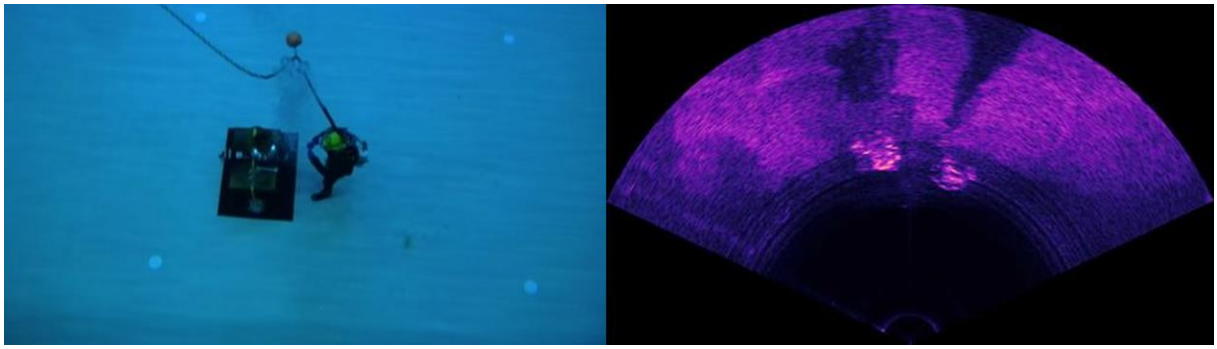

Figure 5: Left RGB image from optical Camera. Right: Sonar image from Oculus M1200d @ low frequency

Similar tasks were carried out by the divers on the second lab Session, and another work was added in order to increase the variation of the recorded activities. The divers processed aluminum profiles with various electric hand tools and used them to build a chair.

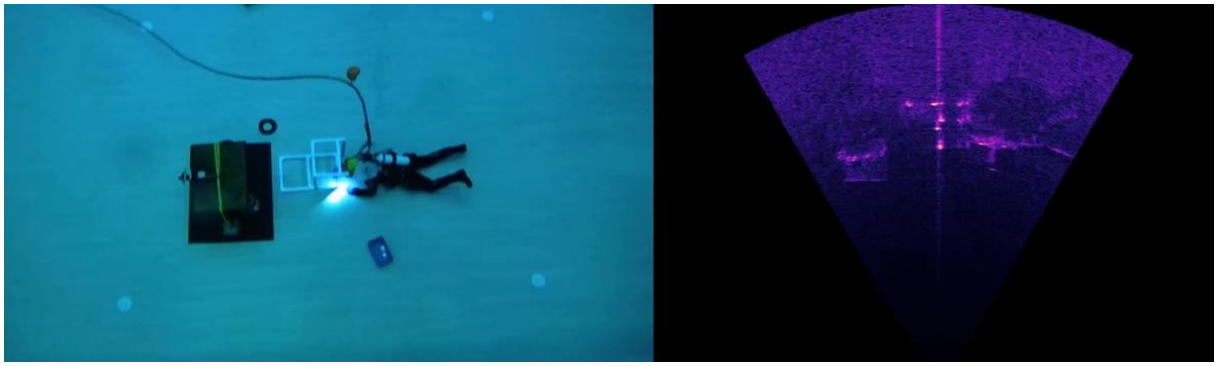

Figure 6: Left RGB image from optical Camera. Right: Sonar image from Oculus M1200d @ high frequency

The Oculus M1200d was operated in two frequencies, a low frequency mode (1200 kHz) that covers a wider field of view and provides a longer range but with low resolution (Figure 5), and a high frequency mode (2000kHz) that offers a narrower field of view but higher resolution. Figure 6 shows an example of the sonar working in high frequency mode, yielding a very detailed picture of the scene, where some details can be witnessed in the raw sonar picture, such as the strong reflections from the aluminum chair and the diver lying on the ground. A total of 9 hours of footage was created during the second session.

For both sessions, Seavision laser scanner was used to collect 3D point clouds of the scene. Figure 7 shows on the left the camera images from Seavision and on the right point cloud created as the result of a triangulation of a laser line project onto the scene.

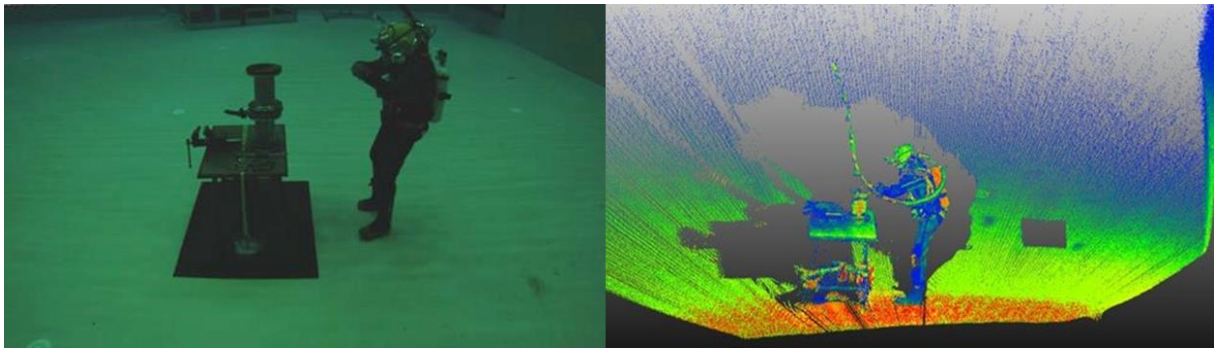

Figure 7: Point cloud created by Seavision underwater laser scanner

In addition to the point clouds, Seavision cameras were also synchronized with the other sensors and were used to collect optical images with a different point of view from the stereo camera system. In this work the point clouds were used to validate the ground truth data.

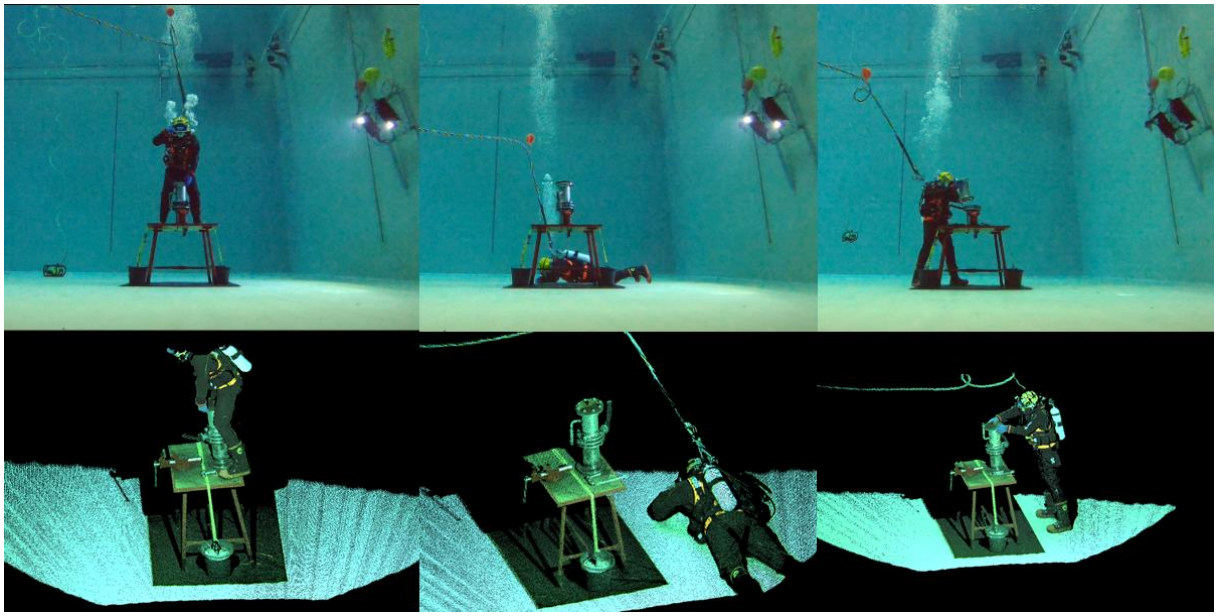

Figure 8: RGB point cloud of different body poses created by the Seavision laser scanner

Field Session 1 at Kreidesee

The first field session was conducted in a lake called Chalk Lake (Kreidesee) which is located in an abandoned chalk mine. It was selected because the visual range underwater is usually up to 25m, what gives us the opportunity to record high quality ground truth data in a more application-oriented environment using the optical camera. The THW divers set up their workbench and equipment on a submerged wooden platform with a 3m × 3m surface area that was fixed 50cm above the floor of the lake at a depth of 5m. A picture of the lake and the submerged platform is shown in Figure 9.

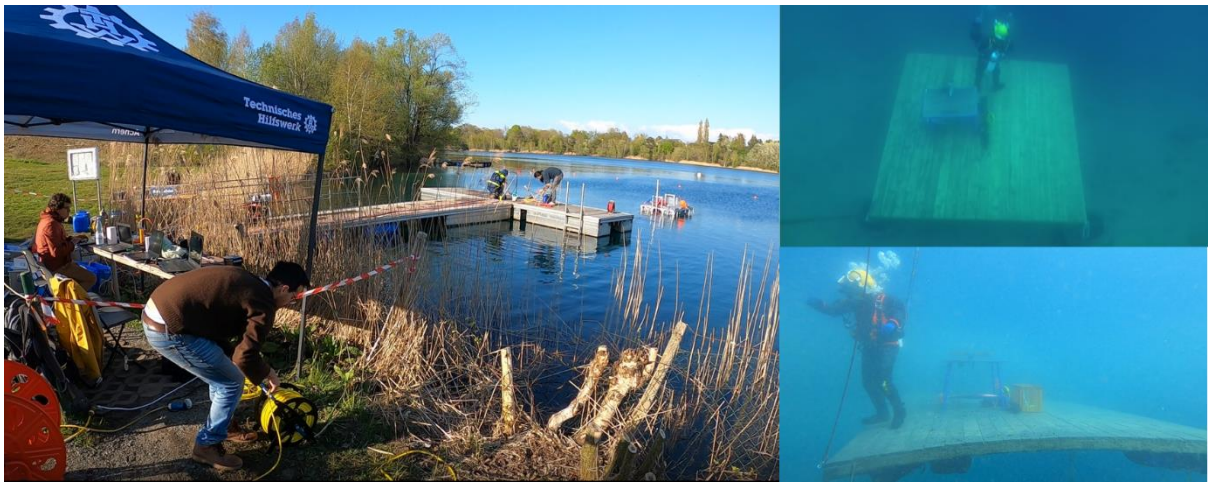

Figure 9: Test setup at Kreidesee in Hemmoor with wooden platform

The sensor setup for the lake was similar to the setup in the lab. The sensors were mounted on a rig that was fixed to a floating platform which was positioned to focus the wooden platform and held in place by ropes. An example of the sensor views is shown in Figure 10 where the wooden platform is clearly visible on the sonar image, which was set to low frequency mode as the sonar could not be moved close enough to the scene in order to be in range for the high frequency mode.

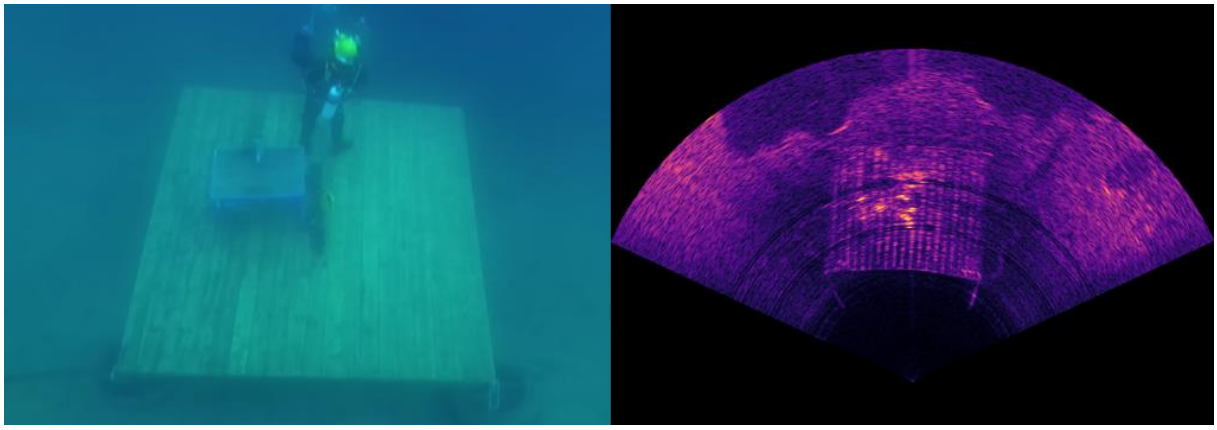

Figure 10: Left RGB image from optical Camera. Right: Sonar image from Oculus M1200d @ low frequency

During several dives, a total of 14 hours of video and sonar imagery was collected including activities like mounting of pipe flanges and woodworking with several tools as shown in Figure 11. Several woodworking tools were used, hand tools such as wood chisels and hand drills as well as electrically and pneumatically operated tools such as drills, chainsaws and saber saws. The variety of tools was put into use in order to get as diverse a range of input data as possible and to identify any interference between the different tool noises and the sonar image.

Figure 11: Tools used in Hemmoor session

Aside from the workbench actions, special activities were performed such as swinging weights, and walking on and around the platform. It was intended to also generate some example material for works that generate dirt and increase the turbidity of the water. Therefore, the diver shoveled some dirt around the platform to increase the turbidity and also a bucket of mud was emptied over the diver to create a situation in which the diver's visibility was suddenly reduced to zero as shown in the picture sequence in Figure 12.

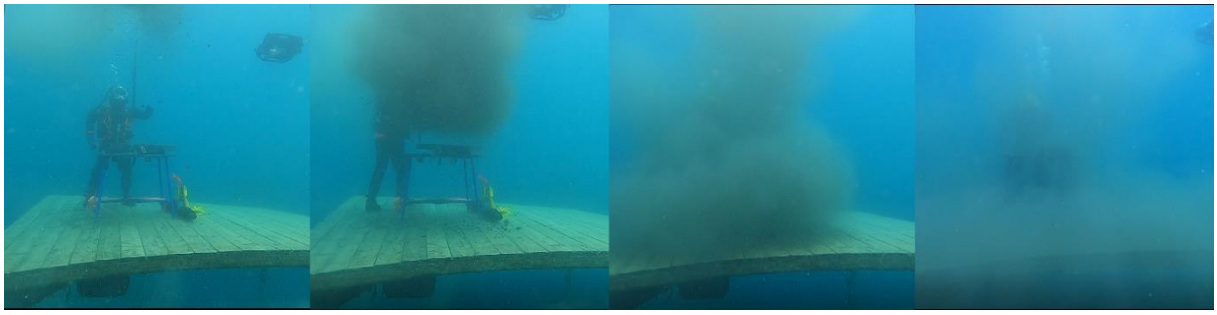

Figure 12: Image sequence from left to right: Emptying a bucket of mud on top of the diver

During the field session, some scans were also taken with Seavision underwater laser scanner. To maximize the signal to noise in shallow waters, the laser operation was performed at night. The distance from the Seavision cameras to the table was approximately 5m. Due to the distance to the target, the laser remission is better suitable for colorizing the point cloud. Figure 13 presents point clouds generated by Seavision with divers at two different body poses.

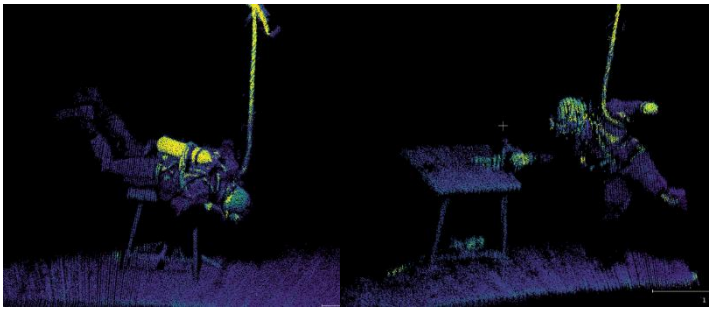

Figure 13: Point cloud generated by Seavision at approximately 5m distance from target.

Field Session 2 in a diver training basin

The second field diving session was conducted in a training basin in Neu-Ulm, Germany. The basin has the outer dimensions of 22m × 20m and a depth of 2,75m and the sensors were mounted to the basin wall as shown in Figure 14.

The pool was chosen because the underwater visibility is about 2m to 3m and thus significantly lower than in the previous tests and represents a more realistic working environment for the technical divers, but still provides good visibility compared to what the divers usually encounter. Furthermore, the filtration system of the pool allows the introduction of debris through the work of the divers, allowing us to use the full range of underwater tools available.

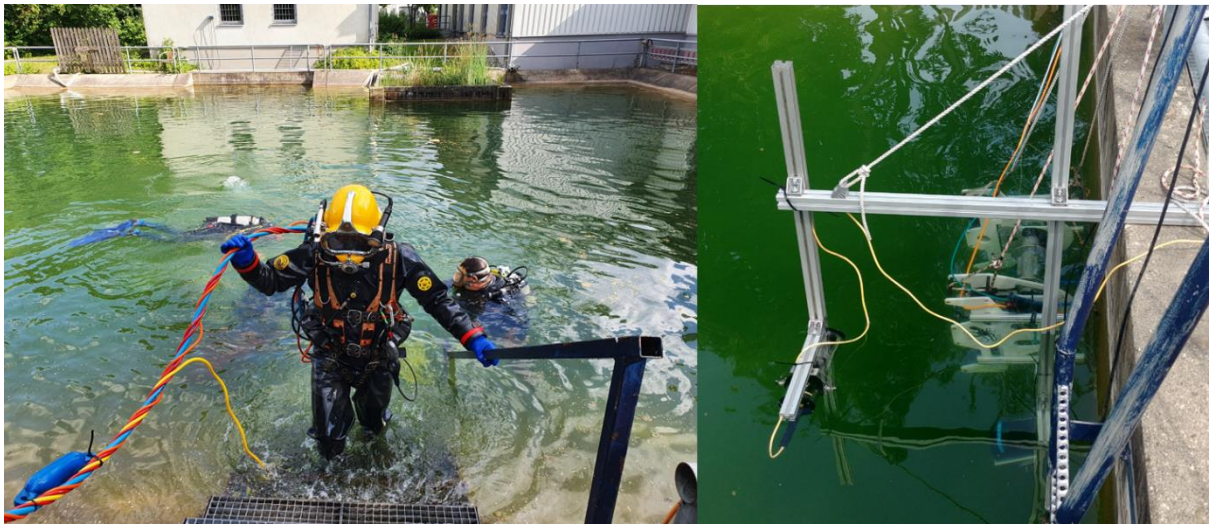

Figure 14: Test setup at Neu-Ulm

Some of the tools that were put into action are shown in Figure 15, from bottom left to right:

1. Drilling stone with hydraulic drill & core drill & electric drill
2. Cutting wood with chainsaw
3. Cutting concrete block with ring saw
4. Cutting metal with PrimeCut

Figure 15: Tools used in Neu-Ulm

Since the scope of session was mainly geared towards data with low visibility, there was no data collection performed with Seavision laser scanner.

PRELIMINARY RESULTS

In this section we present preliminary results using the algorithm described earlier. We used two datasets to train the aforementioned model: (1) an indoor dataset that was collected in the test basin of DFKI, and (2) an outdoor dataset which was collected at Hemmoor lake.

For the first dataset, we used 3078 pairs of sonar and camera images that were temporally aligned. Here the high-frequency mode of the Oculus sonar was used, and both the sonar and camera images were resized to 512 x 512 pixels. To simulate the effect of turbidity or darkness, the input camera images were disturbed by applying a gaussian blur and darkness by varying the intensity factor from 0.2 to 1 (1 equals completely black). The model was then trained for 100 epochs using back propagation with a batch size of 8, and the ADAptive Moments (ADAM) optimizer with a learning rate of $2x10^{-4}$. The model was then evaluated on a separate test set from the same experiment. The data was split into a train/test sets, with 80% used for training and 20% for testing. Few examples of the results can be shown in Figure 16, where the first column from the left shows the sonar image, and the second column shows the disturbed camera image that was used as input. The third column shows the original ground truth image compared to the generated output of the model that is shown in the right most column. In this figure we demonstrate the performance of the trained model with increasing the effect of darkness and blur in the input camera image. One may notice that at darkness levels of 50% to 90%, the model is able to reproduce the original camera image with a high degree of detail. With the darkness pushed to the max, we can notice that some of the details in the generated image are lost, however the diver and the work bench are very clearly identifiable.

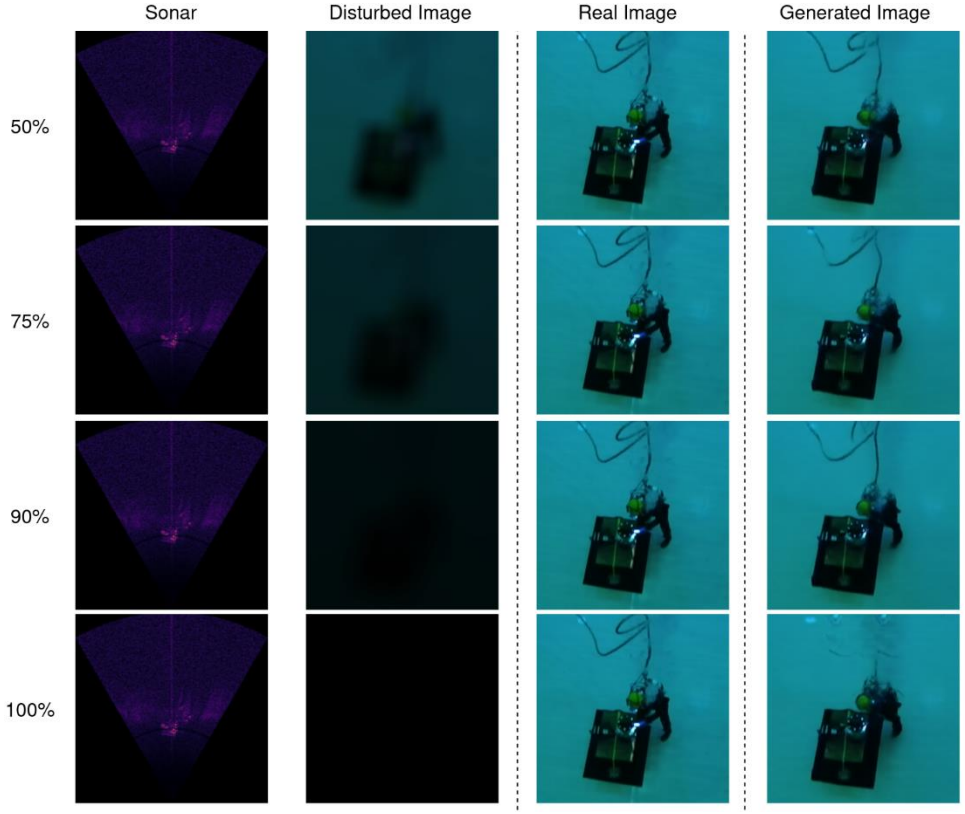

Figure 16: Comparison between the real images and generated images evaluated on the indoor basin dataset

For the second dataset, the imaging sonar was used with the low frequency mode which covers more range compared to the high frequency, but at the cost of reduced resolution. This was necessary as the distance to the diver was greater than in the pool. In this case we used pairs of 3328 training samples where both sonar and camera images were rescaled to 1024x512 pixels. The same procedure was used to train the model as in the case of the first dataset. Figure 17 shows the results of the model, where a noticeable degradation in the performance can be observed when compared to the indoor results. At 50% and 75% darkness, the generated image still highly resembles the original image, and the diver can be identified visually, however the generated image still suffered from blurring. At higher darkness levels, the generated image loses a lot of details and texture.

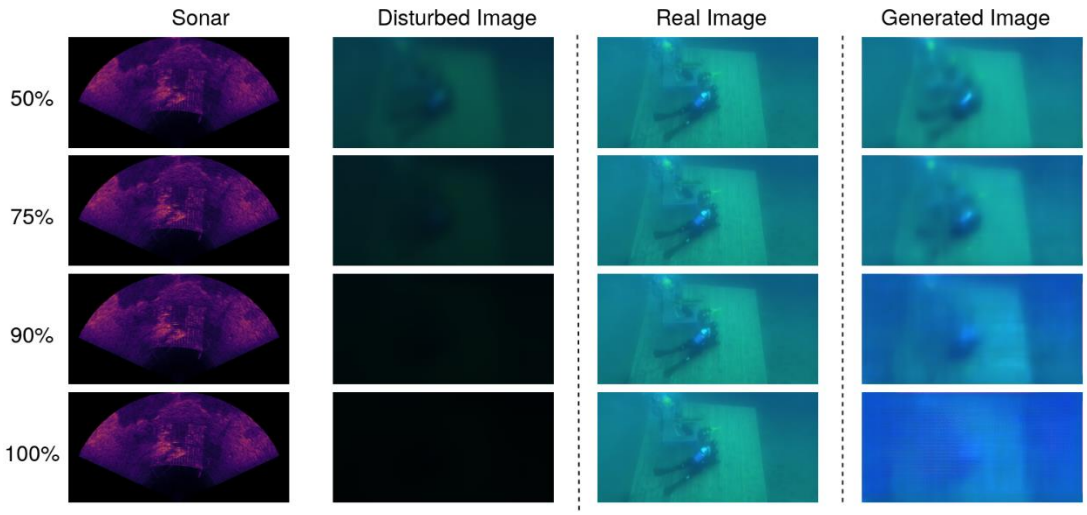

Figure 17: Comparison between the real images and generated images evaluated on the outdoor lake dataset

We report the structural similarity index measure for both experiments in Table 1, which is a measure of the quality of the predicted images based on the original one.

Table 1: SSIM indices for both indoor and outdoor experiments with varying darkness and blurring factor of the camera image

Darkness/Blurring factor	Basin dataset	Lake dataset
50%	0.9676	0.936
75%	0.9715	0.9278
90%	0.9680	0.9225

Further investigation of these results is currently being undertaken to identify the reason behind the discrepancy between the lab and field data. One of the possible justifications might be contributed to the lower level of detail the low frequency mode of the sonar provides compared to the high frequency mode. As a next step, we plan to train on a larger dataset in an attempt to help improve the accuracy of the predicted images.

CONCLUSION & OUTLOOK

In this paper we describe the development of a system that enhances the possibilities for diver monitoring in turbid waters. For this purpose, the steps of data generation are described, which were carried out in great detail and care was taken to ensure that very heterogeneous data was obtained in different environments. It was shown that the C-GAN algorithm used is able to generate a visual image from the sonar data for data generated in a laboratory environment. Nevertheless, it was also shown that it is difficult to produce a good image when the sonar shows little detail in a field environment using the latest version of the algorithm.

Therefore, in the further course of the DeeperSense project, we will focus on using the system in the optimal range of its capabilities. For the next tests, a ROV will be used as a sensor carrier so that the sonar can always be positioned at the optimal distance from the scene.

With this measure and with an extended algorithm training, we aim to achieve positive results for field use by the end of the project. In addition, the data collected will be made available to the research community at the end of the project.

ACKNOWLEDGMENTS

The DeeperSense project is a European research project funded by the EU Horizon 2020 research and innovation program under grant agreement No 101016958. Within the three-year project a common platform for inter sensory learning will be created and three use-cases will be addressed by the 7 partners from Germany, Israel and Spain. The opinions expressed in this document reflect only the author's view and reflect in no way the European Commission's opinions. The European Commission is not responsible for any use that may be made of the information it contains.

REFERENCES

Akkaynak, D. et al., 2014. Use of commercial off-the-shelf digital cameras for scientific data acquisition and scene-specific color calibration. *JOSA A, 31(2), 312-321*.

Arnold-Bos, A., Malkasse, J.-P. & Kervern, G., 2005. A preprocessing framework for automatic underwater images denoising. *Proc. Eur. Conf. Propag. Syst., Brest, France*, Mar, pp. 15-18.

Bazeille, S., Quidu, I., Jaulin, L. & Malkasse, J.-P., 2006. Automatic underwater image pre-processing. *Proc. Caracterisation Du Milieu Marin*, Oct, pp. 16-19.

Berman, D., D. Levy, S., Avidan & Treibitz, T., 2020. Underwater single image color restoration using haze-lines and a new quantitative dataset. *IEEE transactions on pattern analysis and machine intelligence, 43(8), 2822-2837*.

Berman, D., Treibitz, T. & Avidan, S., 2016. *Non-local image dehazing.* s.l., s.n.

J. Li, K. A. S., Eustice, R. M. & Johnson-Roberson, M., 2018. WaterGAN: Unsupervised Generative Network to Enable Real-Time Color Correction of Monocular Underwater Images. *IEEE Robotics and Automation Letters, vol. 3, no. 1, doi: 10.1109/LRA.2017.2730363.*, Jan, pp. 387-394.

Jia, D.-X. & Ge, Y.-R., 2012. *Underwater image de-noising algorithm based on nonsubsampled contourlet transform and total variation.* s.l., s.n., pp. 76-80.

Mirza, M. & Osindero, S., 2014. Conditional generative adversarial nets. *arXiv preprint arXiv:1411.1784*.

P. Isola, J.-Y. Z., Zhou, T. & Efros, A. A., 2017. *Image-to-image translation with conditional adversarial networks.* s.l., s.n., p. 1125–1134.

Ronneberger, O., Fischer, P. & Brox, T., 2015. *U-net: Convolutional networks for biomedical image segmentation.* s.l., Springer, p. 234–241.

Terayama, K., Shin, K., Mizuno, K. & Tsuda, K., 2019. Integration of sonar and optical camera images using deep neural network for fish monitoring. *Aquacultural Engineering, vol. 86*, p. 102000.

Treibitz, T. & Schechner, Y. Y., 2009. Active Polarization Descattering. *EEE Transactions on Pattern Analysis and Machine Intelligence, vol. 31, no. 3*, Mar, pp. 385-399, doi: 10.1109/TPAMI.2008.85.

Wang, T.-C.et al., 2018. *High-resolution image synthesis and semantic manipulation with conditional gans.* s.l., s.n., p. 8798–8807.